

Is the force with us?

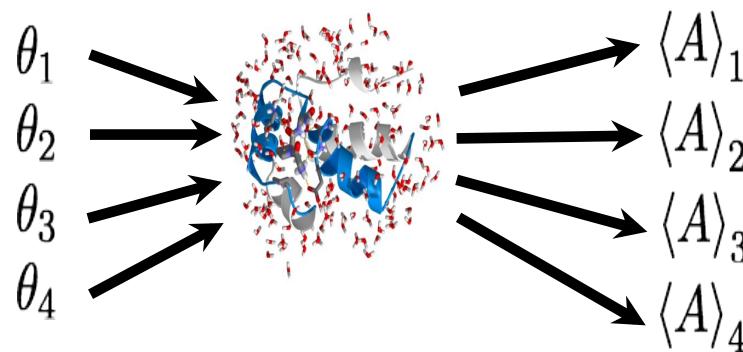
Bayesian inference in computational chemistry.

Chaya D. Stern
PyData NYC
Nov 29, 2017

- How are computational molecular models utilized in drug discovery?



- Bayesian inference of **model parameters** and model selection.



- Reweighting for error propagations.

Biological Macromolecules are the molecular machines of life.

Structural

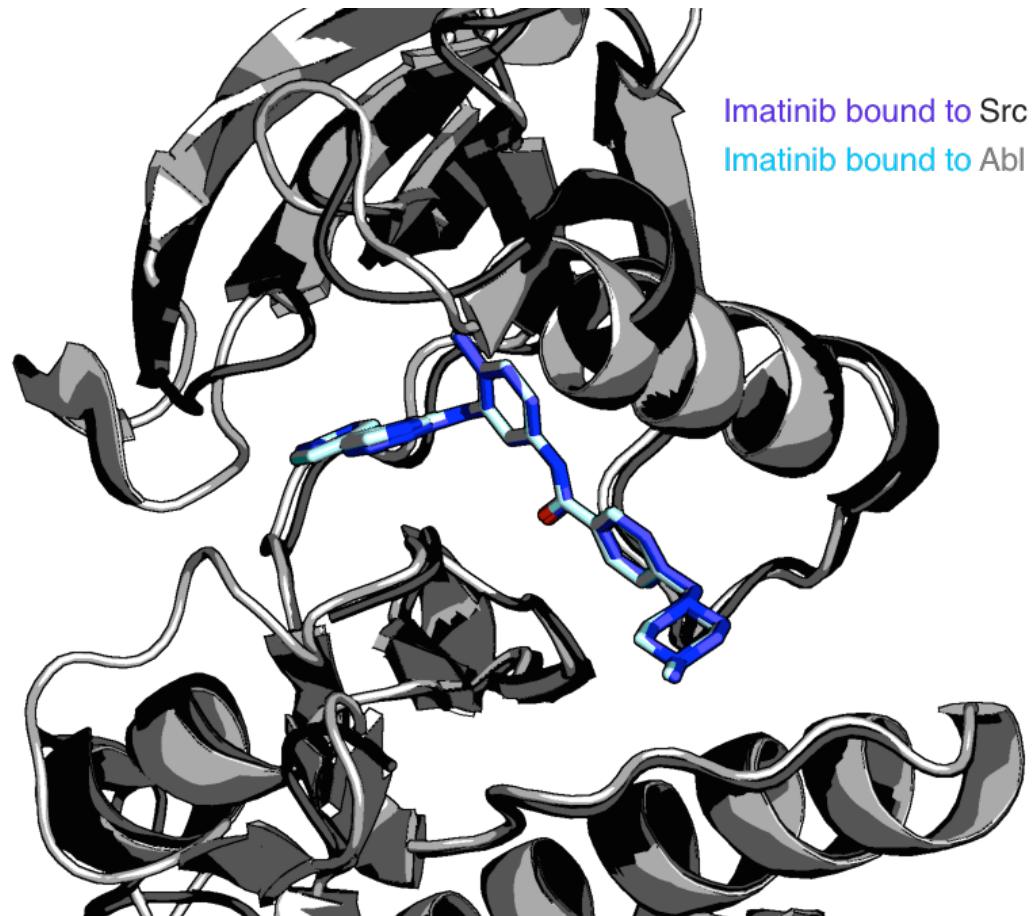
Signaling

