J Comput Aided Mol Des (2010) 24:591–604
DOI 10.1007/s10822-010-9349-1

Towards the comprehensive, rapid, and accurate prediction of the favorable tautomeric states of drug-like molecules in aqueous solution

Jeremy R. Greenwood · David Calkins · Arron P. Sullivan · John C. Shelley

Received: 1 February 2010 / Accepted: 19 March 2010 / Published online: 31 March 2010

Abstract Generating the appropriate protonation states of drug-like molecules in solution is important for success in both ligand- and structure-based virtual screening. Screening collections of millions of compounds requires a method for determining tautomers and their energies that is sufficiently rapid, accurate, and comprehensive. To maximise enrichment, the lowest energy tautomers must be determined from heterogeneous input, without over-enumerating unfavourable states. While computationally expensive, the density functional theory (DFT) method M06-2X/aug-cc-pVTZ(-f) [PB-SCRF] provides accurate energies for enumerated model tautomeric systems. The empirical Hammett–Taft methodology can very rapidly extrapolate substituent effects from model systems to drug-like molecules via the relationship between pK_T and pK_a. Combining the two complementary approaches transforms the tautomer problem from a scientific challenge to one of engineering scale-up, and avoids issues that arise due to the very limited number of measured pK_T values, especially for the complicated heterocycles often favoured by medicinal chemists for their novelty and versatility. Several hundreds of pre-calculated tautomer energies and substituent pK_a effects are tabulated in databases for use in structural adjustment by the program *Epik*, which treats tautomers as a subset of the larger problem of the protonation states in aqueous ensembles and their energy penalties. Accuracy and coverage is continually improved and expanded by parameterizing new systems of interest using DFT and experimental data. Recommendations are made for how to best incorporate tautomers in molecular design and virtual screening workflows.

Keywords Tautomer · Epik · pK_a · Hammett–Taft · DFT · Virtual screening

J. R. Greenwood (✉)
Schrödinger, L.L.C., 120 West 45th St., 17th Floor, Tower 45, New York, NY 10035-4041, USA
e-mail: Jeremy.Greenwood@schrodinger.com

D. Calkins · A. P. Sullivan · J. C. Shelley
Schrödinger, L.L.C., Suite 1300, 101 SW Main Street, Portland, OR 97204, USA

Introduction

Analysis indicates that the majority of the drugs listed in the 1999 World Drug Index are ionisable [1] and it seems reasonable to believe that many of these have the potential to produce multiple tautomeric forms in aqueous solution. Tautomerism takes on many forms and involves a very broad range of chemical entities [2]. Our interest is primarily restricted to pragmatic methods for the treatment of tautomers of ligands in virtual screening, which none the less is subject to a wide range of approaches and preferences, in part because it is still maturing.

These days virtual screening studies routinely involve up to 10^7 molecules and are targeted at dramatically reducing the number of candidates by quickly eliminating those with undesirable calculated properties such as low binding affinities. Despite the fact that each such screen typically involves substantial human and computational resources, they are carried out in many pharmaceutical and biotechnology companies [3]. Whether the screen is protein-based or ligand-based, the tautomeric state (or more generally the protonation state) will often determine the pattern of hydrogen bonding possible between the ligand and the protein, or which pharmacophore features are present in a given spatial orientation to match a 3D hypothesis.

By necessity, we begin by examining the phenomenon of tautomerism in the broadest sense, since there appears to be a diversity of opinion and understanding of the topic among computational chemists, and the available tools vary substantially in their coverage. For ligand preparation and screening, practical considerations are paramount, so we then focus on the kinds of tautomerism that are most important for molecular modellers to include when preparing large libraries of drug-like molecules, and describe various types of tautomerism whose handling we believe we can afford to eliminate or postpone. We consider what the scope and coverage of a high quality tautomerization code for screening applications should be, and when and whether high energy tautomers may be needed. Our emphasis is on how find the best states (including tautomers) which means developing an algorithm that accurately estimates tautomer energies in water, not just enumerates structures, and which can be expanded to cover any desired drug-like molecule, not just those for which experimental data is available. While presenting this perspective is the focus of this article, much of what we describe is put into practice in the ongoing development of Schrödinger's pK_a protonation state prediction application Epik [4, 5] (and a related tool, LigPrep's tautomerizer [6]).

What types of tautomers should be generated for drug-like chemistry in high-throughput virtual screening workflows?

Before considering a method for determining tautomers, it is important to establish exactly what we mean by tautomerism in the context of biological chemistry and drug discovery, and what kinds of tautomeric variations should be generated for virtually screening large numbers of drug-like molecules. Of necessity, we need to consider in some depth not only the thermodynamics of various kinds of tautomeric equilibria, but also the kinetics, as well as drug-likeness, since all three factors influence what substances or mixtures medicinal chemists deliver to pharmacologists for testing as distinct entities. We hope that this triage approach will help shift the focus of the discussion away from subclasses of tautomerism that, although scientifically interesting, are of little importance for virtual screening and that would none the less take considerable time and effort to analyze and support, potentially delaying realization of a practical approach, while yielding little incremental benefit.

In the broadest sense, tautomerism can refer to a wide range of isomeric transformations under different conditions—from the subtle bonding and geometry rearrangements in fluxional molecules such as bullvalenes known as "valence tautomerism", to the heavy-atom bond forming and breaking familiar in the "ring-chain tautomerism" of carbohydrates, to reactions involving the relocation of a hydroxyl via covalent hydration/dehydration, the shift of a methyl group, and so on (Fig. 1). However, in the context of the behavior of biological and drug-like molecules under physiological conditions in solution and when interacting with proteins, the most important and familiar transformations by far are the equilibria established by the simple relocation of one or more protons via sets of protonation/deprotonation reactions, that can occur on a biologically rapid timescale. Thus we advocate predominantly focusing on a somewhat narrow subset of what is known as "prototropic tautomerism", i.e., isomers in which the only change in formal structure involves the relocation of hydrogens and changes in bond orders between heavy atoms (e.g., Fig. 1a) but not including changes to bond orders significantly below 1, i.e., heavy atom bond breaking/forming or ring opening/closing.

On the kinetic side, we are interested in reactions that occur in picoseconds to minutes, not days to years, in water under typical pharmacological assay or biological conditions (ca. 0–40 °C). Note that this includes all prototropic tautomerism reactions that are fast on the NMR timescale, as well as some that are slower (e.g., Fig. 1b). Some authors have suggested an activation energy of 25 kcal/mol as the cutoff between tautomerism and isomerism when designing tautomer tools [7]. Without a rapid and accurate way to estimate the rate of equilibration, we look to the estimated aqueous pK_a values as a proxy for reaction rate. Acid or base catalyzed equilibrations where the required pK_a values lie too far outside the accessible range in water (-1.3 to $+15.7$) can be ruled out. Thus certain kinds of isomers, such as the rearrangement of an allyl to propenyl substituent (Fig. 1c), that meet the broad definition of prototropic tautomerism but only involve proton abstraction from an sp^3 carbon or addition to an sp^2 carbon, and will indeed equilibrate under certain conditions, should not be included because medicinal chemists would be register them as separate chemical entities. In such cases, the interconversion is slow enough for the isomers to be isolated and the biological activity tested separately. At the other extreme, systems like histidine, whose imidazole (Fig. 1a) group has a pK_a as a base of around 6, will extremely rapidly equilibrate between the two neutral tautomers at pH 7 (together with the cationic conjugate acid). Even though we have no practical way of predicting a priori the reaction rate for every possible proton exchange under assay conditions, somewhere between these kinetic extremes we need to delineate between tautomeric variations that should not be automatically generated because a pharmacologist could test as them as separate isomers, while attempting to cover all the chemistry for entities that rapidly establish a tautomeric equilibrium in an assay.

