"Issues in the interpretation, understanding, and use of Drug Discovery data"

Terry Richard Stouch, PhD

Science for Solutions, LLC
Consulting in Drug Discovery and Design
Practice, Technologies, Process
Princeton, NJ
And
Duquesne University

Optibrium User's Meeting
ACS National Meeting
August 2010, Boston

Alternate titles:

Taking Responsibility: Considering The Uncertainty and Context of Our Data

Data: Its more than a number!

… and one other …

Poignant questions

- Can the experimental data be trusted?

Can the experimental data be trusted?

- Доверяй, но проверяй

Can the experimental data be trusted?

- Доверяй, но проверяй
- Trust but verify!

Can the experimental data be trusted?

- Доверяй, но проверяй
- Trust but verify!
- Trust but understand

The data: do we understand our endpoints?

- What is the known experimental error
  - What sort of error? Experimental, repeated measurements? Multiple trials? Multiple lots
- Precision
- Accuracy
- What was the intent of the data?
- How is the data derived?
- How is the data provided by its originators?
- Do we know where the data came from?
- How do we represent and present our data to end users?
- How do users interpret and use the data?
  - What are their needs?
- Data points could be in error by 10's or 100's of nM
  - Yet people will try to interpret them to the level of their *need* and time constraints

The burden of data in Drug Discovery

Potency is always our first concern

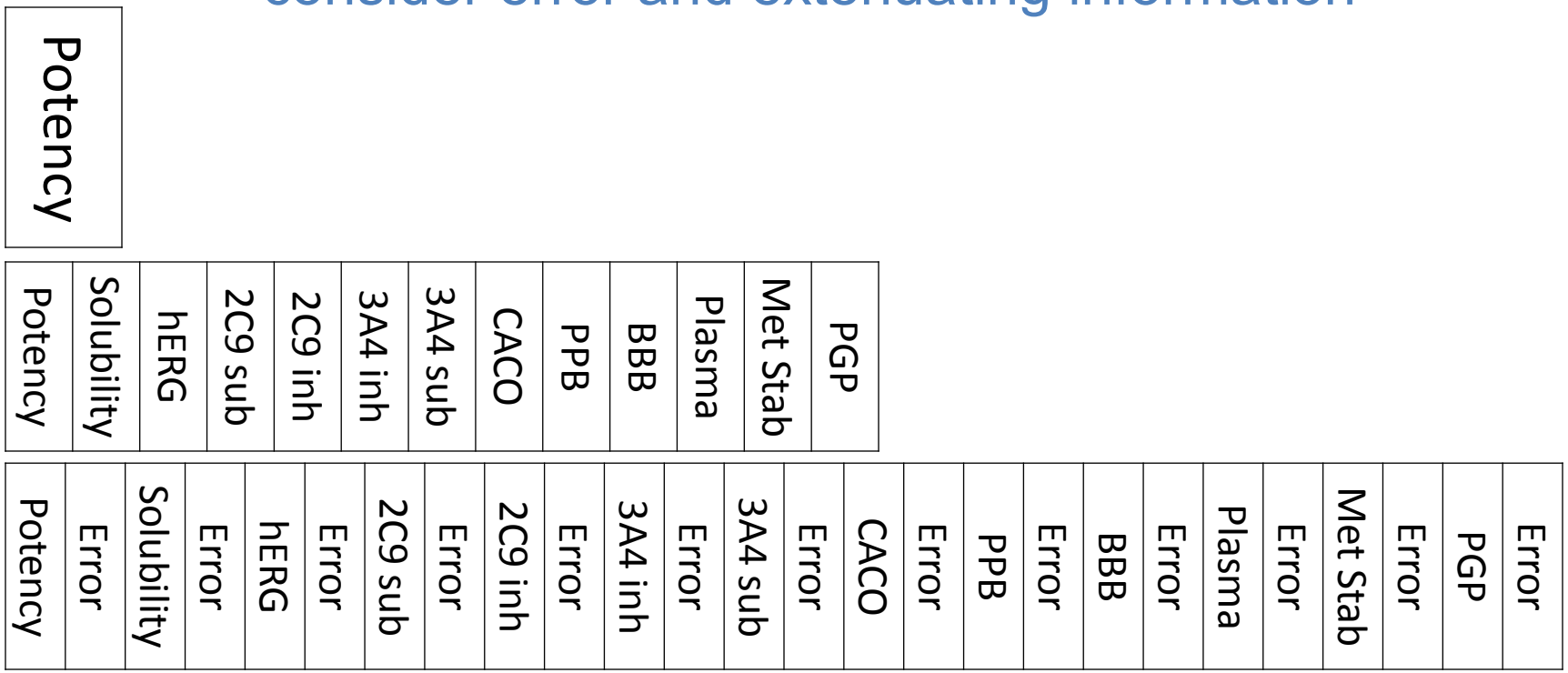
Potency

The burden of data in Drug Discovery

But, there is more than potency to consider

