Experimental (T)error:
Dealing with data in an uncertain world

Terry R. Stouch
Science for Solutions, LLC
Consulting in Drug discovery and design

John D. Chodera
Computational Biology Program, Memorial Sloan-Kettering Cancer Center
http://www.choderalab.org

Slides, references, and IPython tutorials available at http://choderalab.org/cadd-grc-2013

Motivation

How can we, as modelers, deal with data error in a useful and reasonable manner?

How should error affect our interpretation and use of the data?

Our premise:

Realistic assessments of error can save you time, money, and sanity by ensuring you are chasing signal, rather than fitting the noise

The unfortunate truth:

Our field currently lacks standard practices for dealing with error.

Introduction to error

- Endemic
- Ignored
- Hidden
- Substantial
  - Could obviate the use of the data
- High precision / low error is expensive
  - Tremendous care and expertise might be required

Introduction to error

- Sometimes people – users of the data - just don't care!
  - "I don't care if the measurements aren't quantitative, I need to see several significant figures"
  - "I don't care if the answer is right, as long as its fast"
  - ~"If someone tells me they don't trust a number, I'll tell them to get out of my office and not return until they get it right."
  - ~"Just give me the numbers, I'll know how to interpret the data."
  - ~"Give me your data, I can get a model that will fit it to 100%"

"Its all automated now. The QSAR problem has been solved"

Closed-Eye Scientific Software, Inc

The up side of this talk!

- Understanding error helps us avoid and accommodate it resulting in
  - Better decisions
  - Better models
  - More appropriate use of the data and the models
- It might lead to better experiments
- Models can replace high error experiments
- Experimentalist could concentrate on producing high quality and novel data

Outline

What are the sources of experimental error?

How do we measure and report error?

How can we estimate error if not measured or reported?

When is it safe to use literature data?

What level of error is needed for data to be useful?

How do we propagate error?

Can we really get rid of outliers and feel good about ourselves in the morning?

When is it better to stick with a model than to collect data?

Some rules of thumb (for when you have no other information)

Pitfalls, catastrophes, and tales of woe

Slides, references, and IPython tutorials available at http://choderalab.org/cadd-grc-2013

What we aren't going to cover

Errors in structural data (X-ray and NMR)

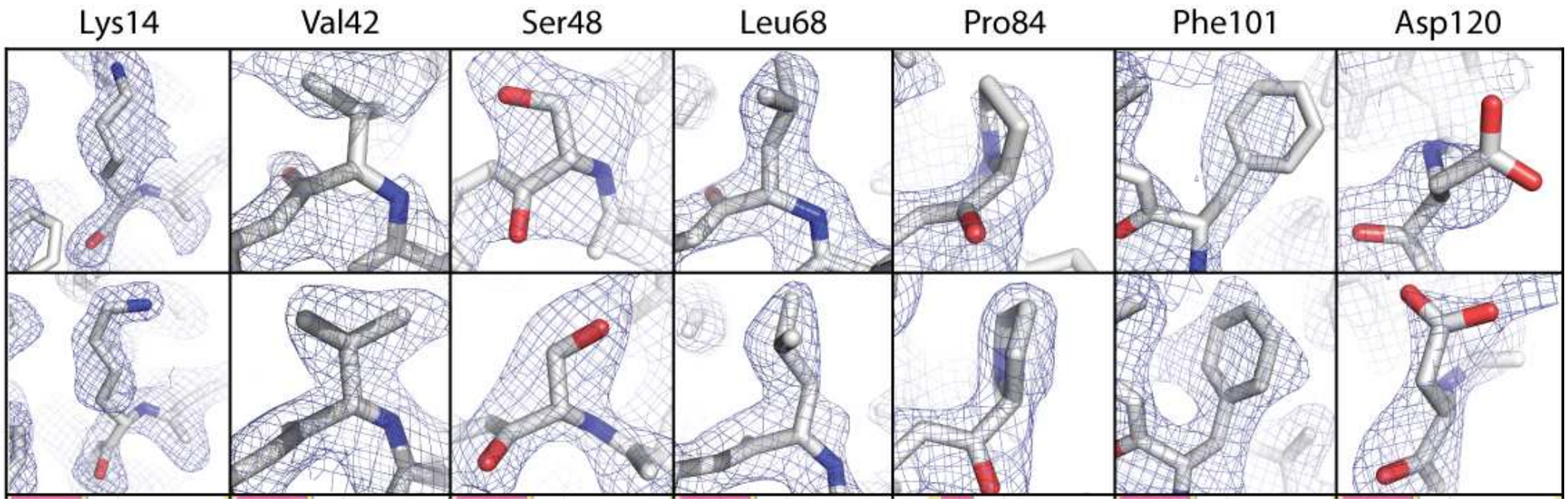
Some brief warnings:

- X-ray and NMR structures are just models derived from data

- often the models don't capture important aspects of the data

- all data-derived models have uncertainty: know how to quantify this

- refinement errors can be more frequent than you might expect



PDB:155c

Bioinformatics 15:3392, 2011

Genesis of the data

- Likely it was not generated for long-range data mining
- Largely generated for a specific need or program at a specific time
  - Consistency might be greater internal to a project and during a particular temporal period

Different sources of error to watch for

- Concentration of *desired* substance
  - Purity
    - Age
    - Stereochemical purity
    - Effects of impure fraction
  - Weighing precision
  - Solubility
  - Reliability of the determination
- Source of material
  - In house, recent synthesis vs. purchased compound collection
- Source of the data
  - HTS
  - “Program” data

Different sources of error to watch for

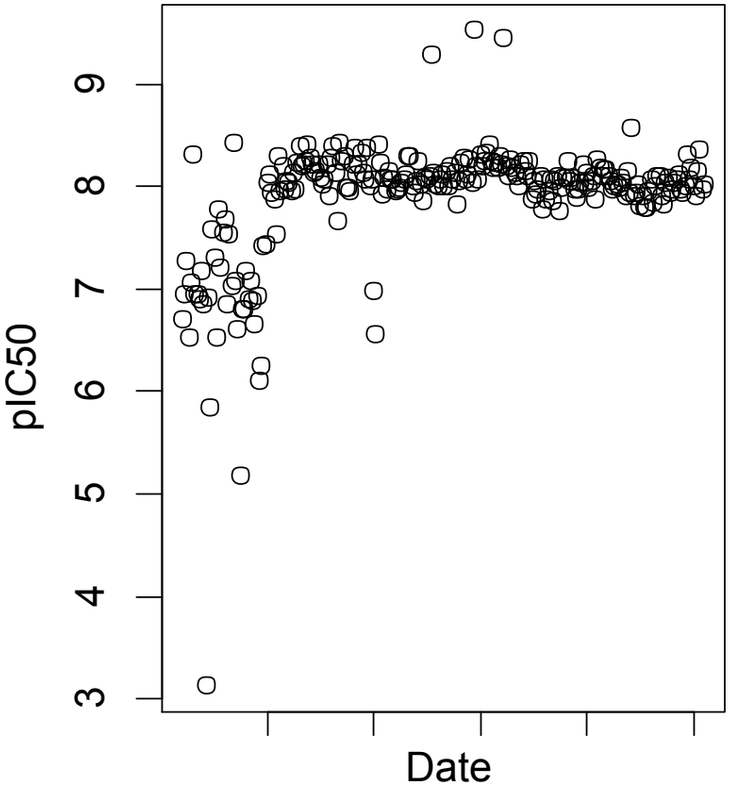
- Different Standard Operating Procedures (SOP) same company
  - The same ultimate endpoint arrived at with different methods
    - Eg. LC/UV vs LC/MS for measurement
    - Weight / volume to determine concentration vs. MS or NMR

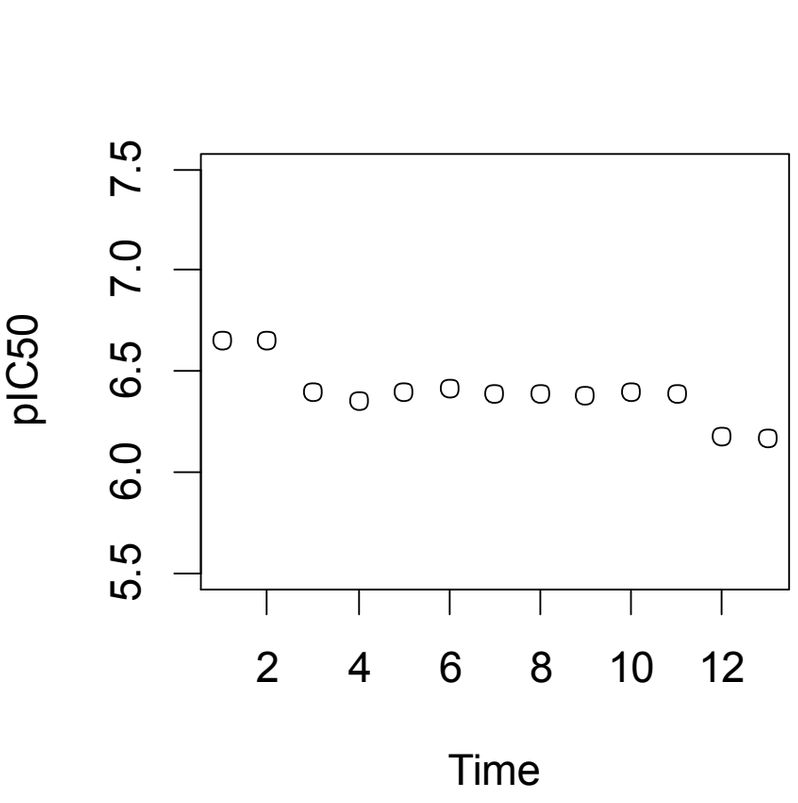
- Same SOPS same company but different: labs, solutions, samples, days, equipment, technicians, water, humidity, temperature, lab coats, time of day, ….
  - Veterans of assays: expect 2X variation at least, 5X not unexpected, 10X no contention

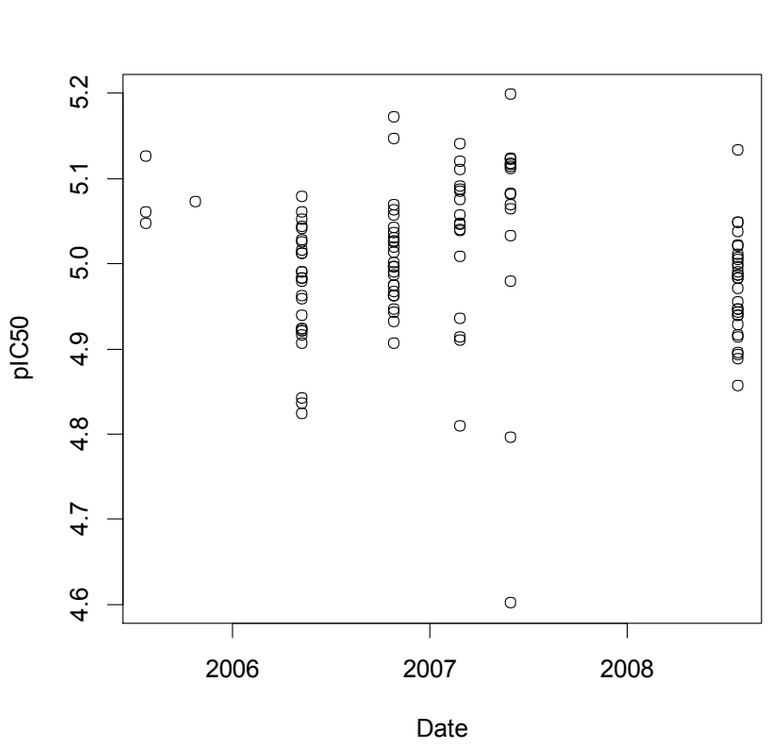
- Different SOPs different institutions
  - Can it be any better than same SOP same company different labs?

Different sources of error to watch for

- How are standards used?
  - When bad values are obtained, is that plate / series ignored? Is the data still saved?
  - If the data is saved, is it flagged? Can data from ‘bad’ plates be identified?



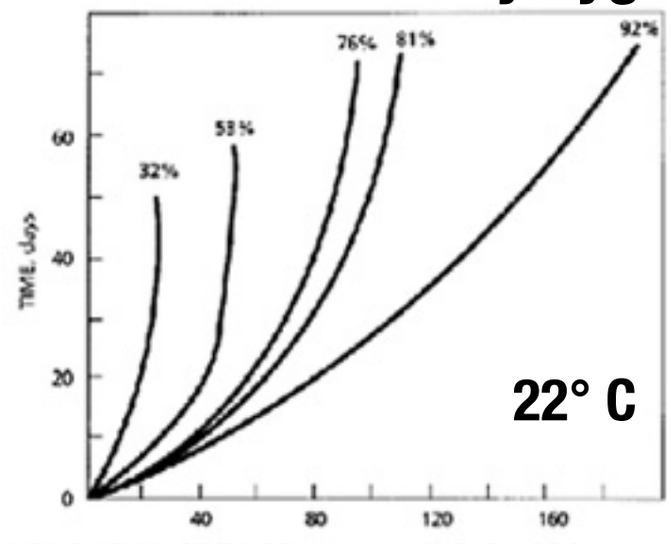




An example source of unanticipated (t)error: DMSOh No!

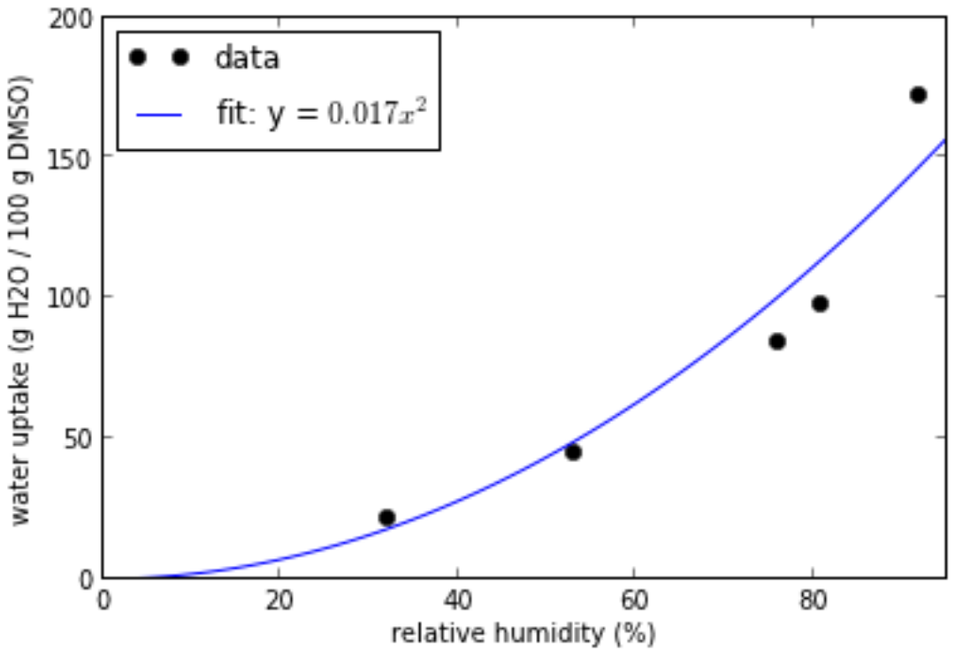
Most compound libraries are stored as stocks in 100% DMSO.

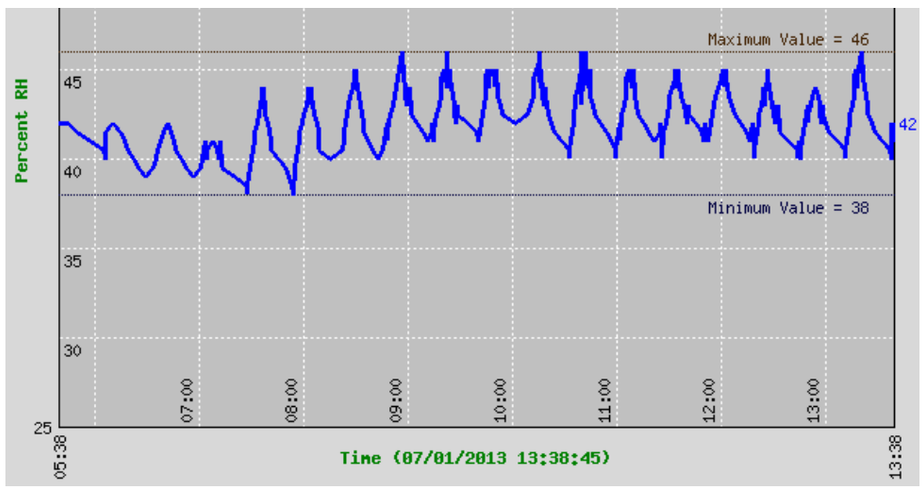
But DMSO is incredibly hygroscopic! Water uptake depends on relative humidity.



WEIGHT INCREASE, grams H 2 O/100 grams DMSO

From http://www.gaylordchemical.com/index.php?page=101b-dmso-physical-properties#101b-4

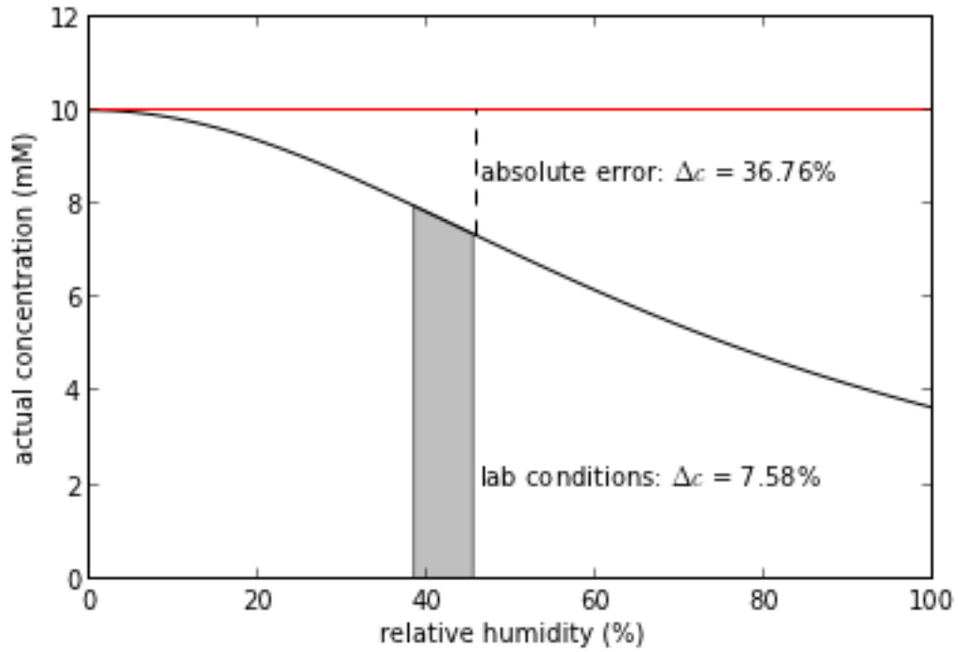




46%

38%

relative humidity fluctuations in my ZRC 17th floor lab



If we left DMSO stock container open, absolute concentration error ~37% and daily fluctuations ~7.6%

(even before temperature-dependence and miscibility is included!)

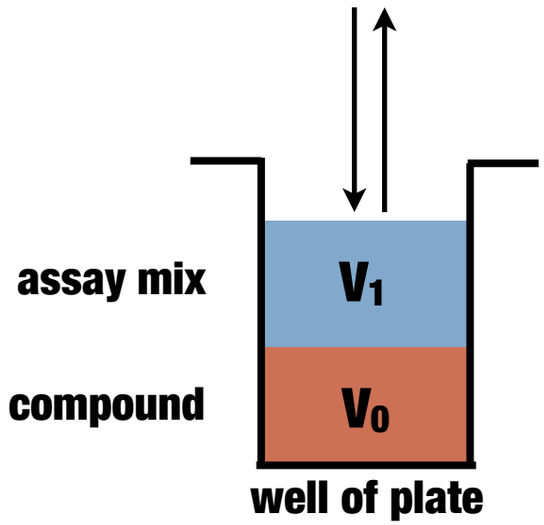
IPython tutorial available at https://github.com/choderalab/cadd-grc-2013

Different sources of error to watch for

- Is the number *derived* (from unseen raw data)
  - What is the error in the derivation?
  - If IC50, then are the diagnostics of the curve available?
- If concentration is involved in the endpoint, how is it determined?
  - Weight and volume? Weighing errors are a problem – typically 2X, often more
  - Mass spec?
  - NMR?
- Omission of metadata or odd metadata (could imply a problem)
  - Hill coefficients, Z-prime, confidence interval ratio, purity, dynamic range qualifiers
  - Odd dates, extreme age (1945 for date of synthesis)
- Inappropriate reporting of results
  - *Don't blame the data, blame the way it was presented*

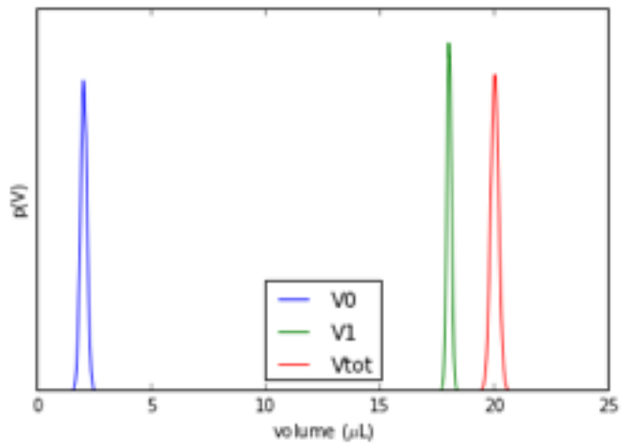
What is experimental error and where does it come from?

fluorescence measurement



$$
\begin{aligned}
V_0 &= 2 \quad \mu L \\
V_1 &= 18 \quad \mu L \\
K_d &= 100 \quad nM \\
C_0 &= 10 \quad mM \\
C_1 &= 1 \quad \mu M
\end{aligned}
$$

finite pipetting precision

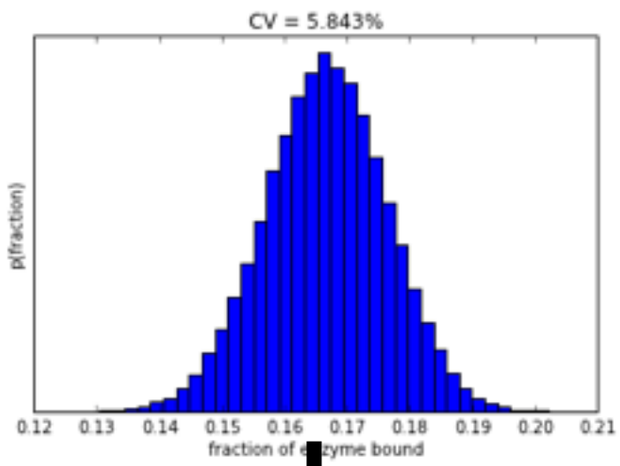


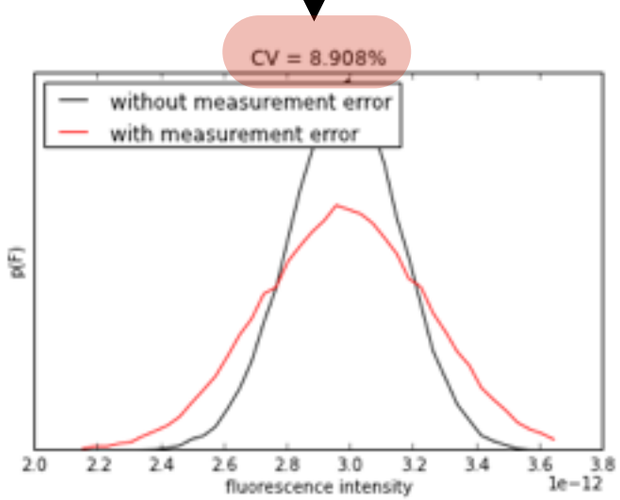
Biomek FX^P Pipetting Performance Specifications

Span-8 Systems

Transfer Volume	Span-8 Syringe Volume	Tip Types	Accuracy ± %	Precision < %
0.5 µL	250 µL	P20, Fixed 60 mm	5	10
1 µL	250, 500, 1000 µL	P20, P50, Fixed 60 mm	3	7
5 µL	250, 500, 1000 µL	P20, Fixed 60 mm	3	5
10 µL	500 µL	P50, P250	3	5

variation in protein:ligand complex concentration

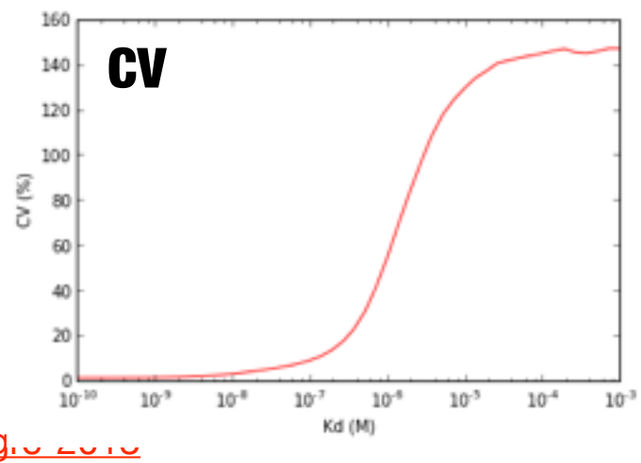




signal broadening due to measurement error

Is the expected CV small enough to be useful?

Modeling an experiment can ensure it will yield useful data or uncover unexpected issues



IPython tutorial available at https://github.com/choderalab/cadd-g...

The central limit theorem is why errors often look Gaussian

Suppose X_n is a random variable from ANY distribution with finite variance.
The sample mean (average) of a few samples will always look Gaussian!

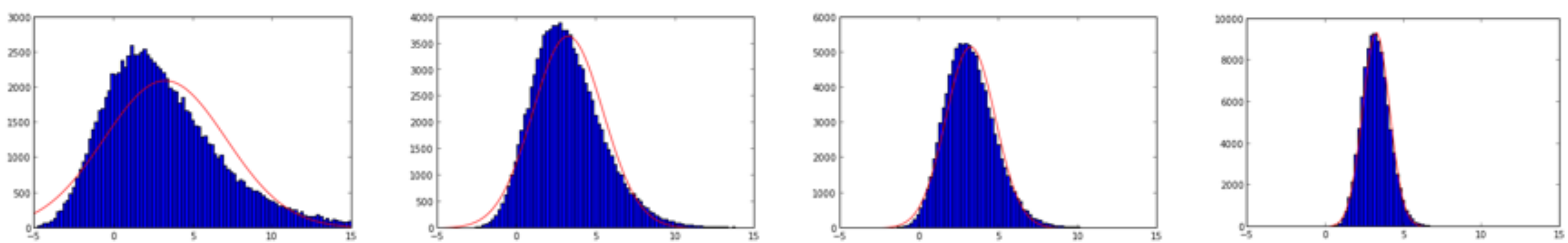
central limit theorem (CLT):

$$\frac{1}{N}\sum_{n=1}^{N} X_n \rightarrow N(\mu, \sigma^2/N)$$

sample mean

$$E[X_n] = \mu$$
$$var[X_n] = \sigma^2$$

This happens with surprisingly few measurements, even for crazy distributions!