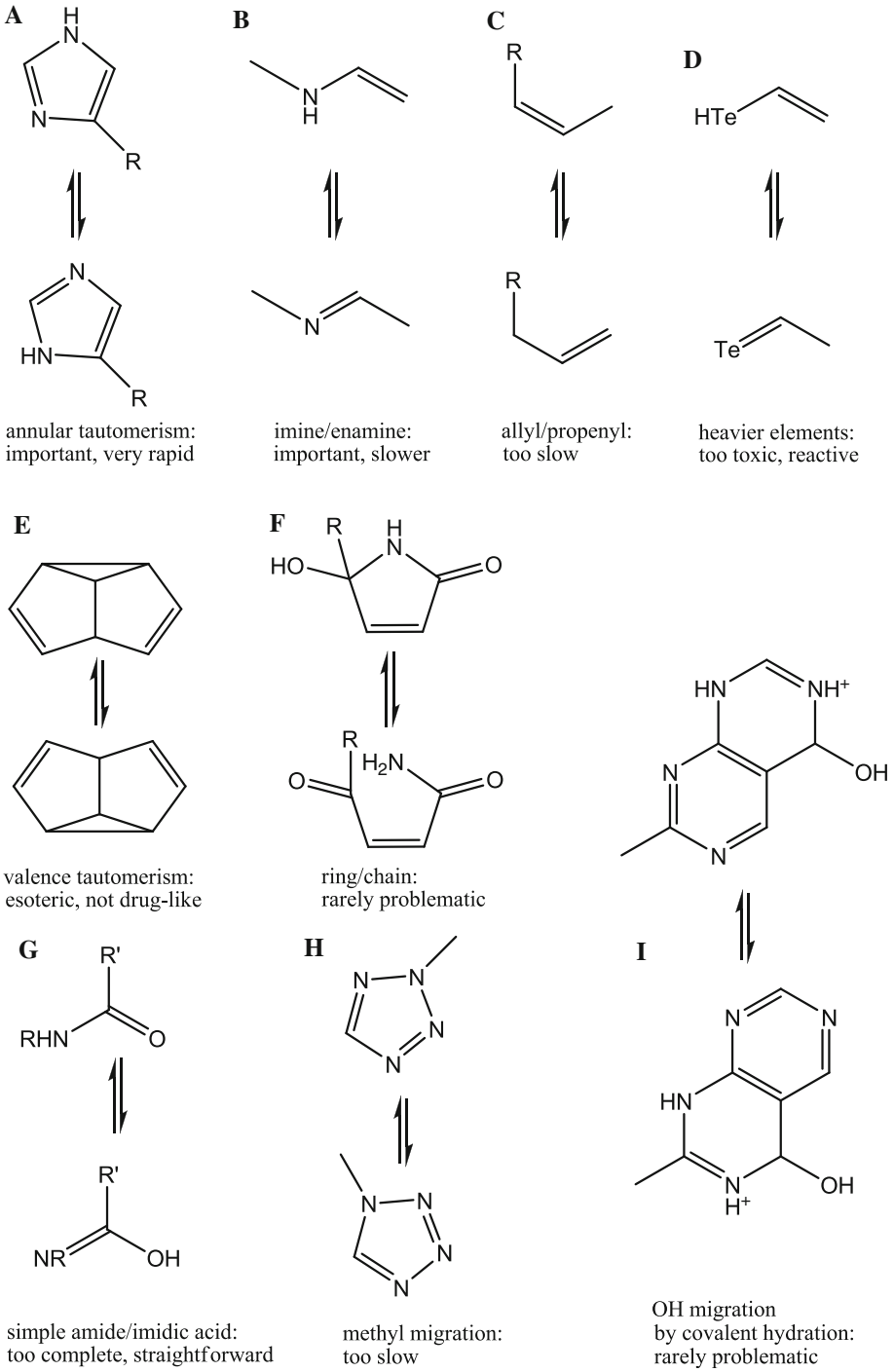


Fig. 1 Different types of tautomerism of greater or lesser importance in computer aided drug discovery with our assessments of their relevance to a fast tautomerization tool suitable for virtual screening

As a shorthand for focusing in on the most kinetically relevant tautomeric equilibria for drug-like chemistry, it is convenient to identify sets of atoms that are connected via conjugation (including aromaticity) and which may involve proton exchange from carbon but must involve at least one heteroatom (specifically nitrogen, oxygen or sulfur). Such species, which may be charged, neutral, or mesionic, typically have a heteroatom that is protonatable or a hydrogen that is labile somewhere in the aqueous pH range, and thus produce a delocalized conjugate acid or base, facilitating tautomeric equilibration.

The third consideration for prioritization is drug-likeness While some tautomer software covers other di/trivalent heteroatoms in low oxidation states and capable of pi-bonding [8], we believe these are rarely of pharmacological interest (e.g., Fig. 1d) and are thus not a priority to parameterize in our opinion. Likewise, coverage of some of the more toxic or reactive types of tautomers can be

postponed (e.g., aci/nitro, oxime/nitroso, and azo/hydrazone). In addition, for virtual screening, we confine our interest to closed shell ground state systems; other tautomer tools exist whose focus is more directed to studying exotic states [9].

Thus, reactions involving keto/enol, thione/thiol, imine/enamine and annular tautomerism reactions are the highest priority for comprehensive and accurate coverage. However, we suggest that the tautomerization infrastructure should allow expansion to cover more classes of tautomerism as the need arises.

Another practical consideration is what to do about moieties that are capable in theory of rapidly reaching tautomeric equilibrium, but where the equilibrium lies so far to one side that there is rarely if ever doubt about what tautomer predominates in solution, no matter what substituents are bonded to the tautomeric group. Thus for example, simple non-aromatic imidic acids convert almost completely to amides (Fig. 1g), and non-aromatic nitroso groups to oximes (which can also have toxicity/reactivity issues). In theory, we could cover all such cases accurately and explicitly. In practice, since there is only one form of interest, and it is highly likely that this is the form recorded by the medicinal chemist who entered the structure (or the machine-generated form in the case of virtual libraries), explicitly encoding tautomerizations of these moieties is unnecessary or at least of low priority. It is possible to treat them in a more general implicit way in the course of acid/base protonation state assignment since the high energy forms will exhibit extreme pK_a values that will unambiguously dictate the correct form during pK_a-based structural adjustment. We do not want to spend extensive parameterization effort or computational time on equilibria where, e.g., 25 kcal/mol separates the commonly depicted state from the next lowest in energy. Again, if specific cases arise where very strongly perturbing substituents cause more than one state to be present, or where a particular type of compound has often traditionally been drawn in a high energy state, tautomerization schemes for these can be added.

We would also assign low priorities to supporting the other kinds of tautomerism involving changes to heavy atom connectivity that organic molecules may undergo in solution, including valence (Fig. 1e), ring-chain (Fig. 1f), alkyl (Fig. 1h) or hydroxyl (Fig. 1i) shifts, amongst others. Indeed these types of tautomerism are not implemented in Epik, although they could in principle be added at some future date. Covering these types of tautomerism would be a much larger task, and with rare exceptions [10] we believe that such reactions not sufficiently common or important in drug-like molecules to create a pressing need for automated methods that handle them rapidly and accurately. Rationalizations for this stance include:

1. the kinds of molecules prone to rapid heavy atom bond breaking and formation rarely make good drugs (outside of specialized applications).
2. when such intramolecular reactions can occur they often have equilibria that strongly favor particular tautomers and thus would usually already be drawn in the appropriate state.
3. some of these reactions are slow enough to be ignored.
4. such molecules are often reactive and thus are at risk of undergoing undesirable intermolecular reactions in vivo (e.g., rapid metabolism, reacting with enzymes, DNA, etc.).

For example, the ring-chain tautomerism of cyclic hemiaminals (Fig. 1f) and hemiketals, often have equilibria that will have been determined to lie heavily to one side (e.g., by spectroscopy, before the compound reaches the modeller or the pharmacologist), and will almost always have been drawn with the appropriate structure. For the time being, in the context of computer aided drug discovery, we argue that it is both pragmatic and adequate for the great majority of drug-like molecules to rely on the heavy atom scaffold assigned by medicinal chemists and to concentrate on improving the coverage and predictions for protonation states in general, of which the assignment of prototropic tautomerism is an important component. When that effort matures it may make sense to include also cover some other types of types tautomerization including ring-chain tautomerism.