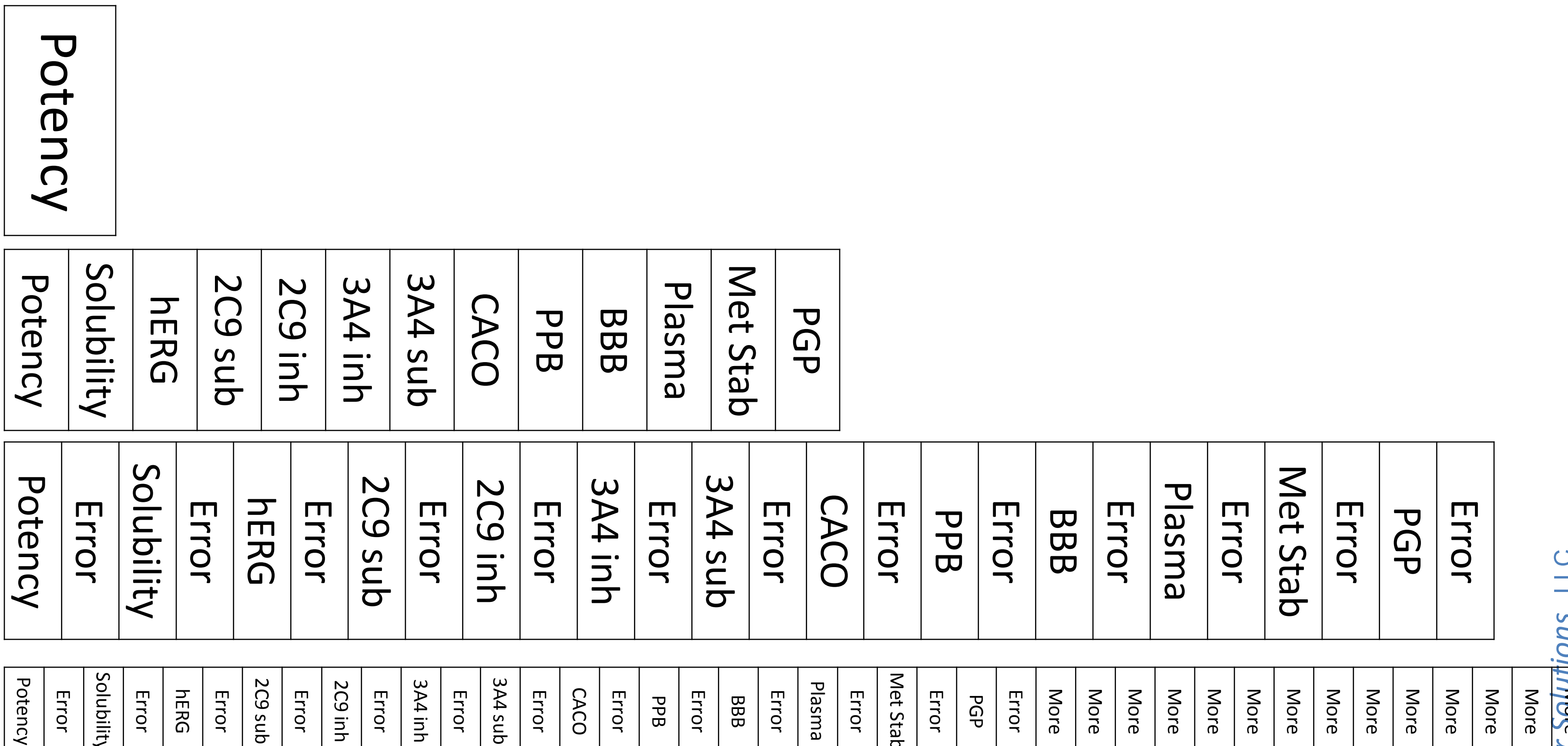
The burden of data in Drug Discovery

Lots more, if we truly deal with our data sensibly and consider error and extenuating information



The burden of data in Drug Discovery

Lots and lots more ….



Science for Solutions, LLC

The Sociology of Drug Discovery

- How do we deal with data, uncertainty, extenuating information, timescales, timelines, goals, inter-relationships?
- Potency might ultimately be of lesser priority than other properties for a marketed drug

Potency	Solubility	hERG	2C9 sub	2C9 inh	3A4 inh	3A4 sub	CACO	PPB	BBB	Plasma	Met Stab	PGP

Potency	Error	Solubility	Error	hERG	Error	2C9 sub	Error	2C9 inh	Error	3A4 inh	Error	3A4 sub	Error	CACO	Error	PPB	Error	BBB	Error	Plasma	Error	Met Stab	Error	PGP	Error

Potency	Error	Solubility	Error	hERG	Error	2C9 sub	Error	2C9 inh	Error	3A4 inh	Error	3A4 sub	Error	CACO	Error	PPB	Error	BBB	Error	Plasma	Error	Met Stab	Error	PGP	Error	More	More	More	More	More	More	More	More	More	More	More	More	More	More

Importance of meta and extenuating data

- Cyp 2C9 inhibition data: triaged by expert
- 25,000 data points in the corporate database
    - 10 years, over 100 drug discovery projects
  - Minus fluorescent complications
  - Minus poor solubility
  - Minus time span where instruments were finicky
  - Minus "difficult" programs
- Was reduced to 5000 *irrefutable* data points
  - 4 significant figures in range of 0 – 100%
  - 'Expected' error of 5 – 10%

Expert knowledge – non-databased – was required to truly understand the data

- *In silico* model: 75 – 78% correct prediction of +/- @ 5 uM
  - As good as experiment ! (?)