Transport

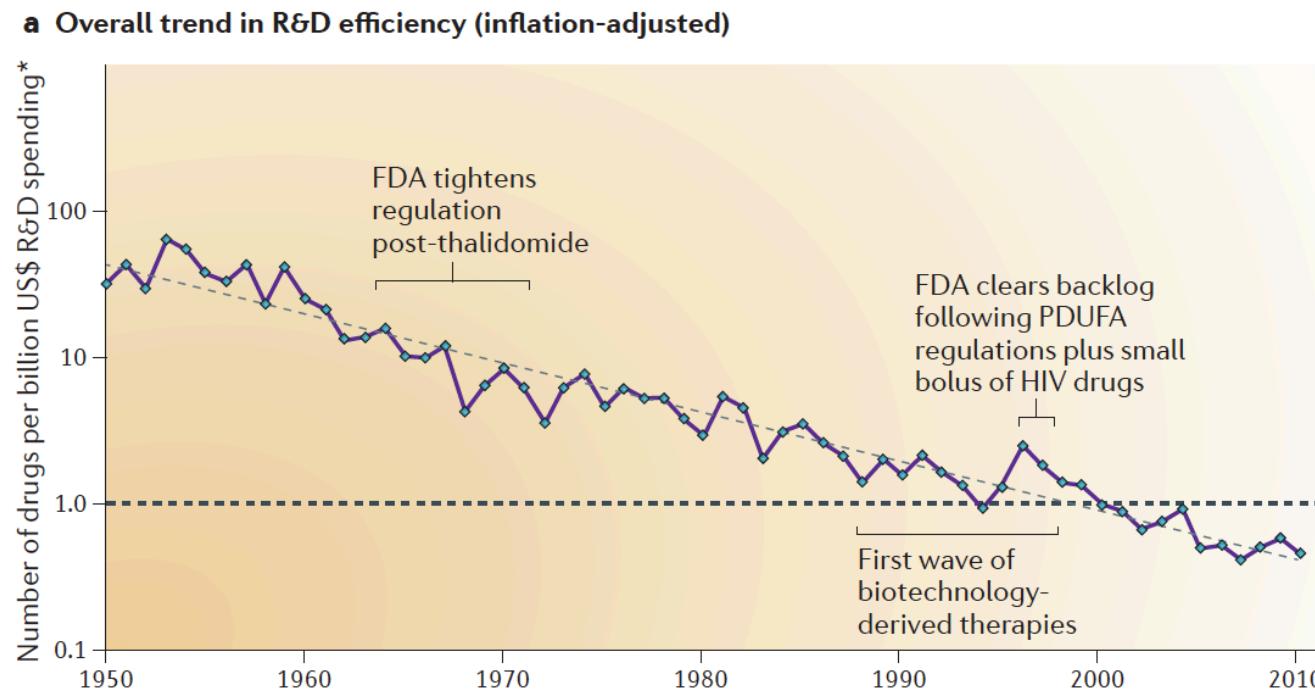
Catalysts

Regulation

Biosynthetic factories



Drug discovery usually ends in failure

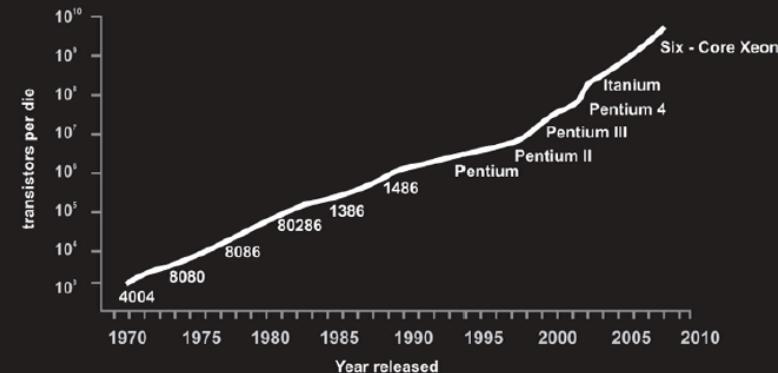
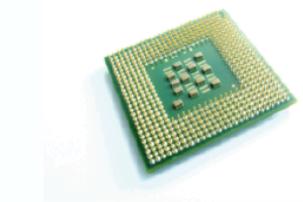
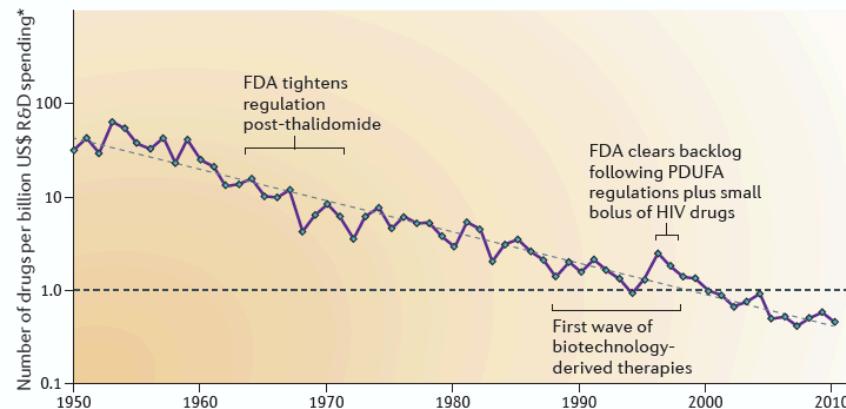


Scannell et al. Nature Rev Drug Disc 11:191, 2012

Drug discovery usually ends in failure



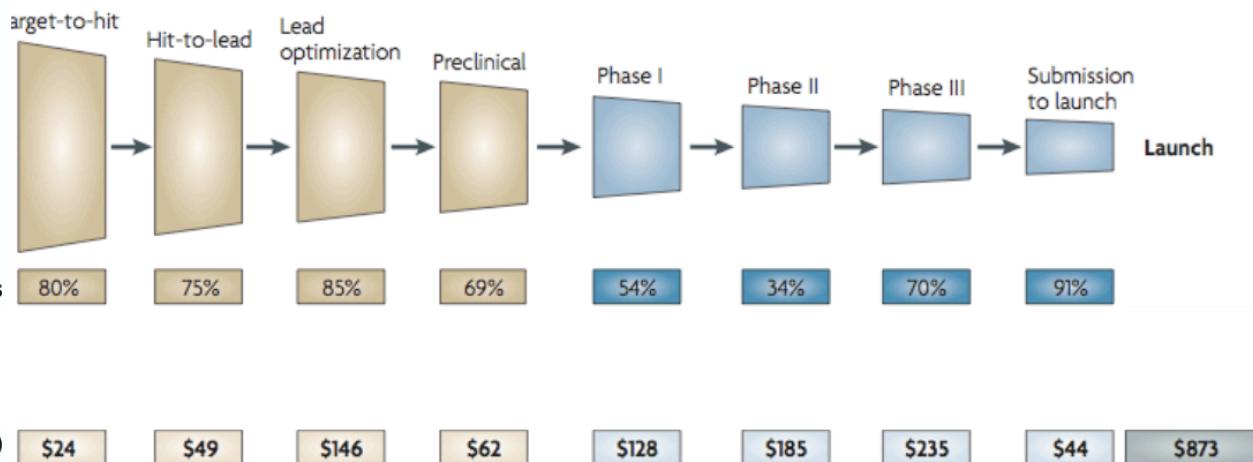
a Overall trend in R&D efficiency (inflation-adjusted)