X_n S_3 S_6 S_{20}

not Gaussian ··········→ **pretty damn Gaussian**

Many instruments operate like this (e.g. plate reader that uses 10 flashes of Xe lamp)

IPython tutorial available at https://github.com/choderalab/cadd-grc-2013

Reporting the error

Rules of error reporting:

- Report only one significant digit in the error estimate
- Report the measurement only to this decimal place
- Report what kind of error estimate you are reporting for numbers or error bars

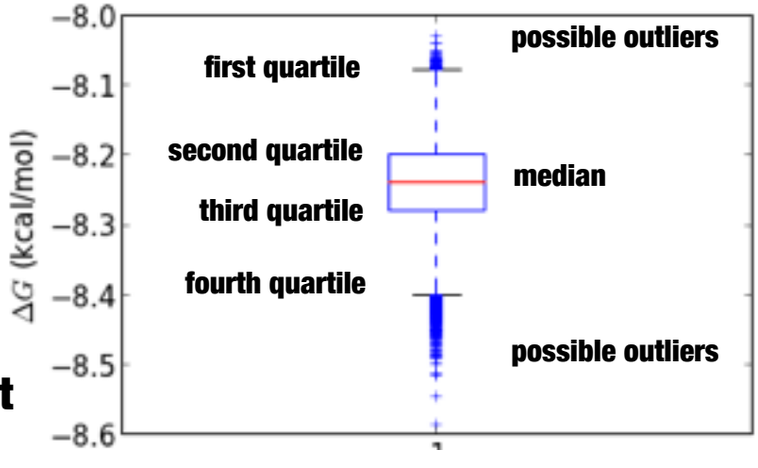
Examples

YES: 4.1±0.1 or 4.1(1)
YES: 4.123±0.002 or 4.123(2)
NO: 4.123±0.1 or 4.123±0.123

YES: standard error of the mean
YES: 95% confidence interval
NO: meaning of error not specified

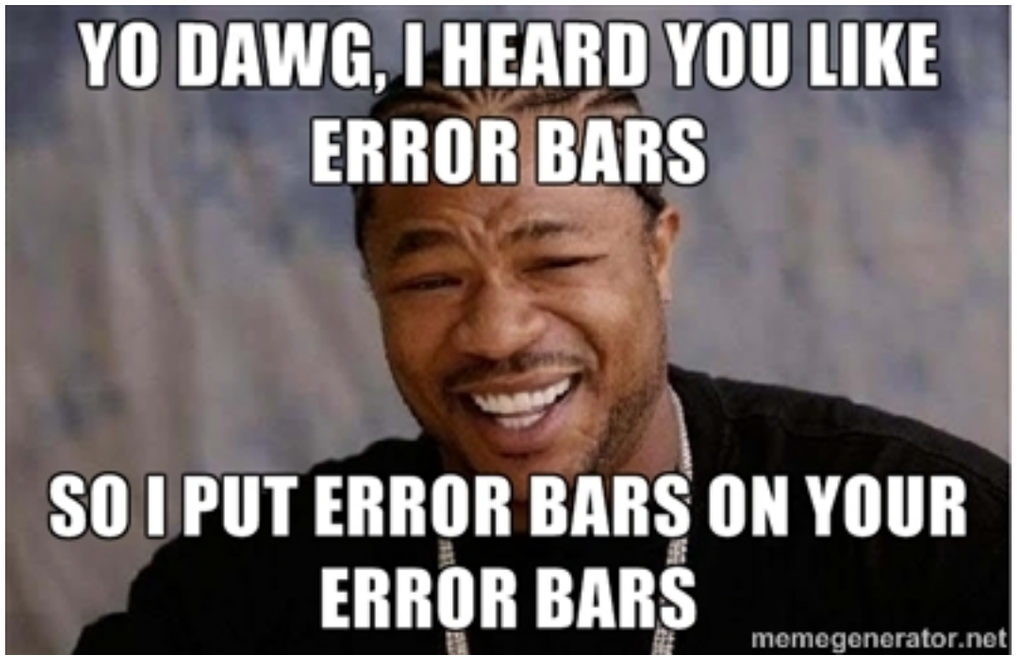
Confidence intervals [low,high] or box-and-whisker plots are also acceptable, and preferred when error is highly asymmetric (e.g. probabilities).

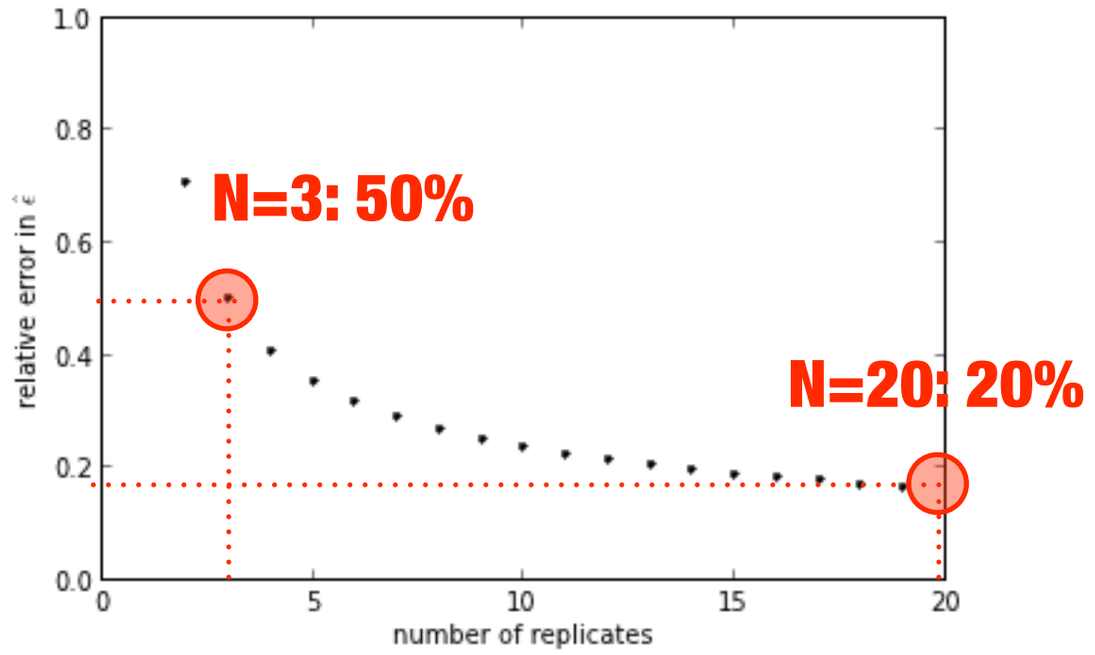
YES: 95% confidence interval is [4.08, 4.15]
YES: box-and-whisker plot showing quartiles



box-and-whisker plot

What is the error in the error?





For large N, relative error in standard error estimate is $\frac{\delta\hat{\epsilon}}{\hat{\epsilon}} = \frac{1}{\sqrt{2(N-1)}}$

For 3 replicates, the uncertainty in the error is over 50%!

Only the order of magnitude of error is known; not even a single digit is certain!

Even for 20 replicates, uncertainty is still close to within 20%!

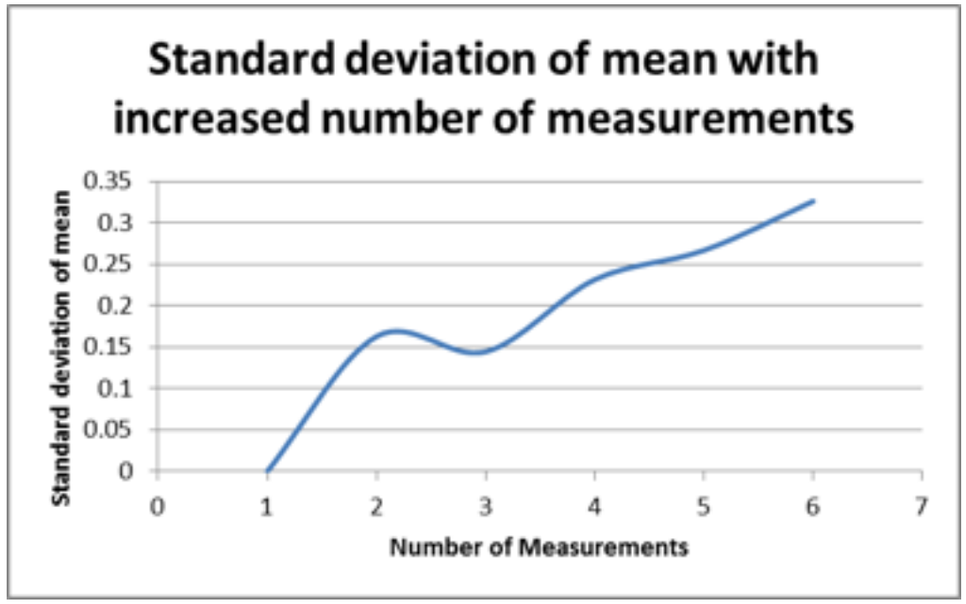
IPython tutorial available at https://github.com/choderalab/cadd-grc-2013

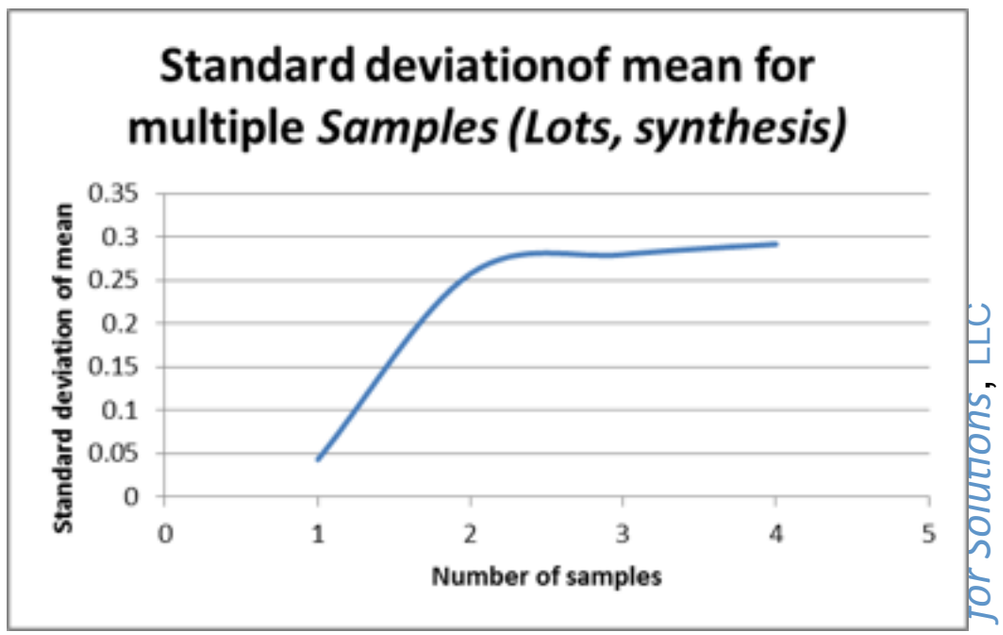
What accuracy do we need to be useful?

- Depends on the need and use
  - Efficacy optimization
    - Factor of 2
  - HTS hit finding campaign
    - Factor of 10
  - Alerts
    - Factor of 100?
- Proper presentation of results is vital
  - "Cutoffs" and classifications can make interpretation tricky for quantitative data
    - E.g.: Issuing a hERG alert at < $1\ \mu M$
      - Propagate error into rational decision making (*a la* Optibrium)
    - E.g." A qualitative solubility assay (i.e. soluble/insoluble DMSO/water precipitate) reported to 3 significant figures

Estimating error when none is reported

- Previous experience
- Variation in
  - Standards
  - Samples of the same compound
  - Multiple measurements of the same compound
  - Between days / equipment
- If in doubt suspect a value of 5 – 10 fold
  - Experimentalists all note for same SOP in same company
    - Different labs, instruments, technicians
      - 2X difference expected
      - 5X easy to find
      - 10X not unexpected





What is the expected error for different procedures

- From a screening veteran:
  - We are routinely using LC-MS measurement for identity check and purity of library compounds.
  - Of course there are major limitations: which method do you use, which gradient, which wavelength….
  - And even when you have settled on one method you can get two different results if the compound degrades in the solvent while it is sitting in the plate, the compound is not picked in a precise volume….
  - But it is good enough to check if there is *major* degradation going on.
  - To be more precise you would need to use NMR but that`s not feasible for a *classical library compound* – only after re-synthesis.

What is the expected error

- Purity: Corporate archive: 10% of the compounds were found to have degraded to < 75% purity (Kramer and Lewis, 2012)
  - IMHO this might be an optimistic number for many companies
  - At one time, purity was defined at 70% in some 80% in other companies
- Pipetting of serial dilutions: The *mixing* of the compound may be done entirely different using different methods resulting in 10 fold shifts of IC50 when using one method over the other. (Well informed screening veteran)
  - Angle of the pipette affects mixing

Modeling experimental error can be a valuable exercise: Biomek dilution series vs Echo acoustic dispensing

In the Pipeline

« Aveo Gets Bad News on Tivozanib | Main | The Medical Periodic Table »

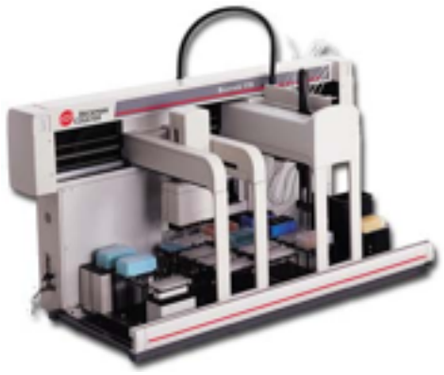
May 3, 2013

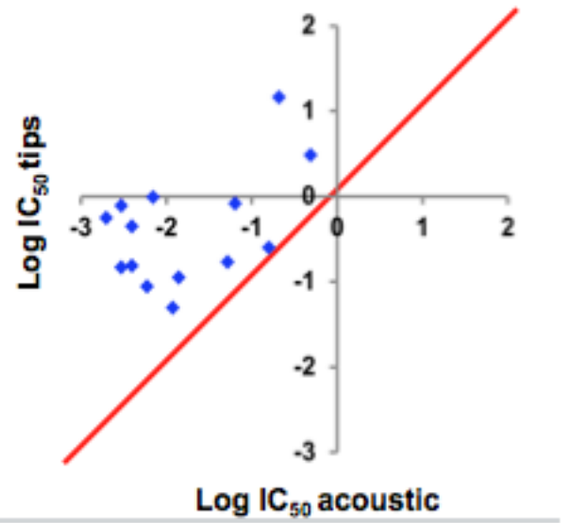
Drug Assay Numbers, All Over the Place

Posted by **Derek**

There's a truly disturbing paper out in *PLoSONE* with potential implications for a lot of assay data out there in the literature. The authors are looking at the results of biochemical assays as a function of how the compounds are dispensed in them, pipet tip versus acoustic, which is the sort of idea that some people might roll their eyes at. But people who've actually done a lot of biological assays may well feel a chill at the thought, because this is just the sort of you're-kidding variable that can make a big difference.

http://pipeline.corante.com/archives/2013/05/03/drug_assay_numbers_all_over_the_place.php





OPEN ACCESS Freely available online

PLOS ONE

Dispensing Processes Impact Apparent Biological Activity as Determined by Computational and Statistical Analyses

Sean Ekins[1]*, Joe Olechno[2], Antony J. Williams[3]

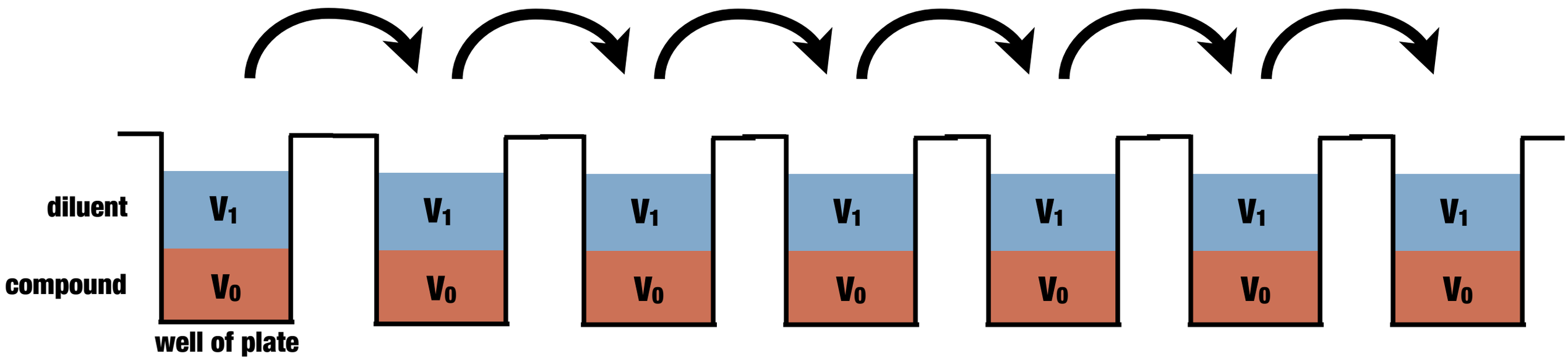
1 Collaborations in Chemistry, Fuquay-Varina, North Carolina, United States of America, **2** Labcyte Inc., Sunnyvale, California, United States of America, **3** Royal Society of Chemistry, Wake Forest, North Carolina, United States of America

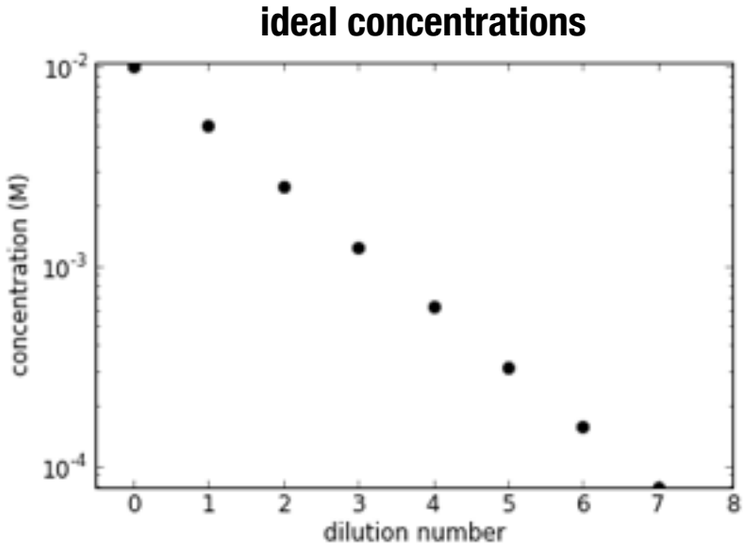
Elkins et al. PLoS One 8:e62325, 2013.

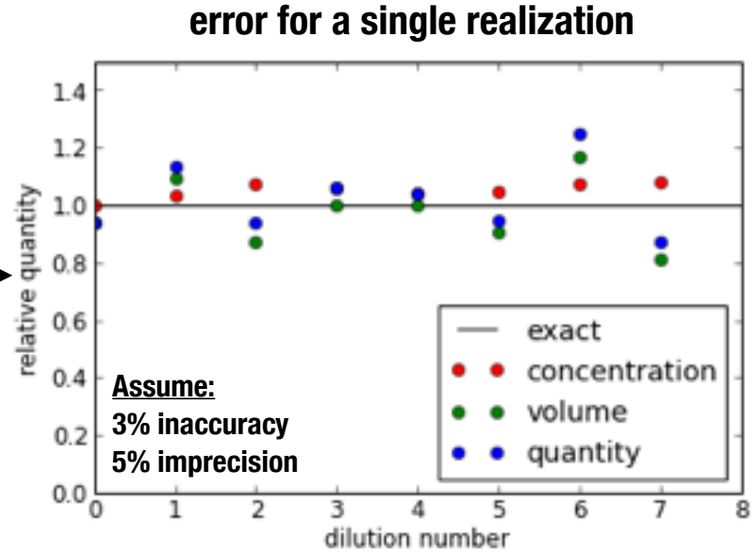
Is this just a result of pipetting accuracy differences?
Let's model the experiment to find out.

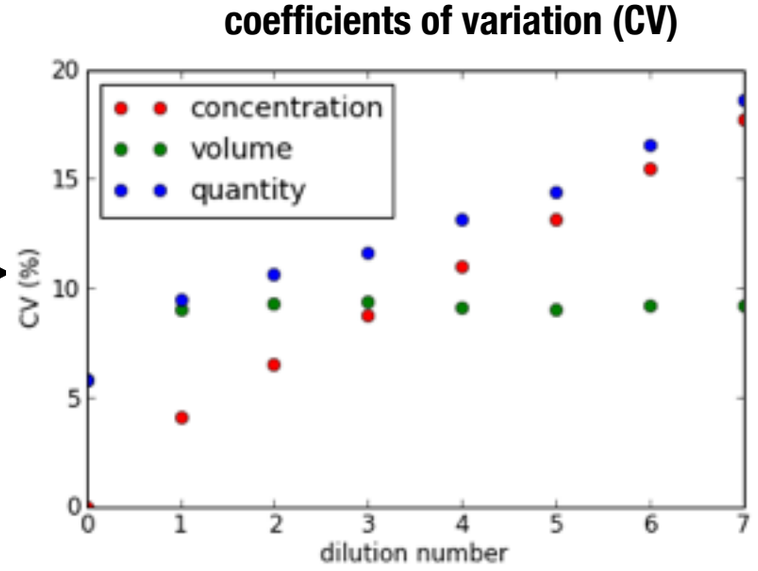
IPython tutorial available at https://github.com/choderalab/cadd-grc-2013

Modeling experimental error can be a valuable exercise: Preparing a dilution series via liquid-handling robot







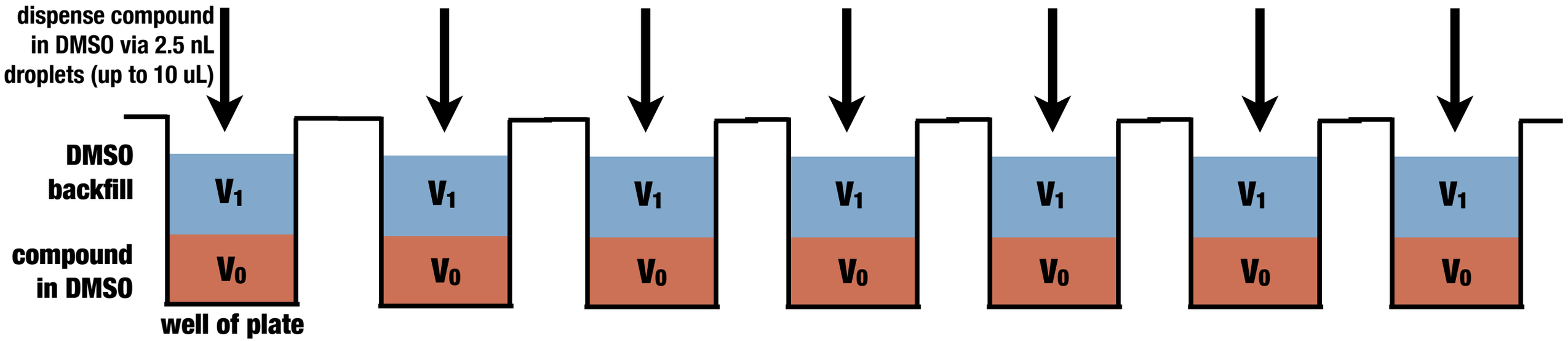


Biomek NX/FX

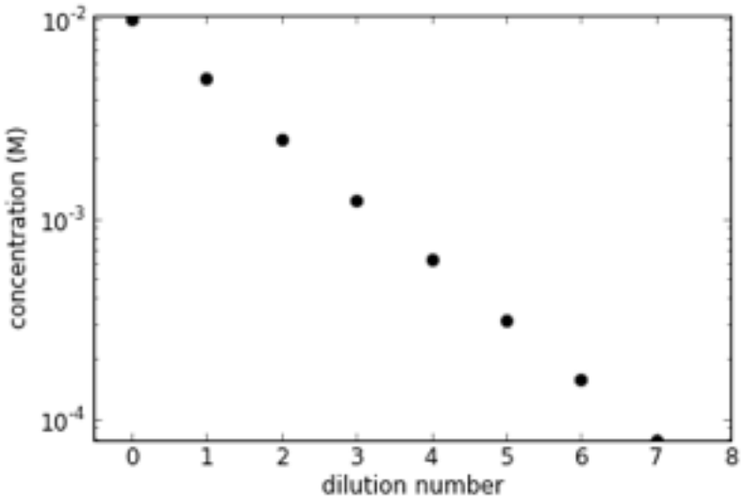
IPython tutorial available at https://github.com/choderalab/cadd-grc-2013


An Inconvenient Truth:
It’s hard to make a good dilution series

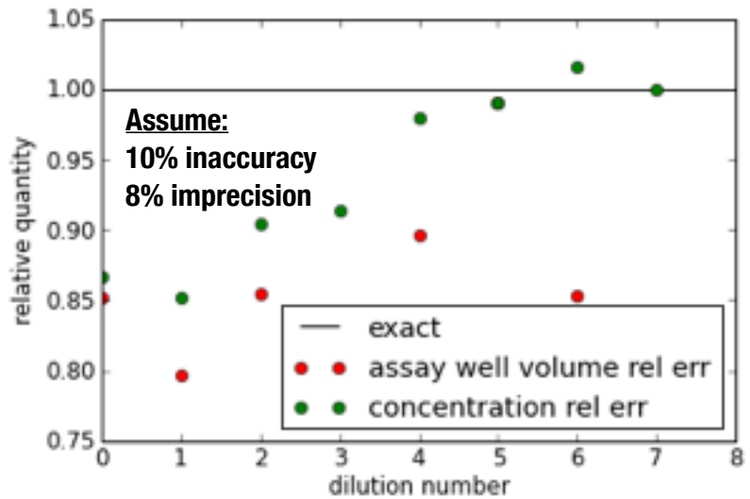
Modeling experimental error can be a valuable exercise: Preparing a dilution series via acoustic dispensing



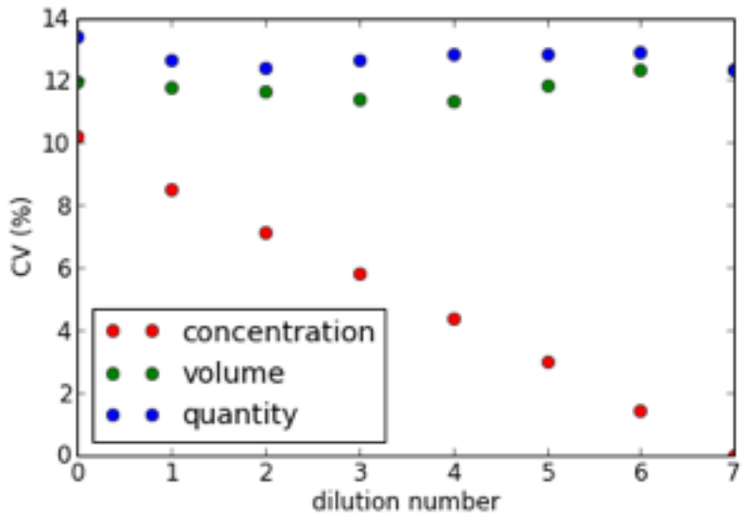
ideal concentrations



error for a single realization



coefficients of variation (CV)

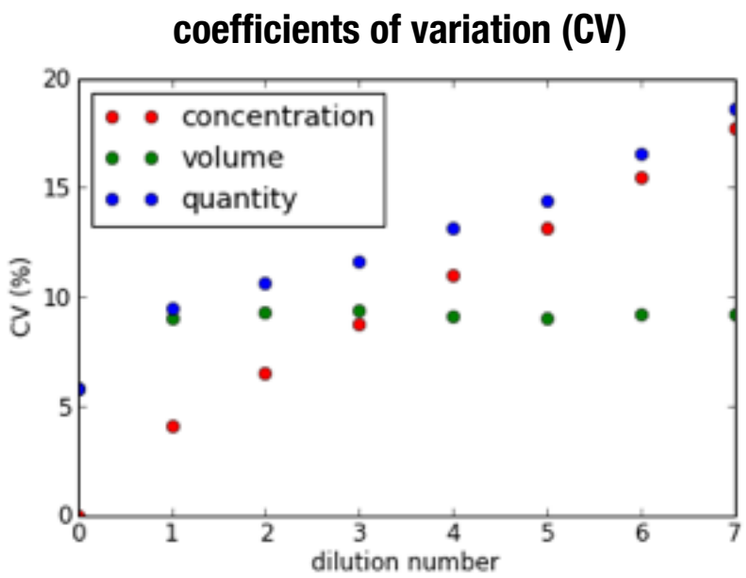


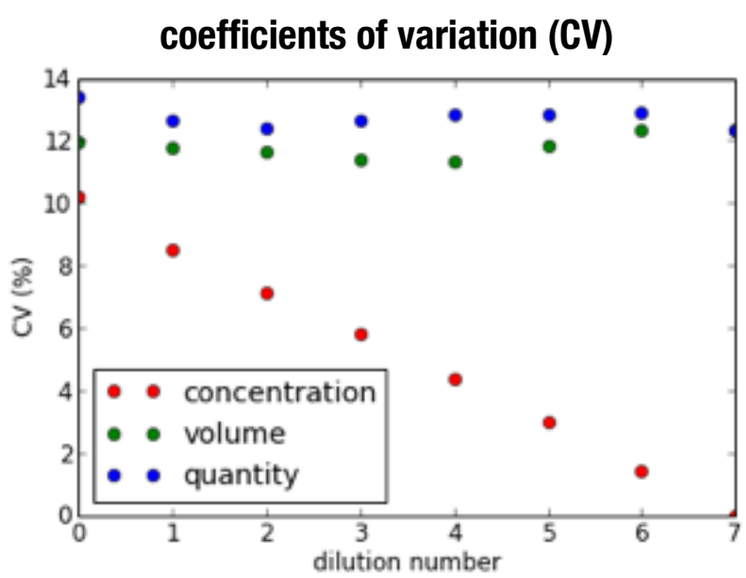
http://www.labcyte.com/sites/default/files/support_docs/Echo%205XX%20Specifications.pdf

