Tautomer Preference in PDB Complexes and its Impact on Structure-Based Drug Discovery

Francesca Milletti[†] and Anna Vulpetti*

CADD, Global Discovery Chemistry, Novartis Institutes for Biomedical Research, CH4002 Basel, Switzerland

Received December 23, 2009

Tautomer enrichment is a key step of ligand preparation prior to virtual screening. In this paper, we have investigated how tautomer preference in various media (water, gas phase, and crystal) compares to tautomer preference at the active site of the protein by analyzing the different possible H-bonding contacts for a set of 13 tautomeric structures. In addition, we have explored the impact of four different protocols for the enumeration of tautomers in virtual screening by using Flap, Glide, and Gold as docking tools on seven targets of the DUD data set. Excluding targets in which the binding does not involve tautomeric atoms (HSP90, p38, and VEGFR2), we found that the average receiver operating characteristic curve enrichment at 10% was 0.25 (Gold), 0.24 (Glide), and 0.50 (Flap) by considering only tautomers predicted to be unstable in water versus 0.41 (Gold), 0.56 (Glide), 0.51 (Flap) by limiting the enumeration process only to the predicted most stable tautomer. The inclusion of all tautomers (stable and unstable) yielded slightly poorer results than considering only the most stable form in water.

INTRODUCTION

Tautomerism, which is a phenomenon whereby a molecule interconverts to other isomers that differ in the position of a double bond and one atom (typically a hydrogen atom),[1] is of special interest in studies of protein−ligand interactions. Since the displacement of hydrogen may convert an acceptor into a donor, a tautomeric rearrangement changes the interaction landscape of a protein−ligand complex. In recent years, tautomerism has received increasing attention within the molecular modeling community, and a number of reviews[2,3] and methods about tautomer enumeration have been published.[4,5]

Tautomerism is especially relevant if one takes into account that in vendor databases ∼30% of the compounds are potentially tautomeric and that ∼10% are represented in a form that is not the most stable in water.[4] However, the exact amount of tautomeric compounds in a chemical library is very difficult to estimate, and different figures have been reported. For example Trepalin et al. found that 0.5% of compounds from vendor screening libraries are tautomeric,[6] whereas Shoichet et al. reported that about 10% of compounds in the ZINC database can be represented in more than one low-energy tautomer.[7]

These differences are linked to the fact that tautomerism is a fuzzy notion: if the energy difference between two tautomers (ΔG) is very high, only one tautomeric form exists, but the maximum ΔG to consider two forms as tautomers is not well-defined. Some authors have suggested a maximum ΔG of 5 kcal mol^{-1},[8] whereas others a value of 28 kcal mol^{-1}.[9] The latter value may be appropriate when also very small concentrations of a tautomer are relevant (e.g., to carry on a chemical reaction). However, if the interest is solely in the physical chemical properties of a tautomeric mixture, the observed properties are mainly those of the major species, and therefore, it is preferable to consider a lower threshold.

The ΔG between two tautomers may change significantly in different media depending on the difference in polarity between the tautomers.[9] For example 2-hydroxypyridine predominates in the hydroxy form in the gas phase and in apolar solvents but in the oxo form in water and in the solid, and this is due to the fact that the oxo tautomer is more polar than the hydroxyl tautomer.[10] The two tautomers of indazole, 1H and 2H, do not differ much in terms of polarity, and therefore, the 1H form is predominant in media of different polarities.[10] For these reasons, the environment around a ligand bound to a protein pocket may or may not cause reversal of tautomer stability compared to other media. However, it would be beneficial to understand possible links between tautomer stability in ligand−protein complexes and in other media, since this information could help in determining the tautomers to use in virtual screening applications.

This is especially important in the light of recent studies which recommend using relevant tautomers only, rather than many possible tautomers. For example, Kalliokoski et al. compared the results of a structure-based virtual screening on different targets from the Database of Useful Decoy (DUD)[11] data set using a tautomer-enriched library and a library containing a single stable tautomer (and protomer) predicted by the program MoKa.[12] The results of this study showed that the use of the predicted more stable tautomer yields similar results to the use of all possible enumerated tautomers, and therefore, it has the advantage of reducing the computational time. Other authors have suggested that the use of all possible tautomers may cause an increase in false positives.[13] For example, in a docking study on the DUD data set, Clark et al.[14] observed that some of the decoys that scored well were in an unlikely tautomeric form.

* Corresponding author. E-mail: anna.vulpetti@novartis.com.

[†] Current address: Hoffmann-La Roche, 340 Kingsland Street, Nutley, New Jersey, 07110.

10.1021/ci900501c © 2010 American Chemical Society
Published on Web 06/01/2010

Other studies, however, have proposed tautomer enrichment as the recommended protocol for virtual screening.[15,16] Oellien at al. studied the effect of tautomer enrichment on pharmacophore-based virtual screening against cyclin-dependent kinase (CDK) and matrix metallopeptidase (MMP8) and found that in the case of tautomer-enriched databases both the discrimination between hits and nonhits and the hit retrieval were higher.[17] On the other hand, other authors have found that the choice of the tautomer has a small impact on the results. For example, Todorov et al. have used dyhydrofolate reductase, transketolase, and α-trichosanthin to investigate to what extent docking allows the retrieval of the correct pose starting from different tautomeric states and found that in most cases the native binding mode could be recovered without much loss of accuracy regardless of the tautomeric state used.[18]

These data show that there is still controversy on the optimal way of using tautomers in virtual screening, but it should be noted that the results obtained were also linked to the relative importance of tautomerism in the different targets used. In this paper, we have addressed two major questions to understand the problem of tautomerism in structure-based drug discovery: (a) Which tautomeric form binds preferentially in the binding site, among common moieties, and how its stability in water reflects its stability in the binding site?; and (b) What is the impact of tautomer enrichment in enhancing the recovery of actives in different targets, and what is the sensitivity of different docking tools in terms of tautomer enrichment?

The paper is organized as follows: first, we have reported experimental data on tautomer stability in different media for 13 common tautomeric compounds and have used different tautomers of these compounds as queries for a substructure search on the Cambridge Structural Database (CSD).[19] The CSD provides a very good reference for tautomer preference of small molecules in a crystal, since hydrogen (H)-atoms are visible, and it allowed us to find how often a compound exists in different tautomeric forms by taking into account also the effect of additional substituents. However, it is important to remember that the prevalent tautomeric form in the crystal structure may be influenced by crystal condition and crystal packing effects and not only by the relative stability of the two tautomers. Then, we have repeated the search on the Protein Data Bank (PDB)[20] database of protein−ligand complexes and have estimated the preferred tautomer in the binding site by analyzing the H-bonding contacts for each tautomer.

In the second part of the paper, we have studied how different docking tools, Flap,[21,22] Glide,[23,24] and Gold,[25−27] perform in increasing the recovery of actives depending on the choice of the tautomers using seven targets of the DUD database. The program MoKa was applied to select the tautomeric form(s) to include for each molecule: (a) only tautomers predicted to have an abundance of 0% in water, (b) only the predicted most stable form, (c) all forms predicted to have an abundance above 5% in water, and (d) all possible tautomeric forms. We found that significantly poorer results were obtained if only tautomers at 0% in water were used, but the other protocols performed similarly.

METHODS

Tautomer Preference in the Gas Phase and Water, Crystal, and PDB Complexes. We selected 13 representative tautomeric compounds (Table 1) sufficiently diverse and common in virtual screening libraries and collected experimental data on their stability in water and in the solid and gas phases (when available). The stability reported in Table 1 in different media refers to the pair of tautomers shown. For some compounds there are more than two tautomers, but for simplicity we have reported in Table 1 only the two lowest energy forms in water.

Tautomers in Table 1 were used as queries for a substructure search in the CSD database (Version 5.30) with the CCDC ConQuest 1.11 program.[28] The search was run by using R factor ≤ 0.10, "three-dimensional (3D) coordinates determined", "not disordered", "no ions", "no errors", "not polymeric", and "only organic" as input settings. For each tautomer of the 13 compounds we annotated the number of hits retrieved from the CSD database.

In addition to this, we extended the substructure search to ligands bound to proteins by using the database of PDB ligands available from Ligand Expo,[29] which contains 10 025 compounds. However, while structures in the CSD include the position of H-atoms, PDB structures are generally at a much lower resolution, and therefore, H-atoms are not reported. However, by analyzing H-bonding interactions of alternative tautomers, it is possible to determine whether a form interacts more favorably than another. This analysis cannot be always conclusive because either the tautomeric region does not interact with any residue of the protein or the interaction is mediated by mixed H-bond donor/acceptors, but it may still provide useful information to understand tautomer preference at the binding site.

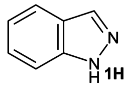
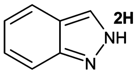
The compounds in PDB complexes available from Ligand Expo were reduced to 9879 after removal of all ions and ligands containing metals, which may affect the tautomeric state of the ligand itself, and then they were processed by TauThor, which is part of the MoKa[30,31] package, to enumerate alternative tautomers. For each pair of tautomers reported in Table 1, a substructure search was run on the Ligand Expo database processed by TauThor, and the most probable tautomeric form at the binding site was assessed by analyzing the H-bonding contacts between the protein and the bound ligand considered in one of the two alternative tautomers reported in Table 1. The polar atoms of the ligand involved in tautomerism can often be involved in contacts with polar atoms of the protein cavity, and thus the number of these polar interactions was used to deduce the preferred tautomeric species in the ligand−protein crystal.

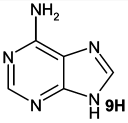
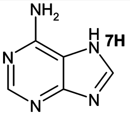
To analyze the H-bonding profile of all tautomers in their crystallographic pose, we selected all the atoms of the protein at less than 3.2 Å from each of the tautomeric atoms of the compound and scored the interaction as favorable ("+1″, for a donor−acceptor pair) or unfavorable ("−1" for a donor−donor or acceptor−acceptor pair). Eventually, for each tautomer a cumulative score was calculated by summing all the interactions obtained. In this analysis, the term "donor" refers to a donor-only group (NH of amides, Trp, Lys, and Arg) and "acceptor" to an acceptor-only group (O atom of carbonyl and carboxylate), whereas mixed donor−acceptor groups (i.e., waters and hydroxyl, either on the ligand or the protein) were disregarded. The list of the PDB codes and corresponding number of H-bonding contacts of the two tautomeric forms are provided in the Supporting Information.