EROOM'S LAW

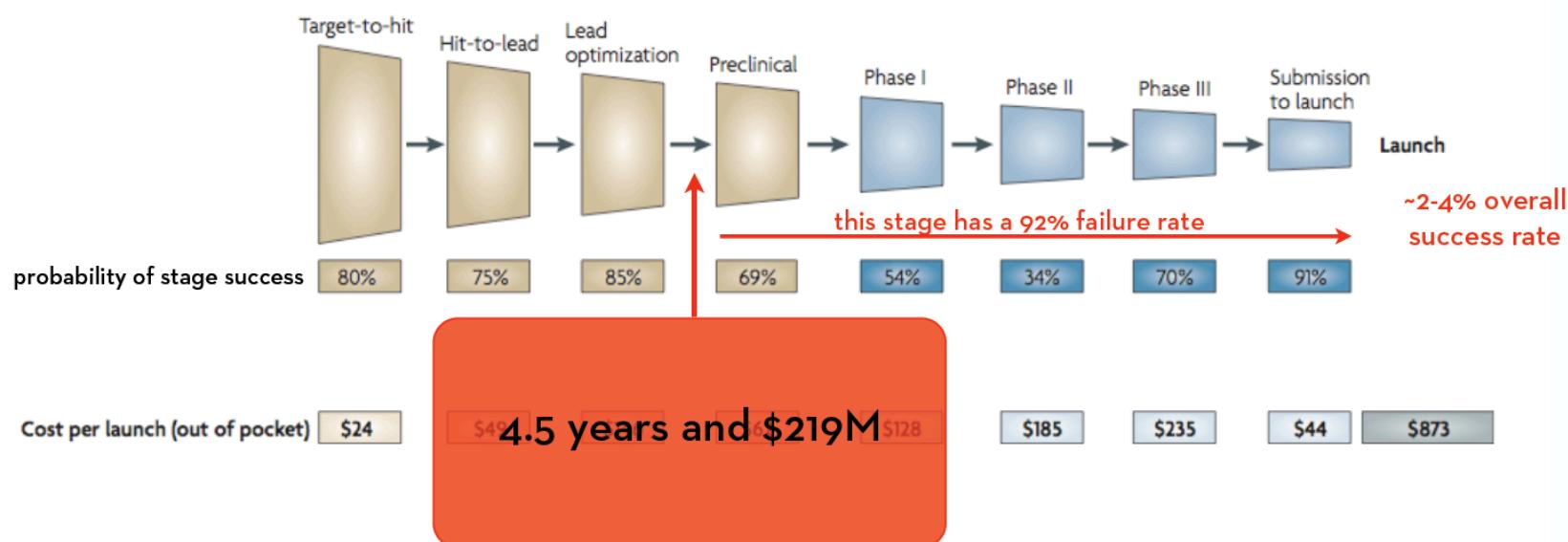
MOORE'S LAW

Each stage of the drug discovery pipeline is highly inefficient.



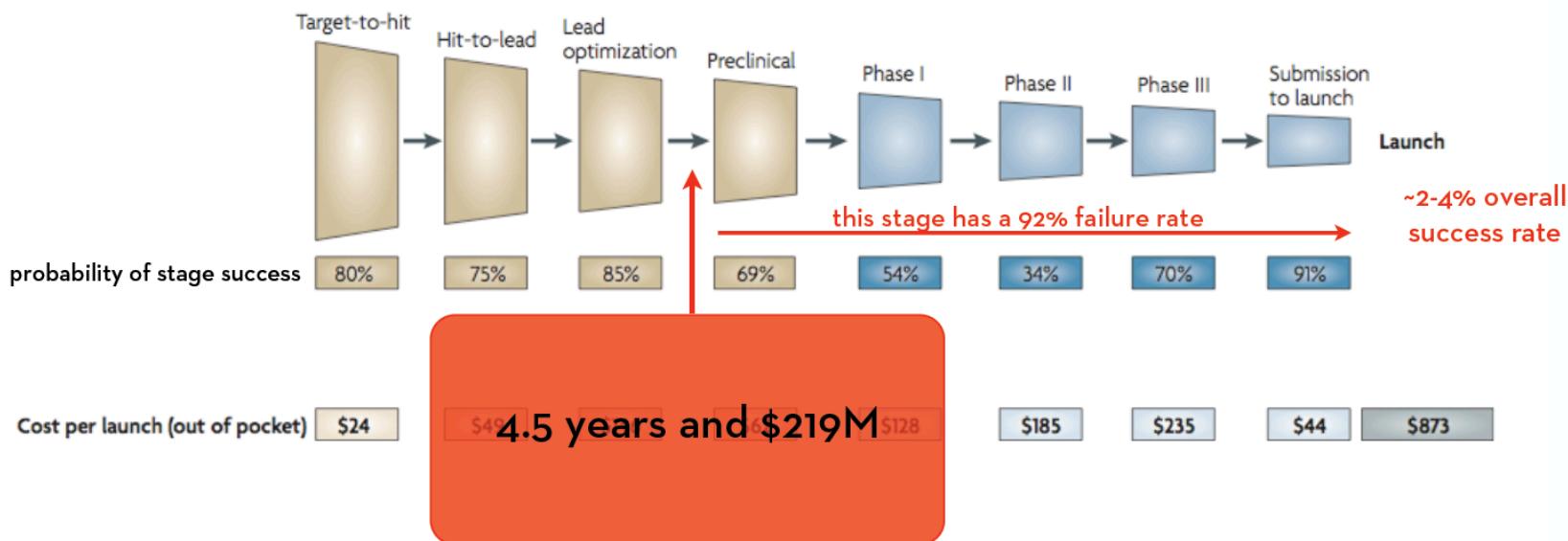
Paul et al. Nat. Rev. Drug Discov. 9:203, 2010.
Chodera et al. Curr. Opin. Struct. Biol., 21:150, 2011.

