

Hi Fidelity Chemistry

Addressing the Concepts of “Structure” and “Compound”

Brian B. Masek,* Robert D. Clark,* Yubin Wu,† Karl Smith* and Robert S. Pearlman†

*Tripos, Inc. and †University of Texas at Austin



The Real World

vs

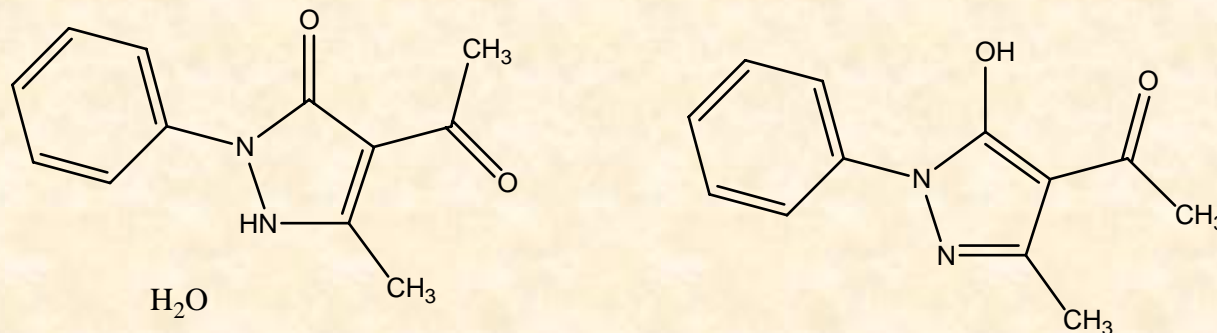
The *in silico* World

- we deal with compounds: real entities
 - cmpd can exist as different structures
 - predominant structure is based on chem. pot. (μ) of environment
 - measured property is a consequence of the structure chosen by Mother Nature
 - typically, we don't know which structure
 - measured data should be associated with cmpd on which it was measured
 - measurements might be wrong due to experimental error
- we deal with structures: CTs
 - cmpd represented by single structure
 - DB curator or chemist determines structure based on “drawing rules”
 - predicted property is a consequence of structure chosen by a human
 - predicted data should be associated with struct on which it was calculated
 - predictions (even from “perfect” SW) will be wrong unless we consider the structure chosen by Mother Nature

Clearly, when using computers to understand or predict properties of compounds, we must address the same range of structures that those compounds can adopt.

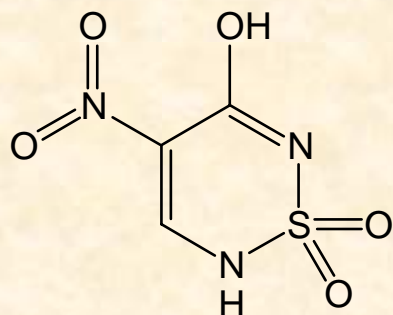
Tautomer Preference – Effects of Environment

- The Cambridge Structural Database* was analyzed to identify “duplicate” compounds.

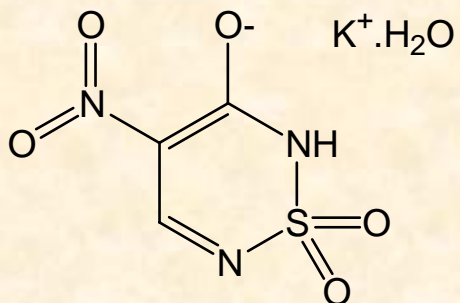


- Examples demonstrate that crystal environment can influence the tautomeric state of a compound.
 - The anhydrous crystal environment favors the enolic pyrazole form
⇒ stabilization by an internal H-bond
 - The crystal environment of the hydrate favors the keto tautomer
⇒ external, not internal H-Bond and COMe group is flipped 180°
 - Bond lengths of both heterocycles support the assigned tautomers