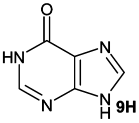
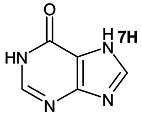
An additional issue in determining the acceptor, donor, or mixed state is the ionization form of a chemical group.

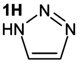
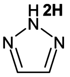
Table 1. Results from the Substructure Search of Alternative Tautomers (M and m) of Common Moieties of Pharmaceutical Interest in the CSD and PDB Databases[a]

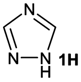
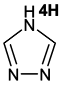
a)

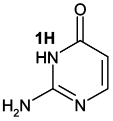
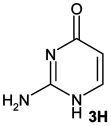
Name	ΔG (kcal mol^{-1}) / medium / relevant form(s)	Major form in Water (M)			Minor form in Water (m)			PDB with identical HB score
		Structure/Name	CSD	PDB	Structure/Name	CSD	PDB	
1. **Indazole**	2.3/water/M; ND/gas (calc.)/M.		45	26		0	0	7

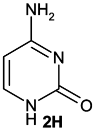
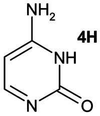
2.
Adenine
	0.8/water/M;
	12/gas (calc.)/M;
	0.8/DMSO/M.
			17	10		3	23	6

3.
Hypoxanthine
	0/water/M and m;
	ND/gas (calc.)/M;
	ND/solid/M.
			2	27		2	21	14

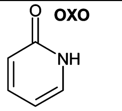
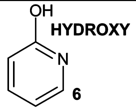
4.
1,2,3-triazole
	0.4/water/M and m;
	ND/gas (calc.)/m;
	ND/solid/M and m.
			22	2		11	0	15

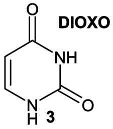
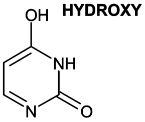
5.
1,2,4-triazole
	ND/water/M;
	ND/solid/M;
	6.3/gas (calc.)/M.
			81	8		3	3	5

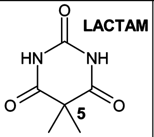
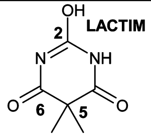
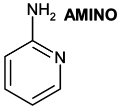
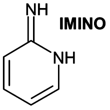
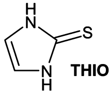
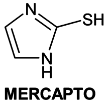
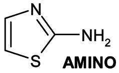
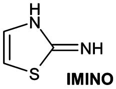
6.
Isocytosine
	0/water/M and m;
	ND/gas (calc.)/
	hydroxyl form (not shown);
	ND/solid/M and m.
			114	930		17	21	147

7. **Cytosine**	3.5/water/M.		27	0		3	0	0

b)

Name	ΔG (kcal mol^{-1}) / medium / relevant form(s)	Major form in Water (M)			Minor form in Water (m)			PDB with identical HB score
		Structure/Name	CSD	PDB	Structure/Name	CSD	PDB	
8. **2-pyridone**	4.2/water/M; -0.65/gas (calc.)/m; ND/solid/m.		172	24		12	2	2

9.
Uracil
	4.9/water/M;
	ND/gas (exp.)/M;
	ND/solid/M.
			91	113		0	3	9

10. **Barbituric Acid** (C5 disubstituted derivatives)	20/water/M.		38	0		0	0	0
11. **2-aminopyrimidine**	8.2/water/M.		120	6		20	1	17
12. **2-thioimidazole**	ND/water/M.		19	4		0	0	0
13. **2-aminothiazole**	5.9/water/M.		138	11		24	0	0

[a] Columns "CSD" and "PDB" report the occurrences of ligands in the corresponding form. Column "PDB with identical HB score" reports the occurrences of ligands in PDB complexes for which it was not possible to determine the most likely tautomer. The second column schematically reports the experimental and calculated energy difference (ΔG) between the two (M and m) tautomeric forms in different media (ND: not determined).

Table 2. Number of Total Tautomeric Actives and Decoys for Each of the 40 Targets in the DUD Database[a]

		no. tautomeric actives			no. tautomeric decoy	
target	no. actives (total)	+5%	all	no. decoys (total)	+5%	all
AR	74	0	18	2630	105	828
ER_AGONIST	67	0	1	2361	121	1036
ER_ANTAGONIST	39	0	15	1399	33	111
GR	78	0	0	2804	25	589
MR	15	0	0	535	18	162
PPAR_GAMMA	81	1	7	2910	34	295
PR	27	0	0	967	16	293
RXR_ALPHA	20	0	0	708	7	138
CDK2	**50**	**16**	**44**	**1780**	**234**	**899**
EGFR	**416**	**30**	**427**	**14 914**	**1032**	**4251**
FGFR1	118	0	116	4216	210	1606
HSP90	**24**	**10**	**23**	**861**	**115**	**418**
p38	**234**	**84**	**166**	**8399**	**392**	**2347**
PDGFRB	157	0	91	5625	277	1214
SRC	**162**	**5**	**140**	**5801**	**383**	**2524**
TK	22	0	22	785	47	462
VEGFR2	**74**	**9**	**57**	**2647**	**172**	**1055**
FXA	142	2	12	5102	755	2440
THROMBIN	65	7	12	2294	110	526
TRYPSIN	43	2	3	1545	79	371
ACE	49	0	2	1728	63	438
ADA	23	1	5	822	47	342
COMT	12	1	1	430	61	219
PDE5	51	4	24	1810	52	307
DHFR	201	1	200	7150	1371	4343
GART	**21**	**21**	**21**	**753**	**115**	**382**
ACHE	105	1	6	3732	45	283
ALR2	26	0	7	920	62	317
AMPC	21	0	0	734	44	215
COX1	25	0	4	850	9	109
COX2	349	3	13	12 491	409	3062
GPB	52	1	4	1851	191	977
HIVPR	53	4	4	1888	187	605
HIVRT	40	0	15	1439	59	434
HMGA	35	1	1	1242	62	356
INHA	85	0	0	3043	95	644
NA	49	0	1	1745	95	498
PARP	33	1	3	1178	100	548
PNP	25	0	25	884	148	628
SAHH	33	0	33	1159	138	664

[a] Targets in bold were selected for this study because of the higher number of tautomeric ligands, predicted to have an abundance greater that 5%.

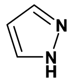
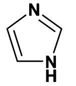
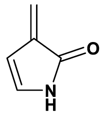
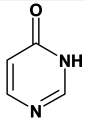
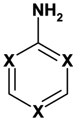
For protein residues we considered Arg/Lys and Glu/Asp in the charged form and all other residues in the neutral form, whereas ligands were considered in the neutral form only. However, because some of the tautomeric moieties in Table 1 are ionizable, we have discussed this effect on a case by case basis in the Results Section.

For each ligand the tautomer with the highest score was selected as the most likely tautomer in the binding site, and this yielded the occurrence of the different tautomers in protein−ligand complexes. When both tautomers yielded the same score, no conclusion was made, but we have still reported these occurrences in Table 1. Last, since the same ligand can be in many different PDB structures and may introduce a source of bias, especially when the environment around these ligands does not change, for each ligand we considered only up to three different PDB complexes.

Tautomer Enrichment in Virtual Screening. *Selection of Appropriate Targets.* The DUD data set was used to investigate how the choice of tautomers affects the recovery of actives using Flap, Glide, and Gold as docking tools. The DUD data set includes 40 targets, and each is associated with a number of actives and with about 36 decoys for each of the actives. The decoys are supposedly inactive compounds with physical chemical properties similar to those of the corresponding actives of a target. For this reason, using the decoys included in the DUD data set is expected to guarantee a more stringent test than using other types of decoys.

For the purpose of this work we reduced the data set so as to include only tautomeric actives and decoys. Table 2 provides detailed statistics about the number of tautomers found for active and decoy molecules for each of the 40 targets. To ensure a sufficient number of relevant tautomers (i.e., with an abundance in water above 5%), we selected only 7 out of the 40 targets: cyclin-dependent kinase 2 (CDK2), epidermal growth factor receptor (EGFR), heat shock protein 90 (HSP90), p38 protein (p38), src protein (SRC), vascular endothelial growth factor 2 (VEGF2), and gart protein (GART). Table 3 summarizes the tautomeric substructures in the inhibitors of the targets considered and the number of tautomers with that substructure. Figure 1 shows key H-bond interactions in representative inhibitor(s)

Table 3. Number of Tautomeric Substructures Retrieved from the Actives Ligand for the Seven Targets Considered in the Study

Tautomeric fragment	CDK2	GART	p38	EGFR	SRC	HSP90	VEGFR2
	9	-	-	25	3	13	17
	9	-	84	11	-	1	3
	11	-	-	-	8	-	4
	-	21	-	-	2	-	-
	22	21	35	412	129	10	38

of each target to understand whether tautomerism is more or less relevant for that particular target.

Ligand Preparation. Ligands extracted from the DUD data set, which were retained in their original 3D conformation, were first neutralized (the original data set contains multiple tautomers and ionization states), and when more than a tautomeric or ionization state was provided for the same compound, only the first structure was kept. Subsequently, both ligand and decoys were processed by TauThor to enumerate tautomers and to predict tautomer stability. The majority of inhibitors of the target considered in this study bind in the neutral state (only GART and SRC inhibitors have deprotonated carboxylic acids), and this was useful to limit the influence of ligand ionization. We deprotonated acids with $pK_a < 5$ and protonated bases with $pK_a > 8$ using MoKa; therefore, only one ionization state was used for any compound.