More examples of meta and extenuating data

- Caco-2 data
  - In house corporate database data not sufficient rigorous to support *in silico* models – (experimentalist who generated the data)
    - Although the assays were performed using industry standard protocols
      - Variable cell lines
      - Too 'high throughput'
  - $2000+ per assay for very rigorous results by CRO
- Solubility
  - High-throughput DMSO precipitation assay, very crude results
    - Essentially shows probability of *insoluble*
    - (FYI Reported as molar solubility to 4 significant figures)
  - Profiling Governance recommended "red flag"
    - Users protest in favor of 4 significant figures based on "need"
    - Interpreting data based on need
  - Assay discontinued and replaced
    - Now very labor intensive, but much more accurate

Interpreting crystallographic density

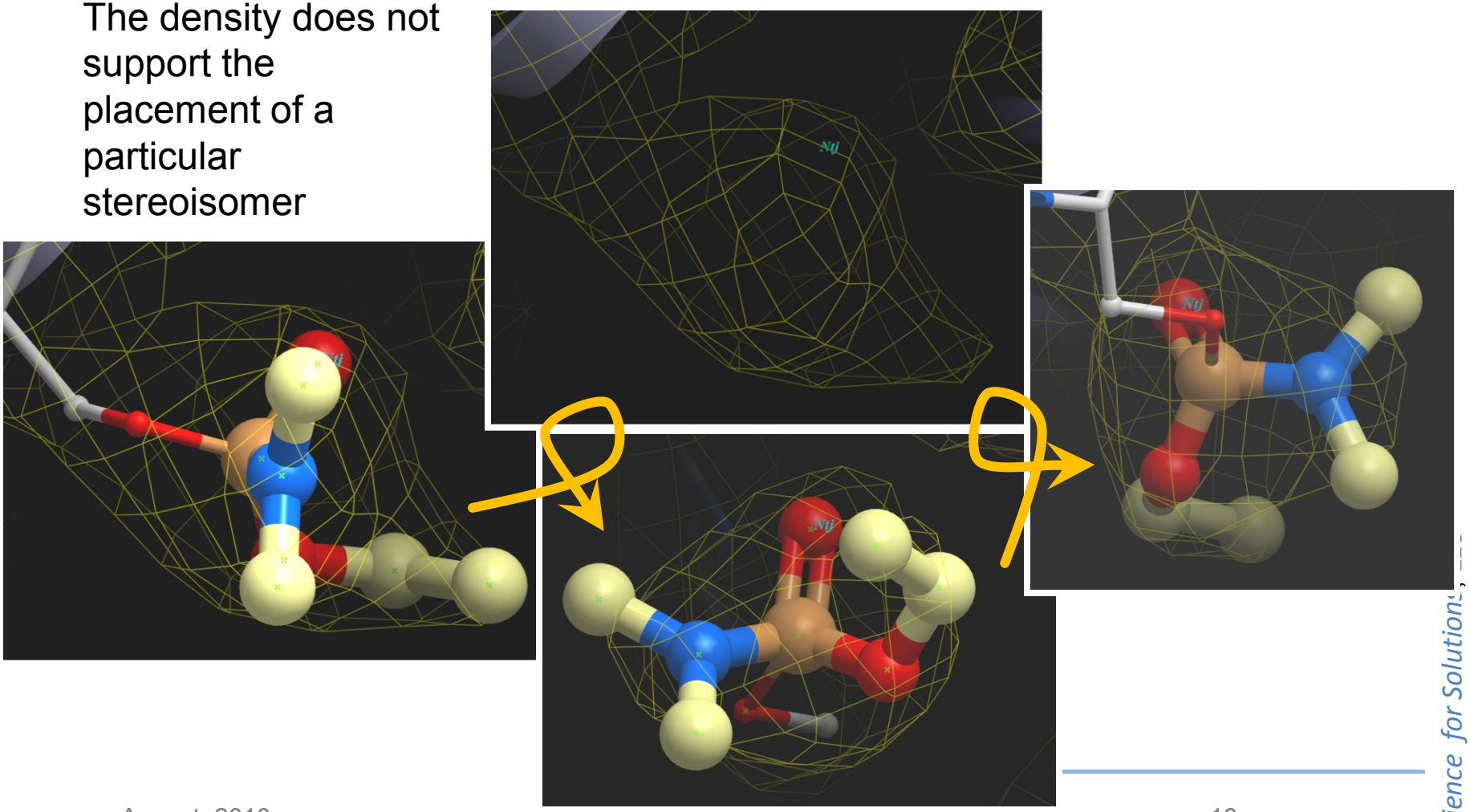
- Placing ‘known’ ligands in density: e.g. 2C0Q, mAChE, tabun inactivated

```
HETATM 8339  O1  NTJ A1543      29.671  15.833  13.782  1.00 39.49           O
HETATM 8340  P1  NTJ A1543      28.716  16.193  12.771  1.00 43.09           P
HETATM 8341  N1  NTJ A1543      29.223  16.110  11.389  1.00 43.95           N
HETATM 8342  C2  NTJ A1543      30.231  15.150  10.956  1.00 44.33           C
HETATM 8343  C1  NTJ A1543      28.682  16.986  10.364  1.00 43.94           C
HETATM 8344  O2  NTJ A1543      28.140  17.492  12.991  1.00 46.49           O
HETATM 8345  C3  NTJ A1543      29.014  18.484  13.584  1.00 47.79           C
HETATM 8346  C4  NTJ A1543      30.150  18.995  12.702  1.00 43.76           C
HETATM 8347  O1  NTJ B1544       9.514   0.380 -38.225  1.00 45.16           O
HETATM 8348  P1  NTJ B1544       9.117   0.166 -36.869  1.00 46.90           P
HETATM 8349  N1  NTJ B1544      10.285   0.016 -35.965  1.00 49.27           N
HETATM 8350  C2  NTJ B1544      11.648   0.356 -36.335  1.00 47.34           C
HETATM 8351  C1  NTJ B1544      10.101  -0.476 -34.611  1.00 50.09           C
HETATM 8352  O2  NTJ B1544       8.214  -0.948 -36.770  1.00 51.80           O
HETATM 8353  C3  NTJ B1544       8.791  -2.220 -37.151  1.00 54.80           C
```