Large amount of time and money are invested in compounds that will never become drugs.



Paul et al. Nat. Rev. Drug Discov. 9:203, 2010.
Chodera et al. Curr. Opin. Struct. Biol., 21:150, 2011.

Large amount of time and money are invested in compounds that will never become drugs.

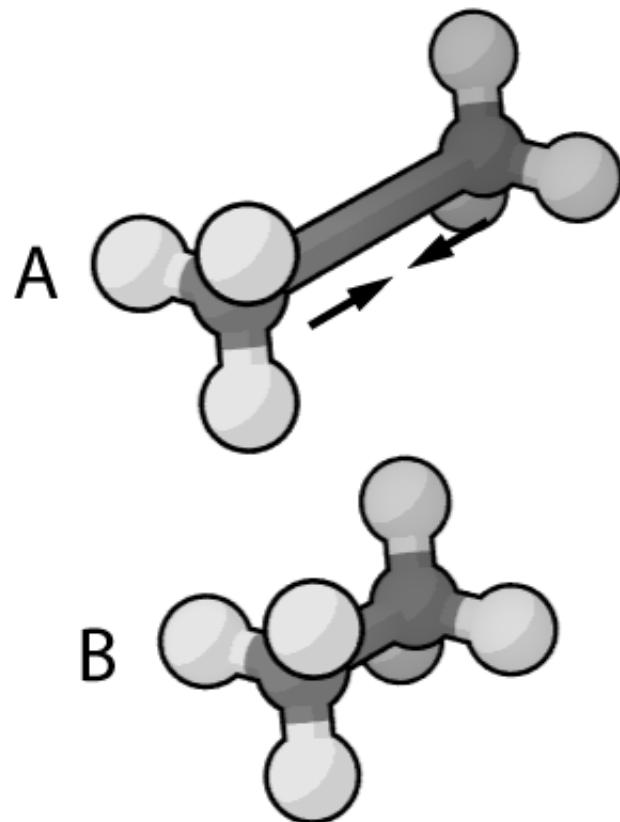


Paul et al. Nat. Rev. Drug Discov. 9:203, 2010.
Chodera et al. Curr. Opin. Struct. Biol., 21:150, 2011.

Computational models can speed up and decrease the cost of this process.

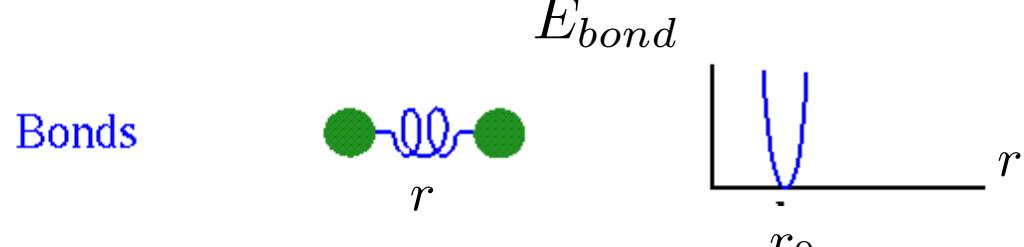
In molecular mechanics simulations, inter and intramolecular forces are modeled with a **forcefield**.

A Newtonian physics **approximation** of the underlying quantum mechanics.



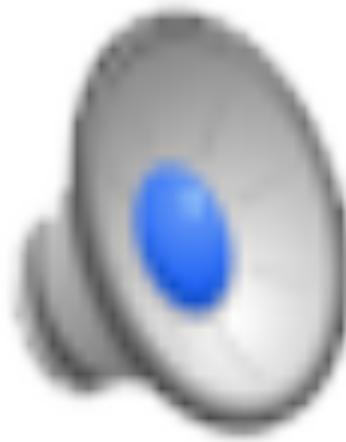
$$E_{bond} = \sum_{bond} k_b(r - r_0)^2$$

k_b : Force constant
 r_0 : equilibrium length



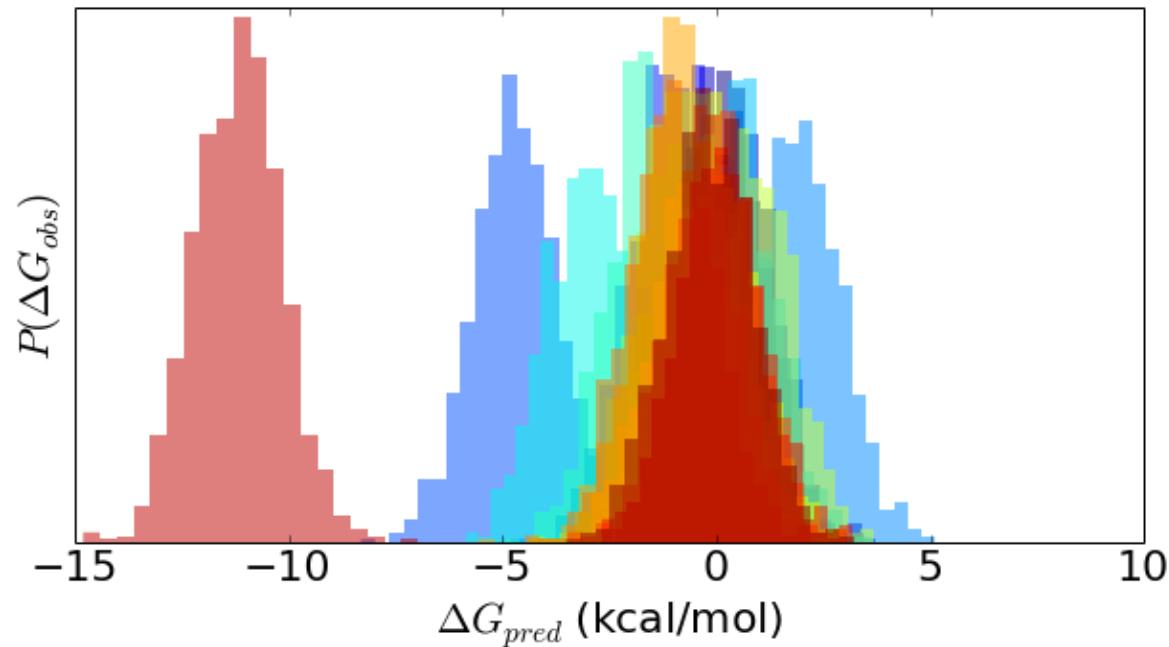
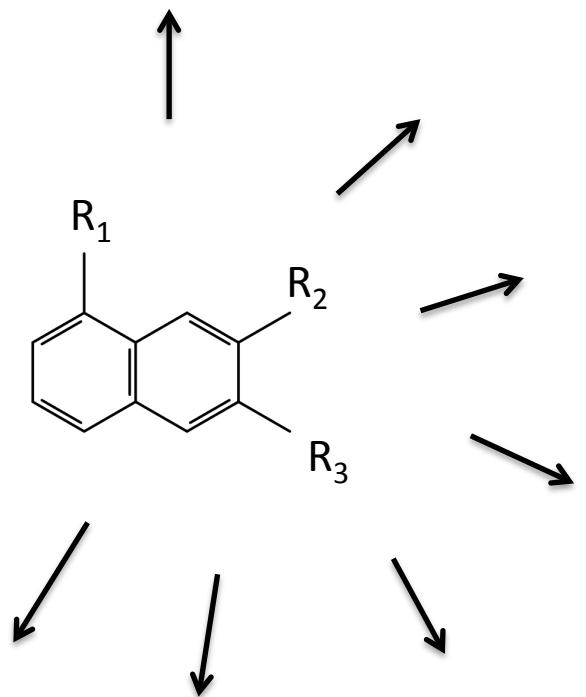
Free parameters need to be fit to reproduce experimental condensed phase data.