Protein Preparation. The starting structures for the targets used for this study were taken from the DUD database, and hydrogen atoms were added separately by the preprocessing programs embedded in the docking tools used. Some of the targets bind to their ligands through water molecules, however, because waters are mixed H-bond donor/acceptor, in principle they are not expected to affect tautomer preference. Nonetheless, waters can be beneficial to correctly place a ligand in the binding site, and having the correct pose is fundamental to study tautomer preference. Therefore, to determine whether it was advantageous to retain water molecules in the different targets, we run a preliminary docking with Glide first without water and then with all the waters that mediate a ligand−protein contact in the DUD targets, as defined by Huang.[32] Eventually, we decided to retain waters as follows: GART WAT 109; p38 none; HSP90 WAT 133, 137, 143, 261, 262, and 56; EGFR WAT 1, 2, 4, 5, 22, and 29; SRC none; CDK2 none; and VEGFR2 none. The 3D coordinates of these water molecules, including the position of the corresponding H atoms, were taken from the data published by Huang (the position of H-atoms was optimized by Huang using the program PLOP).[33]

Flap. Flap[22] performs docking by describing both the protein cavity and the ligands with molecular interaction fields (MIFs) generated by the GRID[34] program. MIFs are obtained by calculating the interaction energy between an atom of the protein/ligand and a "probe" in a point around the protein/ligand. To mimic different types of protein−ligand interactions, it is important to include probes N1 (donor), O (acceptor), and C3 (hydrophobic) in addition to the default probe H, which describes the shape of the protein/ligand. Flap requires two steps for docking, first where MIFs are calculated for different conformers of the ligand, and second where each of these conformers is scored according to the

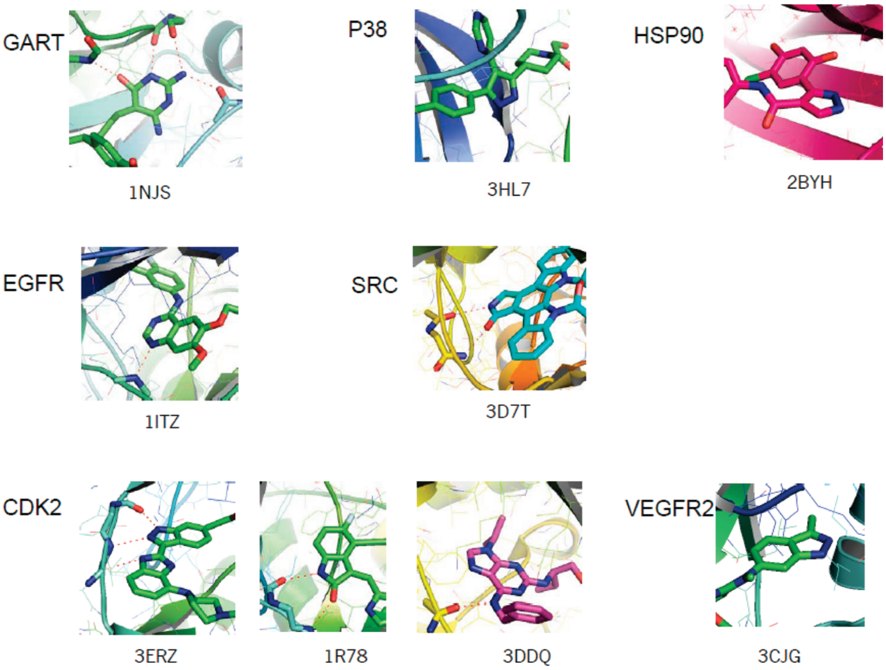


Figure 1. Binding modes of tautomeric inhibitors highlight that tautomeric atoms are sometimes involved in critical contacts with the protein.

match between its MIFs and those of the protein cavity. The protein cavity was defined by the region at less than 3 Å from the cognate ligand. It is important to note that selecting the protein cavity from the position of a known ligand rather than by considering a larger region in the protein pocket can facilitate the selection of the correct pose. However, results may be different if a larger cavity is selected. No experimental data on binding affinity are used in the scoring used by Flap. Flap was set to generate up to 50 conformers/ligand, with the maximum number being reached when the root-mean-square deviation (rmsd) between two conformers is lower than 0.15 Å.

Glide. The first step for docking with Glide[24] was the generation of grids that define the receptor site according to the position of the cognate ligand. For each ligand Glide generates various conformers, places each of them in the receptor site, and minimizes them using the OPLS-AA force field[35] with a distance dependent dielectric. Lowest energy poses are subsequently sampled for nearby torsional minima using a Monte Carlo procedure. We used default van der Waals scaling (1.0 for the receptor and 0.8 for the ligand) and scored the results with Glide SP, which is a modified version of the ChemScore scoring function.

Gold. The program Gold[27] uses a genetic algorithm to explore the conformational flexibility of the ligand and the rotational flexibility of the selected receptor H atoms. The placement of ligands is based on fitting points, which are added to the ligand and to the protein to find a match between acceptor and donor points. In addition to that, Gold uses also hydrophobic fitting points, which are mapped to CH groups of the ligand. In analogy with Flap and Glide, we used the cognate ligand to define the position of the receptor site. Default settings were used in all calculations except for the GA efficiency, which was set at 10% (virtual screening mode). Docking with Gold was performed by using the scoring functions GoldScore and ChemScore, with the intent of evaluating differences between scoring functions parametrized on experimental binding affinity data (Chemscore) and one that was not (GoldScore). For kinases we used the kinase parameter file (instead of the default parameter file) for the ChemScore scoring function. Gold allows to switch on and off water molecules during the docking calculation, however, to have a common protocol in all the tools considered, we considered all waters explicitly.

Analysis of Results. After docking all actives and decoys in alternative tautomeric states generated for each target, the impact of tautomer enrichment was evaluated by using the following protocols: (i) tautomers with predicted abundance of 0% in water; (ii) tautomer with the highest predicted stability; (iii) tautomer(s) with predicted abundance above 5%; and (iv) all possible tautomers. For each of these cases, when more than one tautomer was available for the same compound, we considered the score of the best scoring tautomer. The area under the receiver operating characteristic curve (ROC AUC)[36,37] and the ROC enrichment curves at 10% were used to compare the four tautomer generation protocols in combination with the three docking tools. To obtain ROC enrichment curves we plotted the % of selected actives, Se (sensitivity) versus the % of selected inactives, equal to 1-Sp (specificity), with Se and Sp defined below:

$$\mathrm{Se} = \frac{N\ \text{selected actives}}{N\ \text{total actives}} \quad \mathrm{Sp} = \frac{N\ \text{discarded inactives}}{N\ \text{total inactives}}$$

The ideal ROC curve yields ROC AUC = 1, whereas a random curve yields ROC AUC = 0.5. While ROC AUC is useful to assess the performance of the method across the whole ranked data set, ROC enrichment, e.g., considering to test 10% of the ranked data set, is more suited to assess early enrichment. ROC AUC and ROC enrichment 10% of the ranked data set are provided in our analysis.

RESULTS AND DISCUSSION

Tautomeric Preference in Water, Small Molecule Crystal Structures, and PDB Complexes. To understand possible relationships between tautomer preference in water and crystal structures of small molecules, we have reported, in Table 1, 13 tautomeric compounds with experimental data on their stability in water and the occurrences of alternative tautomers in the CSD database by running a substructure search with the parent compound used as a query. The 13 compounds were divided into two groups: (i) compounds 1−7 in Table 1a undergo annular tautomerism (i.e., imidazole, pyrazole, etc.), and their ΔG is relatively low, whereas compounds 8−13 in Table 1b have a higher ΔG. For each compound we searched in the PDB for the number of times that one of the two tautomers reported in Table 1 yielded an optimal H-bonding interaction, as described in the Methods Section.

Indazole. Indazole has been studied extensively for its tautomerism, both computationally and experimentally, and it has been reported that the 1H tautomer is more stable than the 2H tautomer, both in the gas phase[38] and in water (ΔG in water = 2.3 kcal mol^{-1}).[39] Data from the CSD database show that the 1H tautomer is the only form observed in the structures retrieved, considering also the possible substituents. By analyzing PDB complexes that have indazole as a substructure, we found that the 1H tautomer always gave better interactions than that of the 2H tautomer.

Interestingly, the preference for the 1H form is true also when residues in the protein may potentially favor the 2H tautomer. For example, in the optimization strategy that lead to the discovery of the pyrazole-based VGFR2 inhibitor Pazopanib, it was originally suggested that the 2H tautomer could form a contact with Asp 1046, however, structural data later revealed that this does not occur and that the inhibitor remains in the more stable 1H form.[40] However, it should be noted that in limited cases substituents may reverse the intrinsic stability of indazole, especially considering that the energy difference between 1H and 2H is not very high.[41]

Adenine. The two low-energy tautomers of adenine (7H and 9H) differ for the position of the H atom in the five-membered ring. Experimental data report that the ratio between the 9H and the 7H form is 4:1 in water,[42] and this corresponds to a ΔG = 0.8 kcal mol^{-1}, whereas in the gas phase, the 9H form is largely predominant[43] (calculated ΔG = 12 kcal mol^{-1}).[44] This difference is caused by the much larger dipole moment of the 7H form, which is stabilized in polar media. From data in the CSD database, we observed that 3 hits were represented in the 7H form and 17 in the 9H form. Data from PDB complexes, instead, show that the 7H tautomer is populated more often (23 and 10 cases in the 7H and 9H forms, respectively).

Hypoxanthine. Hypoxanthine may exist as a mixture of numerous tautomers, including keto−enol and amino−imino forms, however only the keto−amino forms are low-energy tautomers. In water similar amounts of the 7H and 9H forms were observed,[45] but the 7H form predominates in the gas phase (calculated)[46] and the 9H form in the solid.[47] The search for hypoxanthine derivatives in the CSD databases returned only 2 hits in both the more stable 9H form and the less stable 7H form, and likewise in the PDB, we found that both forms were populated (27 and 21 structures in the 9H and 7H forms, respectively).

1,2,3-Triazole. 1,2,3-triazole exists as a mixture of two species in water, with the 1H form being predominant (67%) in aqueous solution.[48] However, in the gas phase the 2H form predominates,[49] and in the crystal both forms were observed.[50] From the CSD database 22 hits were retrieved in the 1H form and 11 in the 2H form. In the PDB we could retrieve only 2 hits where the 1H form was favorite and none in the 2H form.