Tautomer Preference – Effects of Environment

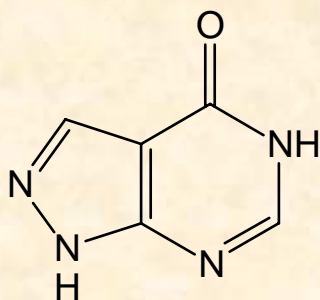


Refcode: HTDZDX10

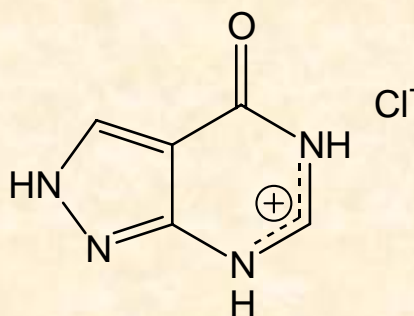


Refcode: BACPUQ10

Deprotonation induces a tautomeric shift



Refcode: ALOPUR



Refcode: FAXFUF10

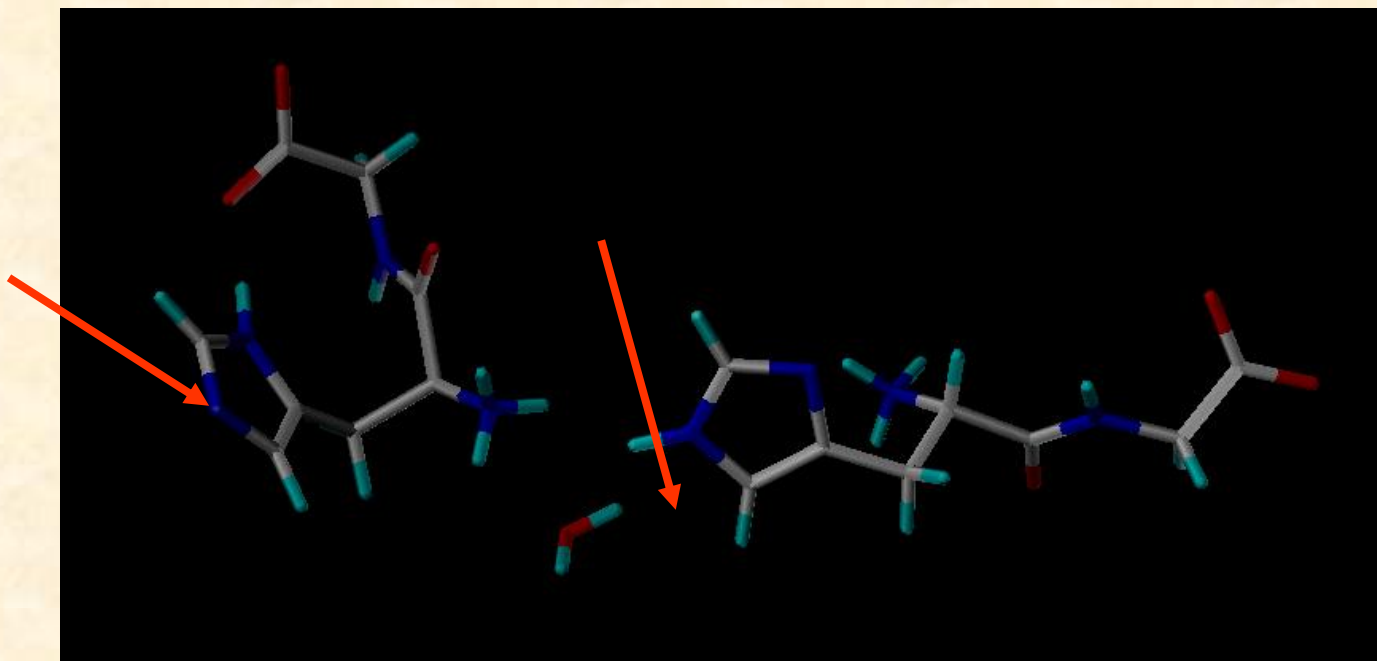
Changing the charge state induces a tautomeric shift in the 5-membered ring

Different tautomers are observed in different environments

Tautomer Preference – Effects of Environment

An Example

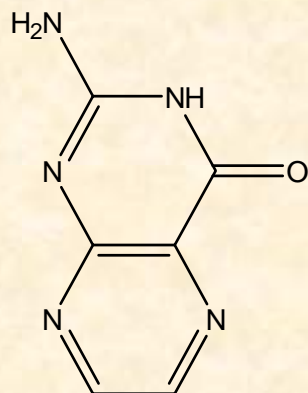
Different tautomers observed in the SAME crystal lattice



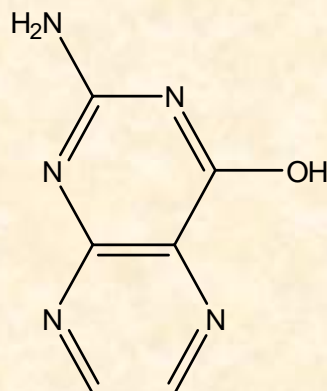
“Coexistence of both histidine tautomers in the solid state and stabilisation of the unfavoured N δ -H form by intramolecular hydrogen bonding: crystalline L-His-Gly hemihydrate” T. Steiner and G. Koellner, *Chem. Commun.*, 1997, 1207.

Example: Ricin Inhibitors - Pterins

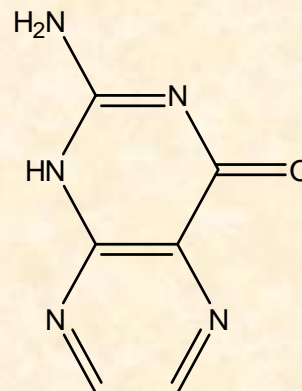
ProtoPlex generates 4 neutral tautomeric forms
(plus additional charged protomers)



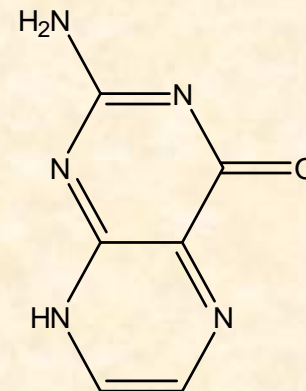
Pterin(1)



Pterin(2)



Pterin(3)

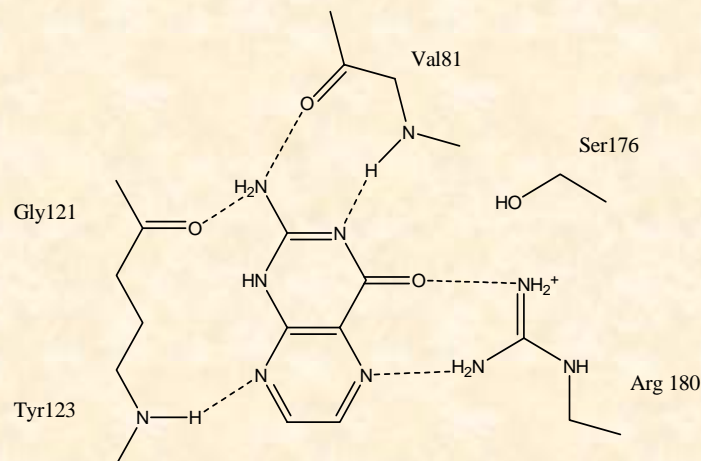


Pterin(4)