Molecular dynamics simulations provide atomic detail to dynamics that is usually difficult to observe.



To design a new drug, we need to predict which small molecule will bind to the target.

Uncertainty in computed properties are needed for designing the next generation of small-molecule drugs

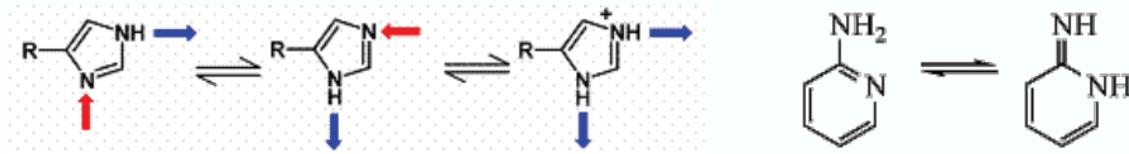


Three main sources of error.

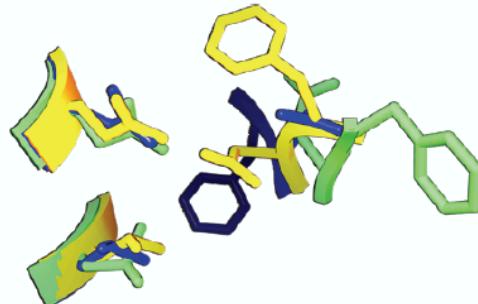
1. The **forcefield** does a poor job of modeling the physics of our system

$$V(\mathbf{q}) = \sum_{\text{bonds}} K_r(r - r_{eq})^2 + \sum_{\text{angles}} K_\theta(\theta - \theta_{eq})^2 + \sum_{\text{dihedrals}} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] + \sum_{i < j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^6} + \frac{q_i q_j}{\epsilon R_{ij}} \right]$$

2. We're missing some **essential chemical** in our simulations
(e.g. protonation states, tautomers, covalent association)



3. We haven't **sampled** all of the relevant conformations



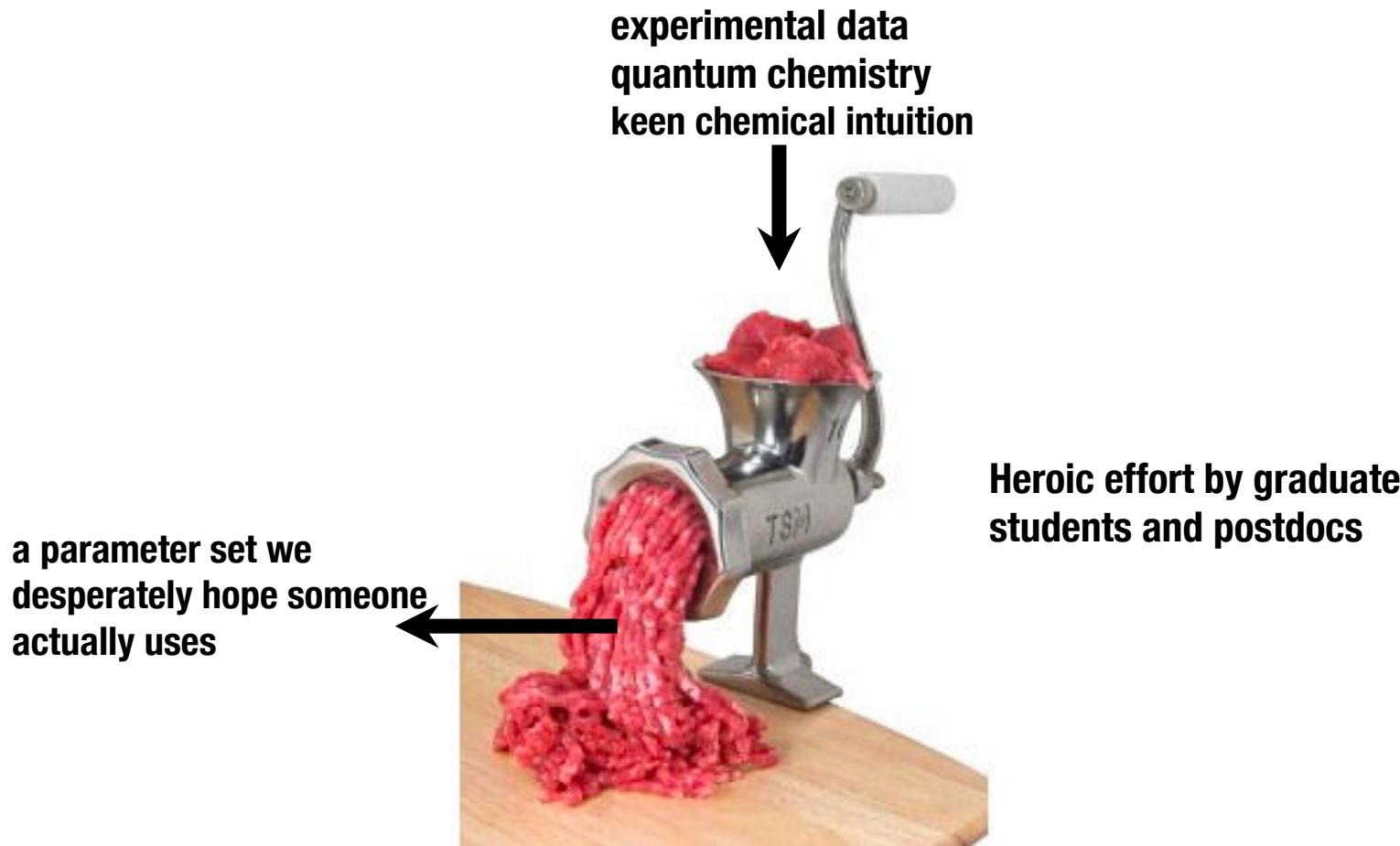
We expect binding free energies to be sensitive to forcefield error.