1,2,4-Triazole. Experimental data in the gas phase and in solid, water, and organic solvents show that 1,2,4-triazole is more stable in the 1H form rather than in the 4H form,[51,52] with ΔG calculated at 6.3 kcal mol in the gas phase. The majority of compounds in the CSD database exist in the 1H form (81 hits versus 3 in the 4H form), however in the PDB, both forms were observed (8 and 3 in the 1H and 4H forms, respectively).

Isocytosine. Isocytosine can be represented in many tautomeric forms (keto−enol, imino−amino, etc.), however here, only the low-energy tautomers 1H and 3H were compared. These two forms exist in equal ratio both in water and in the solid,[53] but data from both the CSD and PDB show a strong preference for the 1H form. It is not clear whether this is caused by the effect of substituents, or if it is the result of intramolecular interactions, which may favor the 1H form by formation of dimers in the crystal and by interaction with the amidic O and N atoms of the backbone in the PDB.

The case of isocytosine is of particular interest given that isocytosine is a substructure of pterin, which is often cited as an example of compound that binds in the less stable form. Although we did retrieve a pterin derivative (neopterin) that binds to toxic ricin in the less stable 3H form, our data show that this event is very rare for a compound based on isocytosine.

Cytosine. In analogy with isocytosine, also cytosine can be represented in various tautomeric forms, but only the 2H and 4H, which are the tautomers with the lowest energies, are considered here. The 2H form is preferred in water by 3.5 kcal mol^{-1} according to experimental determinations[47] and by 2.8 kcal mol^{-1} according to AM1 SCRF calculations.[54] The CSD database returned 27 and 3 hits in the 2H and 4H forms, respectively, and no hits were retrieved from the PDB.

2-Pyridone. 2-pyridone exists in the oxo form in water ($\Delta G = 4.2$ kcal mol^{-1}),[55] but it predominates in the hydroxy form in the gas phase (3:1) and in organic solvents.[56] It is also well documented that substituents in the six-position, that lower substantially the basicity of the nitrogen atom, shift the equilibrium toward the hydroxy form.[57] As a matter of fact, the 12 compounds that in the CSD database were found in the hydroxyl form have a halogen atom in the 6-position, whereas 172 compounds were retrieved in the oxo form. The search in PDB complexes returned 24 hits that interact more favorably in the oxo form, and 2 that interact more favorably in the hydroxyl form. In the two PDB where the hydroxyl form was preferred (PDB codes 3DAG and 3H65), the molecule interacts with a iron atom through the N atom; therefore, the compound is clearly deprotonated in the N atom. Since metals downshift strongly the pK_a of acids, in these two cases the compounds were simply deprotonated, therefore for correct recognition, it is important to set correctly their ionization state rather than their tautomeric state.

Uracil. Uracil predominates in the dioxo form in water as well as in the crystal and in the gas phase,[58] and as opposed to isocytosine and cytosine, its tautomer stability is not much influenced by the medium. It has been reported that the ΔG between the oxo and hydroxyl forms is around 4.9 kcal mol^{-1} in water.[59] Data from the CSD database shows that derivatives of uracil have been observed all the time in the dioxo form. Table 1b does not report the two additional tautomers of uracil (dihydroxy), however, the search for these tautomers in the CSD database did not return any hits. Also in PDB complexes, uracil binds more favorably in the dioxo form in all the examples that were retrieved, with the exception for three cases (2FZ5, 1FVC, and 2EG5). For the PDB structure 2FZ5, determined using nuclear Overhauser enhancement restraints extracted from two-dimensional ^{1}H NMR spectra solved by NMR, the ligand was proposed to be deprotonated at the N atom, and the interaction profile shows that this state yields the optimal H-bonding. Visual inspection of the two other cases was not conclusive to determine the most favorable state because the tautomeric atoms were involved in contacts with more than one H-bond donor/acceptor.

Barbituric Acid. Barbituric acid exhibits two types of tautomerism, lactam−lactim (Table 1b) and keto−enol (not shown). However, for derivatives where both H atoms in C5 are substituted, only the lactim−lactam tautomerism is observed. Experimental and computational data show that both substituted/unsubstituted C5 barbituric acid are stable in the lactam form,[60] which is the only tautomer observed in water and in the gas phase. In the case of unsubstituted C5, the 6-hydroxy tautomer (enolic form, not shown in Table 1b) is the second more stable tautomer, with a calculated ΔG, in polar solvent, of +13 kcal mol^{-1} with respect to the triketo (lactam) form M of Table 1b.[61,62] In the case of disubstituted C5 derivatives, the 2-hydroxy form (lactim form m in Table 1b) is a high-energy tautomer with ΔG of +20 kcal mol^{-1} with respect to the triketo form. In agreement with these data, we have not found compounds in the CSD database in the lactim form.

Many barbituric acid derivatives bind to metalloproteins via a Zn^{2+} cation that interacts with one of the N atoms. Experimental data show that Zn^{2+} binds to weak N-(sulfonamides, imides) and O-acids (hydroxamic acids) in the deprotonated form, although their water pK_a can be as high as 10.[63] This is possible because Zn^{2+} causes an important pK_a downshift, which has been reported to be about 7 for Zn-bound water.[64] Likewise, also barbituric acid derivatives, which have pK_a values between 8 and 9[65,66] for the deprotonation of the N atom, are expected to interact through a deprotonated N atom. This suggests that hypotheses

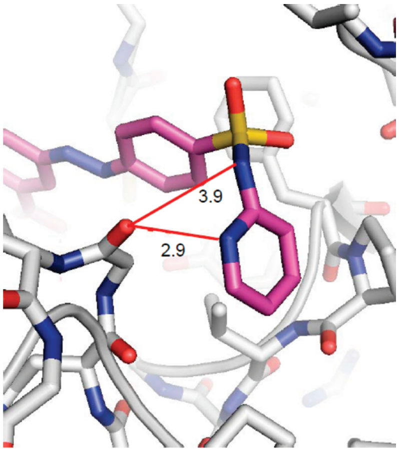


Figure 2. Binding of sulfalazine to glutathione S-transferase (PDB 13GS) suggests that the imino tautomer is preferred, in agreement with data from the CSD database.

where barbituric acid derivatives bind in the lactim form,[67,68] which is not observed in water in relevant concentration, should be reconsidered. This is of particular relevance because the example of barbituric acid derivatives bound to metalloproteins is often reported[2,16] to illustrate that unstable tautomeric forms may be stabilized at the binding site level.

In the PDB we could retrieve only four hits with barbituric acid derivatives: 1YOU, 1G4K, 1JJ9, and 2OVX. The binding mode of barbiturate is the same in all of these structures, with Zn^{2+} coordinating one deprotonated N atom, therefore, no tautomeric interconversion to the hydroxyl form is possible, as discussed above (unless the barbituric acid binds in the neutral state).

2-Aminopyridine. 2-aminopyridine may interconvert to the imino tautomer, however, water's ΔG is estimated to be around 7.3−8.2 kcal mol^{-1},[69] and this makes the amino form the only one virtually present in solution. However, it is important to note that electron-withdrawing groups bonded to the aminic N atom shift the equilibrium in favor of the imino form.[70] As a matter of fact, we did find in the CSD database a large number of 2-aminopyridine in the amino form (120), but we also retrieved 20 compounds in the imino form which have either a nitro, tosil, or sulfonylic group bonded to the exoimine.

2-aminopyridine-based compounds found in the PDB also bind more favorably in the most stable form in water, however as noted also from the study of the CSD database, care must be taken with compounds where the exocyclic N is connected to an electron-withdrawing group. For example, the binding mode of sulfasalazine complexed to glutathione transferase (PDB 13GS, 1.9 Å resolution) suggests that the H atom can be bonded to the pyridinc nitrogen rather than to the sulfamidic N, given that the carbonylic O atom of Gly 205 is 2.9 Å from the pyridinic N but 3.9 Å from the sulfamidic N (Figure 2). Three other structures that contain the same 2-sulfamidopyridine were retrieved from the PDB (3G49, 3CEN, and 1FVV), however from those, it was not possible to draw conclusions on the position of the H atom. The ionization constant of the pyridinic N atom of sulfasalazine is 0.6,[71] therefore, it is unlikely that it is protonated in the protein complex.

As explained for other structures, it is fundamental to take into account also the ionization state of a compound when one studies tautomerism. For example, 2-aminopyridine, which titrates at pH 6.8, is protonated when it interacts with a carboxylate, and among the 14 structures when no conclusion could be made regarding the preferred tautomer, we found cases where the binding occurs through an interaction between the protonated piridinic N atom and a carboxilate (i.e., 3E7M).

2-Thioimidazole. In water, 2-thioimidazole exists predominantly in the thio rather than in the mercapto form,[72] and this is confirmed also by data in the CSD database. Likewise, analysis of the PDB shows that in all cases the optimal H-bonding pattern occurs via the most stable form in water.

2-Aminothiazole. 2-aminothiazole is more stable in the amino rather than in the imino form with a ΔG of 5.9 kcal mol^{-1}.[73] Despite the much larger stability of the amino form in some chemical databases, this structure is reported almost exclusively in the imino tautomer.[4] The CSD database returned 24 hits with the imino form, and these structures show that the 2-imino group is linked to a SO_2 group (19) and a tosyl (1) and to other groups (4), in analogy with 2-aminopyridine. This is in agreement with data by Forlani et al.,[73] who reported that in the presence of a tosil or electron-withdrawing substituent the imino form predominates. However, in the absence of this substituent, the CSD database returns the amino form as the more stable form in water, for a total of 138 hits. In the PDB, no derivative of 2-aminothiazole with an electron-withdrawing group was retrieved, and the all complexes show that the 2-aminothiazole tautomer forms the best interaction (11 cases).

Which Tautomers Are Relevant in a Protein Pocket? The results presented above highlight two major issues to consider for tautomers that bind to a protein pocket. The first is that predicting tautomer stability is challenging because specific substituents change and sometimes reverse the energy barrier between two tautomers (as discussed for 2-aminopyridine and 2-aminothiazole). This calls for tools for virtual screening that are trained more accurately to predict tautomer stability in the presence of different substituents. Although it is virtually impossible to handle correctly any case, improvements would still be valuable for virtual screening applications.