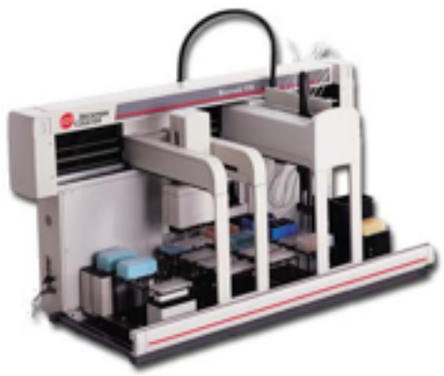
LabCyte Echo

IPython tutorial available at https://github.com/choderalab/cadd-grc-2013

Modeling experimental error can be a valuable exercise: Comparison of tip-based and acoustic dispensing







tip-based

acoustic

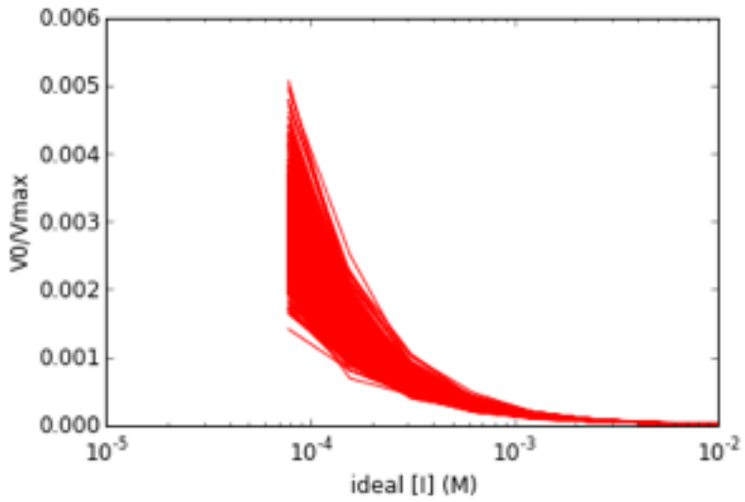
IPython tutorial available at https://github.com/choderalab/cadd-grc-2013

Modeling experimental error can be a valuable exercise: Assay model

<u>Dispense into assay plate:</u>
2 uL compound dilution
10 uL enzyme assay mix

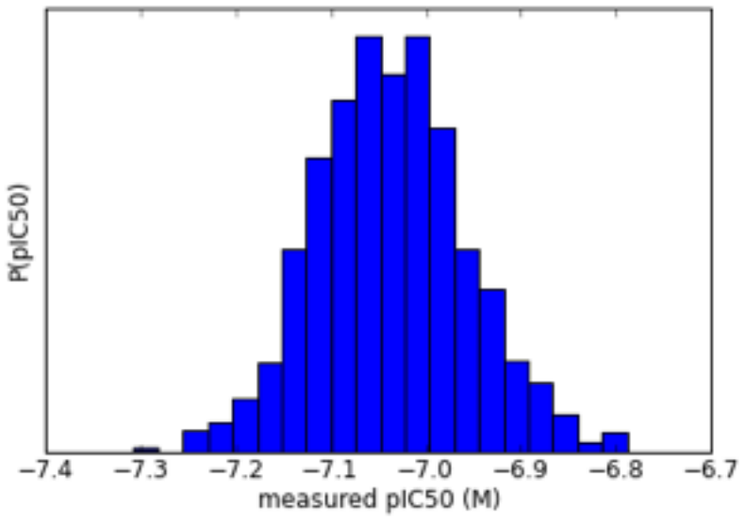
<u>Assume for EphB4 assay:</u>
Michaelis-Menten kinetics
competitive inhibition
[ATP] = 4 uM
[EphB4] = 6 uM
Km(ATP) = 1.71 uM

measure initial reaction velocity



fit it to get IC50

distribution of fit pIC50s



<u>IPython tutorial available at</u> https://github.com/choderalab/cadd-grc-2013

Modeling experimental error can be a valuable exercise: Biomek dilution series vs Echo acoustic dispensing

In the Pipeline

« Aveo Gets Bad News on Tivozanib | Main | The Medical Periodic Table »

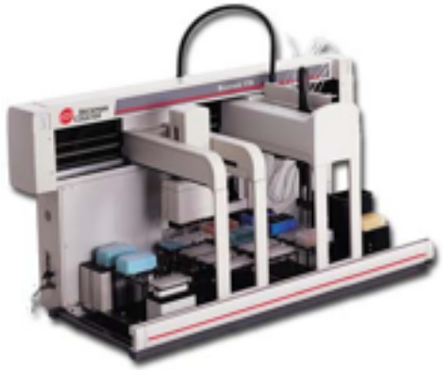
May 3, 2013

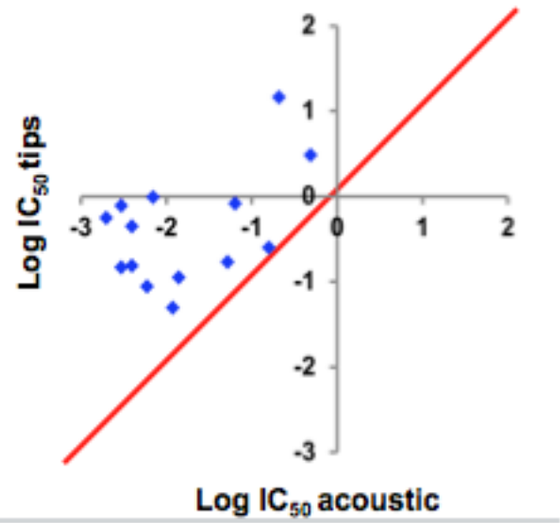
Drug Assay Numbers, All Over the Place

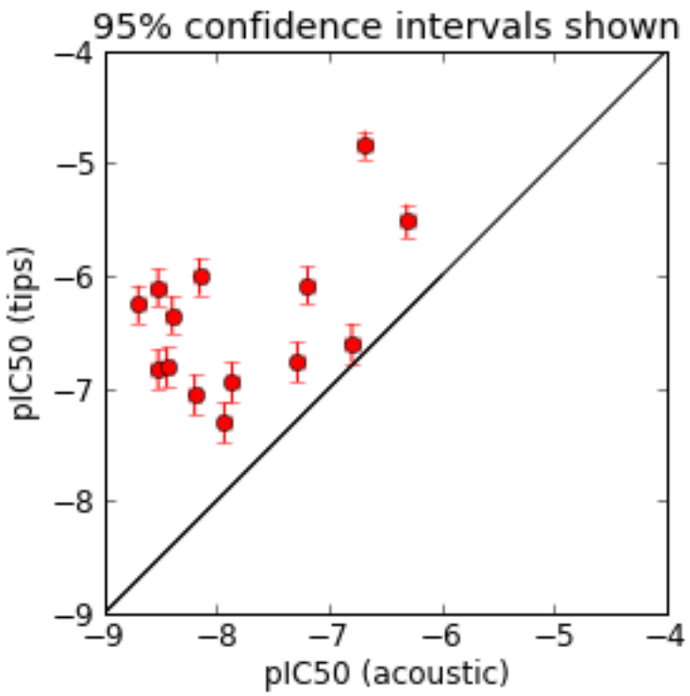
Posted by **Derek**

There's a truly disturbing paper out in *PLoSONE* with potential implications for a lot of assay data out there in the literature. The authors are looking at the results of biochemical assays as a function of how the compounds are dispensed in them, pipet tip versus acoustic, which is the sort of idea that some people might roll their eyes at. But people who've actually done a lot of biological assays may well feel a chill at the thought, because this is just the sort of you're-kidding variable that can make a big difference.

http://pipeline.corante.com/archives/2013/05/03/drug_assay_numbers_all_over_the_place.php







Elkins et al. PLoS One 8:e62325, 2013.

OK, that's not sufficient to explain the discrepancy... What are we missing?

IPython tutorial available at https://github.com/choderalab/cadd-grc-2013

Modeling experimental error can be a valuable exercise: Dilution is a big problem with liquid-based pipetting

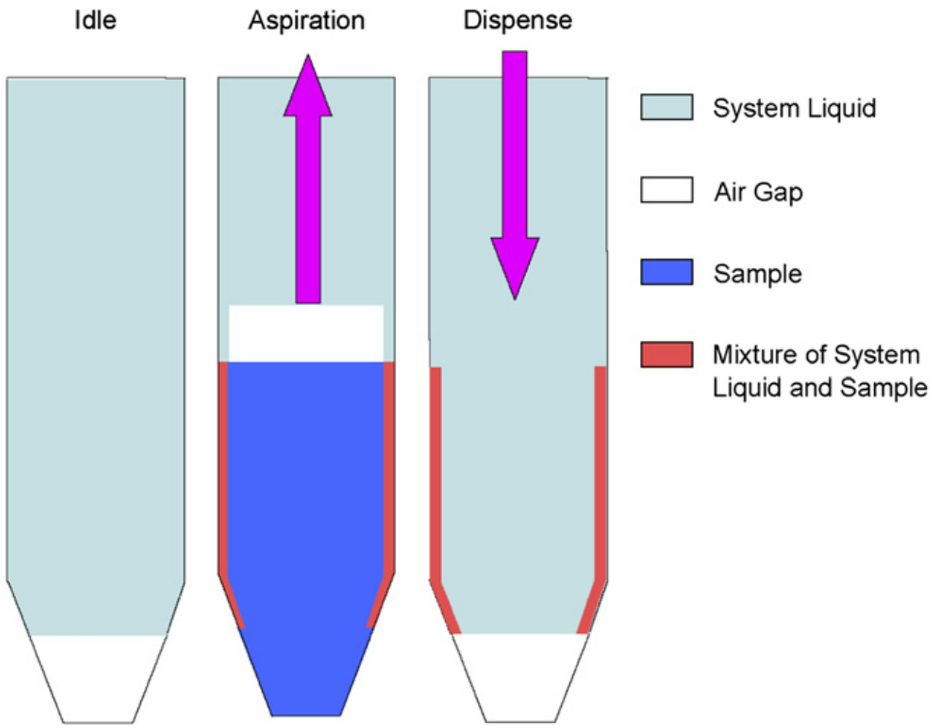


Figure 1. Schematic diagram of the dilution effect.

Table 1. Comparison between the MVS method and the gravimetry method—Tecan ALH using water liquid class

	MVS (μL)		Gravimetry (μL)	
Target volume	20	200	20	200
Mean volume (μL)	18.74	190.08	20.15	199.53
Inaccuracy (%)	−6.30	−4.96	0.75	−0.23
StDev	0.22	1.74	0.6	0.55
CV ($n = 96$) (%)	1.17	0.92	1.94	0.59

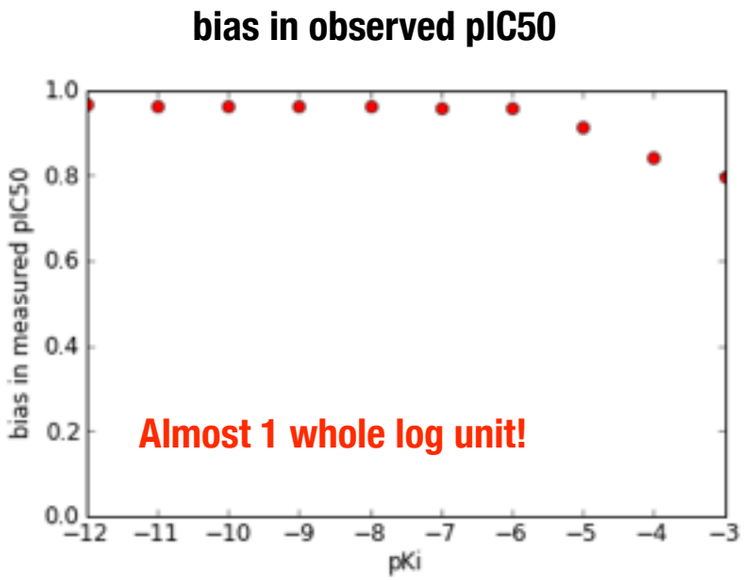
These solutions were delivered by the Tecan ALH using an aspirate/dispense protocol based on a water liquid class. Inaccuracy corresponds to 100 × (Observed volume − Target volume)/Target volume. Coefficient of variation (CV) corresponds to 100 × StDev/mean, where StDev is the standard deviation.

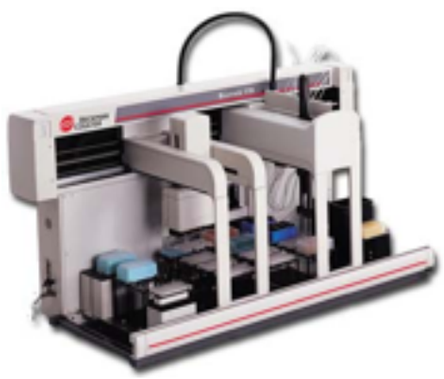
mixture of dyes **weight**

Gu and Deng. JALA 12:355, 2007 (BMS)
Dong, Ouyang, Liu, and Jemal. JALA 11:60, 2006. (BMS)

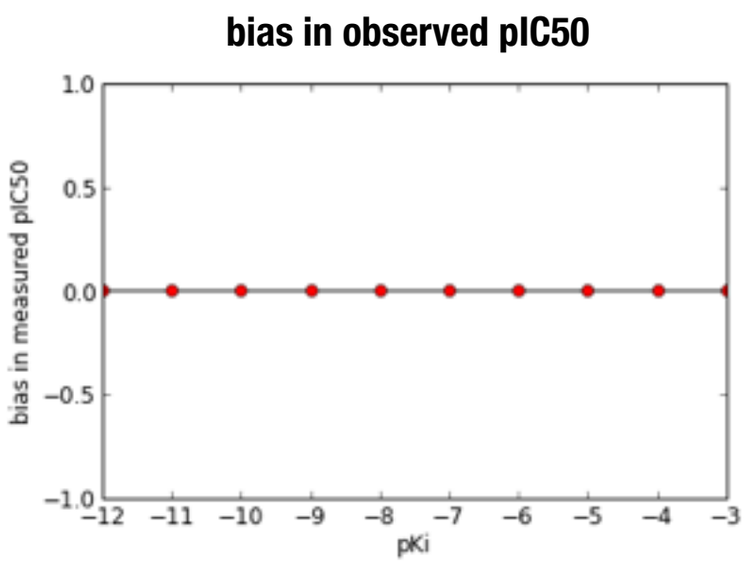
IPython tutorial available at https://github.com/choderalab/cadd-grc-2013

Modeling experimental error can be a valuable exercise: Comparison of tip-based and acoustic dispensing





tip-based



acoustic

IPython tutorial available at https://github.com/choderalab/cadd-grc-2013

Modeling experimental error can be a valuable exercise: Biomek dilution series vs Echo acoustic dispensing

In the Pipeline

« Aveo Gets Bad News on Tivozanib | Main | The Medical Periodic Table »

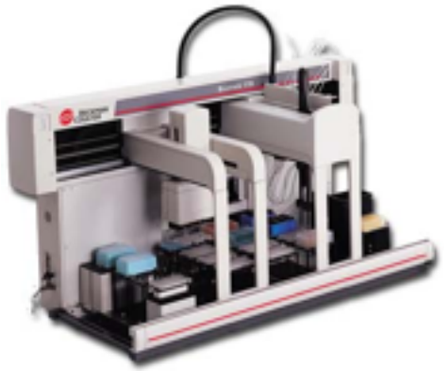
May 3, 2013

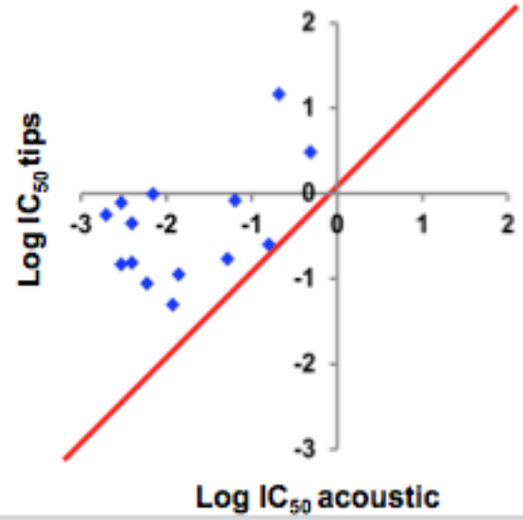
Drug Assay Numbers, All Over the Place

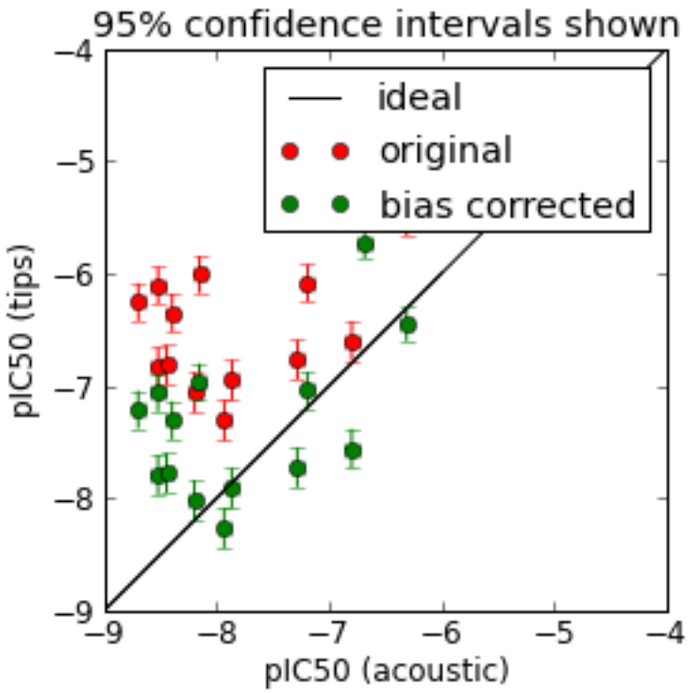
Posted by **Derek**

There's a truly disturbing paper out in *PLoSONE* with potential implications for a lot of assay data out there in the literature. The authors are looking at the results of biochemical assays as a function of how the compounds are dispensed in them, pipet tip versus acoustic, which is the sort of idea that some people might roll their eyes at. But people who've actually done a lot of biological assays may well feel a chill at the thought, because this is just the sort of you're-kidding variable that can make a big difference.

http://pipeline.corante.com/archives/2013/05/03/drug_assay_numbers_all_over_the_place.php







Elkins et al. PLoS One 8:e62325, 2013.

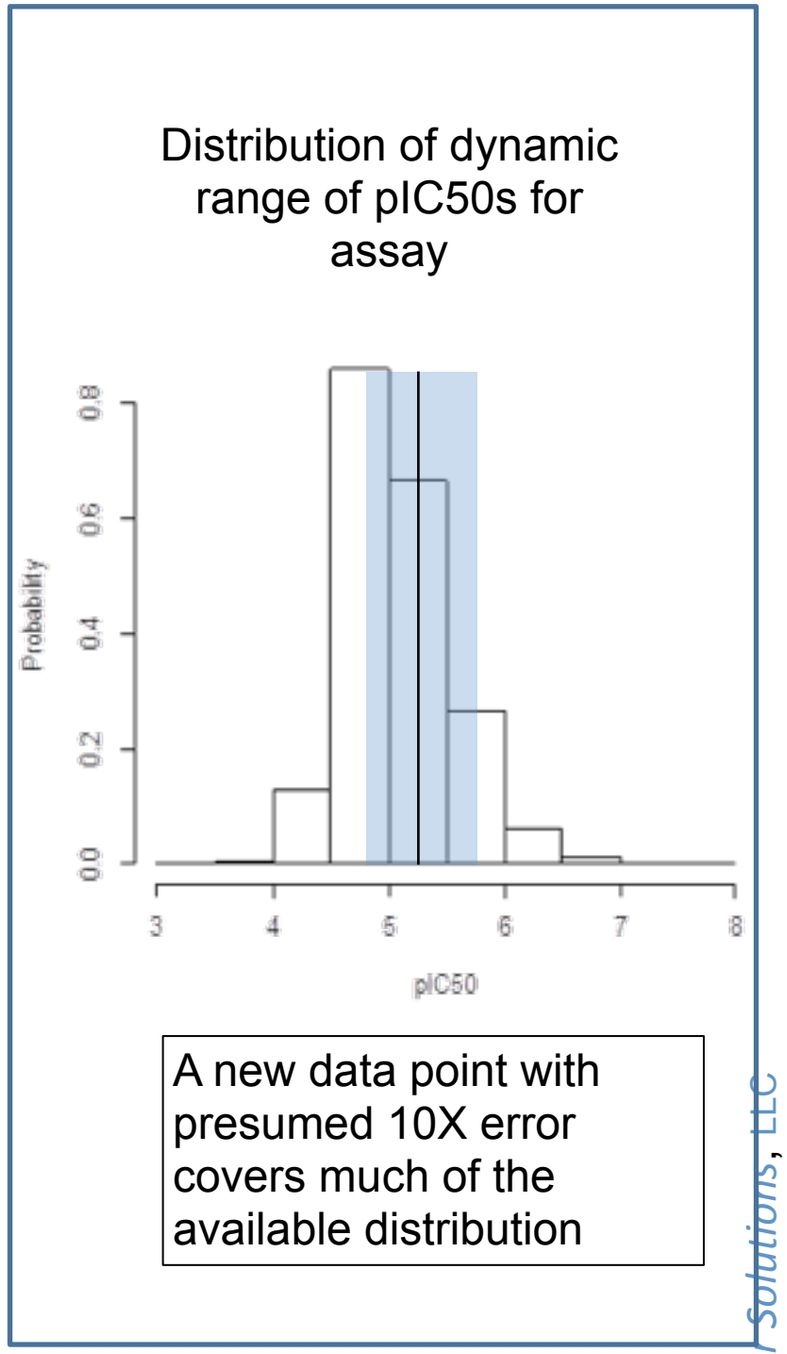
IPython tutorial available at https://github.com/choderalab/cadd-grc-2013

How bad can things really be? (Catastrophic failures)

- Worst case: Misled: The data is useless but does not appear to be so
  - Actual error is greater than the data range
  - Actual error is greater than the needed precision
  - Data is reported with the wrong precision
  - Data is (allowed to be) used incorrectly
- Next worst: Bereft: The data is useless

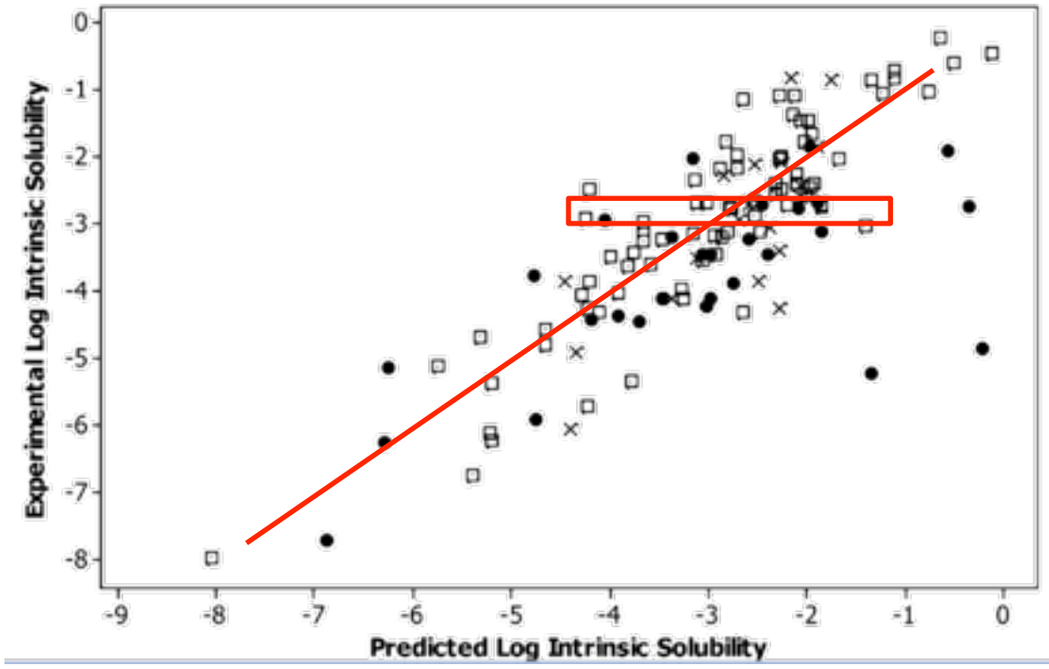
How bad can things really be? (Catastrophic failures)

- Worst case: Misled: The data is useless but does not appear to be so
  - Actual error is greater than the data range
  - Actual error is greater than the needed precision
  - Data is reported with the wrong precision
  - Data is (allowed to be) used incorrectly
- Next worst: Bereft: The data is useless
- Factor of 10 between sites is not uncommon
  - Considering the limited range of much of Pharma data a factor of 10 might swamp the results
  - i.e. The error might be a substantial portion of the dynamic range of the data
  - In particular, the data might not be useful for a desired use
    - E.g. Optimization vs. alert



How safe is it to use collated literature data?

- In 2 words, not very
- Depends on the property as well as how the data and model are used
- Assay conditions can vary widely
- It rarely happens that details of the experiments and standards or common molecules are available
  - Are there common molecules that can be used for calibration?
- Possibly use for approximate, general models of low expected precision
  - Some modest successes can be had
- Expect a factor of 10 error



M. Hewitt, et al 2009, J. Chem. Inf. Model, 49,2572–2587

What if error was not reported or measured? How reliable is public pKi data?

Journal of Medicinal Chemistry

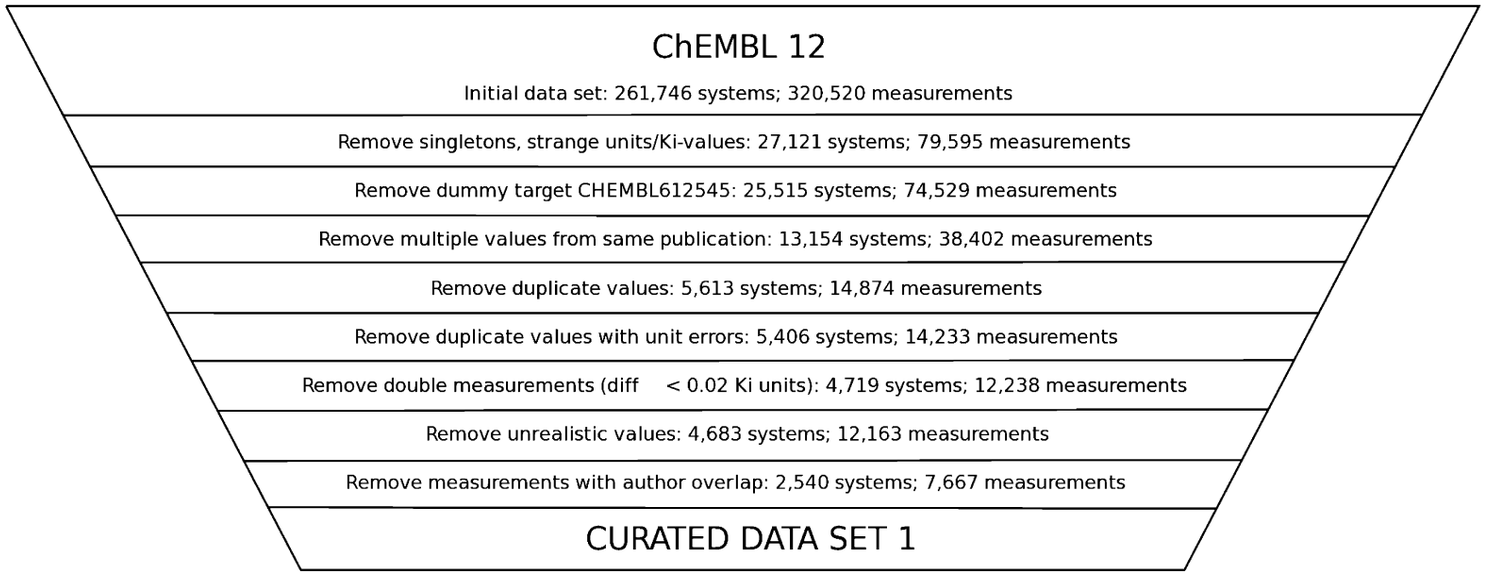
Article

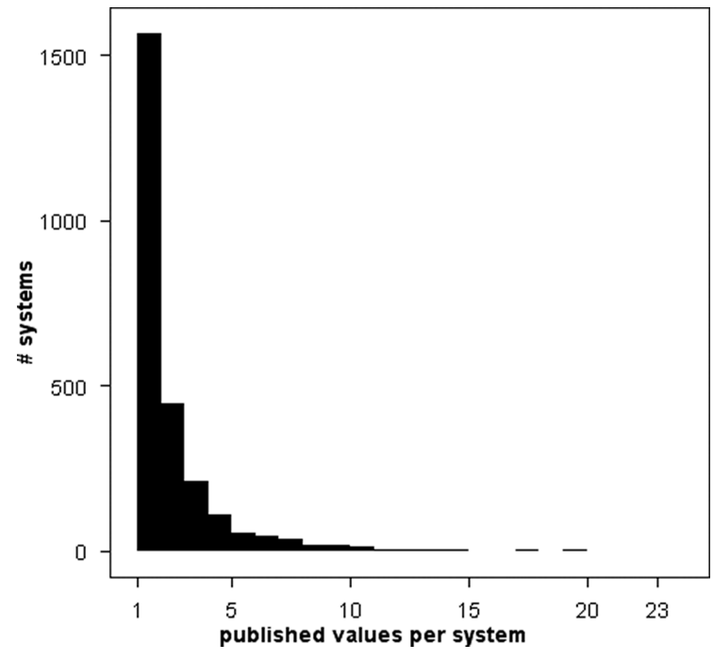
pubs.acs.org/jmc

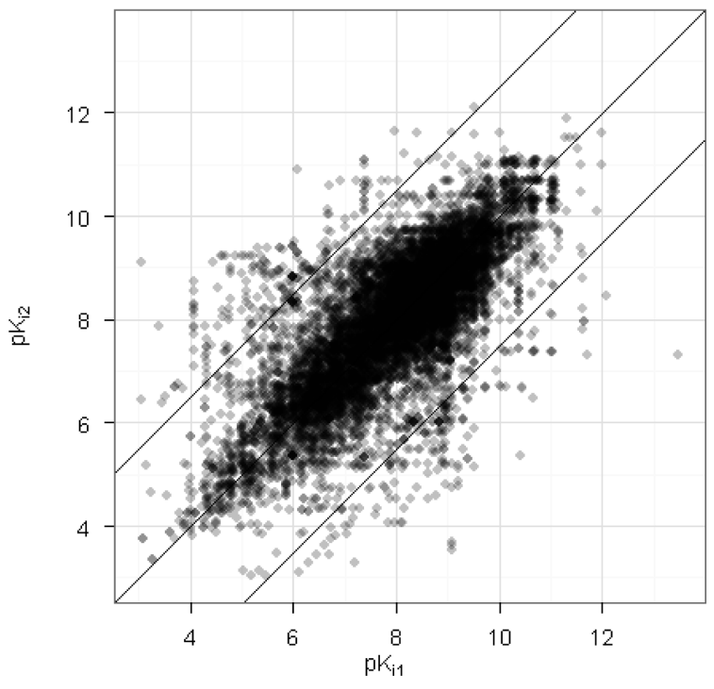
The Experimental Uncertainty of Heterogeneous Public K_i Data

Christian Kramer,*,† Tuomo Kalliokoski,*,† Peter Gedeck, and Anna Vulpetti

Novartis Institutes for BioMedical Research, Novartis Pharma AG, Forum 1, Novartis Campus, CH-4056 Basel, Switzerland





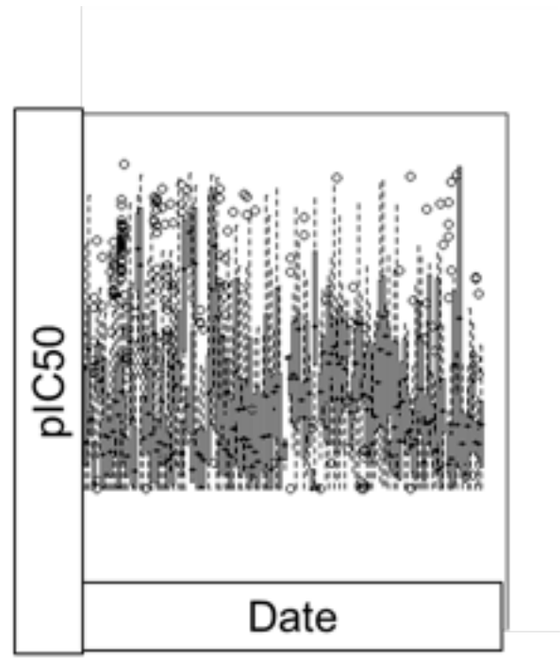


root mean square error (RMSE) ~ 0.54 pKi units ~ 0.75 kcal/mol @ 25 C

Christian Kramer et al. J Med Chem 2012.