Interpreting crystallographic density

- Placing ‘known’ ligands in density: e.g. 2C0Q, mAChE, tabun inactivated

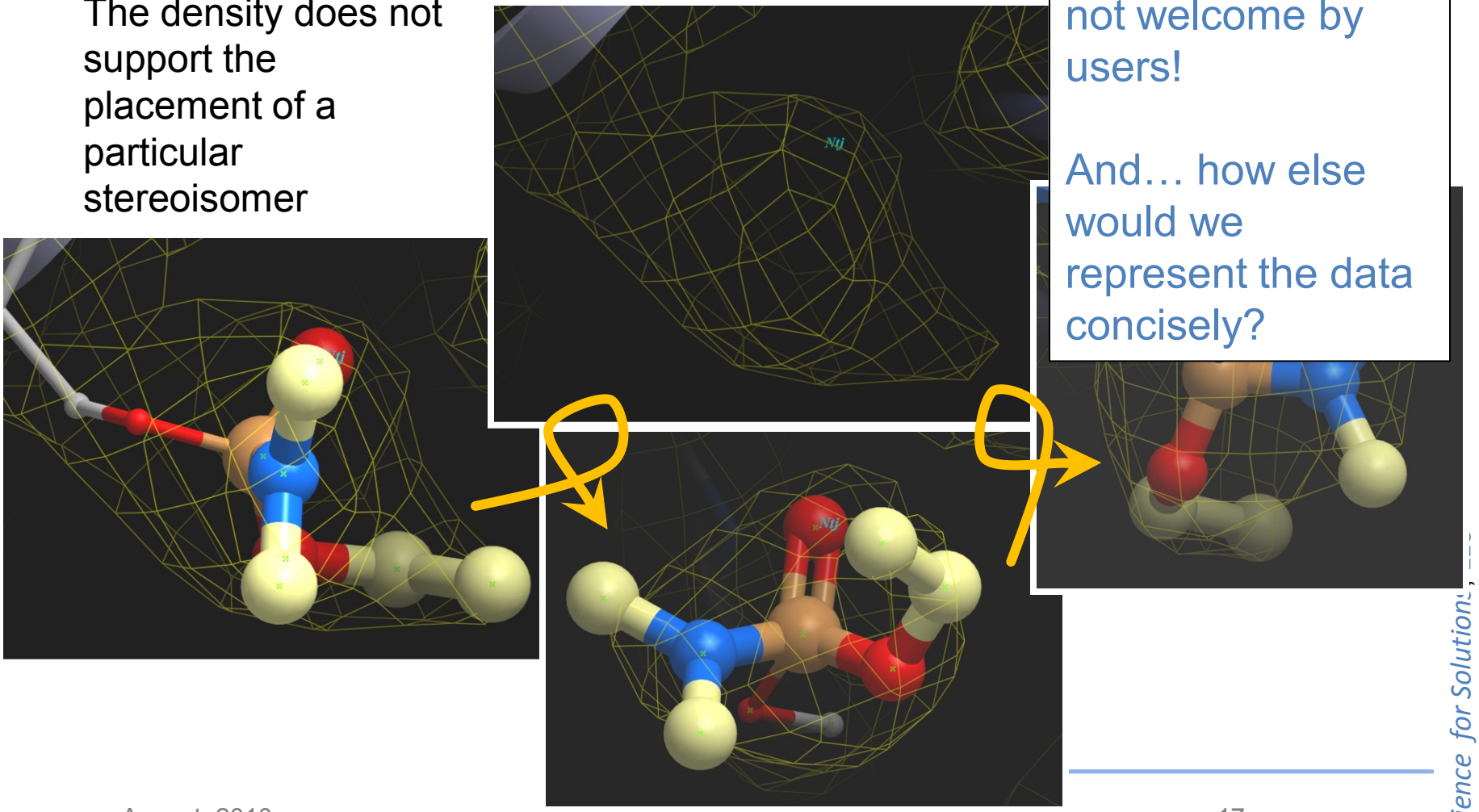
The density does not support the placement of a particular stereoisomer