The second issue is that if the ΔG between two tautomers is known, at least within 1 kcal mol^{-1}, then it is relatively straightforward to predict the tautomer that is most likely to be at the binding site. For example from the 13 cases described above, we observed that compounds with experimental ΔG in water below 1 kcal mol^{-1} (compounds 2−4) tend to be populated in both forms, whereas for the other compounds, only one form is strongly predominant. For example, the 17 tautomers of isocytosine that were retrieved in the PDB in the less stable form correspond to 2% of the total. Therefore, while we recognize that there are cases where the less stable form in water is the one that binds to the protein, it is also important to consider the frequency of these events.

Selecting Tautomers to Enhance the Recovery of Actives in Structure-Based Drug Discovery. In the previous section, we have shown that ligands complexed in PDB crystal structures are most likely to bind in the form that predominates in water if the energy barrier is sufficiently high (>2 kcal mol^{-1}), whereas tautomers that are populated in more equal amounts, particularly tautomers characterized by annular tautomerism, bind in either form. However, compounds are sometimes reported also as high-energy tautomers either because of canonicalization algorithms that convert structures in different formats or because a high-energy form in water is stable in certain organic solvents.[4] This is true also for smiles codes reported in the PDB, where a ligand is sometimes represented in a form that is clearly high energy and cannot interact favorably in the binding site (i.e., PDB 1FVV and 1FVT). For these reasons, tautomer enrichment has been recommended to ensure that the relevant tautomer be available in the screening library. However, to what extent does the tautomer enrichment enhance the recovery of actives? We have evaluated this by using Glide, Flap, and Gold as docking tools.

To understand better the results, we have reported in Table 3 a breakdown of the tautomeric inhibitors of each target, and in Figure 1, we have reported the binding mode of representative compounds to identify targets where the choice of the tautomer is more critical than in others. For example in GART, EGFR, SRC, and CDK2, the tautomeric atoms interact with the hinge region of ATP, and therefore, the choice of tautomer is expected to have an impact in the recovery of actives, but in p38, HSP90, and VEGFR2, the tautomeric atoms are not involved in critical contacts with atoms of the protein that are donors- or acceptors-only.

Protein flexibility can also play an important role in tautomer recognition, as in some cases, it might be possible that lateral chain conformations (e.g., Asn/Gln) may balance diverse configurations assumed by the tautomeric ligand. This critical aspect of conformational changes at the protein level is believed to be minimal in our case studies. As shown in Figure 1, the interaction of the tautomeric atoms of the different ligands under investigation is mainly with the backbone of GART, EGFR, SRC, and CDK2.

It is also important to note that in general tautomeric rearrangements may change the molecular shape between the tautomeric forms in addition to the H-bonding characteristics. For example, this happens in the case of the keto−keto/keto−enol rearrangement, as a result of different possible conformations. In the tautomeric groups considered in this study (Table 3), the change in molecular shape, due to conformational changes, is not significant as the tautomeric rearrangements involve only shifts between atoms in rigid cycles.

GART. GART inhibitors bind to the ATP hinge region via three H-bonds: two H-bond acceptors interact with the amidic NH of Asp 144 and Leu 92 and one H-bond donor to the carbonyl of Thr 140. Because GART inhibitor must compete with ATP, it is fundamental to exploit these interactions to achieve potency. All tautomeric GART inhibitors in the DUD data set have a 2-amino-3H-pyrimidin-4-one ring that interacts with the hinge region, and this moiety can be represented in many tautomeric forms, such as amino−imino and keto−enol or as forms that differ only in the position of the H atom between the two N atoms of the ring.

In analogy with compounds 6−9 of Table 1, for this moiety extensive literature data reports that only the keto−amino form exists in water and that the 2H and 4H tautomers are the low-energy forms, with the 2H tautomer being the most stable.[10] For this particular target, the most stable form is only possible at the binding site, and Figure 3 shows that it is fundamental to include this tautomer to achieve optimal results, since all docking tools would yield poorly especially in the early part of the enrichment curve.

The tautomers generated by MoKa for GART inhibitors include both the high-energy forms (imino and enol), which are estimated at 0% in water, and the 2H and 4H low-energy forms, which are estimated to be at different percentages depending on the specific substituents. MoKa calculates the tautomeric ratio K_T from $pK_T = pK_{aT1} - pK_{aT2}$ and, for all GART inhibitors, predicts that the 2H is most stable form.

In all the docking runs, both ROC AUC and ROC enrichments at 10% show that the selection of tautomers with the lowest abundance—imino and enol forms—yielded significantly poorer results (Figures 3−4) than that of the other selections of tautomers. The largest effect was observed with Glide at 10% ROC enrichment: selecting the high-energy forms yielded only 0.33, as opposed to 0.90 when forms with abundance above 5% were used. Differences between selecting all tautomers versus selecting low-energy forms only are less pronounced, although for this target Flap and Glide perform better if only low-energy forms are considered. The difference in performance between ROC enrichment at 10% for the low-energy forms versus all forms is on average of 0.15 better for all the tools used, except for Gold (Goldscore), which yields the similar results for this target.

p38 and HSP90. Tautomeric inhibitors of p38 contain aminopyrimidine or imidazole groups, whereas for HSP90 they contain aminopyrimidine or pyrazole. In these targets the tautomeric groups are not involved in critical contact with residues of the protein. For each target, 3 out of the 4 docking programs yielded a similar enrichment independently on the choice of tautomer ($\Delta ROC_{10}\% = ROC_{10}\%(+5\%) - ROC_{10}\%(min) \cong 0$). For two cases, p38 in combination with Glide and HSP90 in combination with Gold (Goldscore), a $\Delta ROC_{10}\%$ of 0.15 and 0.1 have been obtained, respectively.

EGFR. As shown in Table 3, the vast majority of EGFR inhibitors contain an amino-pyrimidine that binds to the hinge region of ATP. Analysis of the PDB, the CSD, and the experimental data show that this chemical group is stable as it is, and only electron-withdrawing groups attached to the exocyclic N can change this equilibrium. However, we found that Flap and Gold are not very sensitive to the choice of the tautomer in this case, whereas Glide performs significantly better when low-energy forms are included ($\Delta ROC_{10}\% = 0.55$). In the case of 4-aminoquinazoline, the imino tautomeric form of the 4-aminoquinazoline can also bind to EGFR, maintaining the crystallographic pose. Neither the H-bond at the hinge (with N1) nor the vector of the 4-amino attachment point to the backpocket would be affected.

SRC. A critical interaction of SRC inhibitors is that with the amidic NH and C=O of Met 341. Most SRC inhibitors from the DUD data set have amino-pyrimidine moieties that mediate this contact, with the aromatic nitrogen acting as an acceptor and the aminic N as a donor. Other inhibitors contain tautomeric groups that interact with this motif, such as pyridones and oxindoles. Therefore, it is expected that overall in SRC the choice of the tautomer has an impact in the results. This was the case for all the tools used, however, while Glide and Gold perform better when low-energy forms are included, Flap in this case gives better results if one accounts for high-energy forms, which is surprising.

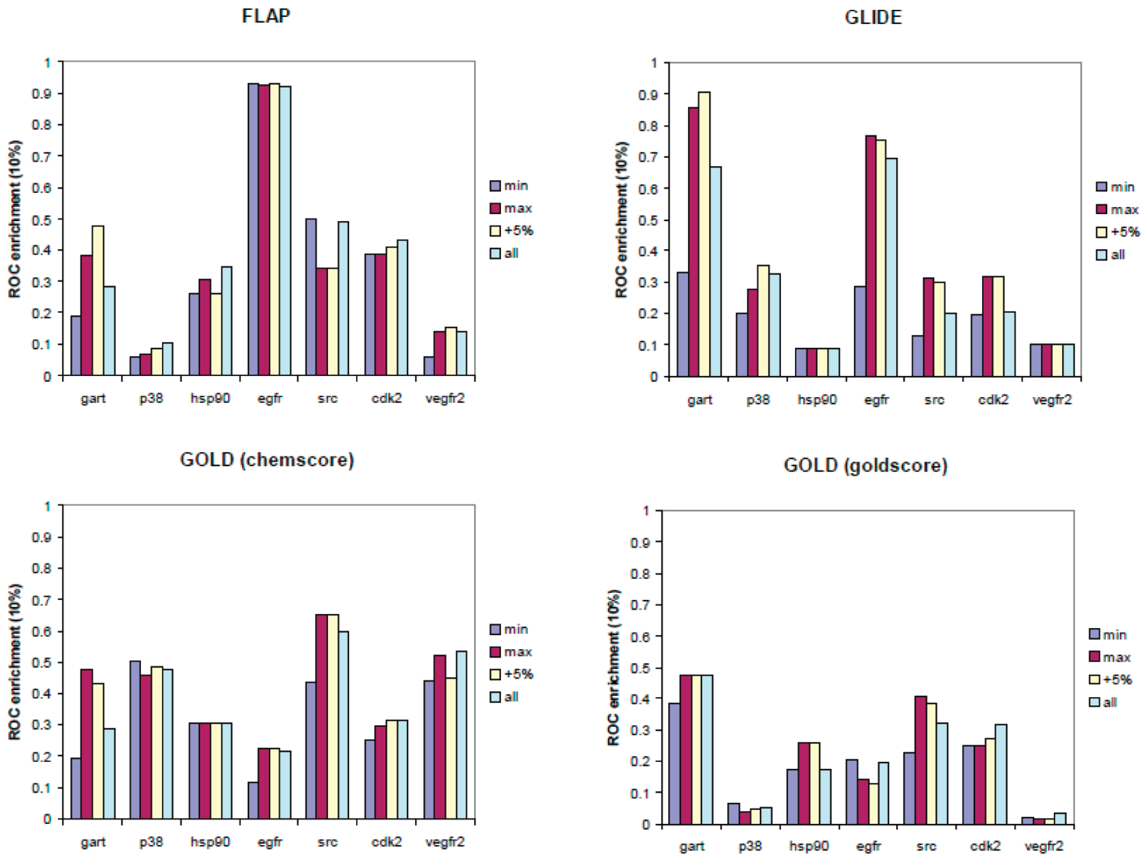


Figure 3. ROC enrichment at 10% obtained by Flap, Glide, and Gold considering different tautomeric form(s): (a) “min” (violet bars): only tautomers predicted to have an abundance of 0% in water; (b) “max” (brown bars): only tautomer predicted as most stable form; (c) “+5%” (yellow bars): all forms predicted to have an abundance above 5% in water; and (d) “all” (cyan bars): all possible tautomeric forms.