In summary, we recommend that a tautomer tool for drug-like molecules should initially focus on developing coverage of proton movements only, cover reactions which reach equilibrium in water in the time it takes to prepare and assay the sample, and place more emphasis on complex or finely balanced cases where there's a reasonable likelihood of more than one state being energetically accessible, particularly those for which the correct distribution is not well known. In addition, there should be particular attention to the tautomers found in biological chemistry and a balance between accurately covering drug-like moieties that are known to be synthesized frequently, versus coverage of rarer or as yet unsynthesized structures.

The comprehensive treatment of heterocyclic tautomerism requires assistance from a first principles approach

The tautomerization of heterocycles is the subset of prototropic tautomerism that is most important to medicinal chemists and gives rise to the richest range of alternative states. However it is also the source of many equilibria that are difficult to predict. The enormous variety, versatility

and ability to fine-tune the molecular properties of heterocycles is one of the features that make them so important for medicinal chemistry. A good example is the set of all heterocycles that mimic how adenine binds to the hinge region of kinases [11]. The subtle competition between aromatic stabilization energy and the relative stability of heteroatoms in alternative hybridization states resulting from the rearrangement of labile protons and electronic structure produces tautomeric repertoires that even the most experienced chemists would have difficulty guessing. Thus, we feel that most of the effort should be devoted to accurately treating substituted heterocycles with at least one potential aromatic tautomer.

We argue that no simple rule-based or purely empirical scheme will ever accurately cover heterocyclic tautomerism beyond the ability to enumerate tautomers. On the one hand, many interesting heterocycles lack accurate experimental data, partly because the large number of experiments needed to fill the gaps have been considered too mundane and repetitive to receive significant funding in recent decades, and partly because novel combinations of rings and substituents are continually being synthesized in the quest for patentable scaffolds with high ligand efficiency. On the other hand, because the aromatic resonance energy invariably affects the tautomeric equilibrium, which is dictated by subtle influences on the $(4n + 2)\pi$ electron density, the outcome may be difficult to predict based on simple rules, as can be quickly seen by comparing how the textbook resonance energies of a variety of common simple heteroaromatics vary substantially (pyridine, furan, thiophene, imidazole, etc.) [12] and the numerous articles, chapters, and volumes devoted to specific examples [13, 14]. Clearly, the precise contribution of delocalization to each tautomer of, e.g., xanthine or pterin will be different and often critical to determining the outcome, thus frustrating efforts based on putting atoms and functional groups from single Lewis structure representations into predictive categories, especially for charged delocalized systems. Unless electronic structure is considered, each heterocycle is effectively a new problem. A comprehensive solution for calculating the tautomeric equilibria of screening collections must therefore rely on some kind of quantum chemistry. To be fast enough to prepare 10^7 ligands, a tautomerization application must either require the development of a very rapid simplified and specialized electronic structure model that is heavily tuned for tautomers, or heavily rely on pre-generated data from much slower quantum chemistry calculations. We favor the latter approach for obtaining our reference data: to pick a popular quantum chemistry method that is already well established and needs no further parameterization, whose generality and accuracy is easily verified by others, and then apply it in advance to a large number of systems.

Choosing a theoretical method in close quantitative agreement with experiment

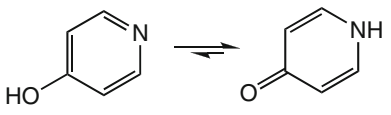
For many years, various types of quantum chemistry have been used to calculate the relative energies of tautomers in gas phase, and as the sophistication of solvation calculations has advanced, also in water [14]. In the last few years one of the most popular and widely used quantum chemistry methods for drug-like molecules of this size is density functional theory (DFT), in combination with continuum solvation models. The best DFT methods for this kind of problem are approaching the accuracy of high level post-Hartree–Fock ab initio methods such as Coupled Cluster calculations [15] which unfortunately are currently still prohibitively expensive for a large scale parameterization campaign for tautomeric systems, due to N^7 scaling despite the advent of efficient parallelization [16]. We note in passing that established NDDO-based semi-empirical theories, while rapid enough to process a large number of tautomers, are not quantitatively or qualitatively reliable enough in our experience (PM3, AM1, MNDO, SAM1). For example, they tend to err for rings with contiguous heteroatoms [17] which we regard as a crucial test of suitability for handling tautomerism (molecules like pyridazinones and isoxazolones). While newer generations of semi-empirical theories claim improvements in mean unsigned errors (e.g., 4.8 kcal/mol for PM6 [18]) this is still a long way from the chemical accuracy where we would be interested in pursuing such methods for tautomeric distributions in water, given that we have tried and abandoned methods with better reported MUEs such as B3LYP [19].

Quantum chemistry is well suited for studying prototropic tautomerism because it involves a simple isogyric ground state comparison of isolated isomers, without complicating factors like basis set superposition error (that need to be considered for intermolecular reactions), and limited influence from difficult to treat dispersion effects. On the other hand, it has long been known that very large basis sets together with post-SCF electron correlation treatment are necessary to obtain convergence and experimental agreement for heteroaromatics, where the tautomer stability depends on subtle changes in the aromaticity and hybridisation of ring atoms [20]. Even with modern computers and software, such calculations are usually too slow for individual flexible drug-sized molecules where the medicinal chemists typically require rapid turnaround for testing ideas, and are out of the question for processing large libraries of drug-like molecules. However, for fairly small test systems, very high quality calculations can be performed exhaustively and routinely, at a level of theory likely to reproduce experiment to within chemical accuracy. While there may not be consensus on the minimum

level of theory that is adequate for studying different kinds of tautomers, we prefer to use a modern Density Functional Theory such as Truhlar's M06-2X [21], with an augmented triple-zeta basis set such as aug-cc-pVTZ(-f), and Poisson-Boltzman self-consistent reaction field continuum model of the free energy of aqueous solvation (e.g., PB-SCRF, implemented in Jaguar [22]) which we believe is more than adequate for this kind of problem, given the other uncertainties in both modeling and experimental techniques encountered in drug design. This hybrid meta density functional has been shown to provide broad accuracy for main group chemistry [23] and we have spot checked the results for tautomers against results from Coupled Cluster Theory and complete basis set extrapolation in gas phase and with a variety of solvation models [24]. Of the many continuum solvation models available, all are generally parameterized against the same experimental free energies of solvation, and PB-SCRF's performance is considered comparable for a variety of applications in independent tests [25–27]. We have looked at more recent alternative solvation models such as SM6 that have been tested with M06-2X [28]. But the local performance for tautomeric equilibria appears to be somewhat better with PB-SCRF, perhaps because the atomic radii have also been fitted to reproduce aqueous pK_a data in the Jaguar pK_a predictor [22]. We continue to monitor developments in quantum chemistry and solvation models, for emerging methods that could improve accuracy without significantly increasing computational cost.

In our hands, this recipe of free energies from M06-2X/aug-cc-pVTZ(-f) [PB-SCRF] has consistently produced excellent agreement with experiment. For example the experimental pK_T of 4-pyridol versus 4-pyridone (Fig. 2) in water is reported to be 3.3 ($\Delta G = 4.5$ kcal/mol) while the corresponding ΔG value from M06-2X/aug-cc-pVTZ(-f) [PB-SCRF] value is 4.2 kcal/mol.

5-Oxazolones are traditionally regarded as a difficult tautomeric system for electronic structure calculations, because a large basis set and high level treatment of electron correlation are necessary to treat sp^3 and sp^2 centres on equal footing, and solvent effects are particularly tricky [20]. Again, the performance of M06-2X with PB-SCRF is in remarkably good agreement with experiment, including the change in tautomeric preference caused by a methyl substituent (Fig. 3) [30]. The only discrepancy is that the presence of the CH tautomer of 4-methyl-3-phenyl-5-isoxazolone at 1.8 kcal/mol (calculated) was not detected in water in this experiment conducted in the early 1960s. We do not know what their lower detection limit was. The method may be slightly overestimating the stability of the CH form, but small amounts are also probably present in water, as indicated by its detection as the second most prevalent tautomer in other solvents, and in qualitative agreement with the relative energies.