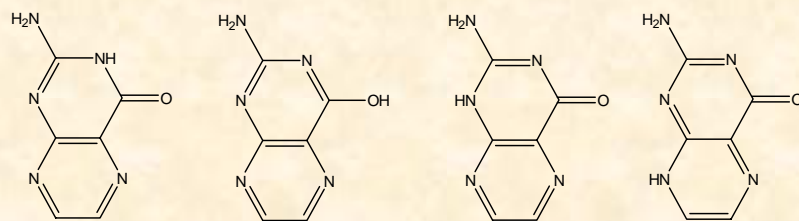
Ionized Protomers not shown

The receptor-bound tautomer (protomer) may not be the preferred tautomer (protomer) in solution

Example: Ricin Inhibitors - Pterins



Redrawn from Wang, et. al, Proteins, 31, 33-41(1998)



Pterin(1)

Pterin(2)

Pterin(3)

Pterin(4)

Ionized Protomers not shown

“A tautomer of pterin that is not in the low energy form in either the gas phase or in aqueous solution has the best interaction with the enzyme.”

S. Wang, et. al., Proteins, 31, 33-41 (1998)

Pterin(1) tautomer was found to be the preferred tautomer in gas and solution

Pterin(3) tautomer was the preferred tautomer in the receptor environment

Why Call It “High-Fidelity Chemistry”?

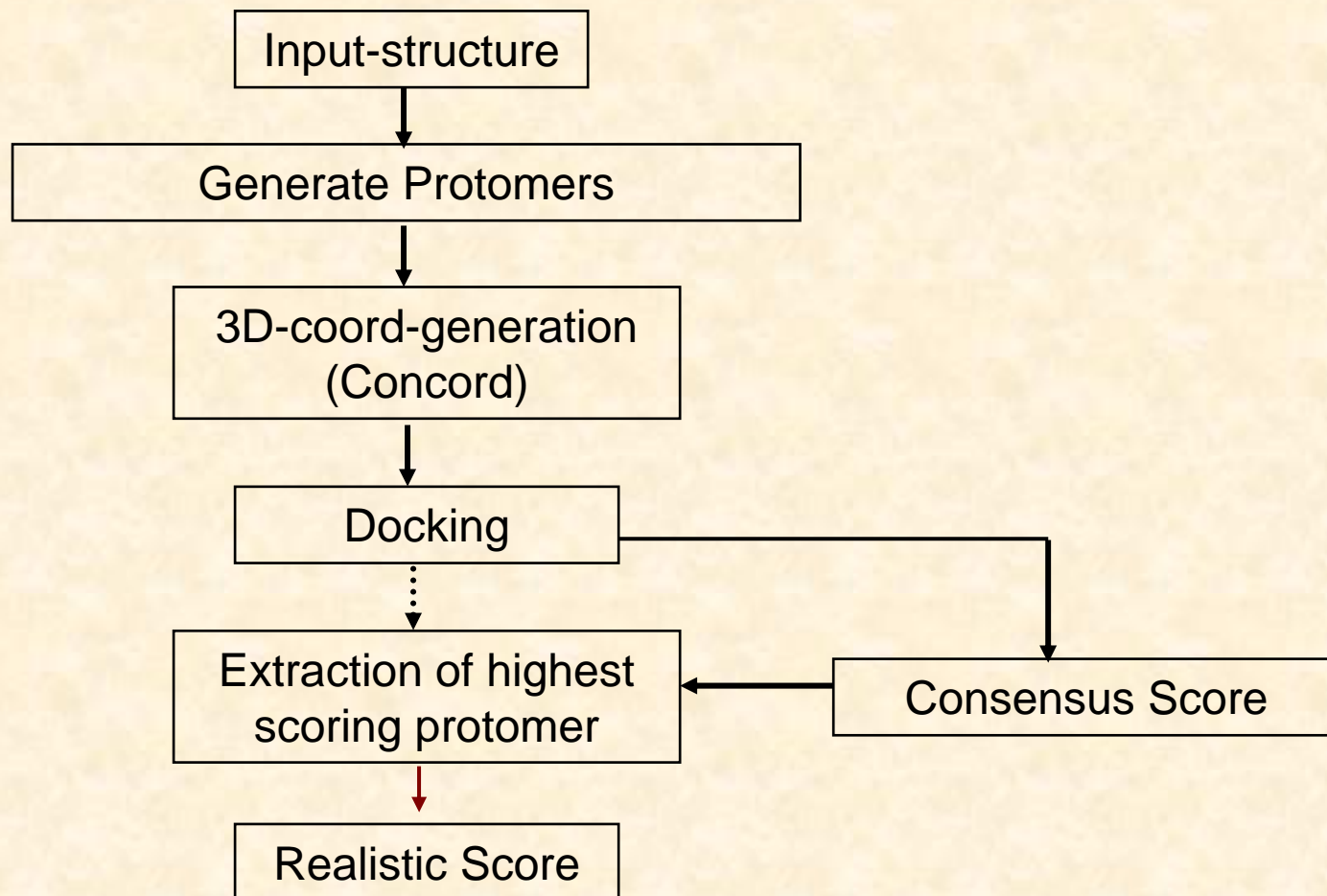
- HiFi *audio* systems reproduce acoustical information with a high degree of accuracy, satisfying the discerning ear of the audiophile.
- High Fidelity *Chemistry* technology is capable of accurately reproducing in chemical information systems all of the various forms of a chemical substance that may be found in a natural environment.