CDK2. CDK2 inhibitors include many pyrazole and imidazole derivatives, which may easily shift the position of the H atom from one nitrogen to the other and the oxyindole derivatives, which are expected to be highly stable in the oxo form only, but because of their potentially tautomeric nature, they are sometimes reported in alternative forms in databases.

Data in Figures 3−4 show that the docking tools used are not very sensitive to the choice of tautomer for CDK2 (except for Glide at 10% enrichment), and even when only minor forms are considered, results are not affected significantly. This is quite surprising, since the binding mode in Figure 1 shows that the tautomeric atoms bind to the hinge region.

Although for CDK2 ROC AUC are only moderately dependent on the choice of the tautomer, we found that the binding mode changes significantly depending on the tautomer, as illustrated in Figure 5, which reports the % of inhibitors whose average rmsd among all the possible tautomers docked by Glide is above or below 3 Å. In CDK2, probably also because of the small size of the inhibitors compared to the size of the binding pocket, 70% of the inhibitors change completely their binding mode depending on the tautomer.

It is important to stress that alternative binding modes for different tautomers are indeed possible. For example, analysis of electron density maps from X-ray determination has shown that for a pyrazole inhibitor of CDK2 (PKF049−365) there are two alternative binding modes, as a result of the low energy to interconvert one tautomer to other.[74] Therefore, this study suggests that, for tautomeric ligands that may be populated in alternative forms, it is particularly important to inspect the experimental electron density map to rule out the possibility of multiple binding modes. Although this is a unique case, other modeling studies of pyrazole-based inhibitors (one focused on a tyrosine kinase inhibitor[75] and another on a KDR kinase inhibitor)[76] have discussed how alternative tautomeric states change the binding mode and how studying tautomer stability may help to select the most suitable form.

VEGFR2. A number of VEGFR2 inhibitors include tautomeric indazoles and imidazoles. However, as explained for HSP90 and p38, the absence of critical contacts involved with tautomeric atoms of VEGFR2 results in a very low dependency of docking results on the choice of tautomer.

To verify that the difference in the results of the docking on the two groups of targets (those with and without a direct interaction with tautomeric atoms) was statistically significant, we performed a one-tailed unpaired t test and obtained a p value of 0.02, which is statistically significant at the 0.05 level. For each group of targets, we calculated the difference in ROC Enr10% between the protocol which used only the minor tautomer (violet ‘min’ bars in Figure 3) and that which uses only the major tautomer (brown ‘max’ bars in Figure 3) by combining results from all docking programs. For those targets in which the interaction with the ligand is mediated by tautomeric atoms (i.e., GART, EGFR, SRC, and CDK2) the mean ROC Enr10% is 13% higher if one uses the major form rather than the minor form, whereas the mean is close to zero for the other targets (i.e., $\Delta ROC_{10}\% = ROC_{10}\%(max) - ROC_{10}\%(min) \cong 0$ for VEGFR, p38, and HSP90).

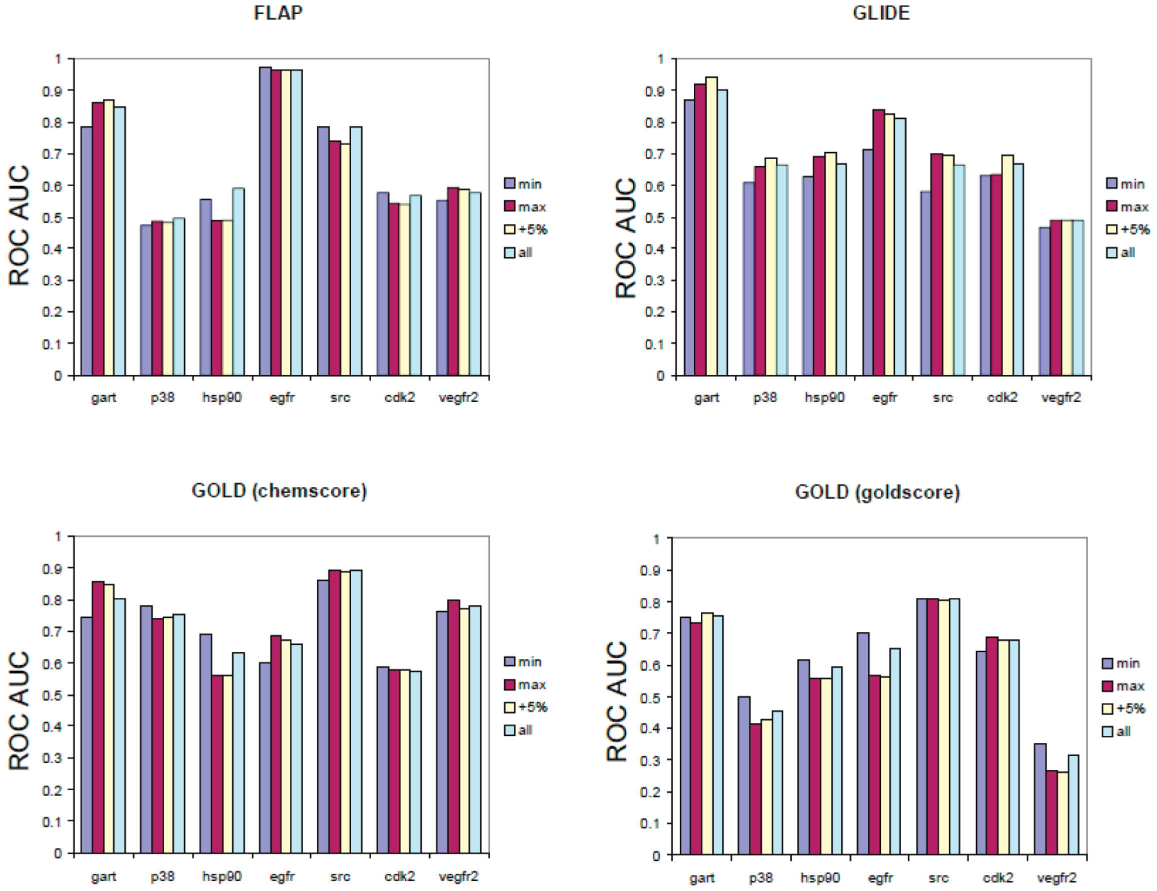


Figure 4. ROC AUC obtained by Flap, Glide, and Gold considering different tautomeric form(s): (a) "min" (violet bars): only tautomers predicted to have an abundance of 0% in water; (b) "max" (brown bars): only tautomer predicted as most stable form; (c) "+5%" (yellow bars): all forms predicted to have an abundance above 5% in water; and (d) "all" (cyan bars): all possible tautomeric forms.

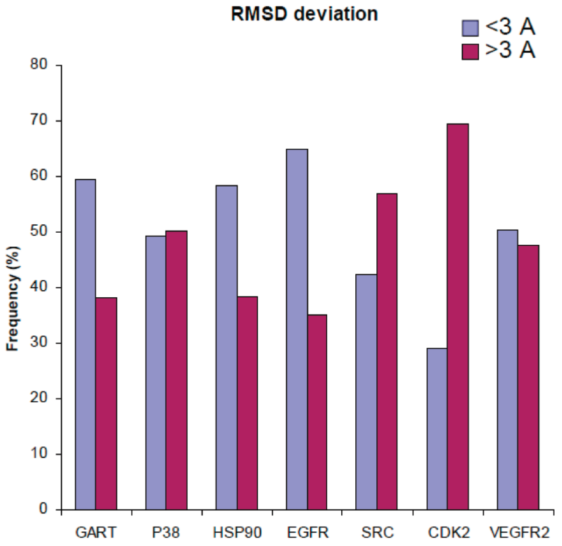


Figure 5. Frequency of active ligands with average rmsd above (brown bars) or below 3 Å (violet bars) for binding modes of alternative tautomers docked by Glide.

The results presented show that significant enhancements can be obtained by docking relevant tautomeric forms, as opposed to unstable forms, and that the presence of only the most stable tautomer may be the optimal strategy, since this minimizes the computational resources and yields results as good as or better than including all possible tautomers. However, the docking of different tautomeric forms is an aspect that should be particularly evaluated on the basis of the molecular system under study.

As a possible next step, the docking score could be corrected by the stability of a tautomer. This idea has been addressed by Schrödinger by using ionization and tautomeric state penalties predicted by Epik.[77] According to Schrödinger, this leads to large improvements in enrichment studies.[78] According to the authors, the improvement comes mainly from penalizing decoy ligands with good Glide scores that are in high-energy (ionization or) tautomeric states. The effect of substituents and of the local protein environment change the relative stability of forms that have $\Delta G < 1-2$ kcal mol^{-1}, and therefore, the correction of the docking score may be, in particular, of some advantage to penalize highly unstable tautomers but likely not to differentiate low energy forms.

CONCLUSIONS

In this paper we have investigated the role of tautomerism in structure-based drug discovery. First, we have tried to study how tautomer preference at the binding site reflects tautomer preference in water, and in which cases tautomer stability in the binding site is reversed compared to tautomer stability in water. We found that the problem of reversed stability can be reduced to specific cases, such as compounds that undergo annular tautomerism and that have low ΔG (<2 kcal mol^{-1}). For tautomers with higher ΔG such as the imino−amino forms of 2-aminopyridine and the keto−enol forms of 2-pyridone derivatives, the stability in the binding site reflects stability in water in the majority of cases.

However, given the dramatic effects caused by specific substituents, care should be taken when assessing tautomer stability in water.