Interpreting crystallographic density

- Placing ‘known’ ligands in density: e.g. 2C0Q, mAChE, tabun inactivated

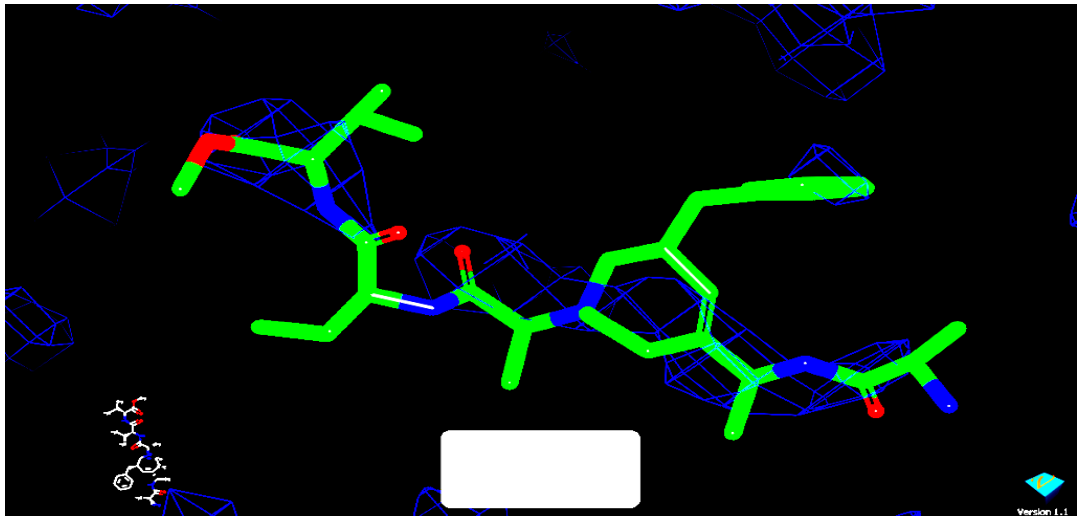
The density does not support the placement of a particular stereoisomer

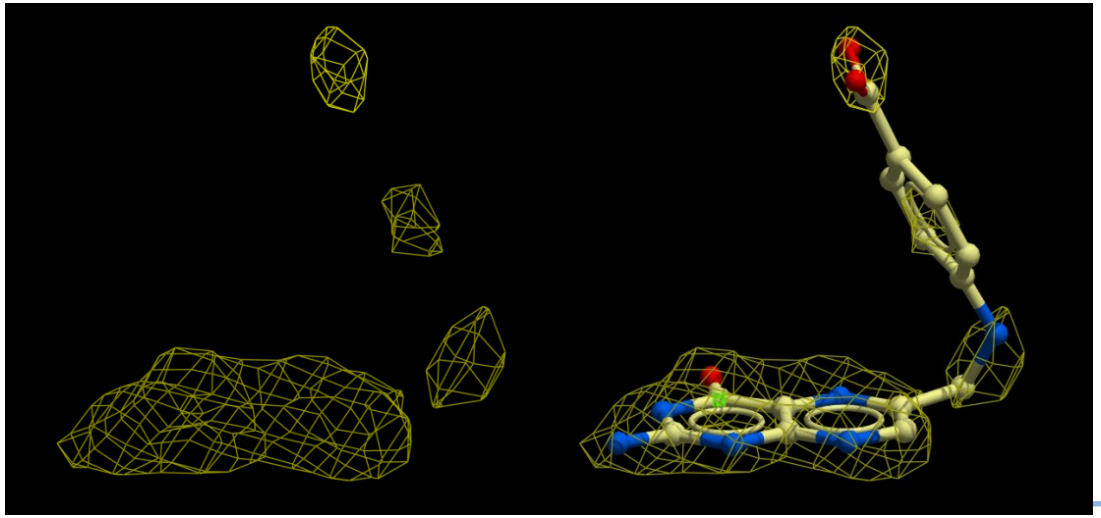


But, ambiguity is not welcome by users!

And… how else would we represent the data concisely?