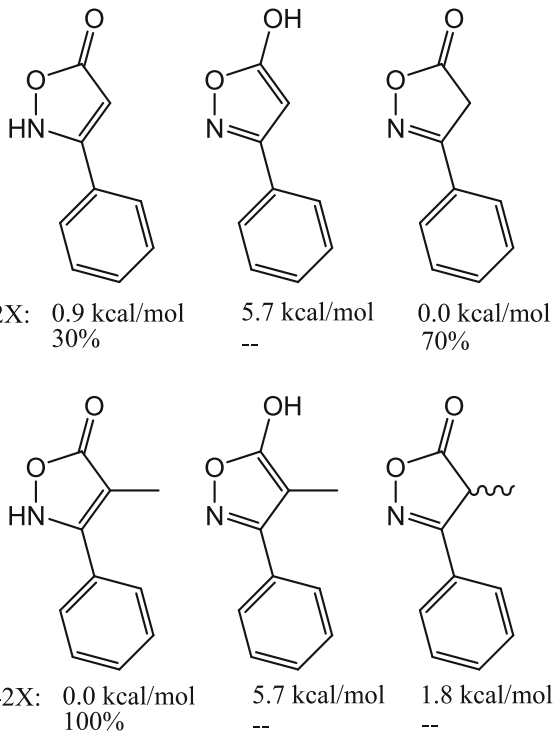
M06-2X: 4.2 kcal/mol
exp: 4.5 kcal/mol

Fig. 2 Close agreement between calculated and experimental aqueous distributions for 4-pyridone [29]

Scale-up and the use of pK_a prediction

Having chosen a quantum chemical method that works sufficiently reliably for individual small cases, the next challenge is scaling-up to a practical application. The aspects requiring automation and scale-up include:

1. obtaining energies for many tautomers of each simple tautomerizable system
2. generating or acquiring many types of relevant, yet simple, tautomerizable systems
3. extrapolating from simple tautomerizable systems to drug-like molecules with complex combinations of substituents



M06-2X: 0.9 kcal/mol 5.7 kcal/mol 0.0 kcal/mol
exp: 30% -- 70%

M06-2X: 0.0 kcal/mol 5.7 kcal/mol 1.8 kcal/mol
exp: 100% -- --

Fig. 3 Close agreement between calculated (M06-2X) and experimental tautomer ratios in water for 5-isoxazolones

4. building an engine that is rapid enough to handle very large databases of molecules.

The first practical issue is how to generate all the potentially relevant tautomers of a given system to submit to quantum chemical calculations; for this we have developed a tautomer enumeration tool that starts by removing all potentially ionizable protons, and works through each feasible charge state of a given system (typically between −2 and +2), considering successive protonation possibilities for most oxygens (sp^3-/sp^2, sp^3/sp^2+), nitrogens (sp^3-/sp^2, sp^3/sp^2+, sp^3+), sulfurs (sp^3-/sp^2, sp^3), and certain carbon protonations (sp^3-/sp^2, sp^3 alpha to heteroatom substituents), including rotamers and *cis*/*trans* isomers where applicable, and reassigning sensible Lewis structures along the way. This kind of scheme is not unique; it resembles for example that of TautGen [9], but the emphasis here is not a completely exhaustive enumeration, but on finding neutral and charged states that can be reached by acid/base reactions alone and which are candidates for significant population at equilibrium under typical assay conditions. In general a large number of tautomers, some of which will be high in energy, will be generated by this script and then evaluated by quantum chemical calculations, to minimize the chances that any low energy tautomers at accessible pHs in water are missed.

There are still an impractically large number of potential tautomeric systems to consider even if one just focuses on simple systems. Thus the next question is how to prioritize which systems to parameterize (seek experimental data for and perform quantum chemical calculations on) and add to the growing database of reference values. We feel that it is a priority to cover not only those systems that are well-described in the literature and compound collections, but also novel and thus potentially patentable scaffolds that may be rare or as yet unknown. It is for such novel structures, where there is typically no experimental data available and there is little experience to guide chemists, that a tautomer tool grounded in accurate quantum chemistry calculations can provide greatest value. Our work in this area has focused on achieving exhaustive coverage for all potentially aromatic monocyclic five- and six-membered heterocycles with up to two heteroatom substituents. Certain bicyclics and monocycles with more tautomerisble substituents, or four or seven members, as well as some acyclic structures, are explicitly covered if they are of known synthetic or biological interest (e.g., a large range of purine bases and their mimetics; acyclics like diketones and enaminones), if they turn up frequently in available compound libraries, or are suggested by internal or external users of our software.

The number of ways of combining heteroatoms and substituents leads to too many possibilities for complicated polycyclic heteroaromatics to examine all such systems by high-level quantum chemistry calculations. Thus pragmatic considerations necessitate the ability to extrapolate from simpler systems to more complex systems (preferably with a degree of redundancy). For most drug-like molecules we have found that describing mono- and bicyclic substructures explicitly is sufficient, although certain tricyclic aromatics will be difficult to describe accurately by extrapolation, due to intramolecular interactions (e.g., heteroatoms or substituents in positions four and five in phenanthrene derivatives). While in some cases these larger systems need their own patterns, fortunately medicinal chemists do not usually focus on such systems since they test the limits of what is normally considered drug-like, for reasons of solubility and toxicity [31, 32]. Our survey suggests the number of types of tautomerizations that need to be explicitly parameterized based upon quantum chemical calculations in order to achieve good coverage for drug-like chemical space will number in the 100s or possibly 1,000s, but not millions, indicating that this task is tractable with current technology and resources.

Thus, the first step is to build the model structures and their energies into an expandable database of minimally substituted cores, with certain rules for what kinds of substituents are allowed that are unlikely to drastically change the tautomeric repertoire of a given system. High energy structures are routinely included for complicated cores, not because they deserve a place in the output equilibrium, but to facilitate correcting poor input structures. In our implementation matching is achieved via SMARTS patterns, and care is taken to standardize canonical mesomeric representations. The process is described in detail in the tautomerizer section of the LigPrep User Manual [6] as well as the primary reference [4]. Note that following validation, the newer M06-2X functional has replaced B3LYP as the method of choice for paramaterizing tautomers in more recent versions of Epik; otherwise the basic methodology remains the same, and the library continues to expand.

While ideal for studying prototypal conjugated ring systems, high level quantum chemistry calculations, including DFT, by themselves are prohibitively expensive for large scale deployment on libraries of drug-like molecules of up to 10^7 chemical entities. Even for single drug-sized molecules, using Jaguar's DFT code which scales well due to the pseudospectral approximation for the SCF, it becomes tedious to disentangle tautomeric energies from conformational effects in solution once substituents with several rotatable bonds are involved. Thus DFT is a valuable but insufficient foundation for covering tautomerism in practice. A comprehensive solution to tautomerism must therefore, in addition to DFT, invoke an empirical scheme for rapidly and fairly accurately extrapolating from

pre-calculated reference values for unsubstituted cores to real molecules of interest.

The crucial technology that allows rapid extrapolation from reference values to molecules of arbitrary size is empirical pK_a prediction combined with structural adjustment. It has long been known that the same Hammett and Taft (H–T) methodology in widespread use for pK_a prediction [33] is equally applicable to the problem of tautomerism, so long as adequate reference data is available. This is no surprise, given that any proton shift can be decomposed into a pair of deprotonation/reprotonation or protonation/deprotonation reactions. As early as 1968, it was shown that substituent effects on the pyridone/pyridol equilibrium are amenable to this treatment [34]. Several general implementations of H–T for microscopic pK_a prediction are available, for example ACD/PhysChem Suite [35] and Pallas pKalc Net [36]. The SPARC implementation is also notable for its augmentation of H–T with a simple molecular orbital theory [37]. The H–T implementation in Epik, along with the extensions we have developed to improve and broaden its applicability, such as charge-spreading and internal mesomer handling, have been previously described in detail [4]. The accuracy of a number of these H–T implementations, including an older version of Epik (v1.6, 2008) were recently independently reviewed [38].