Why Do High Fidelity Chemistry?

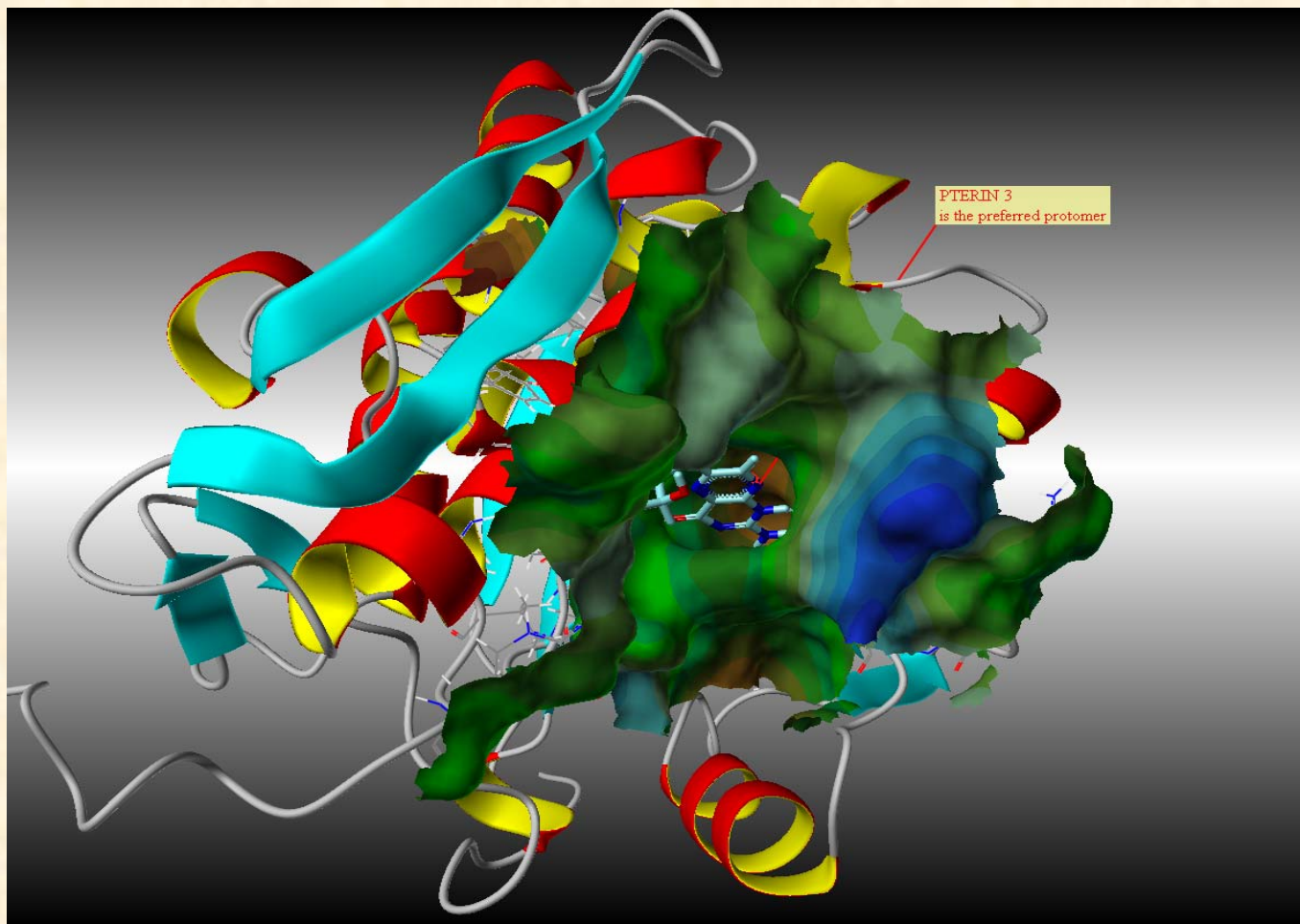
- CADD/QSPR considerations
 - vHTS
 - physical properties
- Cheminformatic reasons
 - data organization
 - lead management
 - avoid duplications
- Intellectual Property considerations
 - prior art
 - avoiding surprises

Application of HiFi Chemistry in CADD

- Example Docking Workflow:



Example: vHTS with FlexX; Input the different tautomers



Other Examples

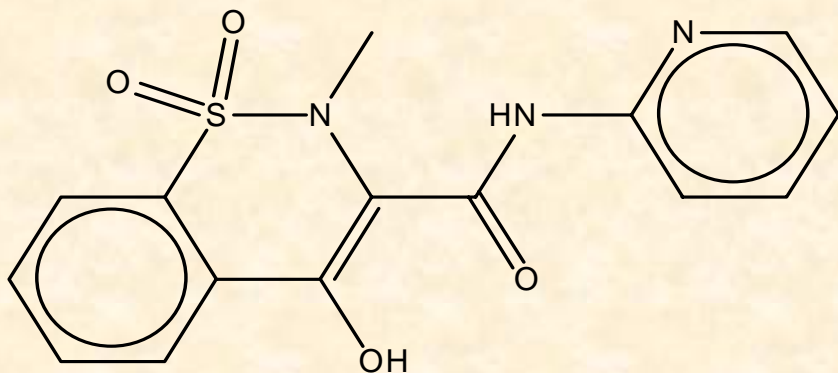
- Kenny and Sadowski looked at several approaches (Permute, Leatherface) for “modifying” structures prior to docking.
 - Several examples of x-ray structure bound to “unexpected” protomers cited.
 - Importance of considering Nitrogen configuration (planar vs. pyramidal)

“There is significant statistical evidence that the chances to reproduce the experimentally observed binding mode increase drastically when taking multiple hydrogen configurations into account.”