Spotting "place holder" values

- (0.0 is really a bad default for most Pharma data)
- Depends on the data form and range
- Suspicious, recurring value – particularly for continuous data
- Values way outside the norm (9999.9999)
- Blank values (depending on how they will be interpreted – set to 0.0?)
- Numbers that occur too often and/or are too ideal
  - Ex: 4.000 as the most frequent response whereas there were few 4 sig fig duplicates otherwise



Spotting "place holder" values

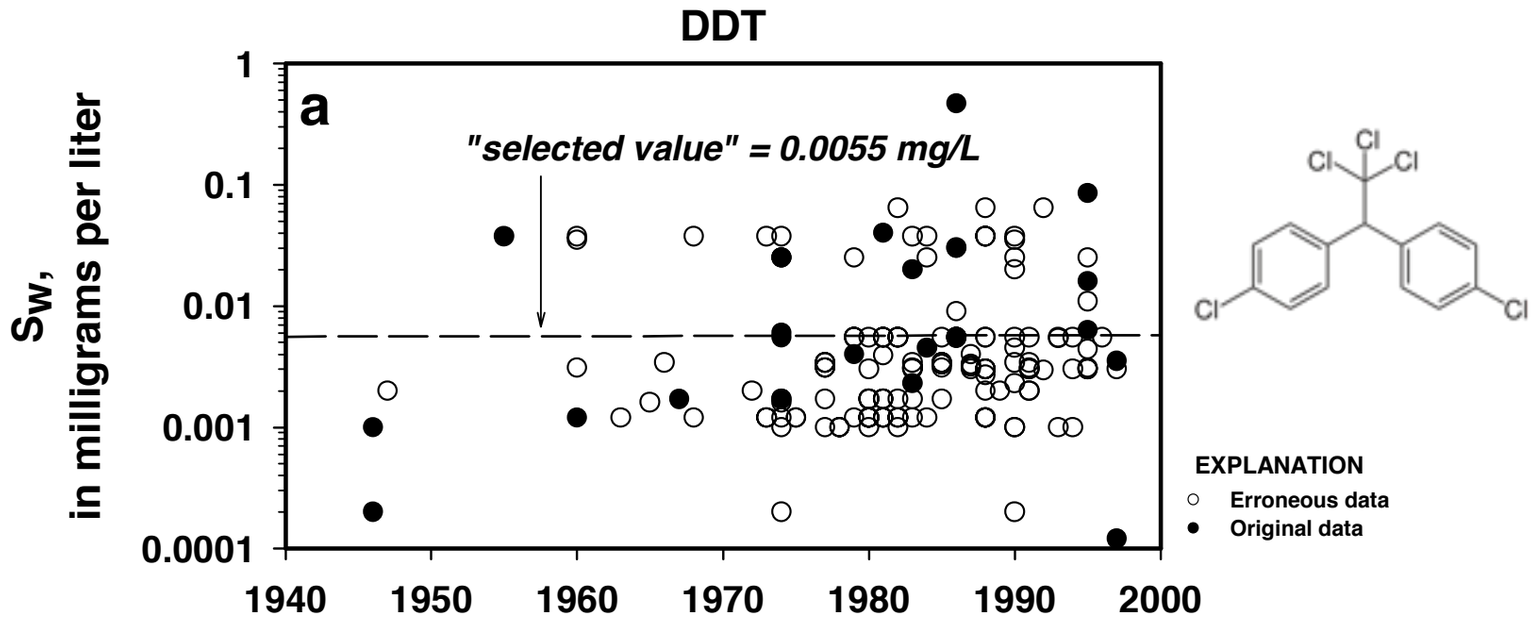
- The danger of “0”
- Is there metadata suggesting that an experiment was not done or interrupted or somehow failed?
  - Example:
  - Mean value for logD of 1.85 (the automatically provided value)
  - 4 measurements
    - 3.7, 0.0, 4.0, 0.0
  - Off in the corner was a free form note that noted for the 0.0 values
    - ~”Insufficient material to complete experiment”
- If one relied on the default average value, the value would be 2 log units in error

Pitfalls with literature data and public databases

The accurate determination of an organic contaminant's physico-chemical properties is essential for predicting its environmental impact and fate. Approximately 700 publications (1944–2001) were reviewed and all known aqueous solubilities (S_W) and octanol-water partition coefficients (K_{OW}) for the organochlorine pesticide, DDT, and its persistent metabolite, DDE were compiled and examined. Two problems are evident with the available database: 1) egregious errors in reporting data and references, and 2) poor data quality and/or inadequate documentation of procedures. The published literature (particularly the collative literature such as compilation articles and handbooks) is characterized by a preponderance of unnecessary data duplication. Numerous data and citation errors are also present in the literature. The percentage of original S_W and K_{OW} data in compilations has decreased with time, and in the most recent publications (1994–97) it composes only 6–26 percent of the reported data. The variability of original DDT/DDE S_W and K_{OW} data spans 2–4 orders of magnitude, and there is little indication that the uncertainty in these properties has declined over the last 5 decades. A criteria-based evaluation of DDT/DDE S_W and K_{OW} data sources shows that 95–100 percent of the database literature is of poor or unevaluatable quality. The accuracy and reliability of the vast majority of the data are unknown due to inadequate documentation of the methods of determination used by the authors. [For example, estimates of precision have been reported for only 20 percent of experimental S_W data and 10 percent of experimental K_{OW} data.] Computational methods for estimating these parameters have been increasingly substituted for direct or indirect experimental determination despite the fact that the data used for model development and validation may be of unknown reliability. Because of the prevalence of errors, the lack of methodological documentation, and unsatisfactory data quality, the reliability of the DDT/ DDE S_W and K_{OW} database is questionable. The nature and extent of the errors documented in this study are probably indicative of a more general problem in the literature of hydrophobic organic compounds. Under these circumstances, estimation of critical environmental parameters on the basis of S_W and K_{OW} (for example, bioconcentration factors, equilibrium partition coefficients) is inadvisable because it will likely lead to incorrect environmental risk assessments. The current state of the database indicates that much greater efforts are needed to: 1) halt the proliferation of erroneous data and references, 2) initiate a coordinated program to develop improved methods of property determination, 3) establish and maintain consistent reporting requirements for physico-chemical property data, and 4) create a mechanism for archiving reliable data for widespread use in the scientific/regulatory community.

USGS Water-Resources Investigations Report 01-4201, 2001.

The literature is filled with erroneous data



USGS Water-Resources Investigations Report 01-4201, 2001.

Data has a habit of being re/misreported

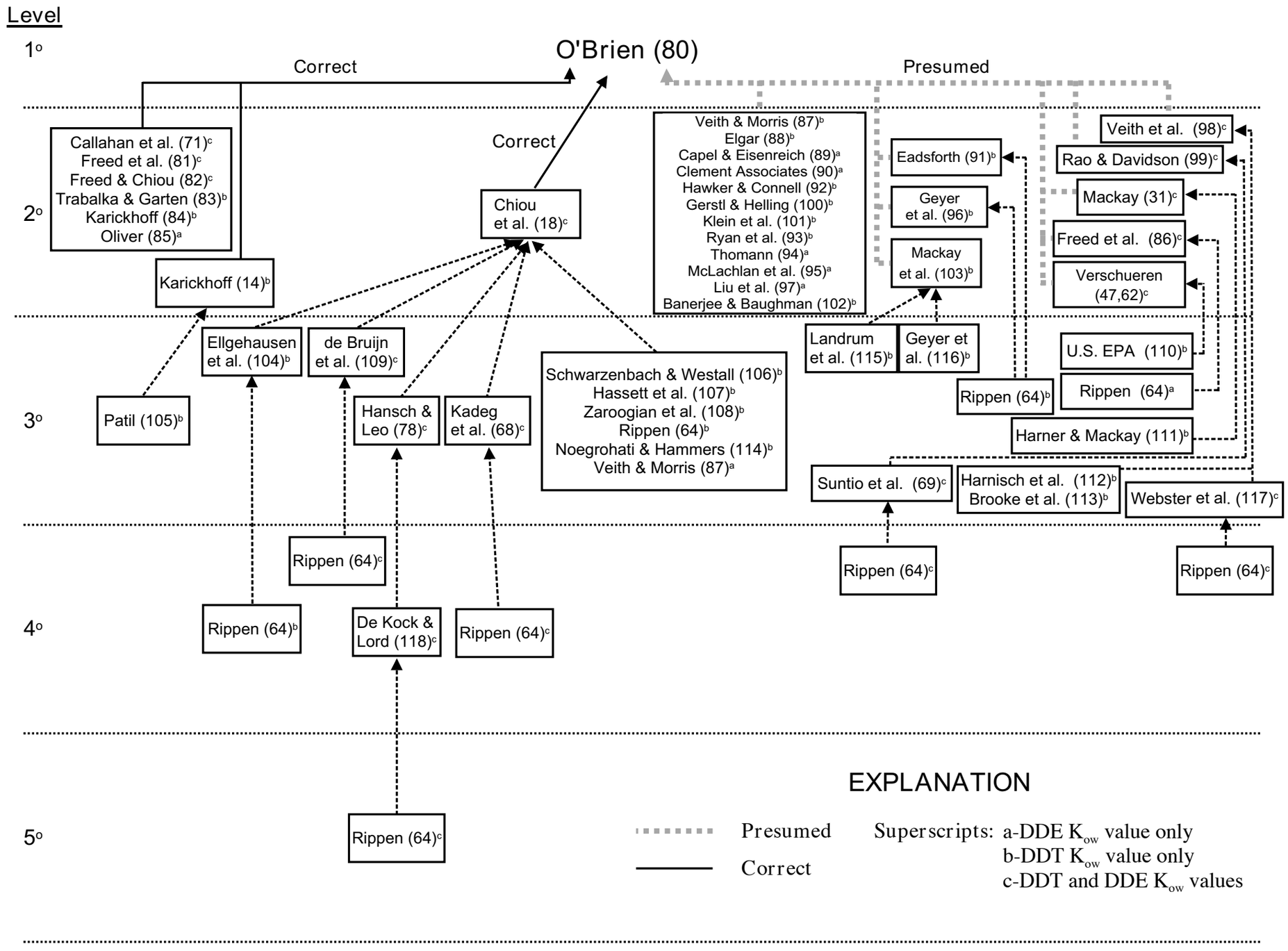


Figure 1. Reference tree showing original O'Brien (*80*) publication and secondary, tertiary and higher references as an example of multi-level referencing.

USGS Water-Resources Investigations Report 01-4201, 2001.

What if error was not reported or measured?
Reported measurements can vary over huge ranges

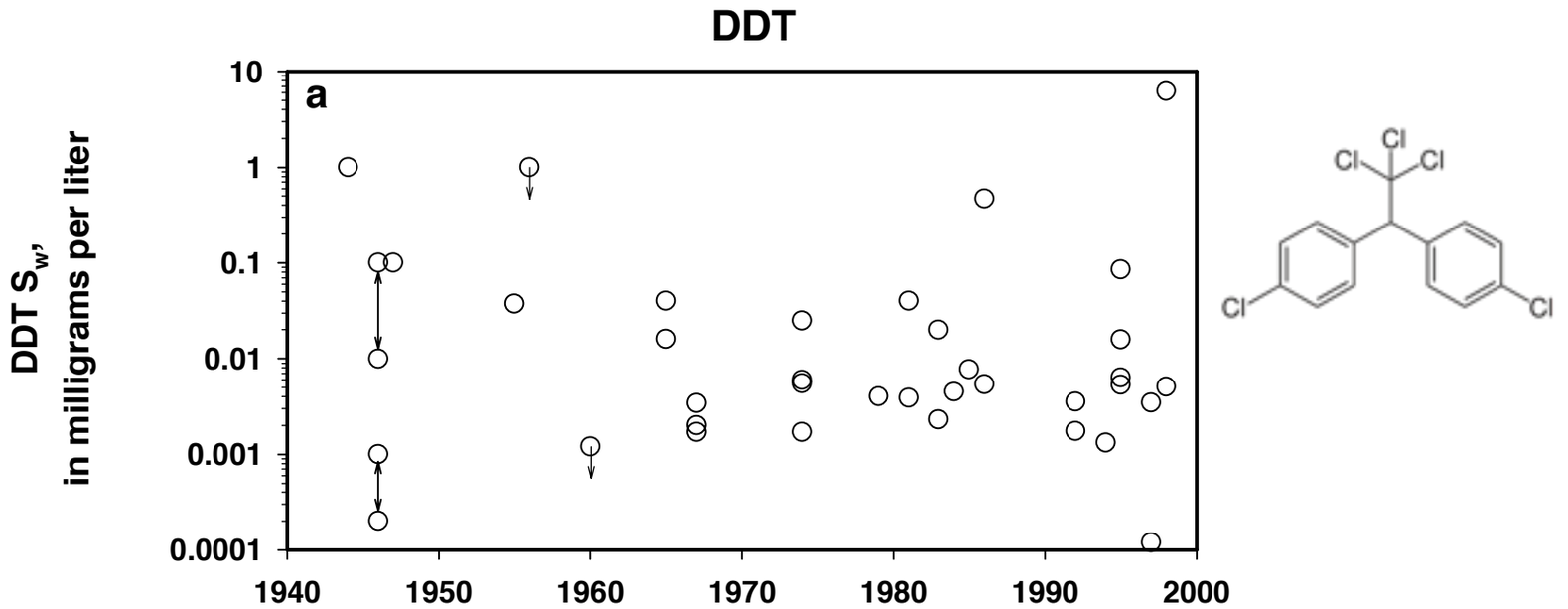


Table 6. Basic Statistics for Uncensored Original S_w and K_{ow} DDT/DDE Data[a]

	DDT	DDE
Statistic	**Aqueous solubility, S_w (milligrams per liter)**	
Mean	0.23	0.12
Standard Deviation	1.00	0.32

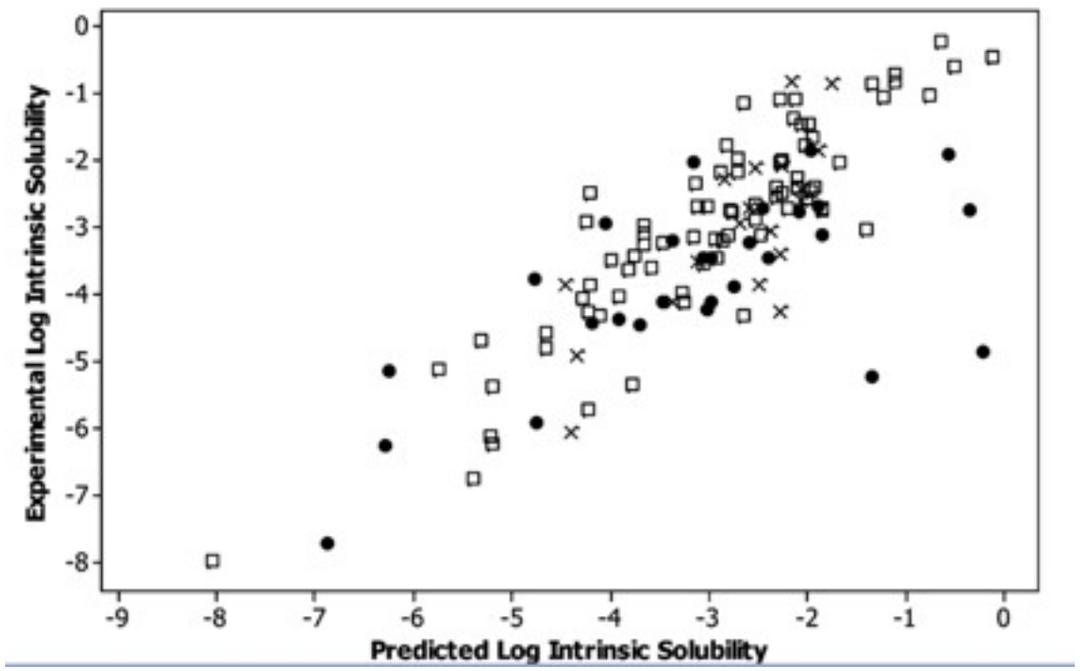
USGS Water-Resources Investigations Report 01-4201, 2001.

Noisy bad data – better than none?

- When is a lot of noisy data worse than much-less but high quality data?
  - Always
  - Always, *unless*, there is a reliable core of high quality data to which the lesser quality data can be compared
- When is NO measurement actually a better option?
  - Often
  - Example: Much high and mid throughput ADME data is not of a precision to help in optimization efforts

When is lots of noisy data worse than less higher quality data?

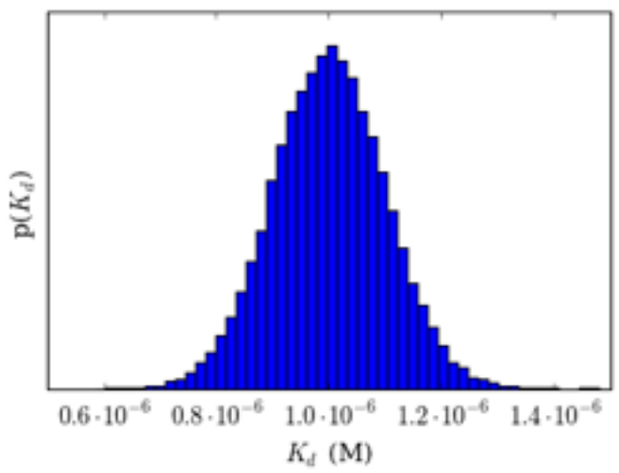
- Depends on the need. Will the data be used for:
  - Alerts
  - Crude general indicator model
  - Optimization
- Perhaps useful when large dynamic range is required and is more important than precision
  - E.g. for crude solubility models



M. Hewitt, et al 2009, J. Chem. Inf. Model, 49,2572–2587

Propagation of (t)error

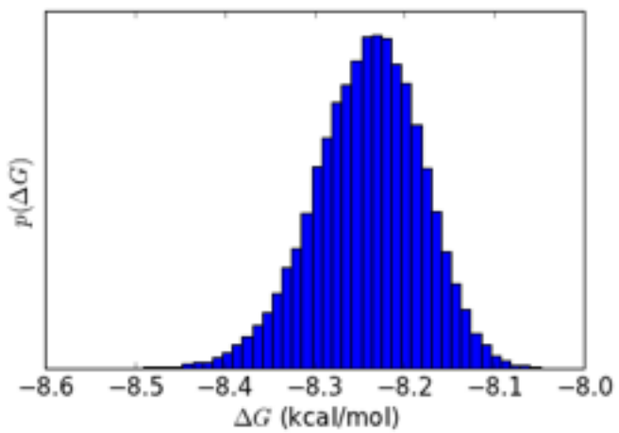
Suppose we make a measurement of the K_d and find 1.0±0.1 uM. (Assume the error in K_d is Gaussian)



Suppose we want to compute the free energy of binding ΔG from the K_d

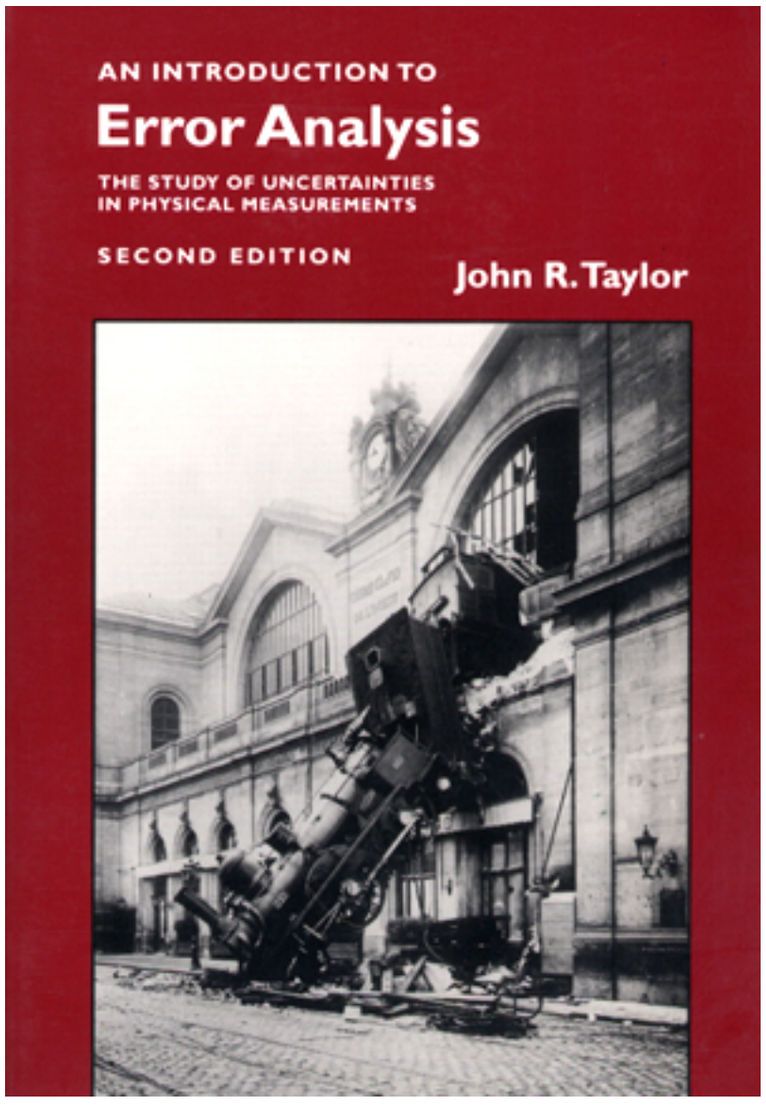
$$\Delta G_{\mathrm{bind}} = k_B T \ln(K_d C_0)$$

How can we propagate the error δK_d into ΔG?



IPython tutorial available at https://github.com/choderalab/cadd-grc-2013

Propagation of (t)error: What the book says



Adding two measurements with error:

$$z = x + y$$

$$\delta^2 z = \delta^2 x + \delta^2 y + \delta x \delta y$$

Multipying two measurements with error:

$$z = x \cdot y$$

$$\delta^2 z = x^2\, \delta^2 y + y^2\, \delta^2 x + xy\, \delta x \delta y$$

Dividing two measurements with error:

$$z = x/y$$

$$\delta^2 z = z^2 \left[\left(\frac{\delta x}{x} \right)^2 + \left(\frac{\delta y}{y} \right)^2 + \left(\frac{\delta x \delta y}{xy} \right) \right]$$

Where does this stuff come from?
Why is it so complicated?
Is there something easier?

IPython tutorial available at https://github.com/choderalab/cadd-grc-2013

Propagation of (t)error: Where does it come from?

Suppose we have a function of a measurement or random variable $y = f(x)$

We make a measurement with some uncertainty $x_0 \pm \delta x$

We can describe the behavior in the vicinity of the measurement using a first-order Taylor series expansion:

$$\Delta f \equiv f(x_0 + \Delta x) - f(x_0) = \left.\frac{\partial f}{\partial x}\right|_{x_0} \Delta x + \mathcal{O}(\Delta x^2)$$

The squared uncertainty is then given by its variance, omitting all but the first-order term:

$$\delta^2 f = E[\Delta f^2] = E\left[\left(\left.\frac{\partial f}{\partial x}\right|_{x_0} \Delta x\right)^2\right] = \left(\left.\frac{\partial f}{\partial x}\right|_{x_0}\right)^2 \delta^2 x$$

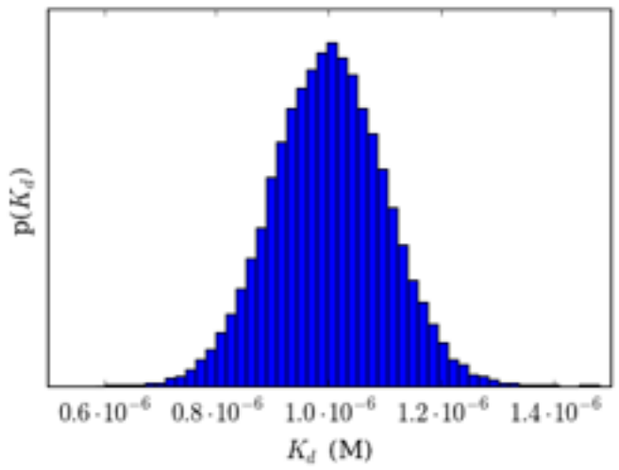
The uncertainty (standard error) is just the square root of the squared uncertainty:

$$\delta f = \left|\left.\frac{\partial f}{\partial x}\right|_{x_0}\right| \delta x$$

IPython tutorial available at https://github.com/choderalab/cadd-grc-2013

Propagation of (t)error: Converting affinity to free energy.

Suppose we make a measurement of the K_d and find 1.0±0.1 uM. (Assume the error in K_d is Gaussian)



Suppose we want to compute the free energy of binding ΔG from the K_d

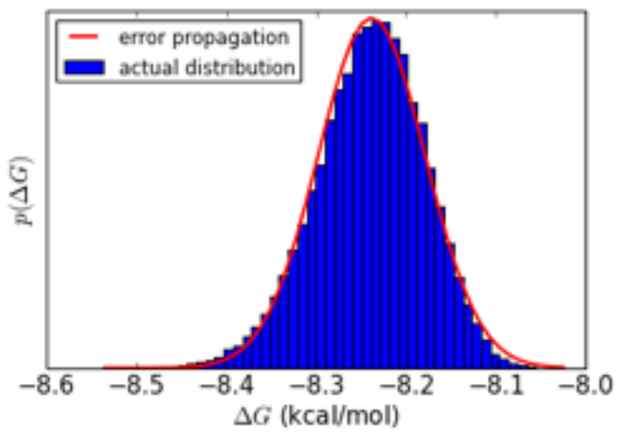
$$\Delta G_{\mathrm{bind}} = k_B T \ln(K_d C_0)$$

How can we propagate the error δK_d into ΔG?

$$\delta f = \left| \frac{\partial f}{\partial x} \right|_{x_0} \delta x$$

Substituting in the free energy definition above:

$$\delta \Delta G = \left| \frac{k_B T}{K_d C_0} \right| C_0 \delta K_d = k_B T \frac{\delta K_d}{K_d}$$



As we can see, a Gaussian with this standard deviation describes the uncertainty quite well!