In theory, if a program were able to give very accurate pKa predictions for all basic heavy atoms and acidic protons in all states over a very wide pH range (e.g. −1.7–15.7), the prototropic tautomerism problem would be solved and structural adjustment would simulate all the acid/base catalyzed tautomeric reactions occurring on a reasonable timescale under biological conditions. And indeed, some tautomer prediction software relies solely on augmenting empirical rules with pK_a estimates, for example TauThor/MoKa [7]. In practice, we have found it to be a large and difficult task to consistently predict all the microscopic pK_a values of all the intermediate states with sufficient accuracy using a purely empirical scheme based on experimental pK_a values, especially for minor contributors to the equilibrium. Likewise, there are only a few tautomers for which the microscopic pK_a values of non-dominant tautomers have been measured, since this generally requires careful spectroscopy rather than potentiometric titration. Experimental measurements also become less reliable at extreme pHs and higher charge states, where side-reactions like hydration or hydrolysis become more prevalent, yet it is often small differences in these extreme values that determine the tautomer ratio. Sometimes comparing experimental data from multiple sources helps. However, it is almost always possible to understand and interpret experimental tautomeric and more generally protonation data with the assistance of quantum chemical calculations. For instance, when interpreting experimental macroscopic pK_a values by which to parameterize Epik's pK_a values, we frequently use DFT (via Jaguar's pK_a predictor [39], as well as tautomer calculations for order of acidity/basicity) to assist in identifying which acid/base reaction is most likely being measured, or on occasion, whether the literature value may be uninterpretable or in error. The comments above, that each new heterocycle is a new story due to the subtle effects of heteroatoms on aromaticity and delocalization, also apply to pK_a values, and so the more novel the scaffold, the more difficulty a purely empirical scheme is likely to encounter.

Nonetheless, the pK_a method has some significant, inherent advantages. By default, Epik performs state adjustment for acids with lower pK_a than 9 and bases with higher pK_a than 5, based primarily on experimental pK_a measurements and the enhanced H–T rules. This takes care of a great many rapid tautomeric reactions and provides a general mechanism for generating tautomers that is complementary to the tautomerizer code. Consider for example, the tautomerism of the conjugate base monoanion of a diacid like a substituted salicylic acid. There is no need to study the pK_T of the phenolate vs. the carboxylate anion from first principles, as predicting these pK_a values is something which pK_a-based structural adjustment excels at.

The layer of redundancy, that is producing tautomers by two independent types of structural adjustment, can serve to greatly expand coverage of drug-like chemistry and improve robustness, as when a reaction has not been specifically parameterized by one method to high accuracy, there's a good chance the other can often compensate to produce a sensible output ensemble. The workflow we have settled upon is to iterate over cycles of direct tautomerization and pK_a-based structural adjustment, keeping track of energy costs and bonuses for each transformation along the way, for inclusion in the final normalized state penalty. To our knowledge, it is this combination of empirical and tabulated first principles data, and the redundant mechanisms for handling tautomerism, that sets our approach apart from other software in this arena.

In this way, the scientific aspects of the tautomer problem should in principle become quite tractable for any drug-like molecule, at least to within a couple of kcal/mol on average, which is commensurate with other uncertainties in the modelling component of molecular design, not to mention the uncertainties in many assays. The problem is thus transformed from a scientific one into an engineering one, with the quality and scope of the database and the pattern recognition becoming at least as important as the basic algorithms. In our own applications, we are devoting considerable resources to continually expanding and updating our tautomer libraries and pK_a coverage, in

concert with feedback from internal and external users of the software. We are focusing on quality and generality, as distinct from sheer quantity, in our attempts to reach full coverage of drug-like chemistry, as simply adding highly specific patterns to tightly reproduce results in standard benchmark compounds is more likely to lead to overfitting than an increase in predictivity or general utility. In our experience, without careful investigation, inaccurate experimental pK_a values may be used or the pK_a value may be ascribed to the wrong proton or functional group. Building a reliable software application for tautomer prediction depends not only on starting from a sound theoretical and empirical basis, but making a major investment in collecting, generating, understanding and entering data, successive rounds of automation, conducting validation, and building a solid user base whose experience and interest in particular systems can be fed back into the expansion of the database.

Do low population tautomers need to be considered in high throughput virtual screening?

Tautomers have not received much considered attention in protein–ligand binding studies until fairly recently. As such there are quite diverse ideas as to what tautomers need to be considered as noted in a recent review [40]. However, on simple thermodynamic grounds, we would expect high energy tautomers to make very poor binders. When considering which states to screen, if a tautomer is present at equilibrium in only a small fraction in aqueous solution (we typically use 1% as a limit) it can be argued that for most purposes there is no need to consider the binding of that tautomer to a receptor. Thermodynamically, if only one tautomer can bind and that tautomer is rare, then the binding affinity will suffer accordingly, compared with an analogue presenting a similar pharmacophore without such an energy penalty. Stated another way, a population of 1% roughly corresponds to having a free energy of cost of 2.76 kcal/mol an energy shift that is often enough to dramatically downgrade the ranking of a candidate ligand in virtual screening studies. While tautomer preferences can change in different environments, and indeed differential desolvation of tautomers upon binding complicates the picture somewhat, as the lower polarity receptor environment can partially reverse the aqueous stabilization of charge separation (as in certain mesionic tautomers and to a lesser extent aromatic lactams) such effects are routinely implicitly included in docking calculations or free energy of binding calculations. For instance most methods for studying protein–ligand binding energies (for example MM-GBSA [41] or MM-PBSA [42]) include surface area and solvation terms that implicitly handle the partial desolvation of tautomers. Thus the tautomer ratio in an arbitrary medium or receptor can be defined using an accurate tautomer ratio in a reference medium (preferably water for biological applications) plus the difference in the free energy of transfer from the reference medium to the receptor. At any rate, such differential desolvation effects are likely to be small (on the order 1 kcal/mol for most tautomers in a mixed water/protein environment, judging by the typical magnitudes of the change in pK_T going from water to a highly non-polar environment which are rarely more than 2–3 kcal/mol [13]) since the receptor is generally more water-like than gas phase-like. With rare exceptions (such as 1jvp.pdb [43]) the difference in the interaction energy dominates (the receptor picks one tautomer from the easily energetically accessible ones in solution). Therefore, to a first approximation for screening purposes, the relative tautomeric energy in solution can be used directly as a penalty when scoring protein–ligand complexes [44]. In any case, there is a balance to be struck since including more tautomers with a lower mole fraction in water (i.e., higher tautomer energies) becomes progressively less likely to yield the most favorable bound state for a given ligand, while it increases processing time and storage requirements. In our experience a mole fraction of 1% is adequate for most virtual screening studies, assuming the energy estimates are reasonably accurate. Virtually screening or synthesizing and testing a heterocycle believed to require a rare tautomer in order to bind is likely to be less productive than redesigning the scaffold to remove the tautomer penalty, or screening alternative scaffolds.

We have analysed thousands of ligand–protein complexes from the PDB, and as of yet we have not found examples where it was necessary to invoke a high energy tautomer in order to understand the binding equilibrium. From time to time, a rare tautomer will be claimed for a particular experimental protein–ligand complex at equilibrium. We will discuss two such examples below and show that the experimental observations can be better explained using low energy tautomers or states. In some cases, it has been necessary to go beyond the deposited coordinates and re-examine the ligand synthesis, or refit the protein structure to the electron density to find the low-energy tautomer consistent with the experimental data. In theory careful use of any crystallographic software plus an understanding of the energetics can reveal the best tautomer, but our preference is to use Prime-X [45], in which hydrogens are explicitly included during forcefield-based refinement, allowing models with alternative hydrogen bonding networks and tautomers to be explicitly considered and compared.