1. The forcefield does a poor job of modeling the physics of our system

$$V(\mathbf{q}) = \sum_{\text{bonds}} K_r(r - r_{eq})^2 + \sum_{\text{angles}} K_\theta(\theta - \theta_{eq})^2 + \sum_{\text{dihedrals}} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] + \sum_{i < j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^6} + \frac{q_i q_j}{\epsilon R_{ij}} \right]$$

Currently we don't know how much uncertainty the forcefield contributes to simulation results.

How are forcefields made?



What would we want out of a forcefield parameterization scheme (while we're dreaming)?

Everything is automatic; don't need to tweak things by hand.

Never need to choose atom types or perform immense feats of chemical intuition.

If we're uncertain about which functional forms are best, automatically chooses.

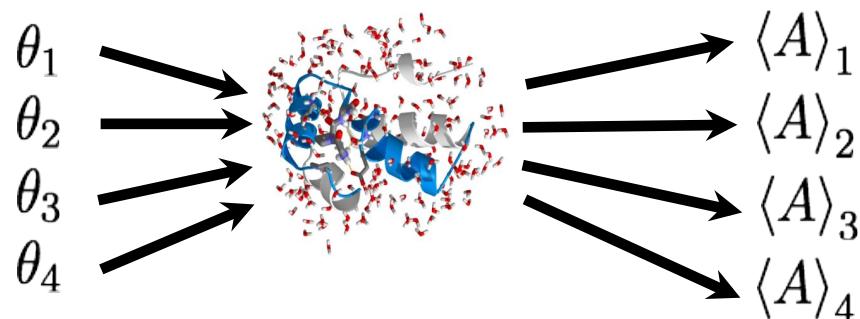
**If we find ourselves in uncharted territory in chemical space, throw in more data.
(Maybe it could give us hints about which data would be useful to obtain?)**

Would be nice if parameterization procedure didn't require me to learn crazy new mathematics or algorithms (mainly because I'm lazy)

Would give us an idea of how reliable its new predictions are expected to be.

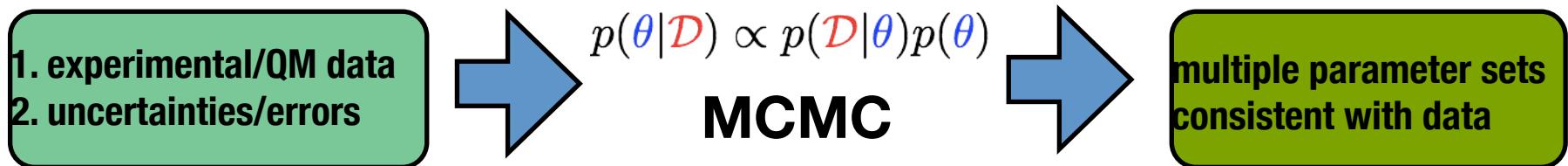
Is there a procedure that could fit these criteria?

Can a **Bayesian** approach provide a better way to fit forcefields?



We cast the parameterization of force fields as a **Bayesian** inference problem.

Bayes rule provides a **probability measure** over unknown parameters given data:



\mathcal{D} Data (QM and experimental)