IPython tutorial available at https://github.com/choderalab/cadd-grc-2013

Propagation of (t)error:
Where does it come from, anyway?

If we have a function of multiple measurements or random variables, $f(x_1, x_2, \ldots, x_N)$

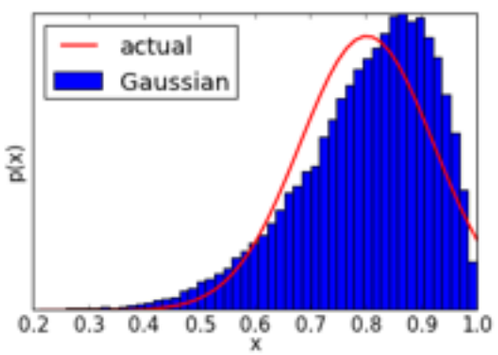
$$\delta^2 f = \sum_{i=1}^{N} \sum_{j=1}^{N} \left. \frac{\partial f}{\partial x_i} \frac{\partial f}{\partial x_j} \right|_{x_1,\ldots,x_N} \delta x_i \delta x_j$$

If independent, $\delta x_i \delta x_j = 0$ **for** $i \neq j$

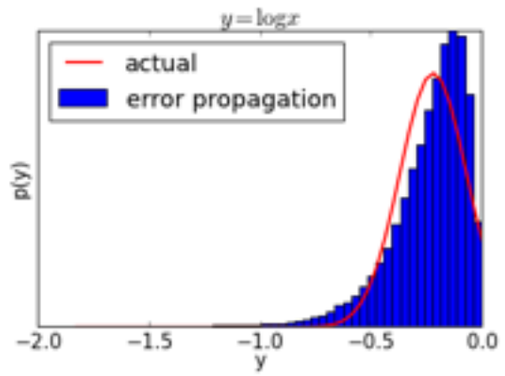
This can get totally messy as functions become complex or involve more variables.

CAUTION!

$x \in (0, 1]$



$y = \log x$



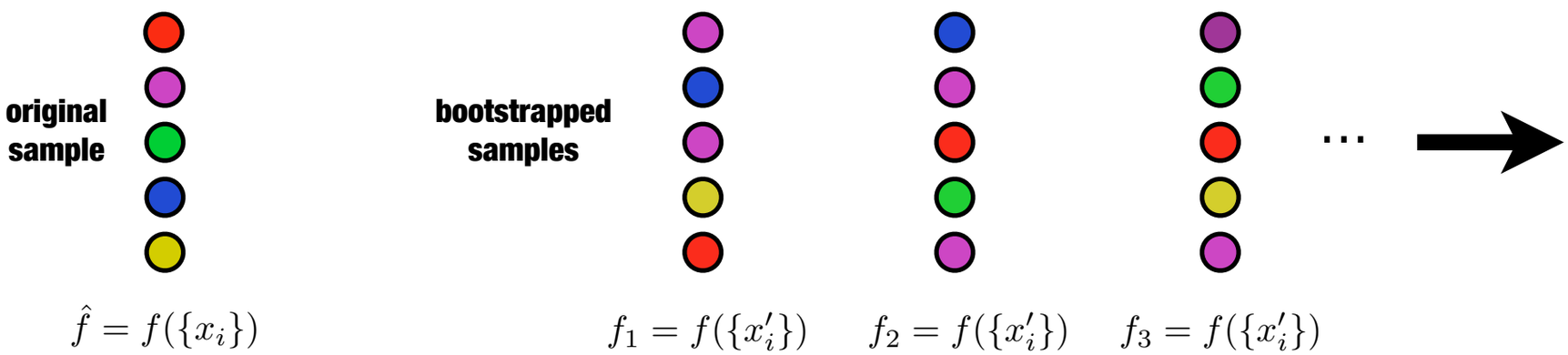
$y \in (-\infty, 0]$

For highly nonlinear functions, the first-order Taylor expansion doesn't do a good job. Especially important for values on intervals, like probabilities (p<0 and p>1 nonsensical).

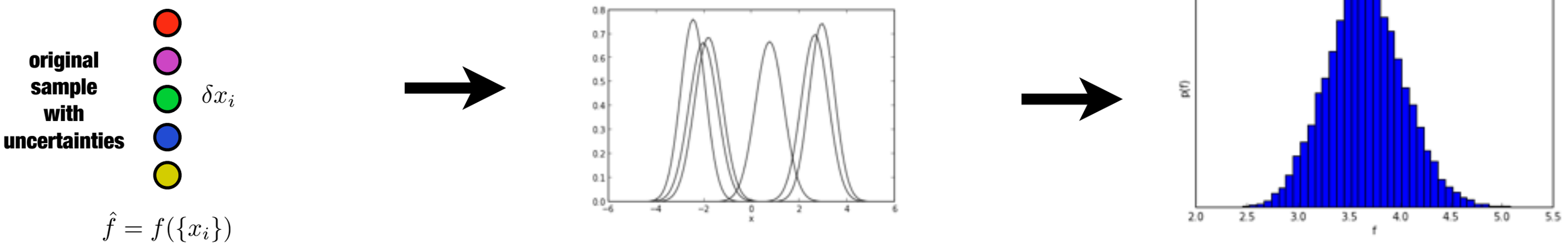
IPython tutorial available at https://github.com/choderalab/cadd-grc-2013

Three simple strategies for propagating error (without losing your mind)

Bootstrapping $f(\{x'_i\})$ **where** $\{x'_i\}$ **drawn with replacement from** $\{x_i\}$



Simulation $f(\{x'_i\})$ where $x'_i \sim p(x|x_i, \delta x_i)$



Bayesian inference

$$\theta \sim p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)\, p(\theta)$$

parameters **posterior** **data likelihood** **prior**

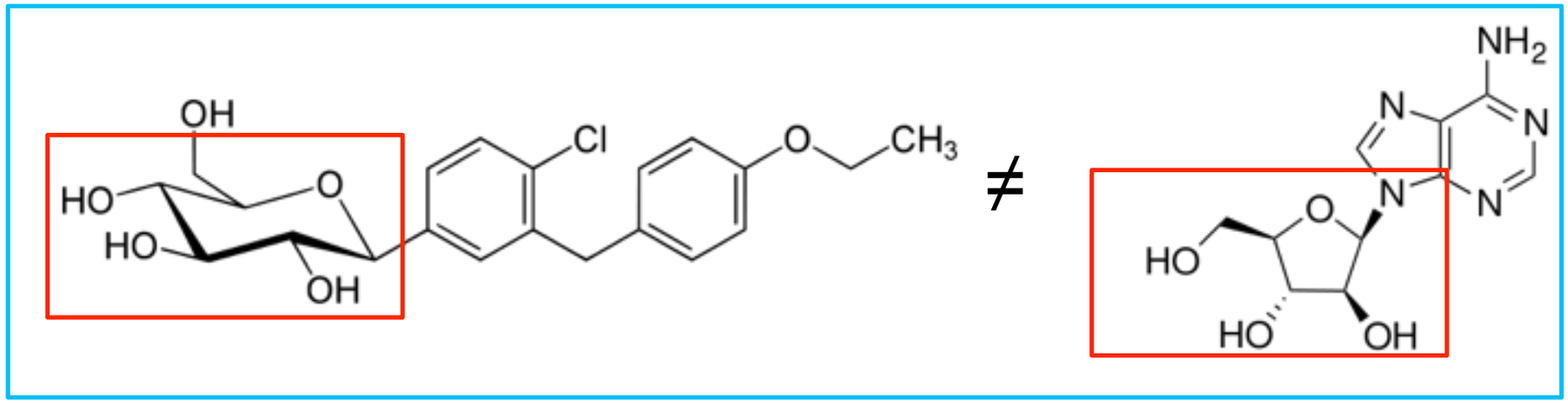
IPython tutorial available at https://github.com/choderalab/cadd-grc-2013

Noisy data: When is NO measurement or prediction actually a better option?

- When data will be over interpreted
  - When data is presented inappropriately
    - E.g. 23.427 ± 254.231
  - Example: Desperate chemists and solubility and PAMPA measurements
    - “I don’t care if the measurements aren’t quantitative, I need several significant figures”
    - “I don’t care if the answer is right, as long as its fast”
- When it will used alone without related data
- When used out of context
- If not reported with sufficient caveats and error estimates

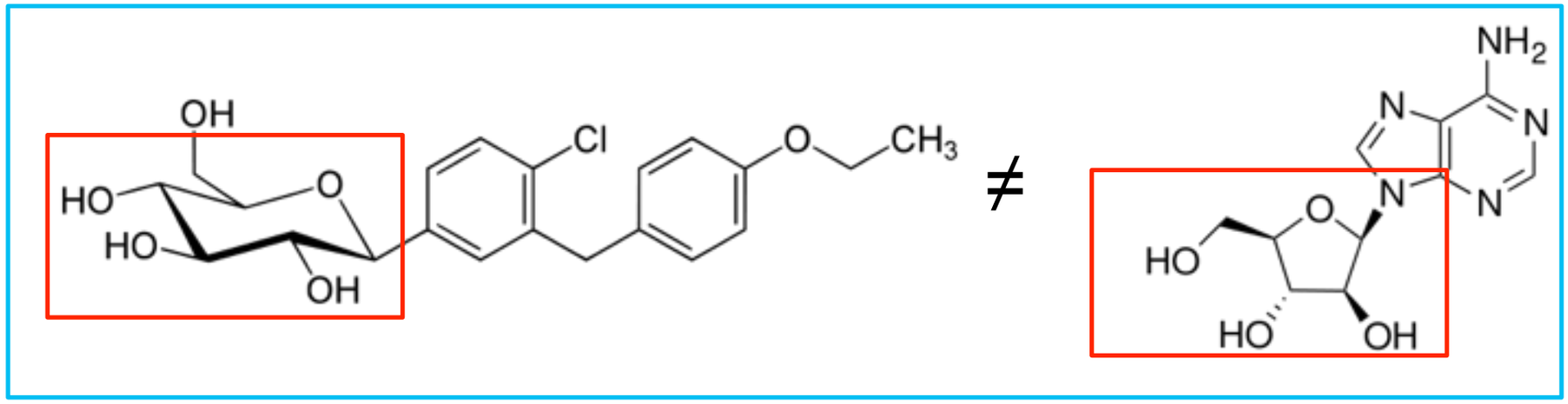
Noisy data: When is NO measurement or prediction actually a better option?

- Example: Over interpreted. without related data, out of context, without sufficient caveats and error estimates
- Blind over-interpretation of a mutagenicity alert from MultiCase toxicity prediction software that stalled a drug discovery program for 2 months



Noisy data: When is NO measurement or prediction actually a better option?

- Example: Over interpreted. without related data, out of context, without sufficient caveats and error estimates
- Blind over-interpretation of a mutagenicity alert from MultiCase toxicity prediction software that stalled a drug discovery program for 2 months



"There is no bad data, only bad data presentation"

"There are no bad soldiers, only bad generals"

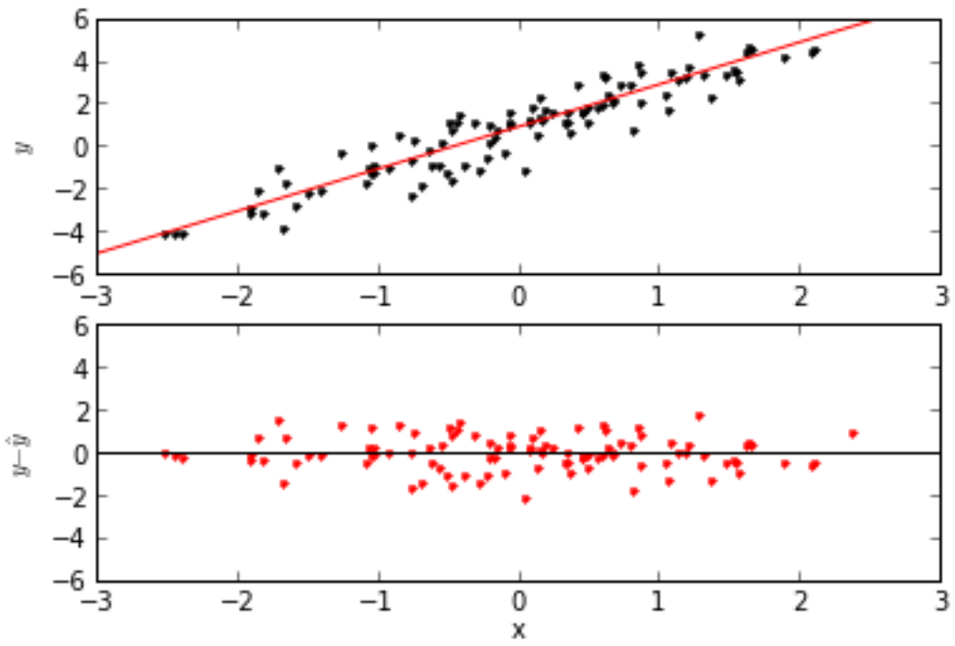
Noteworthy

- Models based on data of high confidence might be as good or better than many medium to high throughput assays
  - Given proper domain
  - Cyp 2C9
  - CACO-2
  - Met Stab
- Experimentalists could devote their time to high quality assays to provide data to enhance and improve models

Propagating error in simple models: A quick note on regression

$$\beta = \frac{\mathrm{cov}(x, y)}{\mathrm{var}(x)}$$

$$\alpha = \bar{y} - \beta \cdot \bar{x}$$



regression

residuals

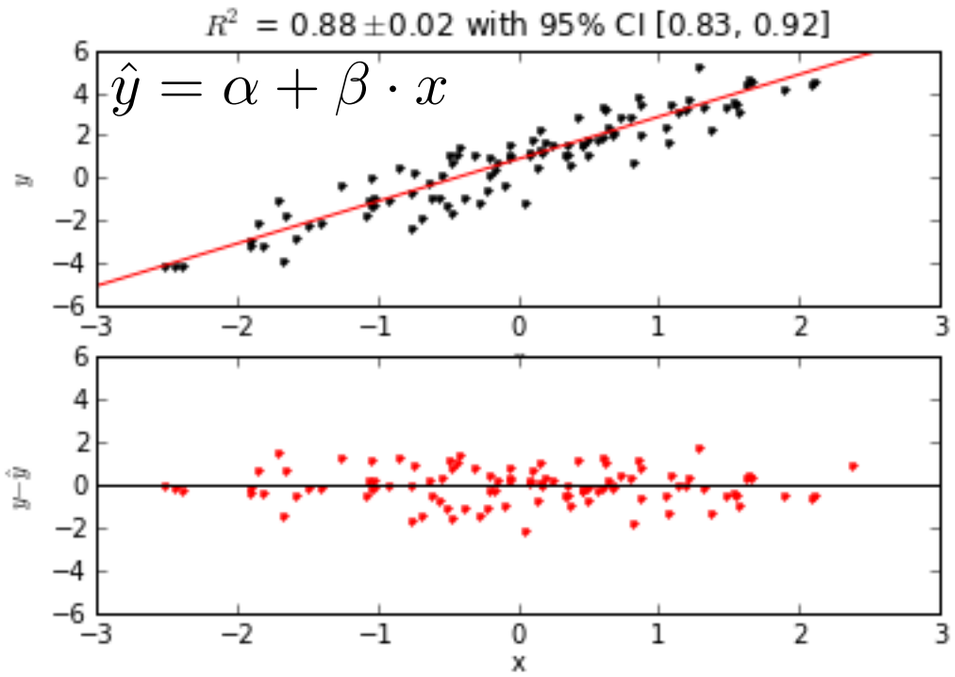
IPython tutorial available at https://github.com/choderalab/cadd-grc-2013

Propagating error in simple models: Bootstrapping can estimate regression uncertainty

$$\beta = \frac{\mathrm{cov}(x, y)}{\mathrm{var}(x)}$$

$$\alpha = \bar{y} - \beta \cdot \bar{x}$$

$$R^2 = 1 - \frac{\hat{\sigma}_r^2}{\hat{\sigma}^2}$$



residual variance

$$\hat{\sigma}_r^2 = \frac{1}{N}\sum_{n=1}^{N}(y_i - \hat{y}_i)^2$$

sample variance

$$\hat{\sigma}_r^2 = \frac{1}{N}\sum_{n=1}^{N}(y_i - \bar{y}_i)^2$$

Recall that uncertainty in coefficient of determination and fit parameters can be easily estimated by bootstrap.

IPython tutorial available at https://github.com/choderalab/cadd-grc-2013

Propagating error in simple models: Generalized R-squared

Cox & Snell pseudo-R^2

$$R^2_{CS} = 1 - \left(\frac{L(0)}{L(\hat{\theta})}\right)^{2/N}$$

Assume Gaussian error in deviations from model:

$$p(y|x) = (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{(y - \hat{y}(x))^2}{2\sigma^2}\right]$$

Write likelihood for whole dataset:

$$L(\theta) = \prod_{i=1}^{N} (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{(y_i - \hat{y}(x_i))^2}{2\sigma^2}\right] = (2\pi\sigma^2)^{-N/2} \exp\left[-\sum_{i=1}^{N} \frac{(y_i - \hat{y}(x_i))^2}{2\sigma^2}\right] = (2\pi\sigma^2)^{-N/2} e^{-N\hat{\sigma}_r^2/2\sigma^2}$$

If we don't know data error, we assume residual variance dominates:

$$\sigma^2 = \sigma_r^2$$

This gives us likelihoods for fit model and null model (where data mean is assumed):

$$L(\hat{\theta}) = (2\pi\hat{\sigma}_r^2)^{-N/2} e^{-N/2}$$
$$L(0) = (2\pi\hat{\sigma}^2)^{-N/2} e^{-N/2}$$

Plugging these in, we recover the standard definition of the coefficient of determination!

$$R^2_{CS} = 1 - \left(\frac{(2\pi\hat{\sigma}^2)^{-N/2} e^{-N/2}}{(2\pi\hat{\sigma}_r^2)^{-N/2} e^{-N/2}}\right)^{2/N} = \left(\frac{\hat{\sigma}_r^2}{\hat{\sigma}^2}\right) = R^2$$

residual variance

$$\hat{\sigma}_r^2 = \frac{1}{N}\sum_{n=1}^{N}(y_i - \hat{y}_i)^2$$

sample variance

$$\hat{\sigma}_r^2 = \frac{1}{N}\sum_{n=1}^{N}(y_i - \bar{y}_i)^2$$

IPython tutorial available at https://github.com/choderalab/cadd-grc-2013

Propagating error in simple models: Fitting data with differing errors

The regression problem can be generalized to the case where each measurement y_i has a corresponding uncertainty σ_i.

The likelihood is then

$$L(\theta) = \prod_{i=1}^{N} (2\pi\sigma_i^2)^{-1/2} \exp\left[-\frac{(y_i - \hat{y}(x_i))^2}{2\sigma_i^2}\right]$$

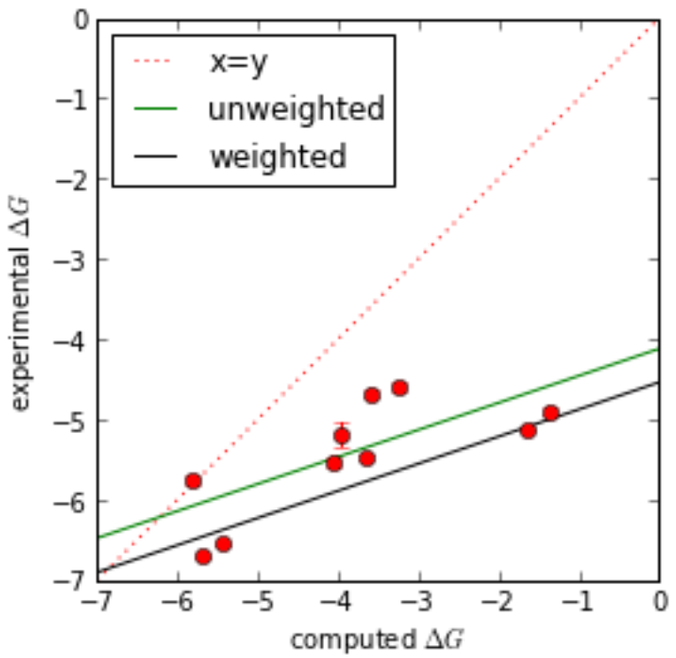
It is generally easier to maximize the *log-likelihood* instead:

$$l(\theta) = \sum_{i=1}^{N}\left[-\frac{1}{2}\log\left(2\pi\sigma_i^2\right) - \frac{(y_i - \hat{y}(x_i))^2}{2\sigma_i^2}\right] = -\frac{N}{2}\log(2\pi) + \sum_{i=1}^{N}\left[-\log\sigma_i^2 - \frac{(y_i - \hat{y}(x_i))^2}{2\sigma_i^2}\right]$$

To find the maximum likelihood estimate $\hat{\theta}$, we use the derivatives:

$$\frac{\partial}{\partial\theta} l(\theta) = \sum_{i=1}^{N} \frac{(y_i - \hat{y}(x_i))}{\sigma_i^2} \left.\frac{\partial\hat{y}}{\partial\theta}\right|_{x_i}$$

T4 lysozyme L99A
experiment vs alchemical free energy



IPython tutorial available at https://github.com/choderalab/cadd-grc-2013

Mobley et al. J Mol Biol 371:1118, 2007.

Outlier detection- when is it safe to drop data?

- Is there some reason other than it does not fit the current model?
  - Suspicious or simply inconvenient?
- Is the data “bad”?
  - Is it warranted by the experimental metadata?
    - Age of compound / Date of assay (vs. that of other compounds) / Purity / Assay conditions / Source
  - Are the physical properties suspicious (calculated or actual)
    - Reactivity / Promiscuity / Solubility
- Is the measurement or compound just different?
  - The compound
    - logP, MW, # rotatable bonds, amount of stereo centers, etc
  - The measurement
    - Leverage – is the value way outside of bounds?
    - Assay variation

What can we do about outliers?
Noise accommodation vs. outlier detection

Outliers are data that are **highly improbable** for the model under consideration.
They often represent real-world anomalies, and may be interesting or useless.

Outliers may be due to:
- change in environment/assay conditions
- unappreciated physical phenomena
- instrumentation error
- human error (mishandling, data entry errors)
- incorrect conversions
- numeric codes to represent missing data

We deal with outliers in one of two ways:
1. If our model can't handle outliers, we use **noise removal** or **noise accommodation** [e.g. robust statistics, robust regression] c.f. Maria Gallardo's dipole moments
2. If the outliers are important phenomena we're trying to identify: **outlier detection** [e.g. uncharacterized effects, significant events, compounds to retest]

Sometimes, frequentists can be handy
A simple statistical tests for outliers: Grubbs' test

Assume normality of data distribution; sequentially identify outliers.

Grubbs' test statistic: $G = \frac{\max_{i=1,\ldots,N} |y_i - \bar{y}|}{s}$

s **sample standard deviation**
$\bar{y}$ **sample mean**
α **significance level (e.g. 0.05)**

There are outliers if we find $G > G_{\text{test}}$

$$G_{\text{test}} = \frac{N-1}{\sqrt{N}} \sqrt{\frac{t^2}{N-2+t^2}}$$

t **is the upper critical value of t-distribution with N-2 dof and significance level** $\alpha/(2N)$

(Bonferroni correction)

```
from scipy import stats
alpha = 0.95 # 95% significance level
t = stats.t.isf(1-alpha/(2*N), N-2)
Gtest = (N-1)/sqrt(N)*sqrt(t**2 / (N-2+t**2))
```

IPython tutorial available at https://github.com/choderalab/cadd-grc-2013

Statistical tests for outliers: Grubbs' test
An example from your nightmares

Technical Note

Does Your Lab Coat Fit to Your Assay?

Journal of Biomolecular Screening
18(6) 744–747

DOI: 10.1177/1087057113481621
jbx.sagepub.com
SAGE

Michael Busch[1], Heinz Bjoern Thoma[1], and Ingo Kober[1]

Abstract
An explanation for randomly occurring spikes on microplates in fluorescence-based assays employing shorter-wavelength readouts is presented. It is demonstrated that lint originating from standard (white cotton) lab coats is most likely to be responsible for such artifacts in assays applying wavelengths at 380 nm excitation and 450 nm emission. The fluorescence properties of this lint are discussed and compared with those of optical brighteners. An alternative to the use of cotton-based lab coats is presented, which led to a reduction of spikes in a high-throughput screening campaign by 90%.

Keywords
HTS, fluorescence, spikes, lab coat, dust, lint

"During assay development...about 5 to 10 randomly distributed spikes (signals up to 8 times higher than a normal assay signal) were found on each 384-well plate. This would have led to 1% to 2% of false-positives in HTS..."