In the second part of the paper, we have investigated to what extent the selection of tautomers affects the recovery of actives. We have selected tautomeric inhibitors and decoys from seven targets of the DUD database and have run docking using Flap, Glide, and Gold. As expected, in the majority of cases, the choice of tautomers has no effect if the tautomeric atom is not involved in contacts with acceptor- or donor-only atoms, as in the case of p38, HSP90, and VEGFR2. However, for the other targets we found that considering only the tautomers predicted with an abundance 0% in water yields the poorest results, whereas including the most stable form, forms with an abundance in water greater than 5%, or all tautomeric (stable and unstable) forms yields comparable results. No advantage is seen if one increases the number of tautomers so as to include also irrelevant high-energy forms. However, the docking of different tautomeric forms is an aspect that should be particularly evaluated on the basis of the molecular system under study.

Note Added after ASAP Publication. This paper was published to the Web on June 1, 2010, with an error to the Supporting Information. The corrected version reposted to the Web on June 8, 2010.

Supporting Information Available: PDB codes of complexes that bind ligands retrieved from the substructure search of Table 1, together with the number of H-bonding contacts of the two tautomeric forms. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) IUPAC Compendium of Chemical Terminology - The Gold BOOK ; International Union of Pure and Applied Chemistry: Research Triangle Park, NC; http://goldbook.iupac.org/T06252.html. Accessed November 1, 2009.

(2) Martin, Y. Let's not forget tautomers. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 693–704.

(3) Raczynska, E. D.; Kosinska, W.; Osmialowski, B. Gawinecki, R.Tautomeric Equilibria in Relation to Pi-Electron Delocalization. *Chem. Rev.* **2005**, *105*, 3561–3612.

(4) Milletti, F.; Storchi, L.; Sforna, 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.

(5) Haranczyk, M.; Gutowski, M. Quantum Mechanical Energy-Based Screening of Combinatorially Generated Library of Tautomers. TauTGen: A Tautomer Generator Program. *J. Chem. Inf. Model.* **2007**, *47*, 686–694.

(6) Trepalin, S. V.; Skorenko, A. V.; Balakin, K. V.; Nasonov, A. F.; Lang, S. A.; Ivashchenko, A. A.; Savchuk, N. P. Advanced Exact Structure Searching in Large Databases of Chemical Compounds. *J. Chem. Inf. Model.* **2003**, *43*, 852–860.

(7) Shoichet, B. Charting chemical space: finding new tools to explore biology. *Session 7: In silico design of biologically active molecules*, Proceedings of the 4th Horizon Symposium, Scarborough, ME , May 20−22, 2004; Nature Publishing Group: London, U.K.; http://www.nature.com/horizon/chemicalspace/kq/7_Shoichet.html. Accessed November 1, 2009.

(8) Minkin, V. I.; Garnovskii, A. D.; Elguero, J.; Katritzky, A. R.; Denisko, O. V. The tautomerism of heterocycles: five-membered rings with two or more heteroatoms. In *Adv. Heterocycl. Chem.*, Academic Press: New York, NY, 2000; Vol. 76, pp 157−323.

(9) Elguero, J.; Katritzky, A. R.; Denisko, O. V. The Prototropic Tautomerism of Heterocycles. In *Advances in Heterocyclic Chemistry*; Academic Press: New York, NY, 2000; Vol. 76, pp 1−64.

(10) Katritzky, A. R.; Pozharskii, A. Tautomerism: Pyridone and Hydropydirines. In *Handbook of Heterocyclic Chemistry*; Pergamon: Amsterdam, The Netherlands, 2000; pp 47−51.

(11) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789–6801.

(12) Kalliokoski, T.; Salo, H. S.; Lahtela-Kakkonen, M.; Poso, A. The Effect of Ligand-Based Tautomer and Protomer Prediction on Structure-Based Virtual Screening. *J. Chem. Inf. Model.* **2009**, *49*, 2742–2748.

(13) ten Brink, T.; Exner, T. E. Influence of Protonation, Tautomeric, and Stereoisomeric States on Protein-Ligand Docking Results. *J. Chem. Inf. Model.* **2009**, *49*, 1535–1546.

(14) Clark, R. D.; Shepphird, J. K.; Holliday, J. The effect of structural redundancy in validation sets on virtual screening performance. *J. Chemom.* **2009**, *23*, 471–478.

(15) Hernández, R.; Orozco, M.; Luque, F. J. Tautomerism of xanthine and alloxanthine: A model for substrate recognition by xanthine oxidase. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 535–544.

(16) Pospisil, P.; Ballmer, P.; Scapozza, L.; Folkers, G. Tautomerism in Computer Aided Drug Design. *J. Recept. Signal Transduction* **2003**, *23*, 361–371.

(17) Oellien, F.; Cramer, J.; Beyer, C.; Ihlenfeldt, W. D.; Selzer, P. M. The Impact of Tautomer Forms on Pharmacophore-Based Virtual Screening. *J. Chem. Inf. Model.* **2006**, *46*, 2342–2354.

(18) Todorov, N. P.; Monthoux, P. H.; Alberts, I. L. The Influence of Variations of Ligand Protonation and Tautomerism on Protein-Ligand Recognition and Binding Energy Landscape. *J. Chem. Inf. Model.* **2006**, *46*, 1134–1142.

(19) Allen, F. H. The Cambridge Structural Database: A Quarter of a Million Crystal Structures and Rising. *Acta Crystallogr., Sect. B: Struct. Sci.* **2002**, *B58*, 380–388.

(20) Berman, H. M.; Henrick, K.; Nakamura, H. Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.* **2003**, *10*, 980.

(21) Baroni, M.; Cruciani, G.; Sciabola, S.; Perruccio, F.; Mason, J. S. A Common Reference Framework for Analyzing/Comparing Proteins and Ligands. Fingerprints for Ligands And Proteins (FLAP): Theory and Application. *J. Chem. Inf. Model.* **2007**, *47*, 279–294.

(22) *Flap*, version May 2009; Molecular Discovery LTD: London, UK; www.moldiscovery.com. Accessed January 12, 2009.

(23) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739–1749.

(24) *Glide*, version 5.5; Schrödinger, Inc: New York, NY; http://www.schrodinger.com/. Accessed January 12, 2009.

(25) Jones, G.; Willett, P.; Glen, R. C. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.* **1995**, *245*, 43–53.

(26) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727–748.

(27) *Gold*; CCDC: Cambridge, UK; www.ccdc.cam.ac.uk; Accessed January 12, 2009.

(28) Bruno, I. J.; Cole, J. C.; Edgington, P. R.; Kessler, M.; Macrae, C. F.; McCabe, P.; Pearson, J.; Taylor, R. New software for searching the Cambridge structural database and visualizing crystal structures. *Acta Crystallogr., Sect. B: Struct. Sci* **2002**, *B58*, 389–397.

(29) Ligand Expo. Protein Data Bank; RCSB PDB: Rutgers, the State University of New Jersey and San Diego Supercomputer Center (SDSC) and Skaggs School of Pharmacy and Pharmaceutical Sciences; http://ligand-expo.rcsb.org/. Accessed August 12, 2009.

(30) Milletti, F.; Storchi, L.; Sforna, G.; Cruciani, G. New and Original pK_a Prediction Method Using Grid Molecular Interaction Fields. *J. Chem. Inf. Model.* **2007**, *47*, 2172–2181.

(31) *MoKa*, version 1.1; Molecular Discovery LTD; London, UK; www.moldiscovery.com. Accessed January 12, 2009.

(32) Huang, N.; Shoichet, B. K. Exploiting Ordered Waters in Molecular Docking. *J. Med. Chem.* **2008**, *51* (16), 4862–4865.

(33) Jacobson, M. P.; Friesner, R. A.; Xiang, Z.; Honig, B. On the Role of Crystal Packing Forces in Determining Protein Sidechain Conformations. *J. Mol. Biol.* **2002**, *320*, 597–608.

(34) Goodford, P. J. A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *J. Med. Chem.* **1985**, *28*, 849–857.

(35) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.

(36) Hanley, J. A.; McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143*, 29–36.

(37) Triballeau, N.; Acher, F.; Ibrabet, I.; Pin, J.-P.; Bertrand, H.-O. Virtual Screening Workflow Development Guided by the "Receiver Operating Characteristic" Curve Approach. Application to High-Throughput Docking on Metabotropic Glutamate Receptor Subtype 4. *J. Med. Chem.* **2005**, *48*, 2534–2547.

(38) Alkorta, I.; Elguero, J. Theoretical estimation of the annular tautomerism of indazoles. *J. Phys. Org. Chem.* **2005**, *18*, 719–724.

(39) Catalan, J.; del Valle, J. C.; Claramunt, R. M.; Boyer, G.; Laynez, J.; Gomez, J.; Jimenez, P.; Tomas, F.; Elguero, J. Acidity and Basicity of Indazole and its N-Methyl Derivatives in the Ground and in the Excited State. *J. Phys. Chem.* **1994**, *98*, 10606–10612.

(40) Harris, P. A.; Boloor, A.; Cheung, M.; Kumar, R.; Crosby, R. M.; Davis-Ward, R. G.; Epperly, A. H.; Hinkle, K. W.; Hunter, R. N.; Johnson, J. H.; Knick, V. B.; Laudeman, C. P.; Luttrell, D. K.; Mook, R. A.; Nolte, R. T.; Rudolph, S. K.; Szewczyk, J. R.; Truesdale, A. T.; Veal, J. M.; Wang, L.; Stafford, J. A. Discovery of 5-[[4-[(2,3-dimethyl-2H-indazol-6-yl)methylamino]-2-pyrimidinyl]amino]-2-methyl-benzenesulfonamide (Pazopanib), a novel and potent vascular endothelial growth factor receptor inhibitor. *J. Med. Chem.* **2008**, *51*, 4632–4640.

(41) Alkorta, I.; Elguero, J. Theoretical estimation of the annular tautomerism of indazoles. *J. Phys. Org. Chem.* **2005**, *18*, 719–724.

(42) Dreyfus, M.; Dodin, G.; Bensaude, O.; Dubois, J. E. Tautomerism of purines. I. N(7)H ⇔ N(9)H equilibrium in adenine. *J. Am. Chem. Soc.* **1975**, *97*, 2369–2376.

(43) Nowak, M. J.; Rostkowska, H.; Lapinski, L.; Kwiatkowski, J. S.; Leszczynski, J. Tautomerism N(9)H ⇔ N(7)H of Purine, Adenine, and 2-Chloroadenine: Combined Experimental IR Matrix Isolation and Ab Initio Quantum Mechanical Studies. *J. Phys. Chem.* **1994**, *98*, 2813–2816.