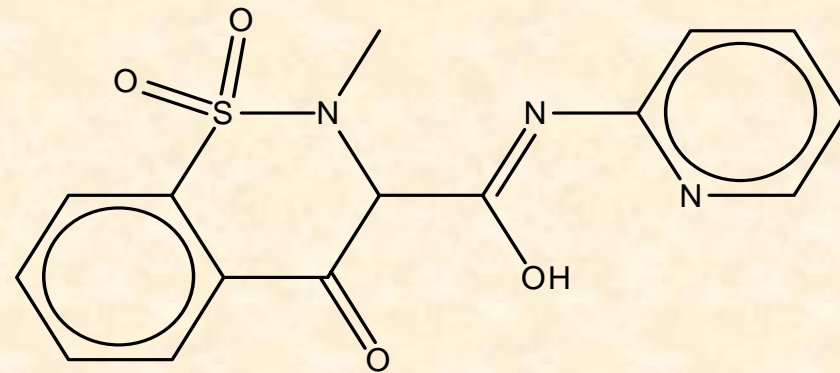
P.W Kenny and J. Sadowski “Structure Modification in Chemical Databases” in Chemoinformatics in Drug Discovery, 2004; Wiley

Improving Property Prediction

Piroxicam



CLogP = 1.90



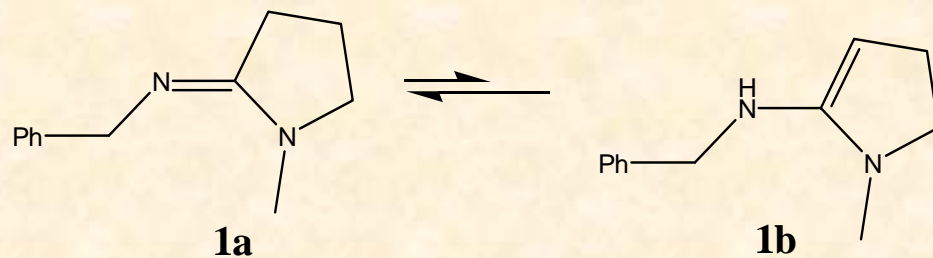
CLogP = 0.60

- which structure is most relevant in water?
 - neither; $pK_a = 6.3$, experimental $\log P$ is 1.0 ± 0.8
- even the best $\log P$ software doesn't treat the unionized species correctly with respect to an easy physical property
- how many existing structure-based models for cuticle penetration, soil mobility, toxicity, environmental fate, etc. adequately address tautomerism?

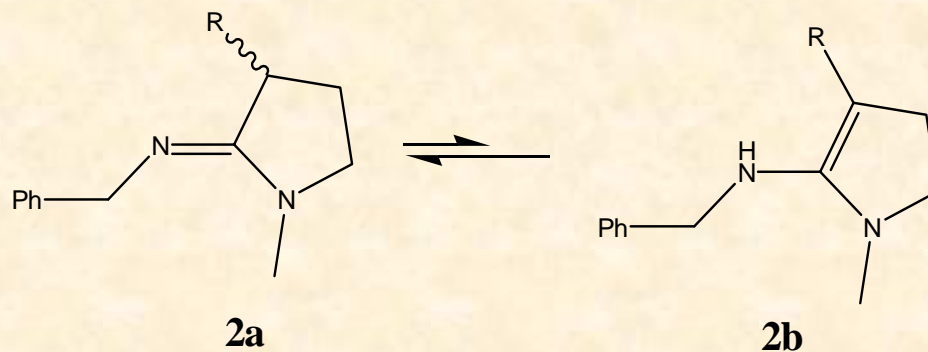
Stereochemical Issues: Proto-Invertible Atoms

- Amidine \leftrightarrow Enediamine tautomerization of **1a** is rapid.

(see M. Pfau, M. Chiriacescu, G. Reviel, *Tet. Lett.*, 34, 327-330, 1993)



- Amidine \leftrightarrow Enediamine tautomerization leads to scrambling of the stereochemistry in **2a**



Stereochemistry

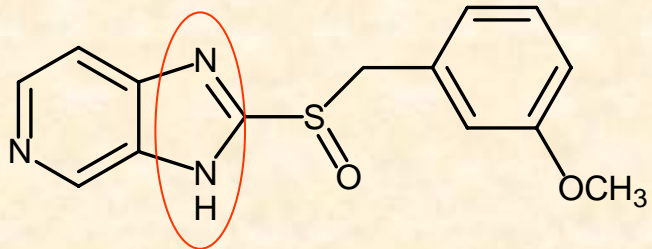
- Tautomeric transforms can change stereochemistry
- Protonation/deprotonation can change stereochemistry
- ProtoPlexing **MUST** be followed by StereoPlexing
- The concepts are inextricably linked \Leftarrow !!

- Cheminformatics and all Computer-Assisted Drug Discovery (CADD) applications should recognize and address the important difference between *truly chiral atoms and bonds* and *protomerically-invertible, pseudo-chiral atoms and bonds*
- Failure to do so will guarantee that computer-based models of chemical behaviors do not accurately reflect the behaviors observed in the real world.