Statistical tests for outliers: Grubbs' test
An example from your nightmares

empty microplate fresh out of packaging

53	62	63	70	77	62	61	58	61	69	70	53	63	82	60	82	63	50	56	61	59	90	66	63
53	40	60	56	70	57	64	64	57	59	73	70	63	59	65	60	67	57	65	46	75	82	83	57
57	56	68	74	66	77	53	44	69	67	60	66	73	67	63	51	62	78	73	44	56	59	62	48
68	47	60	74	75	65	76	60	62	54	65	67	66	58	75	58	54	71	61	61	79	60	74	62
62	55	55	67	66	65	51	51	48	64	82	61	86	78	62	68	59	68	56	55	74	80	61	60
69	64	72	64	63	62	55	53	66	64	89	71	59	65	67	52	67	58	59	60	70	49	74	63
52	60	59	62	71	86	59	55	60	61	76	73	73	68	62	63	71	87	74	66	56	73	82	79
58	72	66	77	86	77	58	64	63	61	73	57	58	57	61	59	55	60	80	52	47	51	68	60
78	75	52	58	87	65	62	66	71	56	66	55	46	57	73	54	80	84	80	54	63	76	58	65
68	70	62	57	50	80	70	68	44	51	65	59	71	77	69	71	68	77	71	64	61	63	44	92
62	63	64	76	57	75	74	55	65	69	62	53	51	72	50	69	65	59	55	65	74	70	69	71
58	65	62	57	54	86	71	61	64	69	63	64	63	50	66	52	44	65	60	60	59	62	64	68
54	48	53	47	59	63	55	55	55	68	53	51	50	73	65	56	80	65	80	63	79	82	57	51
55	60	56	56	61	81	68	60	62	64	59	66	48	75	64	42	68	62	63	55	80	76	60	50
42	67	74	65	53	59	62	71	67	67	63	65	58	48	57	60	53	50	69	67	43	40	46	35
56	54	57	71	60	64	80	76	49	72	62	66	62	60	87	64	72	63	59	67	61	56	58	55

RFU: 10 100 500 1000 5000 10000 20000 30000 40000

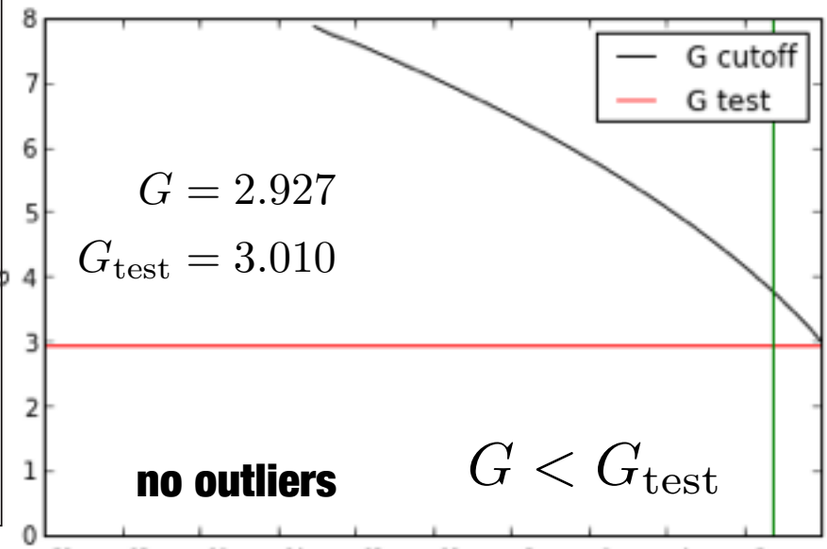


Figure 1. Empty microplate, freshly taken out of a closed bag. Fluorescence readout at 380/450 nm ex/em. The medium signals were less than 70 RFU.

empty microplate exposed to laboratory air for 2 min

53	81	182	58	61	113	69	56	54	63	58	704	66	53	51	60	87	79	87	83	57	56	87	108
63	60	71	63	60	67	69	48	52	38	62	81	54	63	56	57	78	108	64	61	66	68	63	58
69	89	50	67	73	59	56	51	64	51	72	58	269	81	50	59	64	59	54	70	94	55	49	67
66	68	42	91	71	82	63	76	58	50	63	68	60	65	47	60	82	52	75	67	68	61	73	69
59	65	83	72	66	59	47	54	57	76	61	76	84	76	63	47	59	62	74	79	58	88	65	58
61	56	59	74	68	61	71	60	66	49	56	59	56	53	56	50	50	59	63	46	61	70	61	69
57	64	54	59	85	81	50	52	54	66	81	61	62	58	63	59	68	62	49	73	70	71	75	48
61	86	71	59	58	63	69	52	74	66	76	53	89	50	59	52	55	47	48	52	70	68	88	44
59	56	43	146	68	53	68	62	56	57	67	61	58	67	71	52	66	79	60	62	70	63	64	76
56	75	53	69	59	67	240	60	50	66	64	46	61	81	66	43	73	68	57	77	636	60	143	75
47	67	40	57	89	56	44	3448	80	58	56	10122	53	62	65	85	59	70	54	69	66	68	66	62
46	64	90	63	200	81	73	64	54	66	55	50	66	61	48	59	63	47	59	69	67	54	71	65
67	78	61	77	69	49	88	65	76	66	68	74	70	53	55	63	44	55	86	48	58	56	76	58
79	76	56	67	60	75	72	57	65	65	52	59	61	64	53	59	53	52	64	137	58	58	53	73
57	1195	68	61	59	61	61	59	63	56	53	64	57	48	57	57	49	64	48	2171	58	58	60	60
65	61	73	50	51	59	71	80	49	74	228	49	66	52	66	62	78	82	106	56	57	64	69	54

RFU: 10 100 500 1000 5000 10000 20000 30000 40000

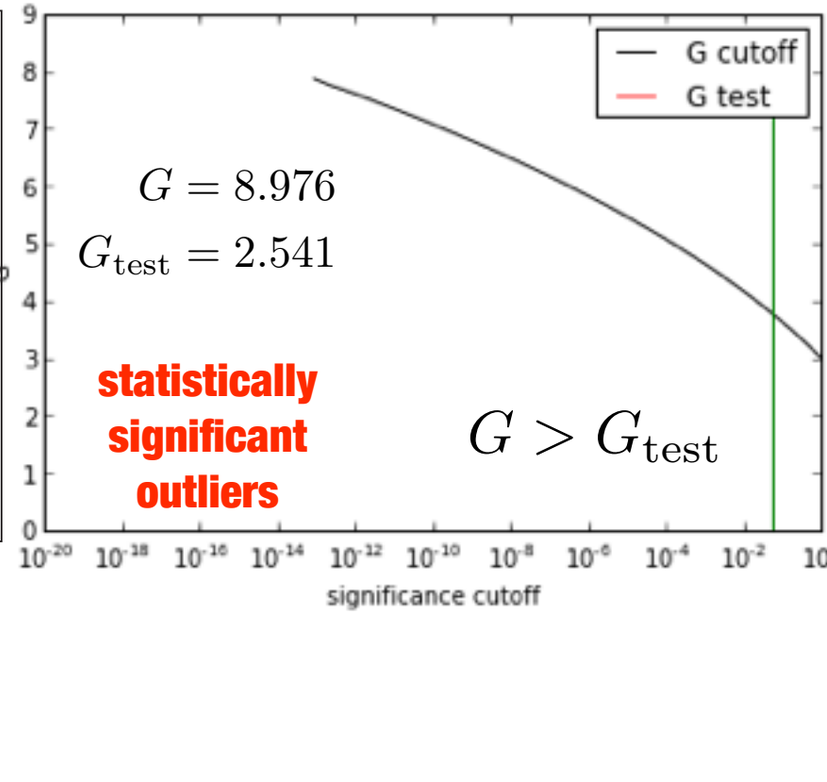


Figure 2. Measurement of the same plate after moving it for 2 min in free laboratory air. Spikes up to 10 000 RFU were observed here.

Statistical tests for outliers: Grubbs' test
An example from your nightmares

empty microplate fresh out of packaging

53	62	63	70	77	62	61	58	61	69	70	53	63	82	60	82	63	50	56	61	59	90	66	63
53	40	60	56	70	57	64	64	57	59	73	70	63	59	65	60	67	57	65	46	75	82	83	57
57	56	68	74	66	77	53	44	69	67	60	66	73	67	63	51	62	78	73	44	56	59	62	48
68	47	60	74	75	65	76	60	62	54	65	67	66	58	75	58	54	71	61	61	79	60	74	62
62	55	55	67	66	65	51	51	48	64	82	61	86	78	62	68	59	68	56	55	74	80	61	60
69	64	72	64	63	62	55	53	66	64	89	71	59	65	67	52	67	58	59	60	70	49	74	63
52	60	59	62	71	86	59	55	60	61	76	73	73	68	62	63	71	87	74	66	56	73	82	79
58	72	66	77	86	77	58	64	63	61	73	57	58	57	61	59	55	60	80	52	47	51	68	60
78	75	52	58	87	65	62	66	71	56	66	55	46	57	73	54	80	84	80	54	63	76	58	65
68	70	62	57	50	80	70	68	44	51	65	59	71	77	69	71	68	77	71	64	61	63	44	92
62	63	64	76	57	75	74	55	65	69	62	53	51	72	50	69	65	59	55	65	74	70	69	71
58	65	62	57	54	86	71	61	64	69	63	64	63	50	66	52	44	65	60	60	59	62	64	68
54	48	53	47	59	63	55	55	55	68	53	51	50	73	65	56	80	65	80	63	79	82	57	51
55	60	56	56	61	81	68	60	62	64	59	66	48	75	64	42	68	62	63	55	80	76	60	50
42	67	74	65	53	59	62	71	67	67	63	65	58	48	57	60	53	50	69	67	43	40	46	35
56	54	57	71	60	64	80	76	49	72	62	66	62	60	87	64	72	63	59	67	61	56	58	55

RFU: 10 100 500 1000 5000 10000 20000 30000 40000

Figure 1. Empty microplate, freshly taken out of a closed bag. Fluorescence readout at 380/450 nm ex/em. The medium signals were less than 70 RFU.

empty microplate exposed to laboratory air for 2 min

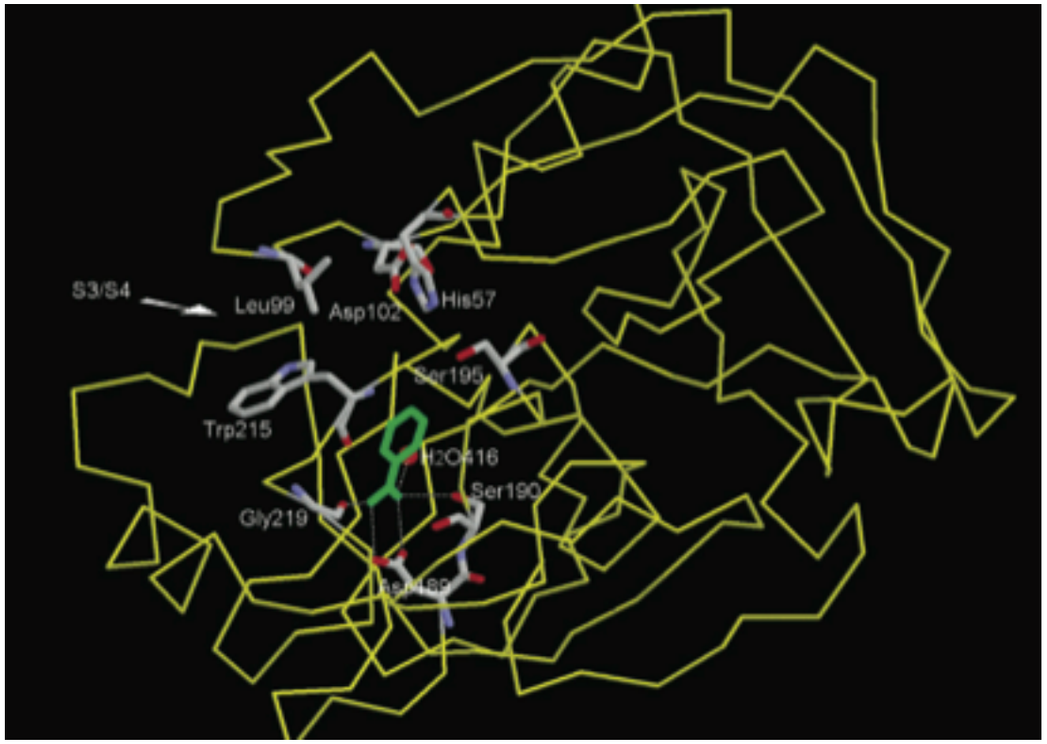
53	81	182	58	61	113	69	56	54	63	58	704	66	53	51	60	87	79	87	83	57	56	87	108
63	60	71	63	60	67	69	48	52	38	62	81	54	63	56	57	78	108	64	61	66	68	63	58
69	89	50	67	73	59	56	51	64	51	72	58	269	81	50	59	64	59	54	70	94	55	49	67
66	68	42	91	71	82	63	76	58	50	63	68	60	65	47	60	82	52	75	67	68	61	73	69
59	65	83	72	66	59	47	54	57	76	61	76	84	76	63	47	59	62	74	79	58	88	65	58
61	56	59	74	68	61	71	60	66	49	56	59	56	53	56	50	50	59	63	46	61	70	61	69
57	64	54	59	85	81	50	52	54	66	81	61	62	58	63	59	68	62	49	73	70	71	75	48
61	86	71	59	58	63	69	52	74	66	76	53	89	50	59	52	55	47	48	52	70	68	88	44
59	56	43	146	68	53	68	62	56	57	67	61	58	67	71	52	66	79	60	62	70	63	64	76
56	75	53	69	59	67	240	60	50	66	64	46	61	81	66	43	73	68	57	77	636	60	143	75
47	67	40	57	89	56	44	3448	80	58	56	10122	53	62	65	85	59	70	54	69	66	68	66	62
46	64	90	63	200	81	73	64	54	66	55	50	66	61	48	59	63	47	59	69	67	54	71	65
67	78	61	77	69	49	88	65	76	66	68	74	70	53	55	63	44	55	86	48	58	56	76	58
79	76	56	67	60	75	72	57	65	65	52	59	61	64	53	59	53	52	64	137	58	58	53	73
57	1195	68	61	59	61	61	59	63	56	53	64	57	48	57	57	49	64	48	2171	58	58	60	60
65	61	73	50	51	59	71	80	49	74	228	49	66	52	66	62	78	82	106	56	57	64	69	54

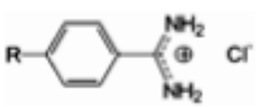
RFU: 10 100 500 1000 5000 10000 20000 30000 40000

Figure 2. Measurement of the same plate after moving it for 2 min in free laboratory air. Spikes up to 10 000 RFU were observed here.

Dark tales of woe:
Entropy-enthalpy compensation: Fact or fiction?

p-alkylbenzamidinium inhibitors of trypsin

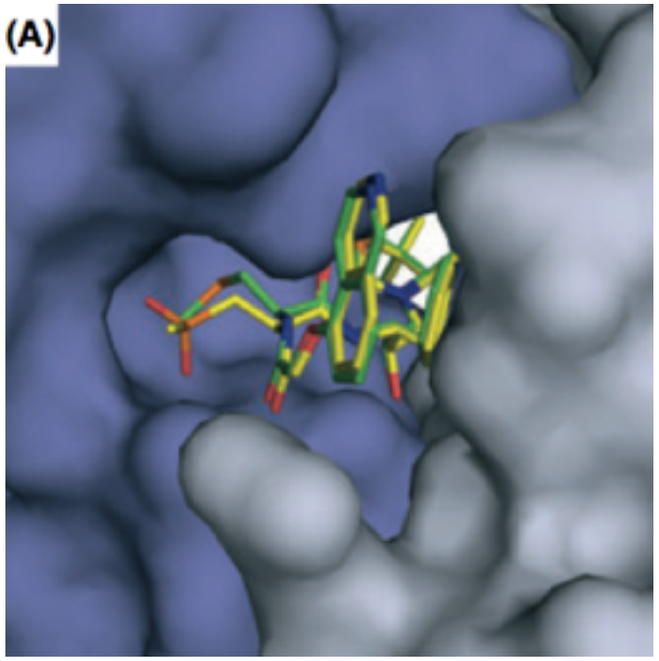


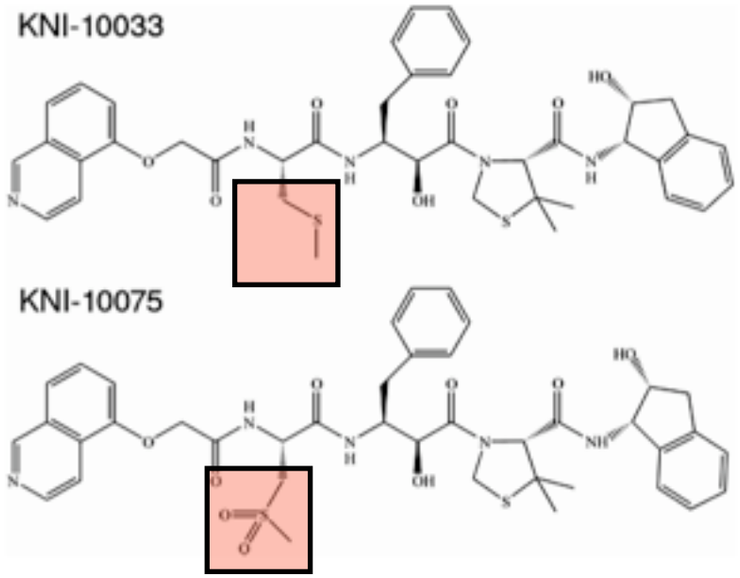


	ΔG (kJ/mol)	ΔH	TΔS
R = H	-26.6±0.3	-18.9±0.4	7.7±0.6
R = *n*-Bu	-26.2±0.3	-9.9±0.5	16.3±0.7

Talhout R, Villa A, Mark AE, Engberts JBFN. JACS 125:10570, 2003.

HIV-1 protease inhibitors





	ΔG (cal/mol)	ΔH (cal/mol)	−TΔS (cal/mol)	K_d (M)
KNI-10033 → pWT	−14 870 ± 90	−8200 ± 230	−6670 ± 90	$1.3^{-11} \pm 2\ 10^{-12}$
KNI-10075 → pWT	−14 620 ± 190	−12 120 ± 610	−2500 ± 190	$2\ 10^{-11} \pm 8\ 10^{-12}$

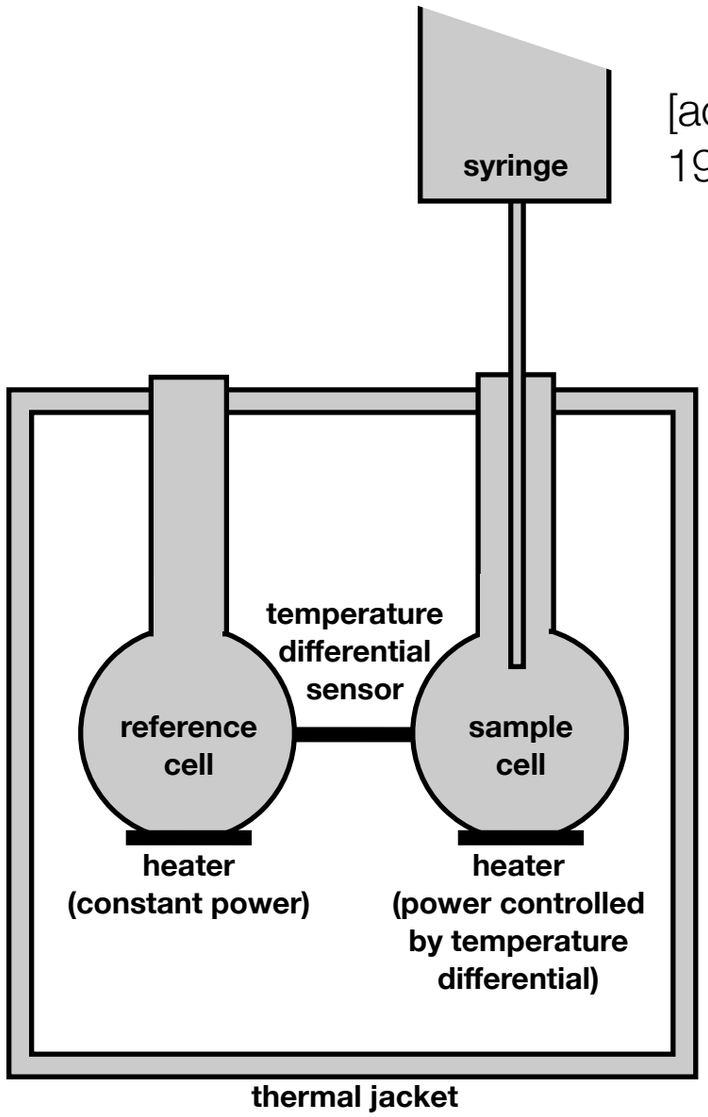
The enthalpy changes were measured at 15, 25 and 35 °C to determine the heat capacity change. (ΔC_p) is equal to −461 ± 34 cal/K/mol for KNI-10033 and −305 ± 41 cal/K/mol for KNI-10075.

Small chemical modifications appear to drastically change entropic and enthalpic components, but have little effect on affinity.

Lafont V, Armstrong AA, Ohtaka H, Kiso Y, Azmel LM, and Freire E.
Compensating enthalpic and entropic changes hinder binding affinity optimization.
Chem Biol. Drug Des. 69:413, 2007.

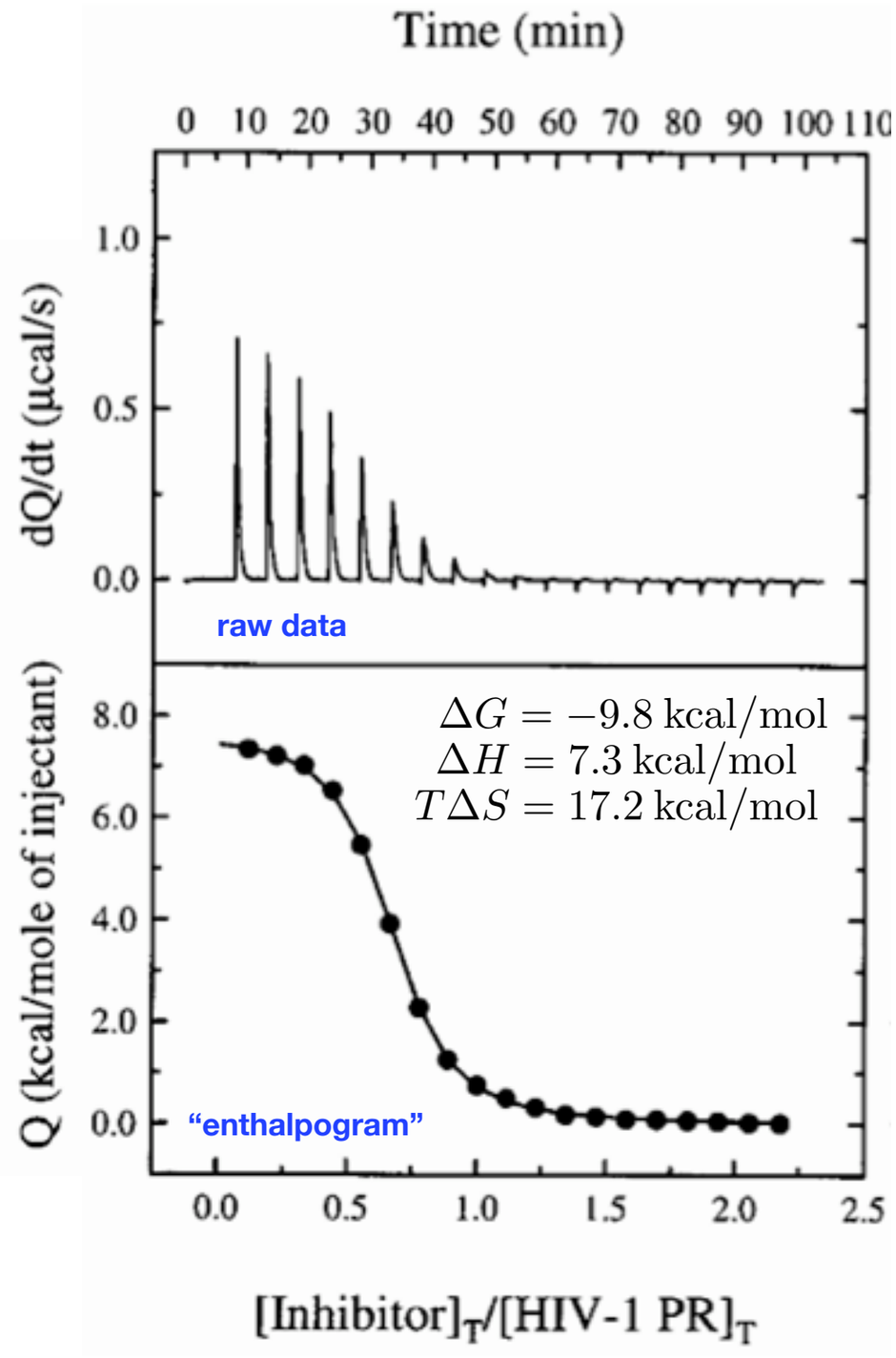
Isothermal titration calorimetry (ITC) can simultaneously interrogate free energies and enthalpies of binding

[illustrative experiment]



[acetyl pepstatin] = 200 uM
19 injections @ 10 uL/injection

[HIV-1 PR] = 20 uM in 1.4 mL
(0.6 mg/cell)



(Note that some reactions have no measurable change in heat, and are not measurable by ITC.)

Velazquez-Campoy A, Kiso Y, and Friere E. Arch. Biochem. Biophys. 390:169, 2001.

How reliable is calorimetric data, really?
A test of variation among truly independent experiments

The ABRF-MIRG'02 study:

Send identical aliquots of the same sample of protein and ligand to 14 core analysis facilities (experts!) and ask them to report the measured ΔG and ΔH by ITC.

The should get the same answer, within error.

The reported errors should match the variation among the reported results.

This experiment is almost never repeated because of the large quantity of protein needed for one ITC experiment, and the undesirability of repeating the experiment from scratch multiple times.

This is pretty much the only dataset of its kind reported in the literature.

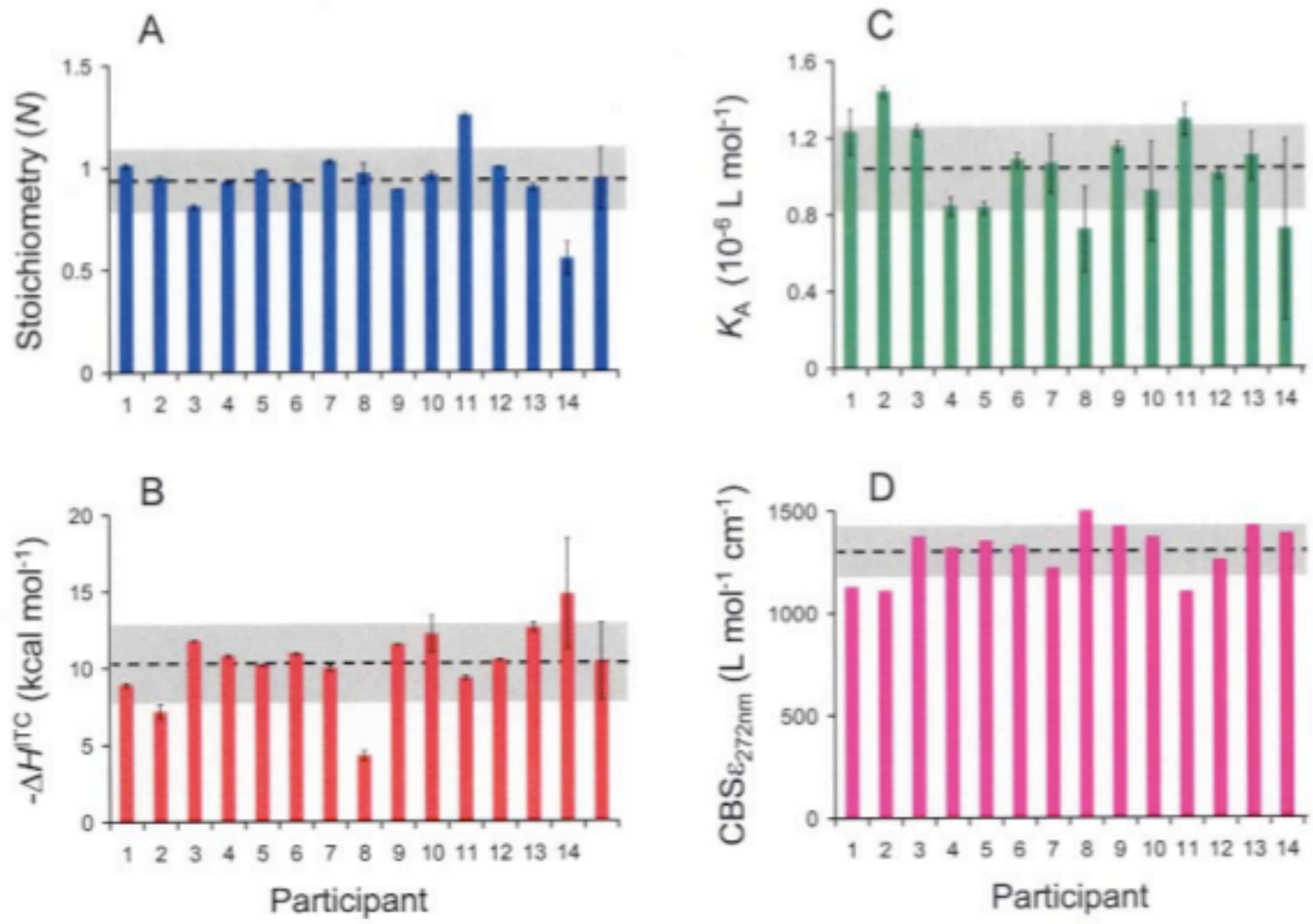


FIGURE 5

ITC characterization of the CBS/CA II interaction (based on Fig. 4 and Table 3). **A:** Stoichiometry (0.94 ± 0.15). **B:** Enthalpy upon binding (−10.4 ± 2.5 kcal mol^{-1}). **C:** Affinity [(1.00 ± 0.22) × 10^6 L mol^{-1}]. **D:** molar extinction coefficient for CBS at 272 nm (1307 ± 126 L mol^{-1} cm^{-1}). Mean values and the standard deviation for 14 determinations are represented by the horizontal *dotted lines* and *gray bands*. Error bars denote the standard deviation of nonlinear least squares analysis, except for participants 10 and 14 who reported standard deviations for replicate analyses. No errors were reported for the extinction coefficient determinations.

Myszka et al. *J. Biomol. Tech.* 14:247, 2003.

How reliable is calorimetric data, really?
Error is significantly underreported

TABLE 3

Summary of Reported Isothermal Titration Calorimetry Results[a]

Participant	Molar Binding Ratio (*N*)	K_A ($10^{-6} \times$ L mol^{-1})	ΔH^{ITC} (kcal mol^{-1})	C-value[b]	Control Titrations
1	1.01 ± 0.01	1.2 ± 0.1	−8.9 ± 0.1	17	CBS into buffer and buffer into CA II. Former subtracted as part of data analysis.
2	0.95 ± 0.01	1.44 ± 0.03	−7.2 ± 0.5	75	None
3	0.81 ± 0.01	1.24 ± 0.03	−11.8 ± 0.02	44	CBS into buffer and buffer into CA II. Former subtracted as part of data analysis.
4	0.929 ± 0.007	0.84 ± 0.05	−10.8 ± 0.1	24	Pilot run.
5	0.987 ± 0.003	0.84 ± 0.03	−10.20 ± 0.04	41	CBS into buffer, subtracted in data analysis.
6	0.921 ± 0.003	1.08 ± 0.04	−10.95 ± 0.05	39	CBS into buffer, subtracted in data analysis.
7	1.03 ± 0.01	1.1 ± 0.2	−10.0 ± 0.2	55	CBS into buffer, subtracted in data analysis.
8	0.97 ± 0.05	0.7 ± 0.2	−4.3 ± 0.3	7.0	CBS into buffer and buffer into CA II. Both subtracted as part of data analysis.
9	0.891 ± 0.002	1.15 ± 0.03	−11.53 ± 0.04	59	CBS into buffer, average value subtracted in analysis.
10[c]	0.96 ± 0.02	0.9 ± 0.2	−12 ± 1	34	CBS into buffer and buffer into CA II. Average of former used in analysis.
11	1.25 ± 0.01	1.3 ± 0.1	−9.3 ± 0.1	9.0	None
12	1.000 ± 0.003	1.01 ± 0.03	−10.51 ± 0.04	29	Buffer into buffer.
13	0.90 ± 0.02	1.1 ± 0.1	−12.6 ± 0.3	12	CBS into buffer, linear regression subtracted in analysis.
14[c]	0.55 ± 0.08	0.7 ± 0.3	−15 ± 4	22	CBS into buffer, subtracted in analysis.