(44) Mezey, P. G.; Ladik, J. J. A non-empirical molecular orbital study on the relative stabilities of adenine and guanine tautomers. *Theor. Chim. Acta* **1979**, *52*, 129–145.

(45) Benoit, R. L.; Fréchette, M. Protonation of hypoxanthine, guanine, xanthine, and caffeine. *Can. J. Chem.* **1985**, *63*, 3053–3056.

(46) Costas, M. E.; Acevedo-Chávez, R. Density Functional Study of the Neutral Hypoxanthine Tautomeric Forms. *J. Phys. Chem. A* **1997**, *101*, 8309–8318.

(47) Shcherbakova, I.; Elguero, J.; Katritzky, A. R. Tautomerism of heterocycles: condensed five-six, five-five, and six-six ring systems with heteroatoms in both rings. In *Advances in Heterocyclic Chemistry*; Academic Press: New York, NY, 2000; Vol. 77, pp 55−113.

(48) Albert, A.; Taylor, P. J. The tautomerism of 1,2,3-triazole in aqueous solution. *J. Chem. Soc., Perkin Trans. 2* **1989**, 1903–1905.

(49) Abboud, J. L. M.; Foces-Foces, C.; Notario, R.; Trifonov, R. E.; Volovodenko, A. P.; Ostrovskii, V. A.; Alkorta, I.; Elguero, J. Basicity of N-H and N-Methyl-1,2,3-triazoles in the Gas Phase, Solution and Solid State: An Experimental and Theoretical Study. *Eur. J. Org. Chem.* **2001**, *16*, 3013–3024.

(50) Goddard, R.; Heinemann, O.; Krüger, C. Pyrrole and a Co-crystal of 1H- and 2H-1,2,3-Triazole. *Acta Crystallogr.* **1997**, *C53*, 1846–1850.

(51) Goldstein, P.; Ladell, J.; Abowitz, G. Refinement of the crystal and molecular structure of 1,2,4-triazole (C2H3N3) at low temperature. *Acta Crystallogr.* **1969**, *B25*, 135–143.

(52) Bojarska-Olejnik, E.; Stefaniak, L.; Witanowski, M.; Webb, G. A. 15N NMR investigation of prototropic equilibria of some triazoles. *Bull. Pol. Acad. Sci. Chem.* **1987**, *35*, 85–90.

(53) Sharma, B. D.; McConnell, J. F. The crystal and molecular structure of isocytosine. *Acta Crystallogr.* **1965**, *19*, 797–806.

(54) Katritzky, A. R.; Karelson, M. M. AM1 calculations of reaction field effects on the tautomeric equilibria of nucleic acid pyrimidine and purine bases and their 1-methyl analogs. *J. Am. Chem. Soc.* **1991**, *113*, 1561–1566.

(55) Karelson, M. M.; Katritzky, A. R.; Szafran, M.; Zerner, M. C. Quantitative predictions of tautomeric equilibria for 2-, 3-, and 4-substituted pyridines in both the gas phase and aqueous solution: combination of AM1 with reaction field theory. *J. Org. Chem.* **1989**, *54*, 6030–6034.

(56) Hatherley, L. D.; Brown, R. D.; Godfrey, P. D.; Pierlot, A. P.; Caminati, W.; Damiani, D.; Melandri, S.; Favero, L. B. Gas-phase tautomeric equilibrium of 2-pyridinone and 2-hydroxypyridine by microwave spectroscopy. *J. Phys. Chem.* **1993**, *46*, 46–51.

(57) Katritzky, A. R.; Rowe, J. D.; Roy, S. K. Potentially tautomeric pyridines. Part IX. The effect of chlorine substituents on pyridine-hydroxypyridine tautomerism. *J. Chem. Soc. B* **1967**, 758–761.

(58) Colarusso, P.; Zhang, K.; Guo, B.; Bernath, P. The infrared spectra of uracil, thymine, and adenine in the gas phase. *Chem. Phys. Lett.* **1997**, *269*, 39–48.

(59) Poulter, C. D.; Frederick, G. D. Uracil and its 4-hydroxy-1(H) and 2-hydroxy-3(H) protomers. pKa's and equilibrium constants. *Tetrahedron Lett.* **1975**, *16*, 2171–2174.

(60) Jeffrey, G. A.; Ghose, S.; Warwicker, J. O. The Crystal Structure of Barbituric Acid Dihydrate. *Acta Crystallogr.* **1961**, *14*, 881–887.

(61) Senthilkumar, K.; Kolandaivel, P. Quantum chemical studies on tautomerism of barbituric acid in gas phase and in solution. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 263–272.

(62) Ralhan, S.; Ray, N. K. Density functional study of barbituric acid and its tautomers. *J. Mol. Struct.: THEOCHEM* **2003**, *634*, 83–88.

(63) King, R. W.; Burgen, A. S. V. Sulphonamide complexes of human carbonic anhydrases. Ultraviolet difference spectroscopy. *Biochim. Biophys. Acta* **1970**, *207*, 278–285.

(64) Lindskog, S.; Ibrahim, S. A.; Jonsson, B. H.; Simonsson, I. Carbonic Anhydrase: Structure, Kinetics, and mechanism. In *The coordination chemistry of metalloenzymes*. Bertini, I., Drago, R. S., Luchinat, C., Eds.; Reidel: Dordrecht, The Netherlands, 1983; Vol. IV, pp 49−61.

(65) McKeown, R. H.; Prankerd, R. J. First thermodynamic dissociation constants of barbituric acid derivatives in water at 25 °C. Part 3. 5,5-Alkylenebarbituric acid derivatives. A comparison with 5,5-dialkylbarbituric acids, and with mono- and di-carboxylic acids. *J. Chem. Soc. Perkin Trans. 2* **1981**, *3*, 481–487.

(66) Krahl, M. E. The Effect of Variation in Ionic Strength and Temperature on the Apparent Dissociation Constants of Thirty Substituted Barbituric Acids. *J. Phys. Chem.* **1940**, *44*, 449–463.

(67) Brandstetter, H.; Grams, F.; Glitz, D.; Lang, A.; Huber, R.; Bode, W.; Krell, H. W.; Engh, R. A. The 1.8-Å Crystal Structure of a Matrix Metalloproteinase 8-barbiturate Inhibitor Complex Reveals a Previously Unobserved Mechanism for Collagenase Substrate Recognition. *J. Biol. Chem.* **2001**, *276*, 17405–17412.

(68) Dunten, P.; Kammlott, U.; Crowther, R.; Levin, W.; Foley, L. H.; Wang, P.; Palermo, R. X-ray structure of a novel matrix metalloproteinase inhibitor complexed to stromelysin. *Protein Sci.* **2001**, *10*, 923–926.

(69) Angyal, S. J.; Angyal, C. L. Tautomerism of N-heteroaromatic amines I. *J. Chem. Soc.* **1952**, 1461–1466.

(70) Rastelli, A.; De Benedetti, P. G.; Albasini, A.; Pecorari, P. G. Physicochemical behaviour of sulpha drugs. Spectroscopic trends and conjugation in phenylsulphonylguanidine derivatives. *J. Chem. Soc. Perkin Trans. 2* **1975**, *6*, 522–525.

(71) Nygård, B.; Olofsson, J.; Sandberg, M. Some physico-chemical properties of salicylazosulphapyridine, including its solubility, protolytic constants and general spectrochemical and polarographic behaviour. *Acta Pharm. Suec.* **1966**, *3*, 313–342.

(72) Bojarska-Olejnik, E.; Stefaniak, L.; Witanowski, M.; Hamdi, B. T.; Webb, G. A. Applications of 15N NMR to a study of tautomerism in some monocyclic azoles. *Magn. Reson. Chem.* **1985**, *23*, 166–169.

(73) Forlani, L.; De Maria, P. Tautomerism of Aminothiazoles. pK_{BH+} Values of 2-Aminothiazoles and of Some Model Imines. *J. Chem. Soc. Perkin Trans. 2* **1982**, *5*, 535–537.

(74) Furet, P.; Meyer, T.; Strauss, A.; Raccuglia, S.; Rondeau, J. M. Structure-based design and protein X-ray analysis of a protein kinase inhibitor. *Biorg. Med. Chem. Lett.* **2002**, *12*, 221–224.

(75) Ho, C. Y.; Ludovici, D. W.; Maharoof, U. S.; Mei, J.; Sechler, J. L.; Tuman, R. W.; Strobel, E. D.; Andraka, L.; Yen, H. K.; Leo, G.; Li, J.; Almond, H.; Lu, H.; DeVine, A.; Tominovich, R. M.; Baker, J.; Emanuel, S.; Gruninger, R, H; Middleton, S. A.; Johnson, D. L.; Galemmo, R. A. Jr. (6,7-Dimethoxy-2,4-dihydroindeno[1,2-c]pyrazol-3-yl)phenylamines: platelet-derived growth factor receptor tyrosine kinase inhibitors with broad antiproliferative activity against tumor cells. *J. Med. Chem.* **2005**, *48*, 8163–8173.

(76) Dingesa, J.; Akritopoulou-Zanze, I.; Arnold, L. D.; Barlozzari, T.; Bousquet, P. F.; Cunha, G. A.; Ericsson, A. M.; Iwasaki, N.; Michaelides, M. R.; Ogawa, N.; Phelana, K. M.; Rafferty, P.; Sowin, T. J.; Stewart, K. D.; Tokuyama, R.; Xia, Z.; Zhang, H. Q. Hit-to-lead optimization of 1,4-dihydroindeno[1,2-c]pyrazoles as a novel class of KDR kinase inhibitors. *Biorg. Med. Chem. Lett.* **2006**, *16*, 4371–4375.

(77) *Epik*; Schrödinger, Inc: New York, NY; http://www.schrodinger.com/. Accessed January 12, 2009.

(78) *Schrödinger Newsletter*; Schrödinger, Inc: New York, NY; http://www.schrodinger.com/newsletter/12/. Accessed September 1, 2009.

CI900501C