θ Forcefield parameters

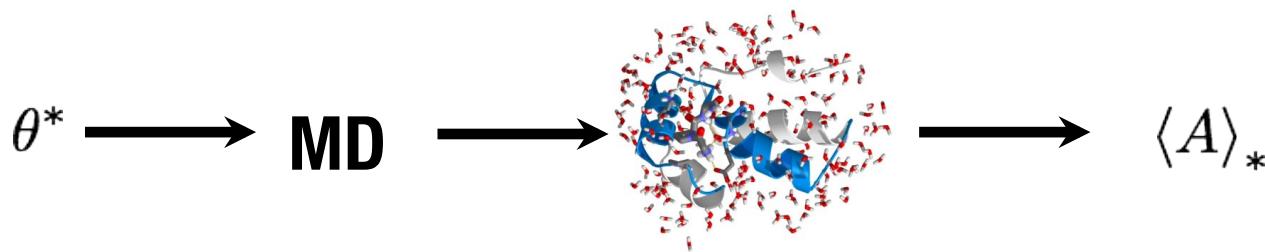
$p(\theta|\mathcal{D})$ posterior

$p(\mathcal{D}|\theta)$ error model for data

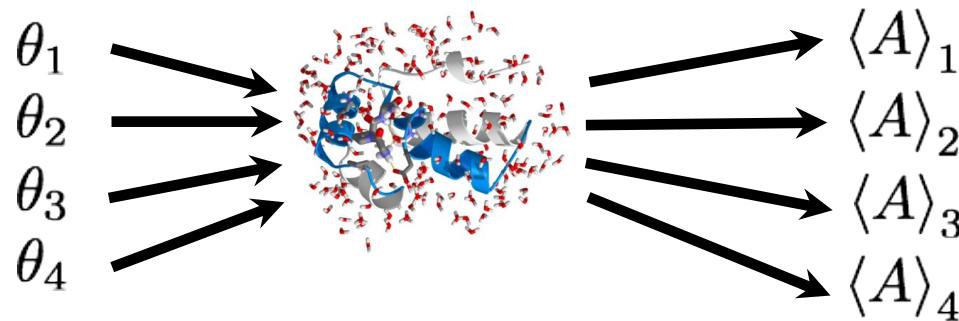
$p(\theta)$ prior on forcefield parameters

A Bayesian approach also provides a straightforward way to propagate uncertainty to computed properties.

Compute properties with one representative parameter set from collection

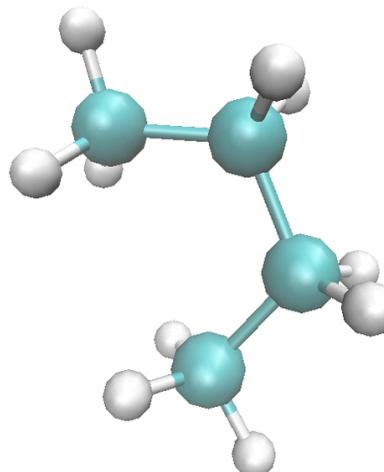
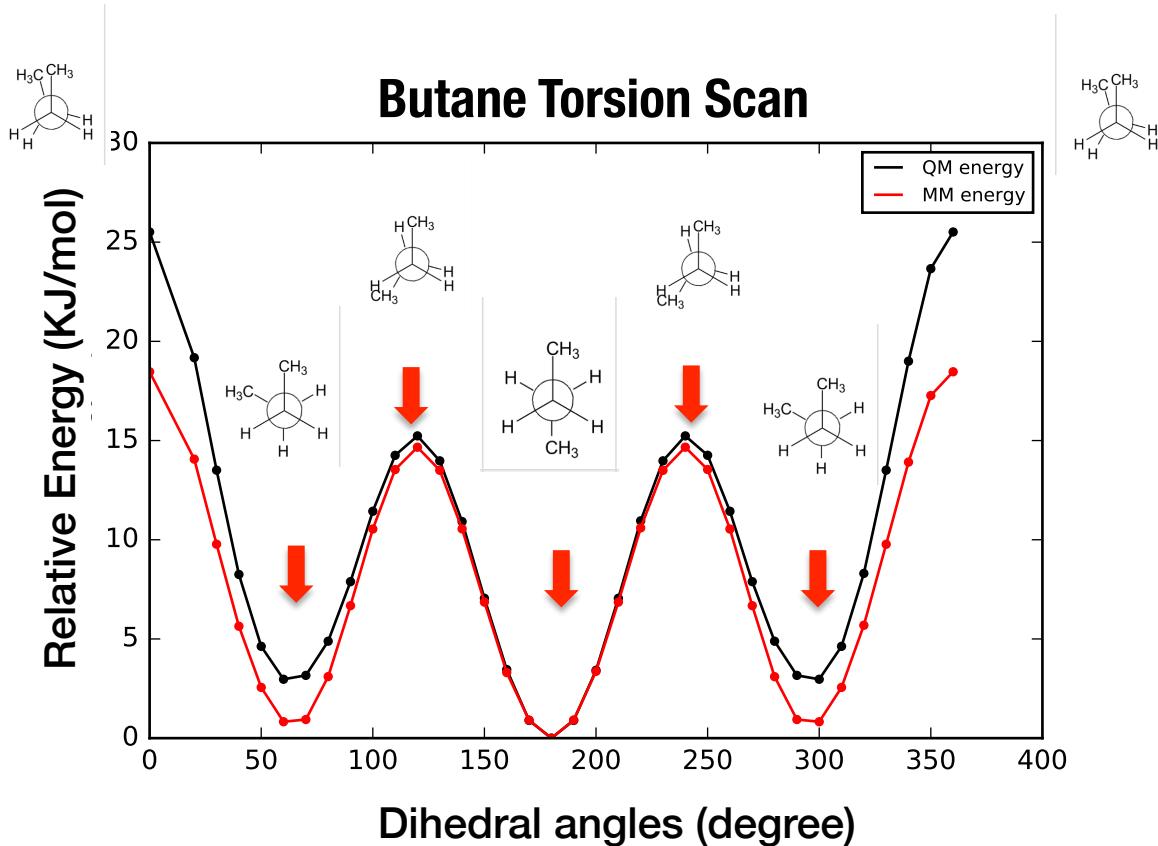


Estimate computed properties for other parameter sets from collection by reweighting (e.g. Zwanzig)



Can estimate both **statistical** and **systematic** components of computed property!

The **dihedral** angle energy is usually modeled with a Fourier series.



$$E_i^{dih} = \sum_m \sum_n = K_{m,n} (1 + \cos(n\phi_i))$$

The model is given by a Gaussian likelihood and uninformative priors.