ACTUAL STANDARD DEVIATION **± 0.2 µM** **± 2 kcal/mol**

± 0.1 kcal/mol

Note that observed 20% error in K_A gives only ± 0.1 kcal/mol error in ΔG, while 20% error in ΔH directly impacts ΔH.

This means absolute error in ΔG is actually still small, while absolute error in ΔH is big.

The reported error bars cannot be trusted.
They're often an order of magnitude (or more!) too small.

Myszka et al. *J. Biomol. Tech.* 14:247, 2003.

ITC measurements on the same system by different laboratories exhibits apparent entropy-enthalpy compensation!

TABLE 3

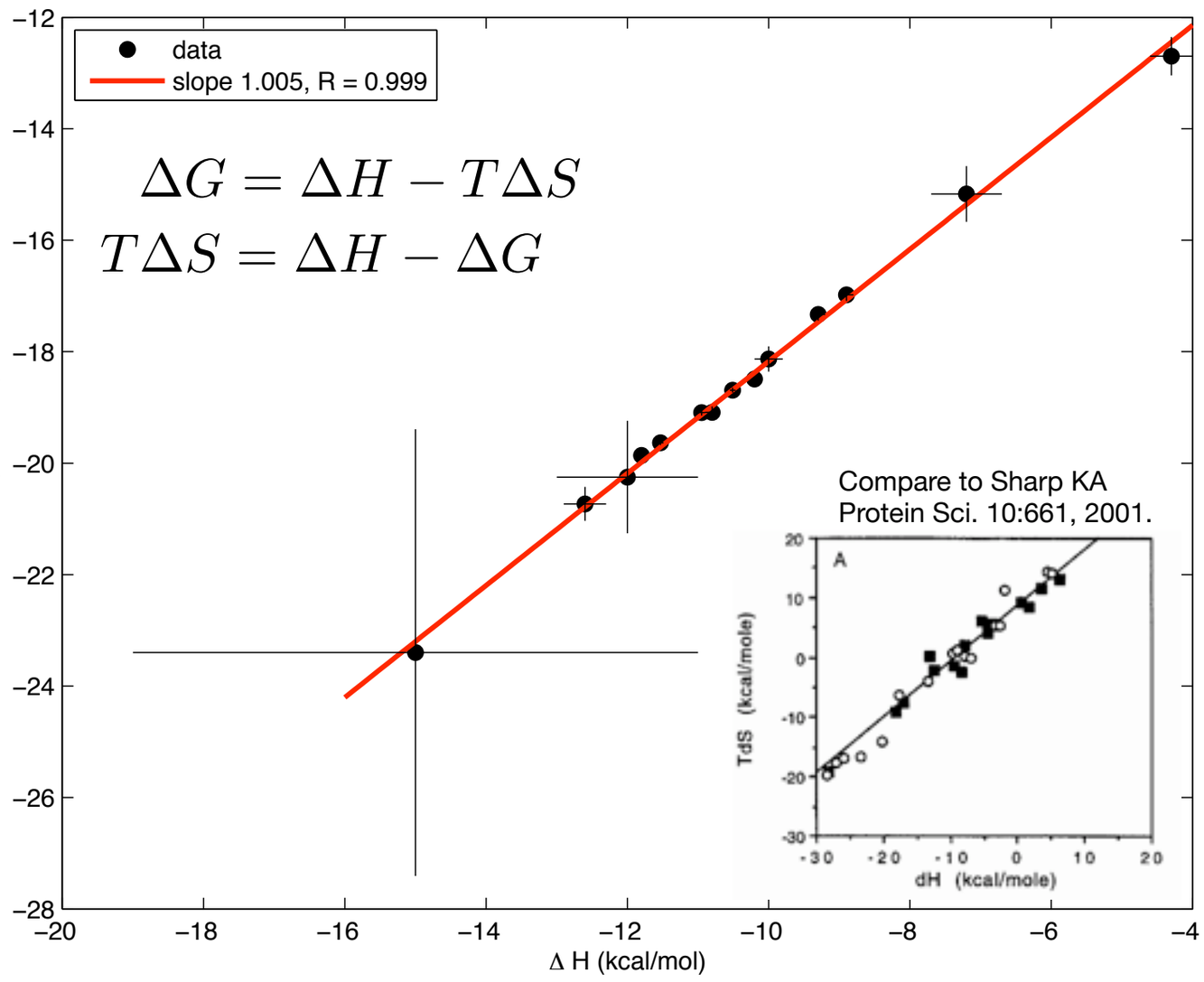
Summary of Reported Isothermal Titration Calorimetry Results[a]

Participant	Molar Binding Ratio (N)	K_A ($10^{-6} \times$ L mol^{-1})	ΔH^{ITC} (kcal mol^{-1})	C-value[b]	Control Titrations
1	1.01 ± 0.01	1.2 ± 0.1	−8.9 ± 0.1	17	CBS into buffer and buffer into CA II. Former subtracted as part of data analysi
2	0.95 ± 0.01	1.44 ± 0.03	−7.2 ± 0.5	75	None
3	0.81 ± 0.01	1.24 ± 0.03	−11.8 ± 0.02	44	CBS into buffer and buffer into CA II. Former subtracted as part of data analysi
4	0.929 ± 0.007	0.84 ± 0.05	−10.8 ± 0.1	24	Pilot run.
5	0.987 ± 0.003	0.84 ± 0.03	−10.20 ± 0.04	41	CBS into buffer, subtracted analysi
6	0.921 ± 0.003	1.08 ± 0.04	−10.95 ± 0.05	39	CBS int lysi
7	1.03 ± 0.01	1.1 ± 0.2	−10.0 ± 0.2	55	CBS int
8	0.97 ± 0.05	0.7 ± 0.2	−4.3 ± 0.3	7.0	CBS int Both subtracted as part of lysis.
9	0.891 ± 0.002	1.15 ± 0.03	−11.53 ± 0.04	59	CBS into buffer, average val btracted in analysis.
10[c]	0.96 ± 0.02	0.9 ± 0.2	−12 ± 1	34	CBS into buffer and buffer into CA II. Average of former used in analysis.
11	1.25 ± 0.01	1.3 ± 0.1	−9.3 ± 0.1	9.0	None
12	1.000 ± 0.003	1.01 ± 0.03	−10.51 ± 0.04	29	Buffer into buffer.
13	0.90 ± 0.02	1.1 ± 0.1	−12.6 ± 0.3	12	CBS into buffer, linear regression subtracted in analysis.
14[c]	0.55 ± 0.08	0.7 ± 0.3	−15 ± 4	22	CBS into buffer, subtracted in analysis.

plot

Myszka et al. J. Biomol. Tech. 14:247, 2003.

CBS/CA II measurements from different laboratories



How can these plots be real evidence of compensation if we can generate the same plot from different measurements on same system?

Analysis of ITC experiments: The Bayesian way

$$p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta)$$

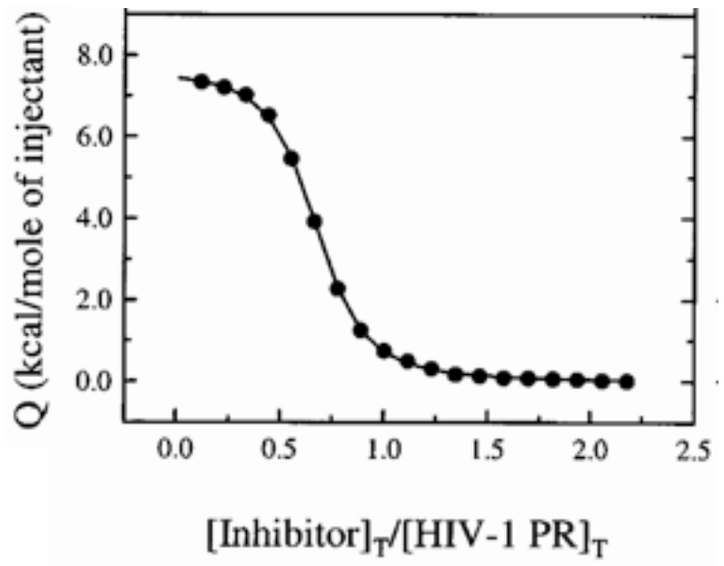
$\mathcal{D}$	data
θ	model parameters
$p(\theta\|\mathcal{D})$	posterior
$p(\mathcal{D}\|\theta)$	sampling distribution (model)
$p(\theta)$	prior

Analysis of ITC experiments: The Bayesian way

$$p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta)$$

$\mathcal{D}$	data
θ	model parameters
$p(\theta\|\mathcal{D})$	posterior
$p(\mathcal{D}\|\theta)$	sampling distribution (model)
$p(\theta)$	prior

$\mathcal{D} = \{q_1, q_2, \ldots, q_N\}$ measurements of evolved heat



Analysis of ITC experiments: The Bayesian way

$$p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta)$$

$\mathcal{D}$	data
θ	model parameters
$p(\theta\|\mathcal{D})$	posterior
$p(\mathcal{D}\|\theta)$	sampling distribution (model)
$p(\theta)$	prior

$\theta = \{\Delta G, \Delta H, T\Delta S, \Delta H_0\}$ thermodynamic parameters

$\Delta G = -kT \ln K_a$	free energy of binding
ΔH	enthalpy of binding
$T\Delta S$	entropic contribution to binding
ΔH_0	heat of dilution

... any additional parameters ...

Analysis of ITC experiments: The Bayesian way

$$p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta)$$

$$p(\mathcal{D}|\theta) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(q_n - q_n^*)^2}{2\sigma^2}\right]$$ Gaussian error model

$\mathcal{D}$	data
θ	model parameters
$p(\theta\|\mathcal{D})$	posterior
$p(\mathcal{D}\|\theta)$	sampling distribution (model)
$p(\theta)$	prior

q_n	measured heat of injection *n*
q_n^*	true heat of injection *n*
σ	std dev of error in measured heat (nuisance parameter)

$$P + L \overset{\Delta H}{\rightarrow} PL$$

$$q_n^* = Q_n - Q_{n-1}$$

$$Q_n = \Delta H \cdot V_n[PL]_n + n\Delta H_0$$ heat potential

Analysis of ITC experiments: The Bayesian way

$$p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta)$$

$\mathcal{D}$	data
θ	model parameters
$p(\theta\|\mathcal{D})$	posterior
$p(\mathcal{D}\|\theta)$	sampling distribution (model)
$p(\theta)$	prior

$$p(\Delta G, \Delta H, \Delta H_0, \sigma) \propto \sigma^{-1}$$ prior on measurement noise

$$[M]_0^* \sim \mathcal{N}([M]_0, \sigma_M^2)$$
$$[L]_s^* \sim \mathcal{N}([L]_s, \sigma_L^2)$$
prior on cell and syringe concentrations

$\Delta G, \Delta H, \Delta H_0$ can be of any sign and value

$\sigma > 0$ scale parameter; can be of any magnitude
(Later, could build in some *a priori* knowledge of instrument error or calibration runs.)

Important points:

*** We don't know the size of the measurement error**

*** We don't know the actual ligand and protein concentrations**

No problem: Just make them nuisance parameters!

Analysis of ITC experiments: The Bayesian way

$$p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta)$$

$$p(\theta|\mathcal{D}) = (2\pi)^{-N/2}\sigma^{-(N+1)} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(q_n - q_n^*)^2\right]$$ posterior

$\mathcal{D}$ data

θ model parameters

$p(\theta|\mathcal{D})$ posterior

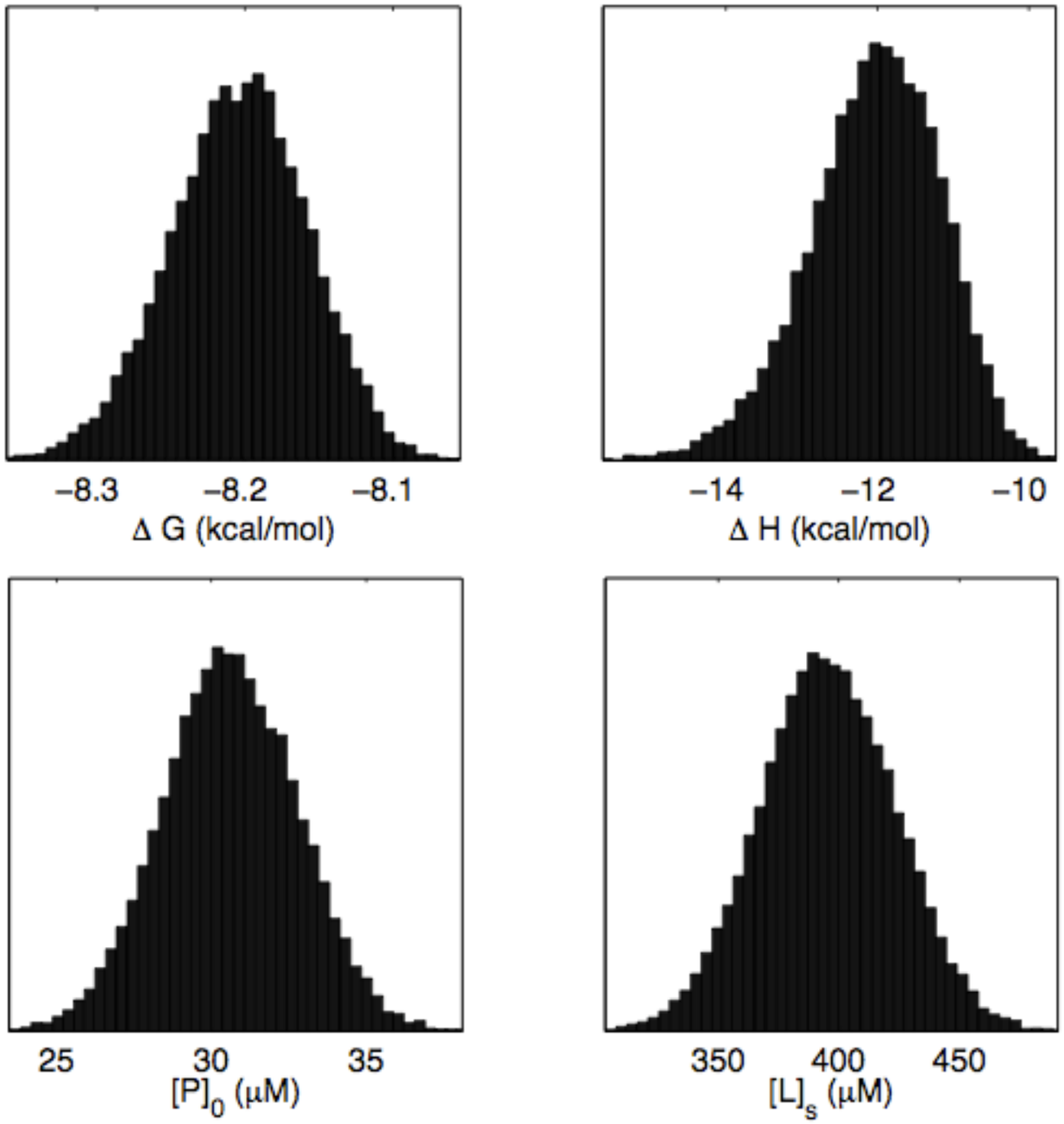
$p(\mathcal{D}|\theta)$ sampling distribution (model)

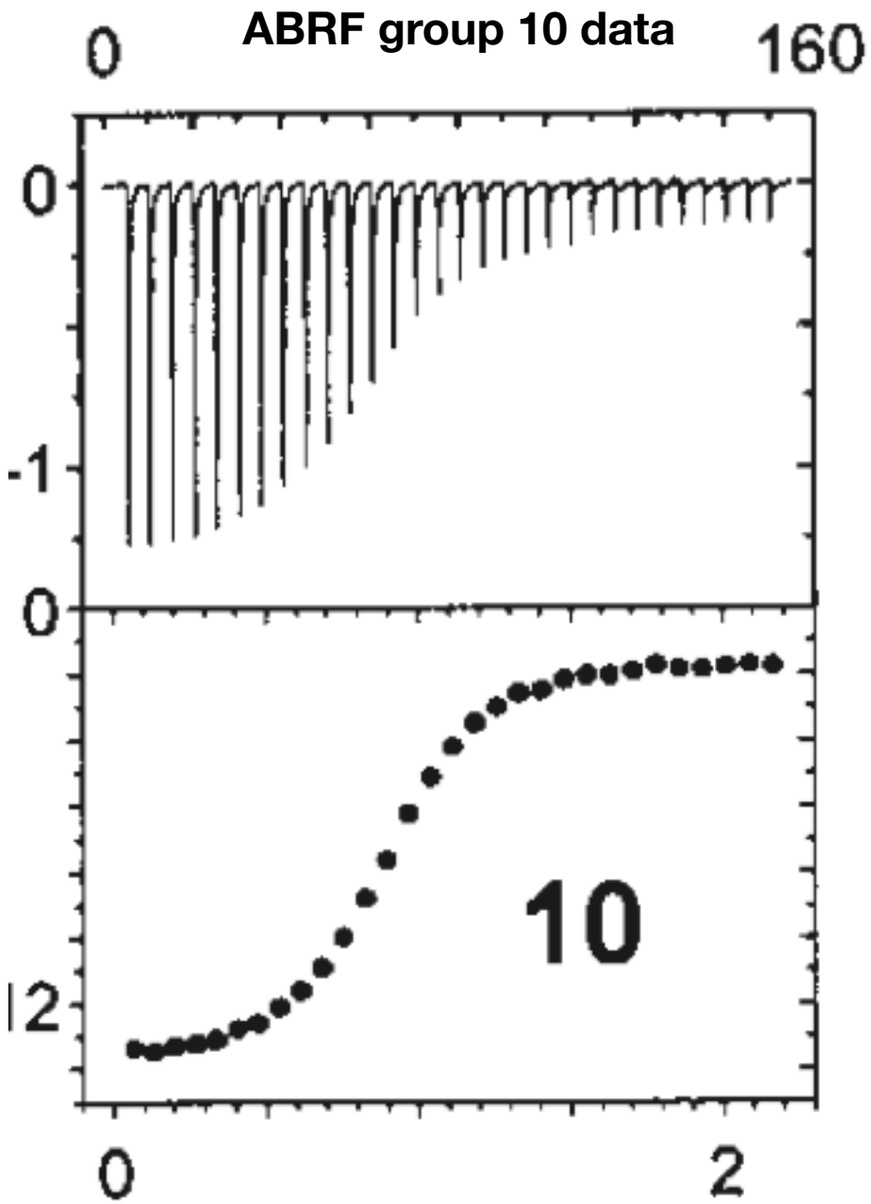
$p(\theta)$ prior

q_n^* nonlinear function of thermodynamic parameters

Analysis of ITC experiments: The Bayesian way

Marginal posterior probability distributions describe the uncertainty in each thermodynamic parameter inferred from the data.

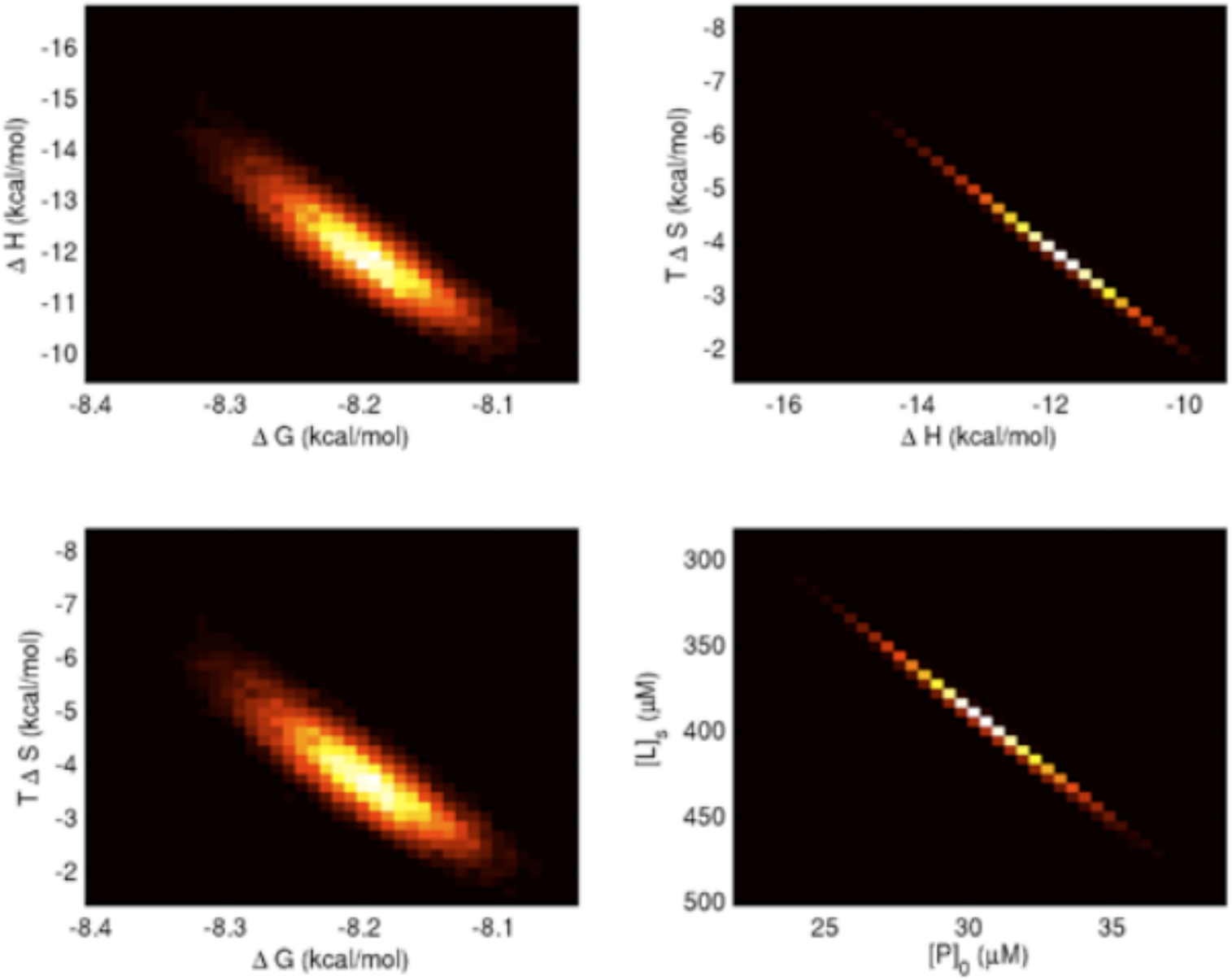




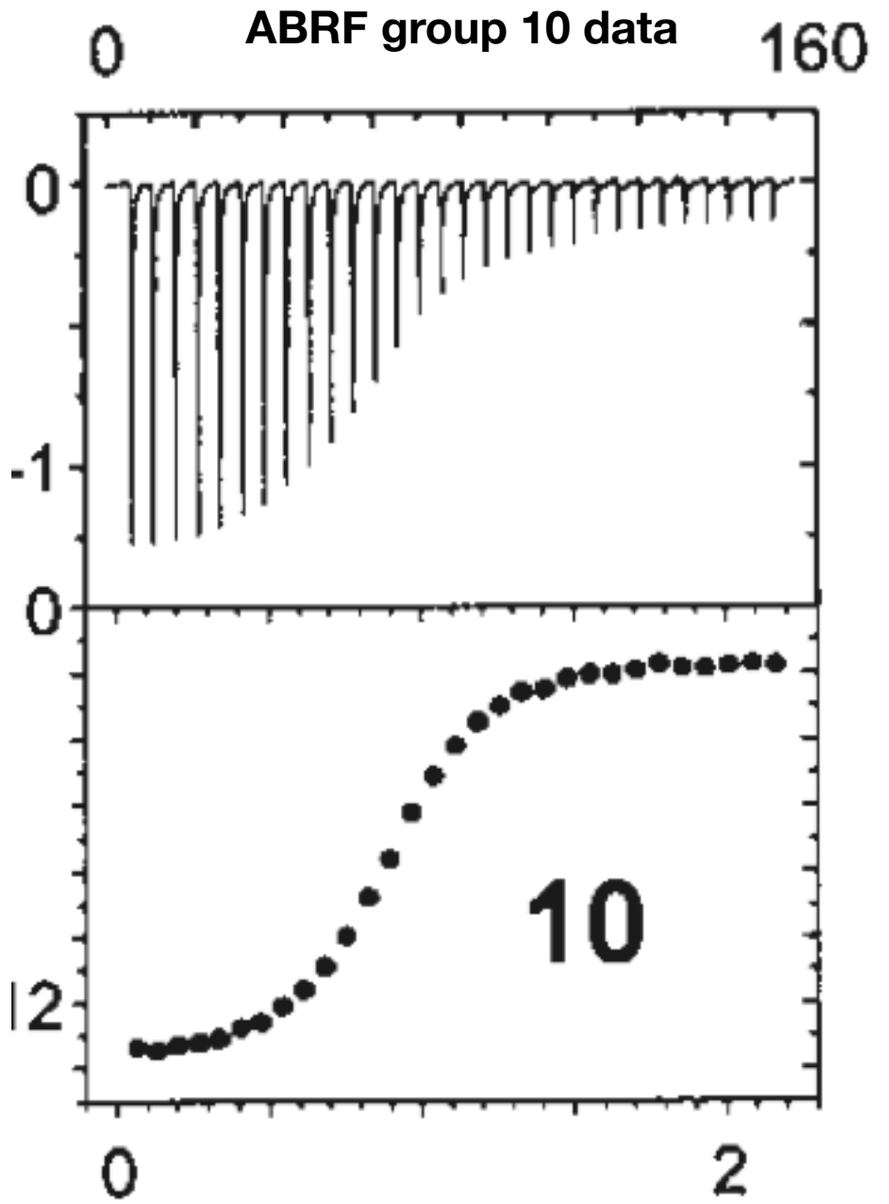
Myszka et al. J. Biomol. Tech. 14:247, 2003.

Analysis of ITC experiments: The Bayesian way

Joint distribution functions describe the correlated uncertainty between any set of parameters.



Correlation in error in ΔH and TΔS explains much of apparent entropy-enthalpy compensation behavior.



Myszka et al. J. Biomol. Tech. 14:247, 2003.

Bayesian approach has numerous advantages

Provides true posterior joint distribution of all thermodynamic parameters

Asymmetric confidence intervals and non-normal marginal distributions

Easy to “plug in” new binding models.

Can eliminate baseline “blank” experiment through titration in excess

Make joint inferences from data from multiple experiments

Instrument parameters can be conditioned on calibration data (e.g. NaCl titrations)

Expected information content of new experiments can be estimated for protocol design

Experimental design:

> “Will experiment X give me enough information to make it worthwhile?”
> “What is the best experimental design to reduce the uncertainty in Z?”
> “Do I have to run a baseline for sample X?”

Code will be made available at http://simtk.org/home/bayesian-itc

Other explanations for apparent compensation?
The “free energy window” effect: Instrumental limits?