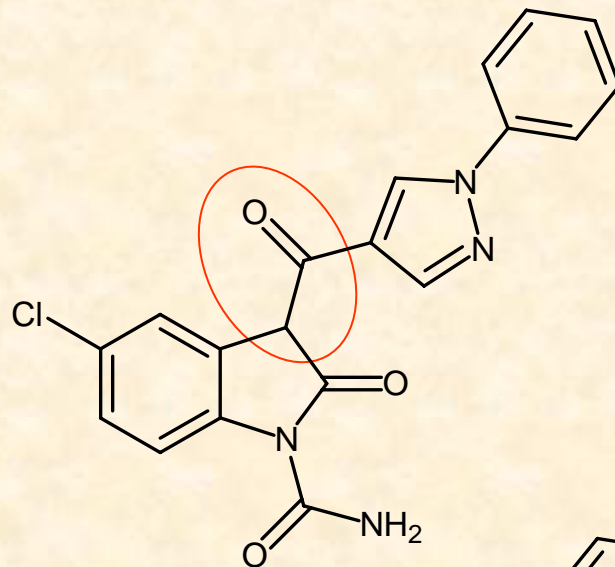
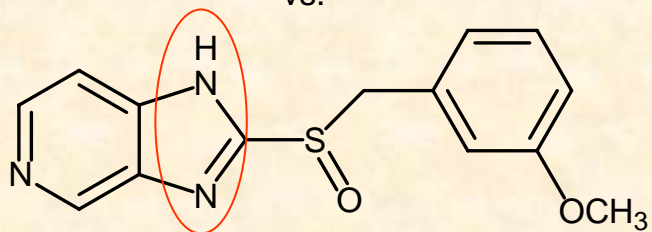
Cheminformatic Reasons for Adopting HiFi Chemistry

- Better storage of data
 - **measured** properties of compound should be associated with the **compound** (with notations re: experimental conditions)
 - **predicted** properties “of a compound” should be associated with (stored under) the particular **structure** used for the prediction
 - each **structure**, in turn, should be associated with a **compound**
 - ⇒ need a **unique identifier** to connect the two
- Better (more robust) results when searching for compounds, data, structures, and substructures
 - *e.g.*, tautomeric substructure searching in database software

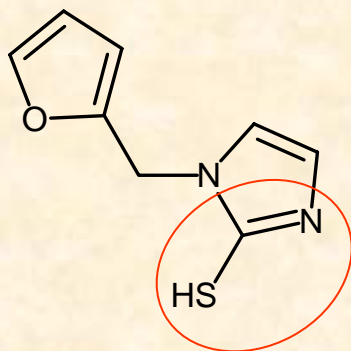
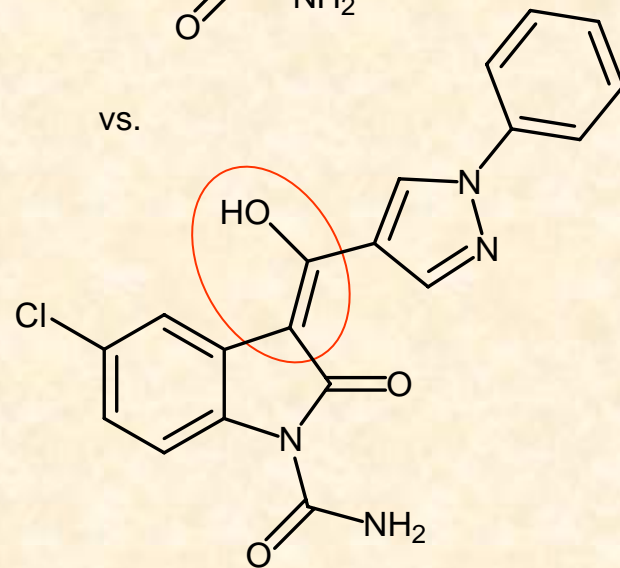
Duplicate Entries Identified in MDDR using Canonical ID



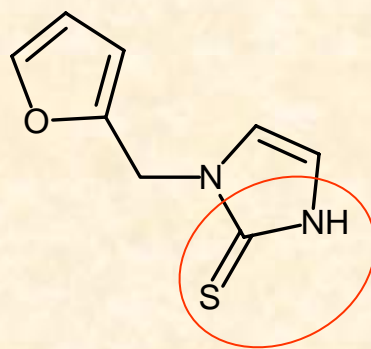
vs.



vs.

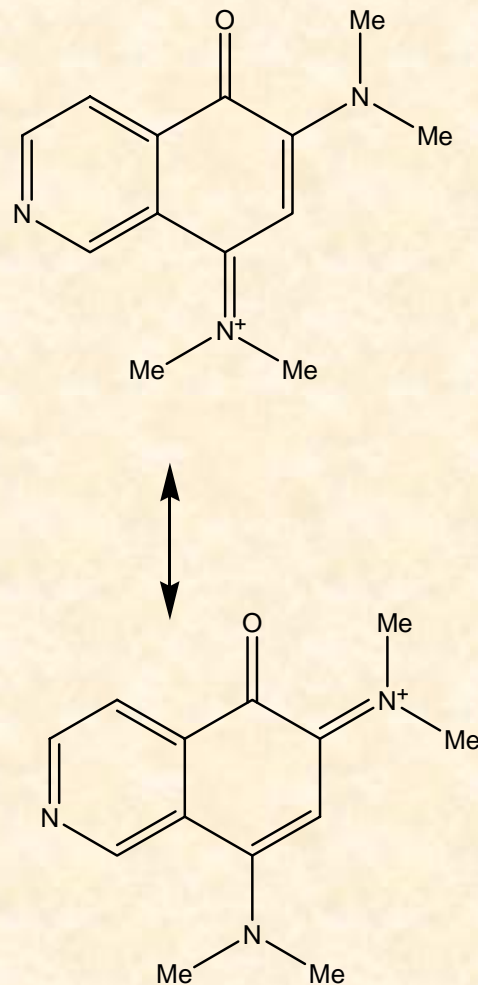


vs.