In the first example, the barbiturate bound to MMP8, PDB code 1jj9 [46], has been claimed to be in a neutral

hydroxy form which has an energy 16.9 kcal/mol higher than the lowest energy tautomer according to M06-2X calculations. Two important reviews [2, 39] have cited this structure as an example of why high energy tautomers need to be considered. A better explanation is that the metal-ligating barbiturate group is anionic when bound (Fig. 4)—a state which is also a favorable state in water at pH 7. Glu198 (pK_a roughly 4.4) may be partially or fully neutralized when barbiturate is bound, to complete the hydrogen bond network, whereas protonating the oxygen (pK_a likely less than 0) is unfavorable. It is not immediately clear whether the mildly basic piperidine nitrogen is (partly) protonated so the zwitterion may be present in appreciable amounts in solution and perhaps favored in the receptor, a less critical question that could be addressed by a QM/MM study. The barbiturate anion explanation is consistent with other anionic MMP ligands such as hydroxamates (e.g., 1mmb.pdb) and phosphonates (e.g., 1i73.pdb). Careful receptor and ligand preparation including metal-preferring states for metalloproteins such as MMP8 (as implemented for NH acids among other groups in Epik) allows this kind of ligand to be modelled and docked in a way that is consistent with the experiment, without resorting to high energy tautomers.

In the second example, Chlorthalidone bound to Carbonic Anhydrase II (PDB code 3f4x), a lactim rather than lactam tautomer is claimed for the 3-hydroxyisooxindole cyclic hemiamidal [47], and has also been reviewed in the context of the need for high energy tautomers [39]. The lactim tautomer has a very high energy (15.8 kcal/mol) relative to the lowest energy tautomer in water according to M06-2X/6-311+G(d,p) [PB-SCRF] as shown on the left hand side of Fig. 5. Ring-chain tautomerism can be neglected, as the ring-opened state has an even higher energy. Contrary to the original authors' interpretation, it is not at all clear from the X-ray structure that the receptor requires the lactim tautomer, because an alternative low energy H-bonded network with chi-flipped Asn67 and the much more likely lactam form bound can be constructed to fit the X-ray diffraction data (right hand side of Fig. 5). This model can be conveniently generated using the all-atom forcefield-assisted density fitting in Prime-X [44] that takes account of H-bonding, but for which other crystallography packages plus forcefield approaches or QM//MM calculations could be used. Unfortunately the electron density is somewhat weaker in the neighborhood of the isoindole than the surrounding protein and phenylsulfonamide, making it difficult to draw a definitive conclusion about the ligand state and conformation based on the X-ray diffraction experiment alone; understanding of the chemistry of the ligand is needed. Given that this ligand is weaker at CA-II than some of the other carbonic anhydrase isoforms, it would be interesting to compare the H-bonding networks in this region across isoforms; a receptor preference for the lactim H-bonding pattern would be expected to drop the potency. Again, careful ligand and receptor preparation can obviate the need for considering a rare tautomer, and to the extent that a rare tautomer may be involved, it is likely to be detrimental to activity, and thus best avoided in the course of routine design and screening.

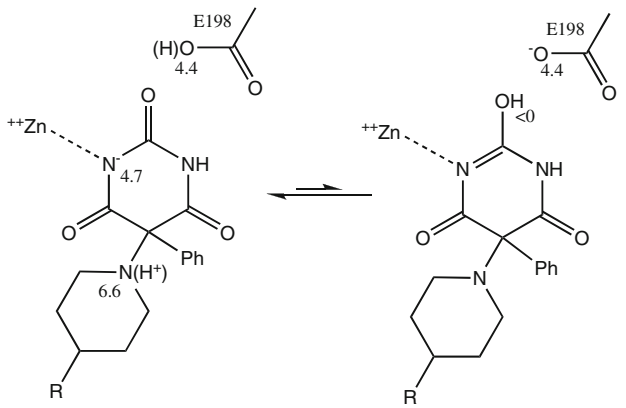


Fig. 4 Proposed ligand structures for 1jj9. *Right* a structure presented by Brandstetter et al. [46] and used in Posposil et al. [2, 40] to justify the need to consider high energy tautomers in protein–ligand complexes. *Left* a more likely low energy form in which the barbiturate moiety binds to the zinc in MMP8 as an anion and not as a high energy lactim tautomer. pK_a values are provided for key atoms. Given these pK_a values if a hydrogen bond bridge is present between the ligand O and E198 then the hydrogen would reside on E198 the vast majority of the time and thus should be considered part of the receptor

At what level should a tautomer be considered rare enough to be discarded as uninteresting for screening or design? As shown in Fig. 6, the structure of neopterin bound to Ricin A (1br5 [48]) provides an illustration of a borderline case. The receptor unequivocally prefers the 1H tautomer, for which M06-2X accords an energy penalty of 2.1 kcal/mol in water, which is within our recommended default cutoff of 1% (which corresponds to a penalty of 2.76 kcal/mol). Thus, by default in Schrödinger's virtual screening workflow, we would screen this tautomer, and find the correct binding mode, but substantially penalize its score. Note that the K_i of this compound is reported to be >2 mM [47] presumably rendering it of marginal interest, along with other similarly weak pterins at this site. Perhaps an early design goal for this site would be to engineer out the poor tautomer profile from the scaffold.

If rare tautomers are not needed to explain ligand binding in experimental complexes, the question arises as to whether it is more effective to virtually screen only a single tautomer, or an ensemble. Kalliokoski et al. [49] argue that similar enrichments in ligand-based screens can be obtained for a reduction in computer time, by using only a single lowest energy tautomer. Given that electrostatic

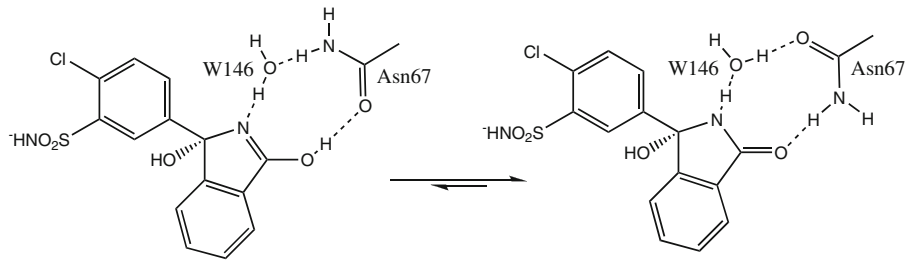


Fig. 5 Alternate tautomeric forms for the binding mode of Chlorthalidone in 3f4x. The high energy tautomer (*left*) has been used to explain binding [47] and cited as an example of the need to consider high energy tautomers in general [40]. However, flipping the assignment of the terminal amide of Asn67 in 3f4x allows carbonic anhydrase II to bind the much lower energy lactam tautomer of Chlorthalidone (*right*)

complementarity (including hydrogen bonding patterns) is important for many high affinity protein–ligand complexes, we prefer to include a small range of accessible tautomers (and protonation states) to maximize the chances of finding the one with the best complementarity, particularly in structure-based screening. For example, in finely balanced cases such as pyrazoles and imidazoles, selecting only one tautomer more or less at random is difficult to justify. So the goal of including all tautomers with a significant mole fraction in solution still seems like a reasonable course. In our internal validation [43] (manuscript in preparation) we find that adding the state penalty (including tautomer penalty) to a scoring function significantly improves enrichment. This approach combined with the inclusion of desolvation terms in the scoring function suggests that multiple protonation states including tautomeric variations can be properly included in virtual screening without a large increase in cpu time due to exhaustive examination of all enumeratable forms. Both of these factors are taken into account automatically in our virtual screening workflow (Epik + Glide [50]) for which our default is to screen an ensemble of states present down to a mole fraction of 1% in water, regardless of whether they are tautomers or different

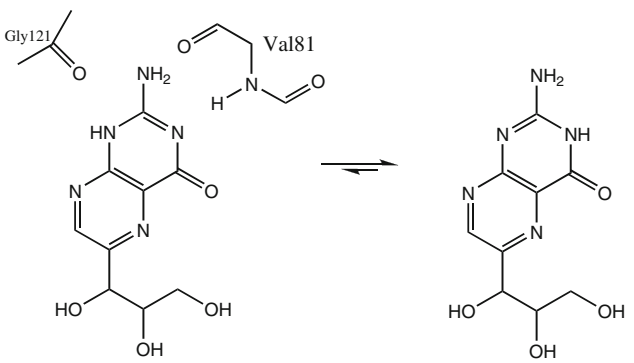


Fig. 6 Neopterin in Ricin A (1br5) [48]. The pterin prefers the 3H tautomer in water (*right*) and pays a modest penalty (ca. 2 kcal/mol) to bind in the less favourable 1H tautomer (*left*), contributing to its very weak inhibition

charge states. Our recommendations could easily be implemented using other energy-aware tautomer generation tools and docking codes, but the key goal is to try to mimic the experimental conditions, i.e., the states present in solution, and incorporate their energies. We note that at the 1% level for most ligands, only one or a very few low energy states need to be considered, so handling tautomer structures and energies in a physically realistic way improves results without greatly increasing cpu time—the expansion rate is typically less than double for large databases of commercially available compounds.