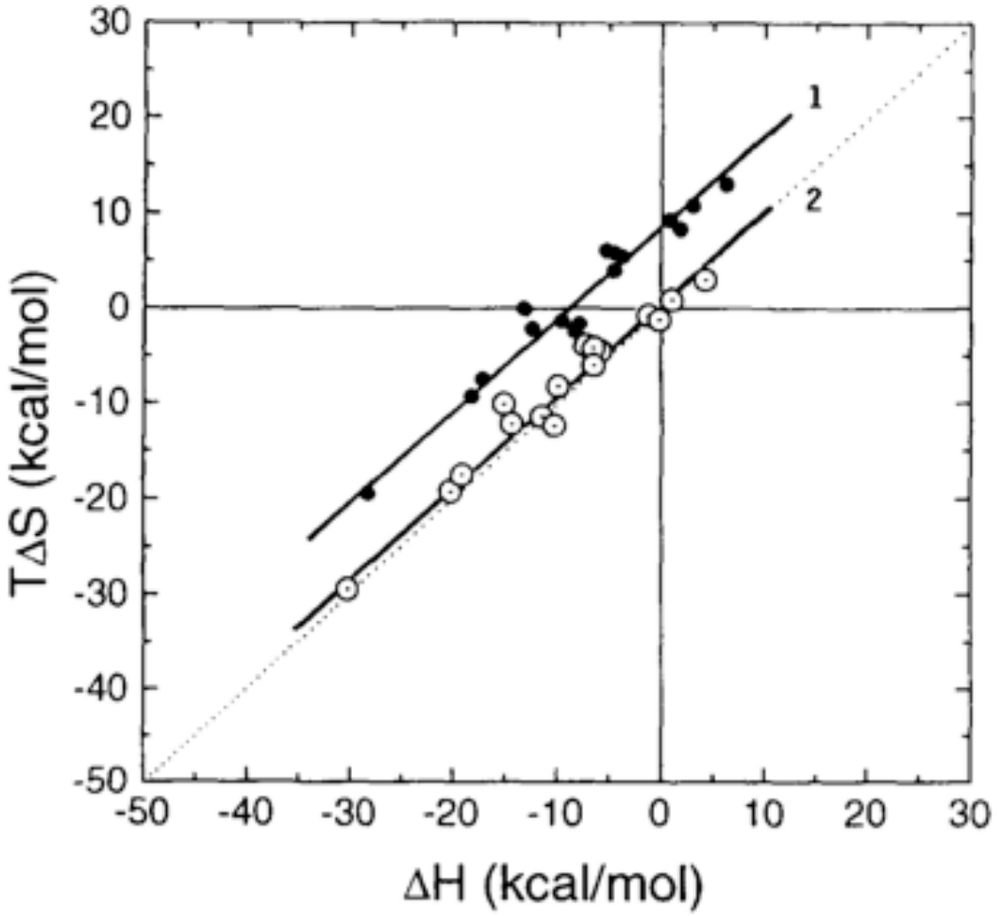
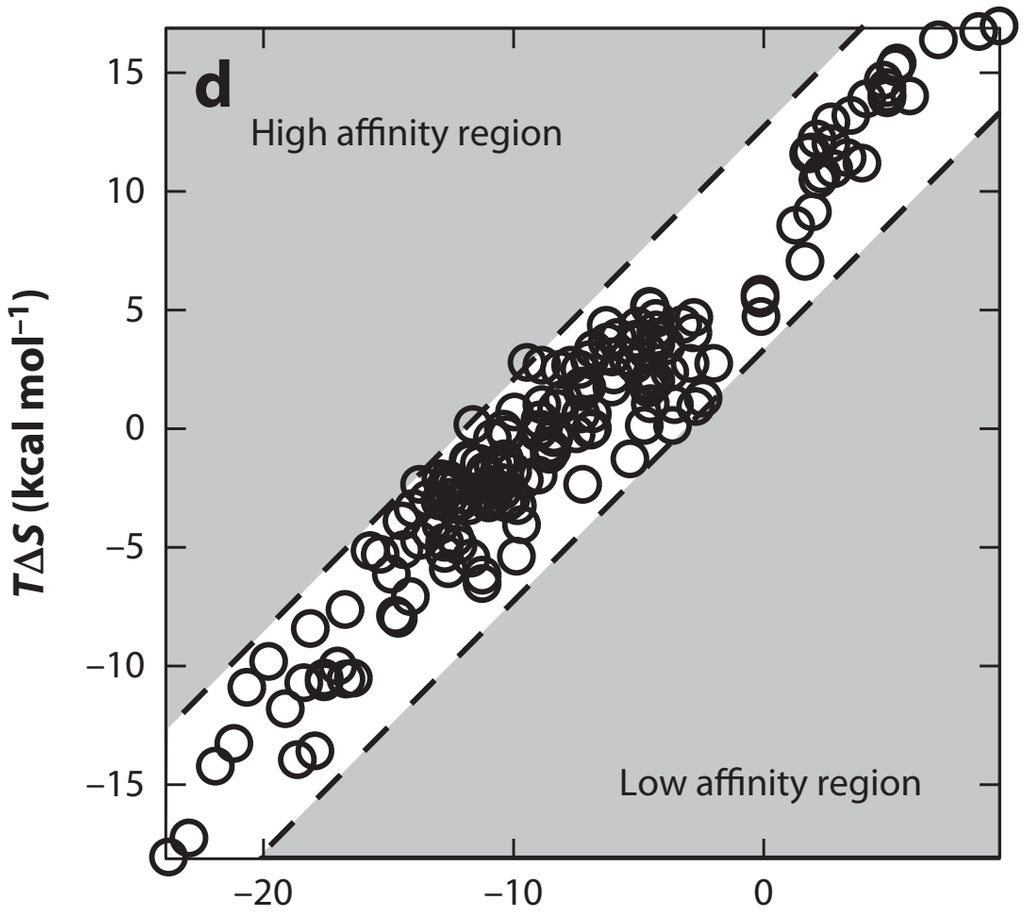


FIG. 3. **The relationship between ΔH and $T\Delta S$ for the binding of Ca^{2+} to the proteins (●).** The relationship between $\Delta H'$ and $T\Delta S'$ after subtracting the contribution of Ca^{2+} binding by itself (⊙). The data were taken from Table II. *Lines 1* and *2* were obtained by a least squares fit of each data set. The *dotted line* represents $\Delta H - T\Delta S = 0$.



$$\Delta G = \Delta H - T\Delta S$$

All ITC measurements fall in a narrow range due to instrumental limitations

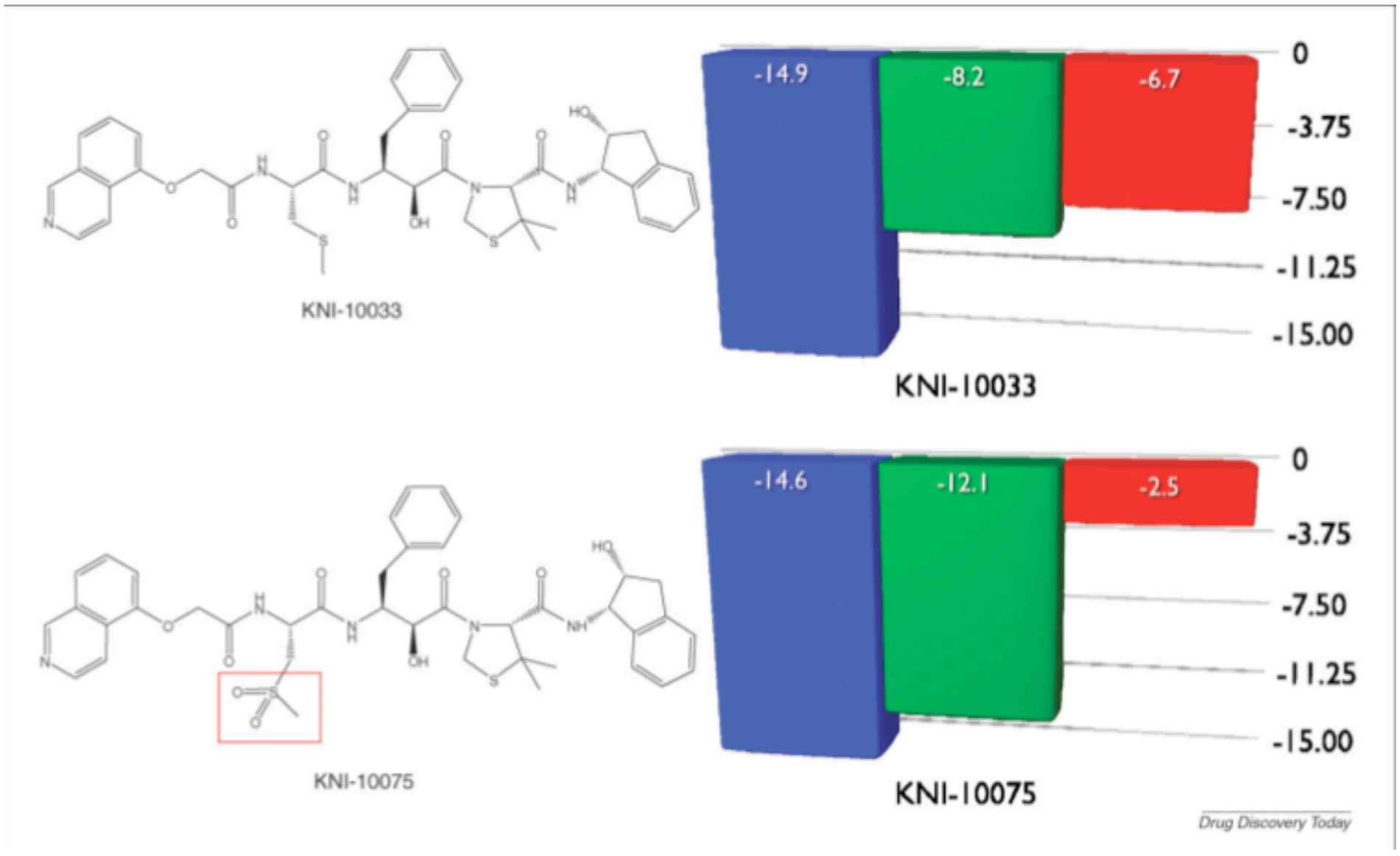
Exner O. 1997. How to get wrong results from good experimental data: a survey of incorrect applications of regression. J. Phys. Org. Chem. 10:797–813

Chodera and Mobley. Annu Rev Biophys 42:121, 2013.

Should we worry about compensation in drug design?

"The thermodynamic signature...provides a unique experimental way of characterizing the binding mode of a drug molecule."

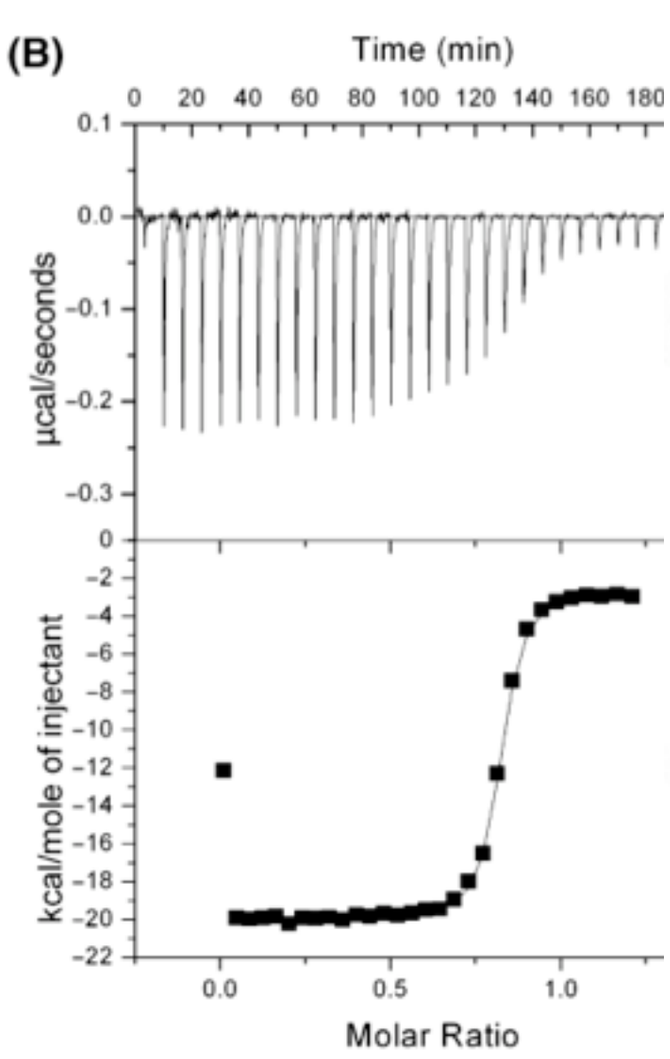
Freire E. *Drug Discovery Today* 13:869, 2009.



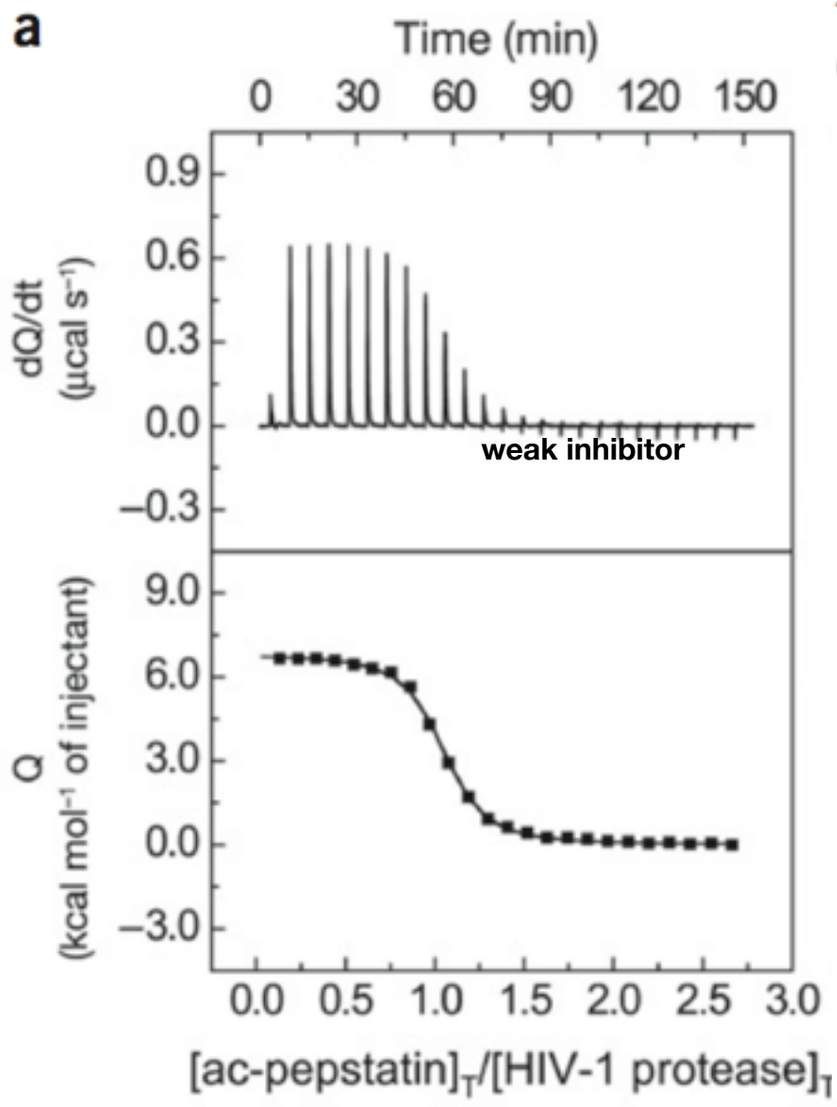
Does this claim hold up to what we've learned about correlated errors in ITC?

A more honest assessment of errors

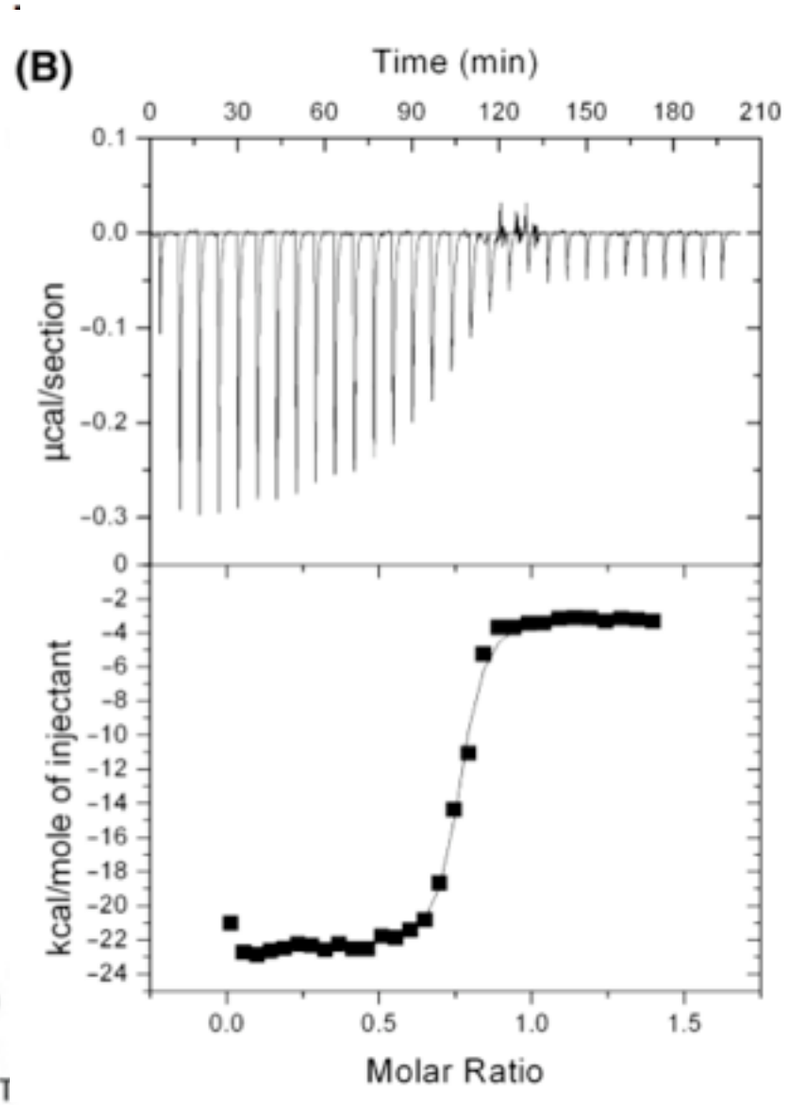
KNI-10033 displacement of acetyl pepstatin



acetyl-pepstatin

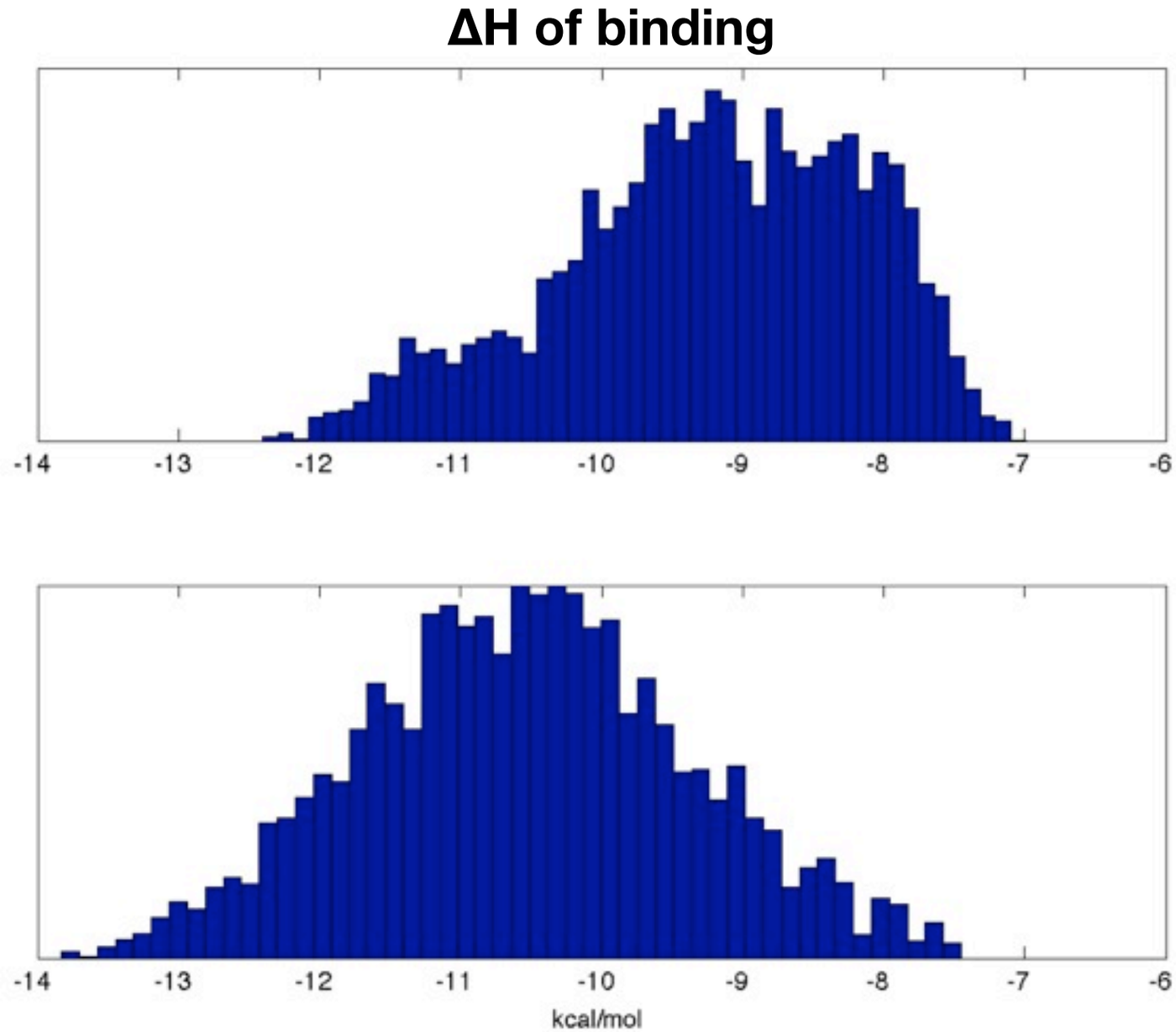


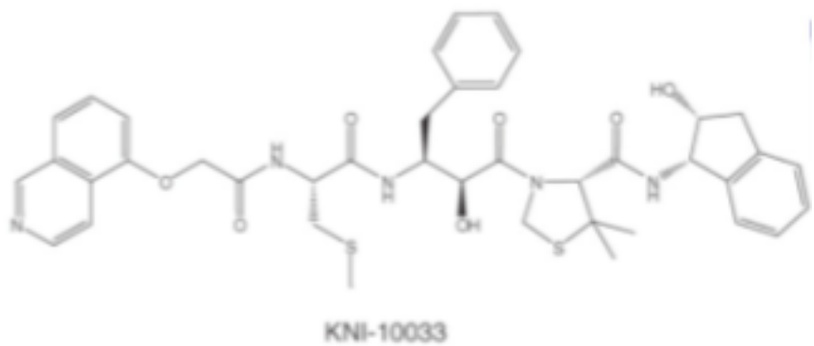
KNI-10075 displacement of acetyl pepstatin

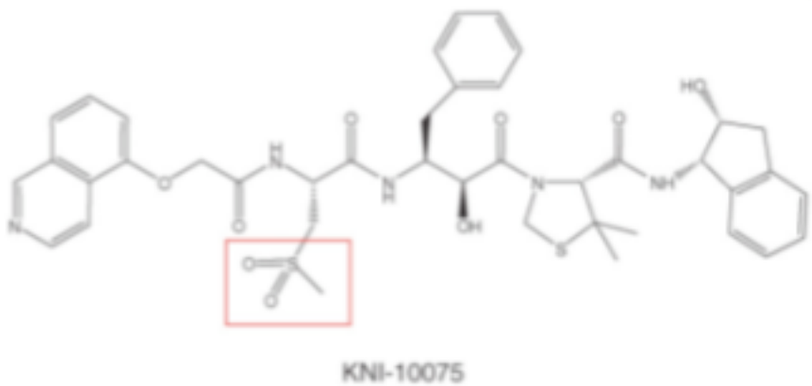


Velazquez-Campoy A, Kiso Y, and Friere E. Arch. *Biochem. Biophys.* 390:169, 2001.
Velazquez-Campoy A and Friere E. *Nature Protocols* 1:186, 2006.

A more honest assessment of errors







Enthalpy change is clearly not the 4 kcal/mol difference the fingerprints would imply

Rules of thumb: Error magnitudes of last resort

Affinity/activity data:

Public Ki data: 0.54 pKi units

Kramer et al. J Med Chem 2012.

Public IC50 data: 0.68 pIC50 units

Kramer et al. PLoS One 8:e61007, 2013.

ITC data: 20% error in Kd and ΔH (unless evidence of propagating ligand concentration error)

Chodera and Mobley. Annu Rev Biophys 42:121, 2013; Myszka et al. J. Biomol. Tech. 14:247, 2003.

SPR data: 25% error in Kd, 20% in k_{on} and k_{off}

Myszka et al. J. Biomol. Tech. 14:247, 2003.

Concentrations:

prepared by a lab tech using mass/volume: ~10%

Myszka et al. J. Biomol. Tech. 14:247, 2003.

measured by NMR: 3%

Agilent technical note: Easy, Precise and Accurate Quantitative NMR [combining accuracy and precision]

An Editor's plea

Editorial policy, Author and Reviewer guidelines

Could we as a community gather what we've discussed this week and formulate it into clear, manageable, realistic guidelines that could be universally distributed, endorsed, and enforced?

Would this require some aspect of education or links to the same?

Corresponding Author Name: _____

Manuscript Number: _____

Reporting Checklist For Life Sciences Articles

This checklist is used to ensure good reporting standards and to improve the reproducibility of published results. For more information, please read Reporting Life Sciences Research.

Figure legends

Each figure legend should contain, for each panel where they are relevant:

- the **exact sample size (*n*)** for each experimental group/condition, given as a number, not a range;
- a **description of the sample collection** allowing the reader to understand whether the samples represent **technical or biological replicates** (including how many animals, litters, cultures, etc.);
- a **statement of how many times the experiment shown was replicated in the laboratory**;
- **definitions of statistical methods and measures**:
  - very common tests, such as *t*-test, simple χ^2 tests, Wilcoxon and Mann-Whitney tests, can be unambiguously identified by name only, but more complex techniques should be described in the methods section;
  - are tests one-sided or two-sided?
  - are there adjustments for multiple comparisons?
  - **statistical test results**, e.g., ***P* values**;
  - definition of '**center values**' as **median or average**;
  - definition of **error bars as s.d. or s.e.m.**

Any descriptions too long for the figure legend should be included in the methods section.

Please ensure that the answers to the following questions are reported in the manuscript itself. We encourage you to include a specific subsection in the methods section for statistics, reagents and animal models. Below, provide the page number(s) or figure legend(s) where the information can be located.

Statistics and general methods

Reported on page(s) or figure legend(s):

1. How was the sample size chosen to ensure adequate power to detect a pre-specified effect size?

   For animal studies, include a statement about sample size estimate even if no statistical methods were used.

2. Describe inclusion/exclusion criteria if samples or animals were excluded from the analysis. Were the criteria pre-established?

3. If a method of randomization was used to determine how samples/animals were allocated to experimental groups and processed, describe it.

   For animal studies, include a statement about randomization even if no randomization was used.

4. If the investigator was blinded to the group allocation during the experiment and/or when assessing the outcome, state the extent of blinding.

   For animal studies, include a statement about blinding even if no blinding was done.

5. For every figure, are statistical tests justified as appropriate?

   Do the data meet the assumptions of the tests (e.g., normal distribution)?

   Is there an estimate of variation within each group of data? Is the variance similar between the groups that are being statistically compared?

(Continues on following page)

Reagents

Reported on page(s) or figure legend(s):

6. To show that antibodies were profiled for use in the system under study (assay and species), provide a citation, catalog number and/or clone number, supplementary information or reference to an antibody validation profile (e.g., Antibodypedia, 1DegreeBio).

7. Identify the source of cell lines and report if they were recently authenticated (e.g., by STR profiling) and tested for mycoplasma contamination.

Animal models

Reported on page(s) or figure legend(s):

8. Report species, strain, sex and age of animals.

9. For experiments involving live vertebrates, include a statement of compliance with ethical regulations and identify the committee(s) approving the experiments.

10. We recommend consulting the ARRIVE guidelines (*PLoS Biol.* **8**(6), e1000412, 2010) to ensure that other relevant aspects of animal studies a adequately reported.

Human subjects

Reported on page(s) or figure legend(s):

11. Identify the committee(s) approving the study protocol.

12. Include a statement confirming that informed consent was obtained from all subjects.

13. For publication of patient photos, include a statement confirming that consent to publish was obtained.

14. Report the clinical trial registration number (at ClinicalTrials.gov or equivalent).

15. For phase II and III randomized controlled trials, please refer to the CONSORT statement and submit the CONSORT checklist with your submission.

16. For tumor marker prognostic studies, we recommend that you follow the REMARK reporting guidelines.

Data deposition

17. Provide accession codes for deposited data.

    Reported on page(s) or figure legend(s):

    Data deposition in a public repository is mandatory for:
    a. Protein, DNA and RNA sequences
    b. Macromolecular structures
    c. Crystallographic data for small molecules
    d. Microarray data

    Deposition is strongly recommended for many other datasets for which structured public repositories exist; more details on our data policy are available here. We encourage the provision of other source data in supplementary information or in unstructured repositories such as Figshare and Dryad.

18. Is computer source code provided with the paper or deposited in a public repository? If so, indicate how it can be obtained.

References

- Christian Kramer and Richard Lewis, “QSARs, Data and Error in the Modern Age of Drug Discovery,” Current Topics in Medicinal Chemistry, 2012, 12, 1896-1902
- M. Hewitt, et al 2009, J. Chem. Inf. Model, 49,2572–2587

Acknowledgments

- All those experimentalists who are slaving in the trenches against the forces of nature and time to provide knowledge and insight to the ungrateful hoards.
- Especially those who have spent hours explaining to me what they do.
- Ant

More Acknowledgments

Ant (“I’m going to make you an offer you can’t refuse”) Nicholls

ITC and entropy-enthalpy compensation
Kim Branson, Sarah Boyce, Paul Novick, Vijay Pande, David Minh, David Mobley

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

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

AAAS Fellow

IUPAC Fellow

tstouch@gmail.com

1-609-275-7234