Resonance Canonicalization

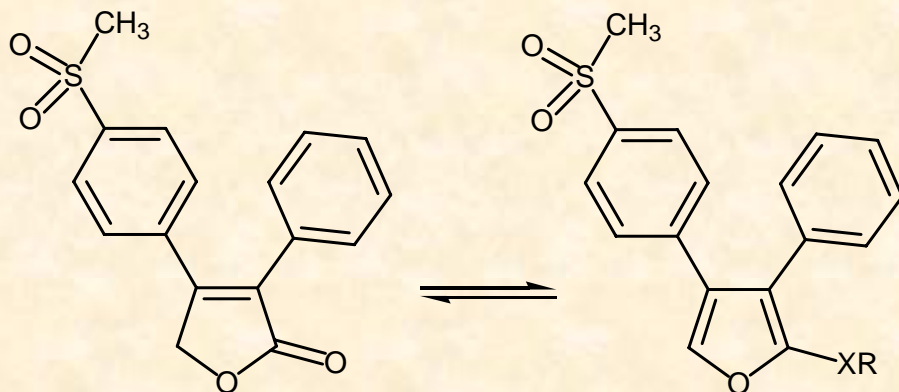
- Which structure did the chemist draw?
- Which structure is in the database?
- A substructure search query for one form will only hit one form and miss the other.
- Which form will be used as the “canonical form?”
- Incompletely handled by High Fidelity Chemistry at this time



IP Reasons for Adopting HiFi Chemistry

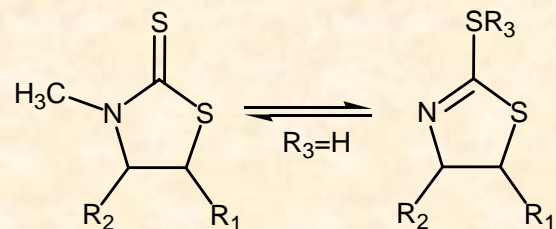
- Substantial R&D investment demands “high fidelity” results when evaluating prior art. Companies simply can’t afford to risk being surprised.
- Many patent claims for chemical compounds to be used as drugs or agrochemicals are open to (mis)interpretation.
- Current methods for identifying whether a compound is covered by an existing patent are incomplete and leave companies open to allegations of infringement.

Intellectual Property Examples



Vioxx (Merck)
EPO 705 254

Pfizer prior art
EPO 679 157



Bluestone
fungicide application
2,860,962 (1958)

Alvord
prior art
1,962,109 (1934)

Three issues of law:

- Infringement by the active principle
- *in situ* (biological) conversion (e.g., metabolism or excretion)
- Infringement during manufacture &or distribution

The Science of High Fidelity Chemistry

- Heuristics need to be fine-tuneable and validated
 - acidic carbons (dicarbonyls, enamines, nitromethylenes, ...)
 - aromaticity (pyridones, hydroxyazoles, xanthenes, ...)
- Need to explore applications to real-world R&D
 - HTS (and vHTS)
 - QSAR => QCAR
- Exploit the scientific synergies with pK_a prediction
 - need to improve pK_a prediction technology
 - ⇒ tautomer generation is critical
 - longer term: improve HiFi Chemistry
 - ⇒ rank protomers by prevalence (quantitatively?)
- Resolve ambiguities inherent in delocalized systems
 - resonance canonicalization

Hi Fidelity Chemistry

Addressing the Concepts of “Structure” and “Compound”

Brian B. Masek,* Robert D. Clark,* Yubin Wu,† Karl Smith* and Robert S. Pearlman†

*Tripos, Inc. and †University of Texas at Austin



High Fidelity Chemistry Requirements

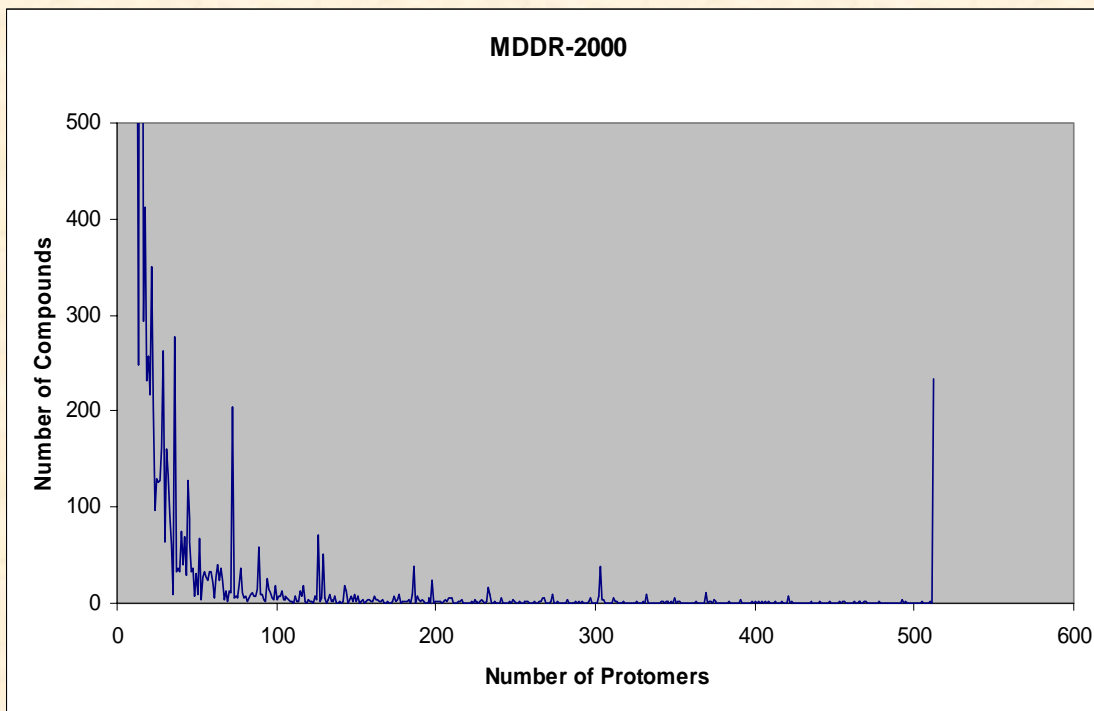
1. Need to **enumerate** the full range of *structures* which a *compound* can adopt
2. Need to **normalize** structures according to user specifications
3. Need a **unique identifier** to tie any *structure* to the *compound* to which it corresponds