Ligand placement and conformation in crystal structures

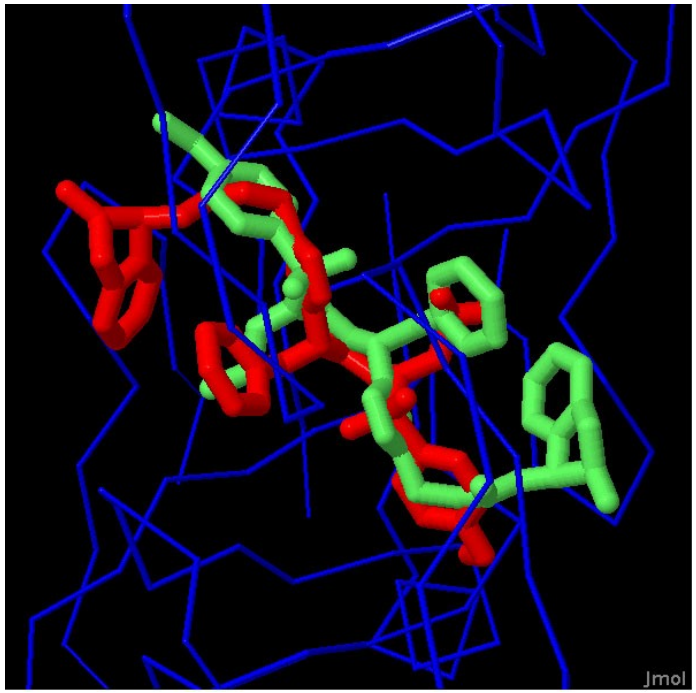




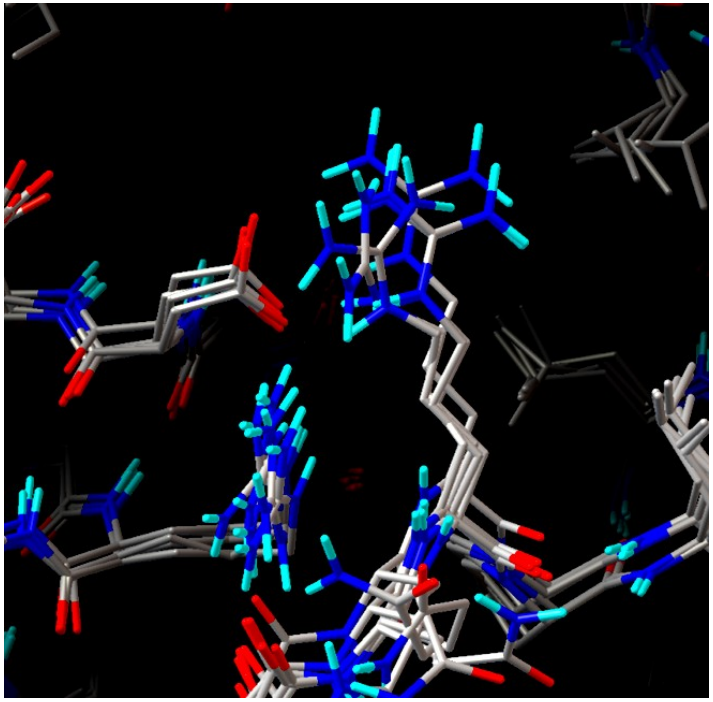
Insufficient density means that the final structure is determined by modeling and the weight it is assigned during crystallographic refinement.

Probably these are good approximations of reality. But there could be alternate interpretations.

Multiple conformations and interpretations



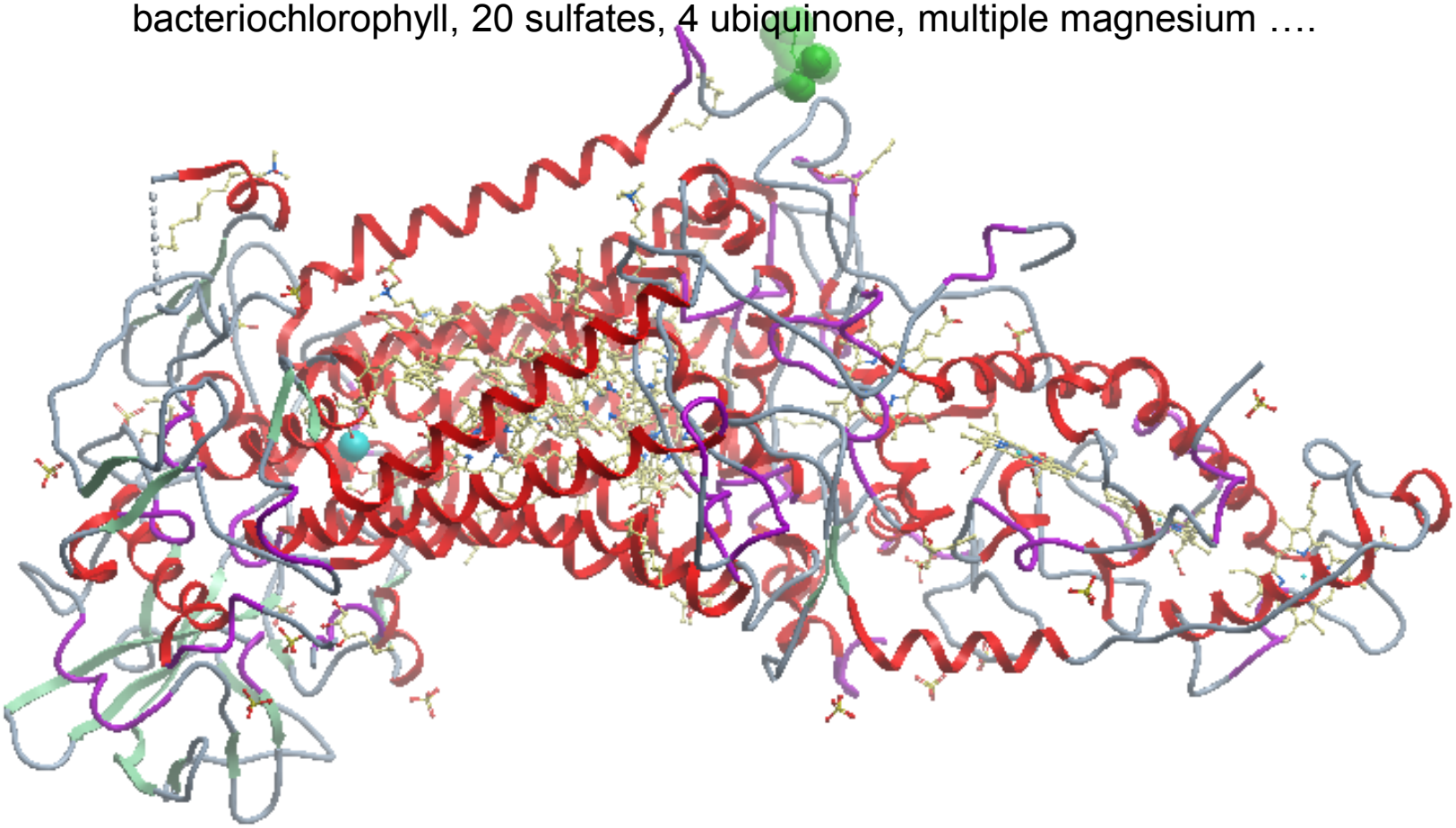
HIV Protease Inhibitor: Dual occupancy: 2 copies, one structure