The value of 1% mole fraction for screening is of course arbitrary. Increasing this can save cpu time in time critical projects while potentially eliminating some likely weaker binders from consideration. Given achieving perfect coverage and accuracy will be a challenge for some time to come for all fast ligand preparation codes, a smaller cutoff than 1% might provide some insurance for catching states whose aqueous stability is underestimated. But so long as the tautomer prediction is reasonably accurate, increased cpu time and presumably more false positives make it hard to justify reducing the default cutoff for routine screening in our opinion. Additional flexibility is provided in Epik, since one can adjust the target pH, the pH window, and minimum tautomer probability independently, as well as choose whether to add and flag extra states that may be needed for screening at metalloproteins, where strong metal ligation is better able to remove certain protons than water (such states typically display fewer tautomers, as in 1jj9 above). No matter what tautomer preparation code, cutoffs and screening protocols are used, in our experience using the prevalence in solution to decide whether a state should be screened, and adding a term to the scoring function or free energy estimation relating to that prevalence, produces the best results.

Naturally, outside of virtual screening, other imperatives may apply. For research into enzyme reaction mechanisms, rare tautomers may need to be considered since they can certainly be formed or required along a reaction path and

sometimes significantly influence the reaction rate. For example, the production of a high-energy tautomer could decrease the affinity of the enzyme for the product, increasing the rate of forward reaction as the product is ejected and the enzyme active state regenerated. Likewise, quite rare tautomers could have an effect on the properties of a nucleic acid, given the large number of copies of bases present [51], and exotic states have also been implicated in the context of radiation-induced damage to DNA [52]. For studying such species and mechanisms, more accurate and time consuming treatments involving large-scale quantum mechanics (QM) or mixed quantum mechanics/molecular mechanics (QM–MM) calculations are appropriate for use on specific substrates. However these are quite specialized applications which are distinct from virtual screening for good binders to a receptor, and conversely the computational requirements of such enzyme or DNA mechanistic research are incompatible with the large scale on which virtual screening is typically conducted.

Sometimes molecular structures are recorded in source datasets in rare tautomeric forms having been drawn in an unlikely state due to lack of experience, by mistake, according to tradition, lack of experimental data, or by an automated tool. In our opinion the main reason for supporting rare tautomers in fast tautomerization tools is not that they should be included in the prepared output ensemble for routine use, but rather to enable the software tool to recognize them and automatically transform them into lower energy tautomers. For example, folic acid is frequently depicted as hydroxy tautomer (see for example http://images.google.com/images?q=folic+acid) which has quite a high energy according to M06-2X calculations (7.7 kcal/mol) and thus is expected to represent an extremely small mole fraction in solution. It is clear from numerous X-ray structures that the lower energy keto forms of folates and antifolates bind to thymidylate synthase and dihydrofolate reductase. Thus, a comprehensive ligand preparation tool needs to be able to convert arbitrary high energy input tautomers into low energy output tautomers, while keeping track of the energies of the contributors to the aqueous ensemble.

Cheminformatics considerations

Some cheminformatics solutions place emphasis on representing each registered compound as a single cannonical tautomer, regardless of its energy. While this makes identifying duplicate input structures drawn in different states simpler, only using one tautomer means that the properties of each of the tautomers present in an energetically reasonable ensemble need to be mapped back into the cannonical database entry. It also requires a robust and comprehensive canonicalization code, which is a more difficult problem than is generally appreciated. For example, conversion between neutral and mesionic tautomers is not amenable to schemes involving just the movement of protons and double bonds. Another issue is that it is not clear that it is helpful to convert all potential cases of a given type of tautomerism to one representative form, when the equilibrium may lie heavily to one side or the other depending on the context, for example aliphatic imines versus conjugation-stabilized enamines. Another pragmatic consideration is whether the input structure (which can typically originate with organic chemists) can be recovered when substantial changes are introduced to canonicalize a structure.

Our preference involves tracking each of the output low energy states using a chemoinformatics tool (in our case with the Canvas cheminformatics package [53]), together with their State Penalties, for each input structure, thus avoiding canonicalization and de-canonicalization. This does not preclude the identification of duplicate chemical entities arising from different input structures, which can still be identified from sharing an output fingerprint in common. Carrying small ensembles of aqueous states in the database instead of single idealized representations means that structures are always ready to screen, without having to go back and forth from a canonicalized structure. This is even more pertinent in the case of pharmacophore-based screening or substructure searching, where speed is particularly desirable, and which are believed to be quite sensitive to the tautomeric state used [8]. By using the workflow we recommend here, explicit queries of ensembles will only return those relevant hits capable of actually achieving a given state without a substantial energy penalty.

Maintaining ensembles also provides a route to physically appealing ways of dealing with bulk properties, such as logP: rather than relying on a single value for an arbitrary cannonical structure, it could be estimated as a weighted mean of the partial logPs of a relevant distribution of states. This is an ongoing area of research, and we encourage those developing tools for physical properties to consider the utilization of low energy ensembles instead of single formal structures.

Concluding remarks

We have outlined the themes and thinking about tautomers that have gone in to our own work and which shape the ongoing development of Epik and LigPrep's tautomerizer tool, components of Schrödinger's ligand preparation suite. Though many of the individual components are not novel, it is the way in which first principles (DFT) results are

complementary to and can be combined with empirical pK_a measurements and estimation that sets this approach apart. Key insights include that receptors only bind tautomers with fairly low energy in water, and that best results can be obtained in virtual screening by considering small ensembles of reasonably low energy protonation/tautomeric states, and including the relative free energies of those states in the scoring protocol. While the approach is neither perfect nor fully mature, the errors are acceptably small for the great many systems which have already been parameterized, and work is ongoing, based upon feedback from our user base, to continually expand coverage for drug-like molecules.