Technology to address these requirements has already been implemented by coupling two existing programs - ProtoPlex™ and StereoPlex®

Generating Plausible Protomers: MDDR case study

The 2000 version of MDL's MDDR database was "multiplexed" with ProtoPlex

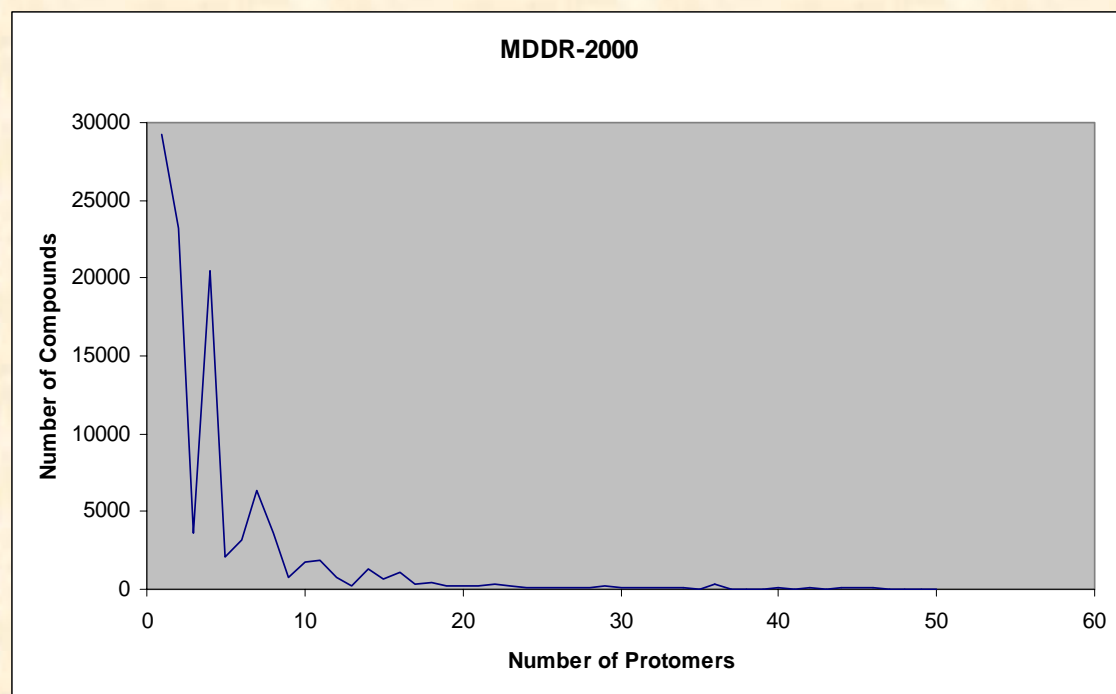
(includes both protonation/de-protonation and tautomerization)



Generating Protomers with ProtoPlex

- 106,591 structures input
- 846,335 protomers generated
- 11 CPU hrs on an SGI Origin 2000
- 0.38 CPU sec/input structure
- 0.05 CPU sec/protomer generated

Generating Plausible Protomers: MDDR case study



- About one-quarter of the compounds have a single protomer
 - About one-quarter of the compounds have two protomeric states
 - The remainder of the compounds have multiple protomeric states

 - Average Protomers per Cmpd = 7.94
 - Average Protomers per Cmpd = 6.83 if the 233 compounds with >512 protomers are excluded.
- Same as previous plot – except
 - Expanded Y-axis range to show large number of structures with <10 protomers
 - Focused on the 1-50 protomers/cmpd region.

Finding Duplicates with HiFi Chem's Canonical Identifier

The 2000 version of MDL®'s MDDR database (106,592 structures) was analyzed using ProtoPlex v2.1

1. Identified "Structural Duplicates" (4680 structures) with DiverseSolutions®

- for example, different salt forms of the same primary fragment
- 2 structures excluded due to errors in the input CT

2. Generated Canonical Identifiers with ProtoPlex v2.1

- 101,910 structures input
- 101,844 Canonical ID's in SMILES Format were generated
- 66 (**0.06%**) – errors producing a Canonical ID (**now fixed!**)
- 30 CPU minutes on an SGI Origin 2000 (63 structures/CPU sec).
- 9 CPU minutes on an HP LC Series with Proliant D140 processor

3. Identified Duplicate Canonical ID's with DiverseSolutions

- 233 pairs of duplicates found
- 43 due to different tautomer structures used for the same compound.
- 190 due to different protomeric (charge) and possibly tautomeric state for the same compound.