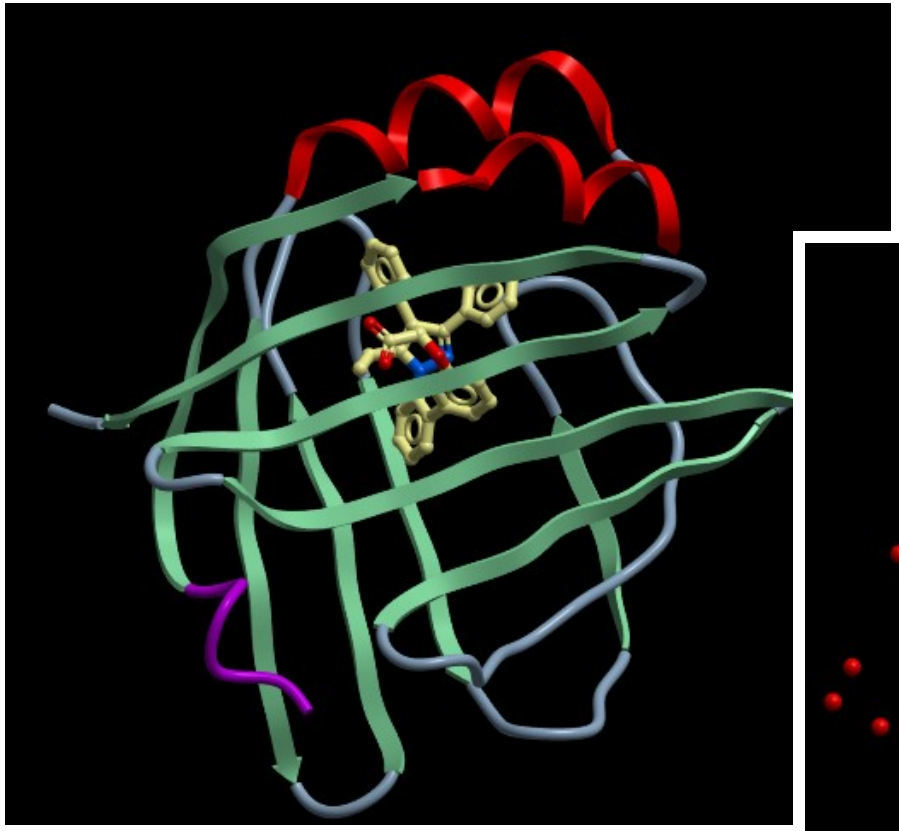
Arg8 in HIV Protease: Multiple copies in the asymmetric unit

Intent and scrutiny

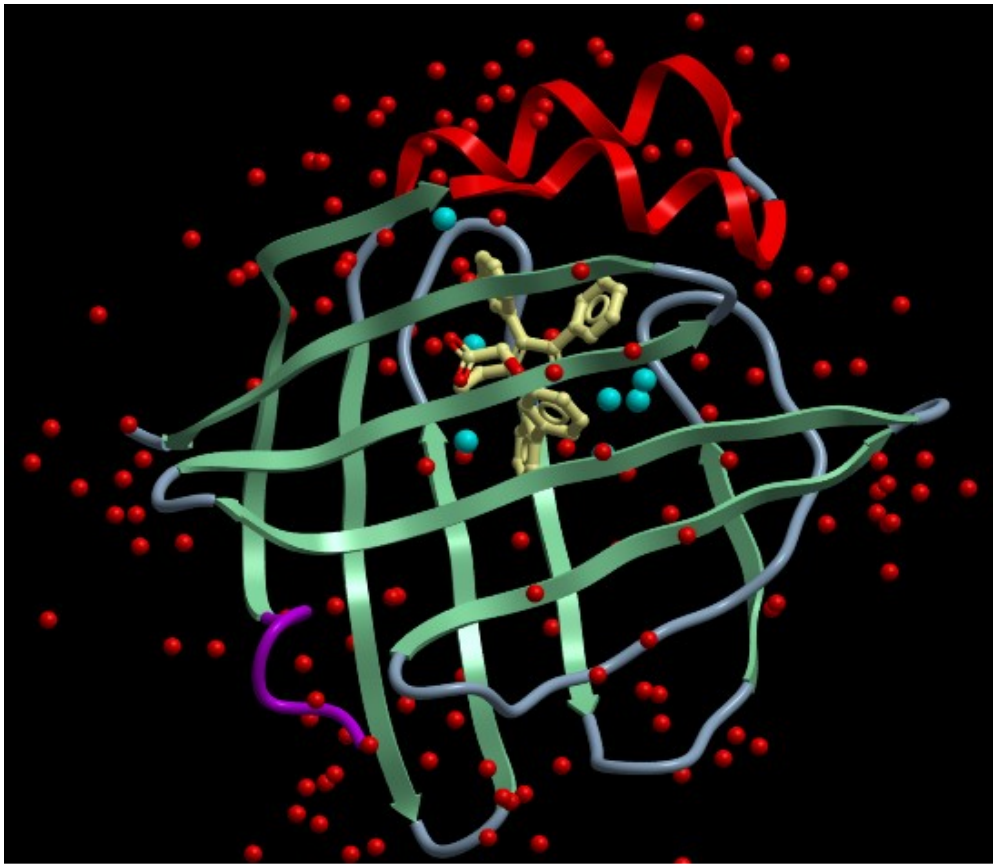
- "Error" in one bond placement
- Intent: One of the first structures of the photosynthetic reaction center!
  - Some perspective: one atom or one bond in 11,000+ atoms on the N-terminal residue of PRC Cytochrome C with 4 protein chains, 4 hemes, 4 bacteriochlorophyll, 20 sulfates, 4 ubiquinone, multiple magnesium ….