References

1. Comer J, Tam K (2001) In: Testa B, van de Waterbeemd H, Folkers G, Guy R (eds) Pharmacokinetic optimization in drug research: biological, physicochemical, and computational strategies. Wiley, Weinheim, pp 275–304
2. Pospisil P, Ballmer P, Scapozza L, Folkers G (2003) Tautomerism in computer-aided drug design. J Recept Signal Transduct Res 23(4):361–371
3. Rester U (2008) From virtuality to reality—virtual screening in lead discovery and lead optimization: a medicinal chemistry perspective. Curr Opin Drug Discov Dev 11(4):559–568
4. Shelley JC, Cholleti A, Frye LL, Greenwood JR, Timlin MR, Uchimaya M (2007) Epik: a software program for pK(a) prediction and protonation state generation for drug-like molecules. J Comput Aided Mol Des 21(12):681–691
5. Epik, v2.0 (2009) Schrödinger, Inc., New York
6. LigPrep, v2.3 (2009) Schrödinger, Inc., New York
7. Milletti F, Storchi L, Sforna G, Cross S, Cruciani G (2009) Tautomer enumeration and stability prediction for virtual screening on large chemical databases. J Chem Inf Model 49(1): 68–75
8. Oellien F, Cramer J, Beyer C, Ihlenfeldt WD, Selzer PM (2006) The impact of tautomer forms on pharmacophore-based virtual screening. J Chem Inf Model 46(6):2342–2354
9. Haranczyk M, Gutowski M (2007) Quantum mechanical energy-based screening of combinatorially generated library of tautomers. TauTGen: a tautomer generator program. J Chem Inf Model 47(2):686–694
10. Pisklak M, Maciejewska D, Herold F, Wawer I (2003) Solid state structure of coumarin anticoagulants: warfarin and sintrom. ^{13}C CPMAS NMR and GIAO DFT calculations. J Mol Struct 649(1–2): 169–176
11. Zhang J, Yang PL, Gray NS (2009) Targeting cancer with small molecule kinase inhibitors. Nat Rev Cancer 9(1):28–39
12. Pauling L (1960) The nature of the chemical bond, 3rd edn. Cornell University Press, Ithaca
13. Elguero J, Katritzky AR, Marzin C, Linda P (1976) The tautomerism of heterocycles. Academic Press, New York
14. Rauhut G (2002) Recent advances in computing heteroatom-rich five and six-membered ring systems. Adv Heterocycl Chem 81: 2–85
15. Bryantsev VS, Diallo MS, van Duin ACT, Goddard III WA (2009) Evaluation of B3LYP, X3LYP, and M06-class density functionals for predicting the binding energies of neutral, protonated, and deprotonated water clusters. J Chem Theory Comput 5(4):1016–1026
16. Harding ME, Metzroth T, Gauss J, Auer AA (2008) Parallel calculation of CCSD and CCSD(T) analytic first and second derivatives. J Chem Theory Comput 4(1):64–74
17. Fabian WMF (1991) Tautomeric equilibria of heterocyclic molecules. A test of the semiempirical AM1 and MNDO-PM3 methods. J Comput Chem 12(1):17–35
18. Stewart JJP (2007) Optimization of parameters for semiempirical methods V: modification of NDDO approximations and application to 70 elements. J Mol Model 13(12):1173–1213
19. Tirado-Rives J, Jorgensen WL (2008) Performance of B3LYP density functional methods for a large set of organic molecules. J Chem Theory Comput 4(2):297–306
20. Cramer CJ, Truhlar DG (1993) Correlation and solvation effects on heterocyclic equilibria in aqueous solution. J Am Chem Soc 115(19):8810–8817
21. Zhao Y, Truhlar DG (2007) The M06 suite of density functionals for main group thermochemistry, kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06 functionals and twelve other functionals. Theor Chem Acc 120:215–241
22. Jaguar, v7.6 (2009) Schrödinger, Inc., New York
23. Zhao Y, Truhlar DG (2008) Exploring the limit of accuracy of the global hybrid meta density functional for main-group thermochemistry, kinetics, and noncovalent interactions. J Chem Theory Comput 4(11):1849–1868
24. Frydenvang K, Greenwood JR, Vogensen SB, Brehm L (2002) Structural features of ATPA and Thio-ATPA—potent and selective GluR5 receptor agonists. Crystal structure determinations and quantum chemical calculations. Struct Chem 13(5–6): 479–490
25. Ahlquist MSG, Kozuch S, Shaik S, Tanner DA, Norrby P-O (2006) On the performance of continuum solvation models for the solvation energy of small anions. Organometallics 25(1): 45–47
26. Nielsen PA, Jaroszewski JW, Norrby PO, Liljefors T (2001) An NMR and ab initio quantum chemical study of acid-base equilibria for conformationally constrained acidic alpha-amino acids in aqueous solution. J Am Chem Soc 123(9):2003–2006
27. Nielsen PA, Jaroszewski JW, Norrby PO, Liljefors T (2002) Conformational analysis of cyclic acidic alpha-amino acids in aqueous solution—an evaluation of different continuum hydration models. Unpublished data and PhD thesis, University of Copenhagen
28. Marenich AV, Cramer CJ, Truhlar DG (2009) Performance of SM6, SM8, and SMD on the SAMPL1 test set for the prediction of small-molecule solvation free energies. J Phys Chem B 113(14):4538–4543
29. Katritzki AR, Lagowski J (1963) Adv Heterocycl Chem 1:339
30. Katritzky AR, Øksne S, Boulton AJ (1962) The tautomerism of heteroaromatic compounds with five-membered rings—III: further isoxazol-5-ones. Tetrahedron 18(6):777–790
31. Le HT, Lamb JG, Franklin MR (1998) Drug metabolizing enzyme induction by benzoquinolines, acridine, and quinacrine; tricyclic aromatic molecules containing a single heterocyclic nitrogen. J Biochem Toxicol 11(6):297–303
32. Harvey RG (1991) Aromatic hydrocarbons: chemistry and carcinogenicity. Cambridge University Press, Cambridge
33. Perrin DD, Dempsy B, Sergeant EP (1981) pKa prediction for organic acids and bases. Chapman and Hall, London
34. Gordon A, Katritzky AR, Roy SK (1968) Tautomeric pyridines. Part X. Effects of substituents on pyridone–hydroxypyridine equilibria and pyridone basicity. J Chem Soc B 1968:556–561
35. ACD/PhysChem Suite, v12.0 (2009) Advanced Chemistry Development, Inc., Toronto
36. Pallas pKalc Net., v2.0 (2009) Compudrug International Inc., Sedona

37. Hilal S, Karickhoff SW, Carreira LA (1995) A rigorous test for SPARC's chemical reactivity models: estimation of more than 4300 ionization pKa's. Quant Struct Act Relatsh 14:348–355
38. Liao C, Nicklaus MC (2009) Comparison of nine programs predicting pK(a) values of pharmaceutical substances. J Chem Inf Model 49(12):2801–2812
39. Klicic J, Friesner RA, Liu S-Y, Guida WC (2002) Accurate prediction of acidity constants in aqueous solution via density functional theory and self-consistent reaction field methods. J Phys Chem A 106(7):1327–1335
40. Martin YC (2009) Let's not forget tautomers. J Comput-Aided Mol Des 23(10):673–704
41. Guimaraes CR, Cardozo M (2008) MM-GB/SA rescoring of docking poses in structure-based lead optimization. J Chem Inf Model 48(5):958–970
42. Fogolari F, Brigo A, Molinari H (2003) Protocol for MM/PBSA molecular dynamics simulations of proteins. Biophys J 85(1): 159–166
43. Furet P, Meyer T, Strauss A, Raccuglia S, Rondeau JM (2002) Structure-based design and protein X-ray analysis of a protein kinase inhibitor. Bioorg Med Chem Lett 12(2):221–224
44. Repasky M (2009) Enhancing Glide Enrichment Using Epik Ionization and Tautomeric State Penalties. Schrodinger Quaterly Newsletter, August
45. Prime-X, v1.6 (2009) Schrödinger, Inc., New York
46. Brandstetter H, Grams F, Glitz D, Lang A, Huber R, Bode W, Krell HW, Engh RA (2001) The 1.8-A crystal structure of a matrix metalloproteinase 8-barbiturate inhibitor complex reveals a previously unobserved mechanism for collagenase substrate recognition. J Biol Chem 276(20):17405–17412
47. Temperini C, Cecchi A, Scozzafava A, Supuran CT (2009) Carbonic anhydrase inhibitors. Comparison of chlorthalidone and indapamide X-ray crystal structures in adducts with isozyme II: when three water molecules and the keto-enol tautomerism make the difference. J Med Chem 52(2):322–328
48. Yan X, Hollis T, Svinth M, Day P, Monzingo AF, Milne GW, Robertus JD (1997) Structure-based identification of a ricin inhibitor. J Mol Biol 266(5):1043–1049
49. Kalliokoski T, Salo HS, Lahtela-Kakkonen M, Poso A (2009) The effect of ligand-based tautomer and protomer prediction on structure-based virtual screening. J Chem Inf Model 49(12): 2742–2748
50. Glide, v5.5 (2009) Schrödinger, Inc., New York
51. Rejnek J, Hobza P (2007) Hydrogen-bonded nucleic acid base pairs containing unusual base tautomers: complete basis set calculations at the MP2 and CCSD(T) levels. J Phys Chem B 111(3):641–645
52. Shukla M, Leszcynski J (2008) Radiation induced molecular phenomena in nucleic acids: a comprehensive theoretical and experimental analysis. Springer, Berlin
53. Canvas, v1.2 (2009) Schrödinger, Inc., New York