Intent: Ligand is well determined; external waters are not



The ligand was the intent and purpose of the crystallography. No benefit to spending time on external waters.



Problems: Training in statistical and data analysis

- Comp chemists are often not formally trained in statistics
- Statistics is and data analysis is not a trivial field
  - Yet it is often treated as such in chemical literature

Current practice of statistics on and modeling of chemical data

- Does *any* value of R**2 signifies variance explained? 0.3, 0.4,…..,
- No mention of error
- Interpretation of coefficients in co-linear data
- Ignoring multi-colinearity
- Development of "new" methods
- Everyone wants to be a hero! (Failed analysis do not get kudos)
- Following (poorly) the rules (of thumb)
  - Regression:
    - EG: study that selected12 variables for 80 compounds only because it was within the d:N =1:5 rule of thumb
      - But: d=N is trivial and the probability of chance increases as d approaches N
      - The minimum number of variables should be used
- "Its not statistically significant, but let's just look at the trends"
- "Let's just see if it works"

How does over-interpretation of data affect analysis?

- As modelers, are we being too hard on ourselves by forgetting the uncertainly in our data?
- Uncertainty might be of a level to obviate the value of the data for the particular need
  - Eg: data of 50, 80, 100 nM but with ± 300 nM error is not significantly different
- We might be weighting some experimental data too highly
- Error bars on experiment can be very large, although inconvenient
- Learn to accommodate a “fuzzy” interpretation of the data
  - But, how is this affected by limited data?

Recommendations for software developers

- Always leave a data field for error/uncertainty
  - Perhaps one for *concern?*
  - Query user for error on input of data
  - Ask for expected precision
- Supply appropriate significant figures
  - Query user when inappropriate
  - Don't "Excel" the data
  - Ex: crude solubility assay data reported12.455782 µM !
- Help the user understand their data
  - Value of 24.752 ± 189.293 µM
- Ligand crystal density as a default view if possible
  - Ligand conformational energy provide with cautionary statement
  - Input all occupancies and all alternates and present to the user to choose
  - Include additional information

Paths forward for chemical data *modeling* (if not statistics)

- Spend time with the data
  - EDA: Exploratory data analysis
  - Explore reasonable data spaces with multiple approaches
    - Cluster analysis, PC plots, RP Trees, on and on
  - Look for structure (SAR) in the data space, examine related data
- Provide value
  - Extra information
  - Table look up is not evil –if you've got the number, show it
  - *Show* how the model works on related compounds
  - Show the training set compounds
  - Estimates of error
- Be aware of error in the data
  - Maybe your model should *not* fit all of the data
  - But it should not be an excuse to through out compounds

Understanding our data

- Uncertainly and error
  - What sort of error?
    - Experimental, repeated measurements? Multiple trials? Multiple lots
- Precision
- Accuracy
- Intent
- Derivation
- Origination: Where did the data came from?
- Represent and presentation
- Interpretation and use
- Need

Can the experimental data be trusted?

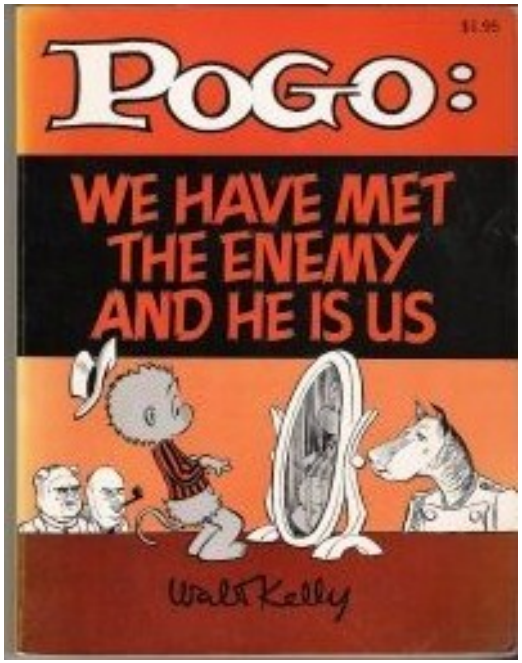
- Alternate title:

The Errors of our Ways

Can the experimental data be trusted?

- Alternate title:

The Errors of our Ways



Acknowledgments

- Cyp 2C9 studies
  - Litai Zhang, BMS
  - Ken Santone, BMS
  - Mike Sinz, BMS (?)
- Crystallographic analysis
  - Greg Warren, OpenEye Scientific Software

End of presentation

Terry Richard Stouch, PhD

President, Science for Solutions, LLC

Consulting in Drug Design; Pharmaceutical Research, Technologies, Process; Molecular Simulation; Computational Sciences; Structural Biology

Duquesne University, Adjunct Professor of Chemistry and Biochemistry, Bayer School of Natural and Environmental Sciences,

The University of Kentucky, Adjunct Professor, Department of Pharmaceutical Sciences, School of Pharmacy

Senior Editor-in-Chief, Journal of Computer-Aided Molecular Design, Springer Publishing

Protein Data Bank, Research Collaboratory on Structural Bioinformatics

AAAS Fellow, IUPAC Fellow

tstouch@gmail.com

1-609-275-7234