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

$$P(D|\Theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\sum_i (E_i^{QM} - (E_i^{MM} + E_i^{dih}) + c)^2}{2\sigma^2}}$$

E_i^{QM} : QM energy for configuration i

E_i^{MM} : MM energy without dih for configuration i

$$E_i^{dih} = \sum_m \sum_n = K_{m,n} (1 + \cos(n\phi_i))$$

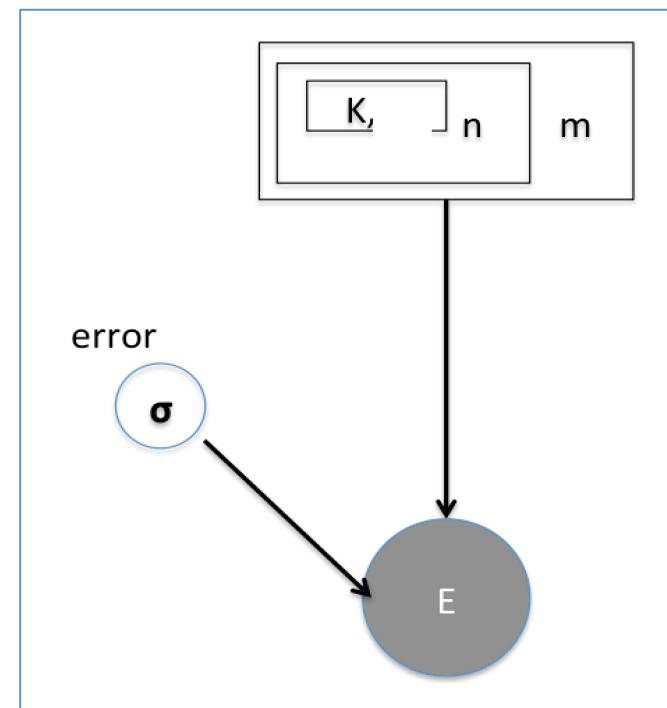
m: torsion type

n: periodicity

Φ_i : torsion angle for the i^{th} torsion.

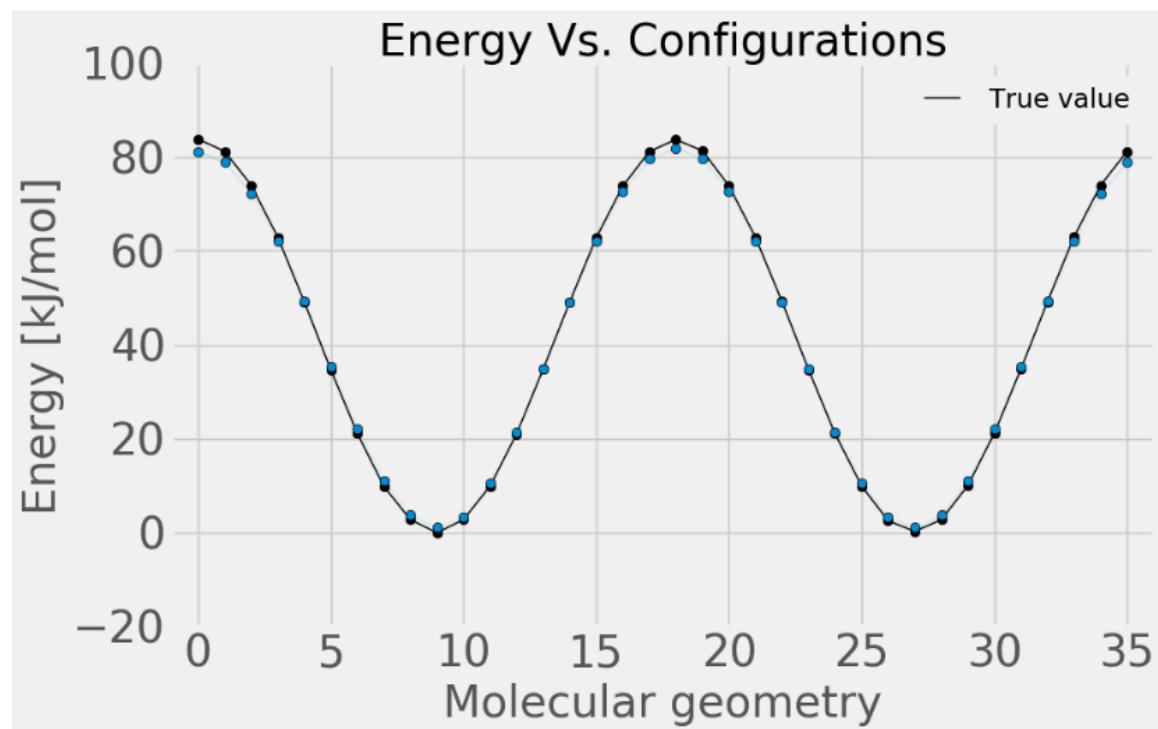
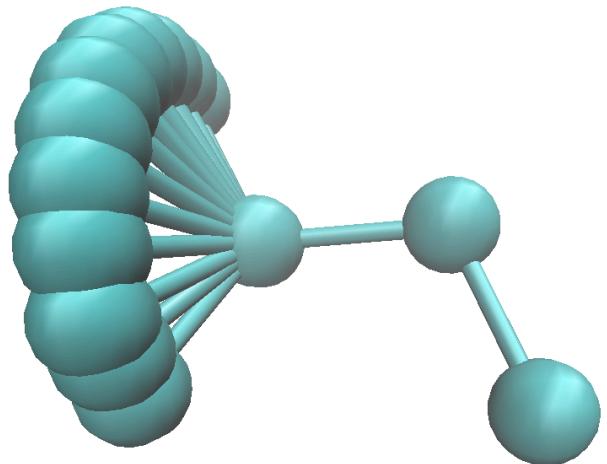
Code (pre-alpha) is available on Github:

<https://github.com/choderalab/torsionfit>



Toy model of 4 carbons

Jupyter notebook demo



Model selection with Pymc

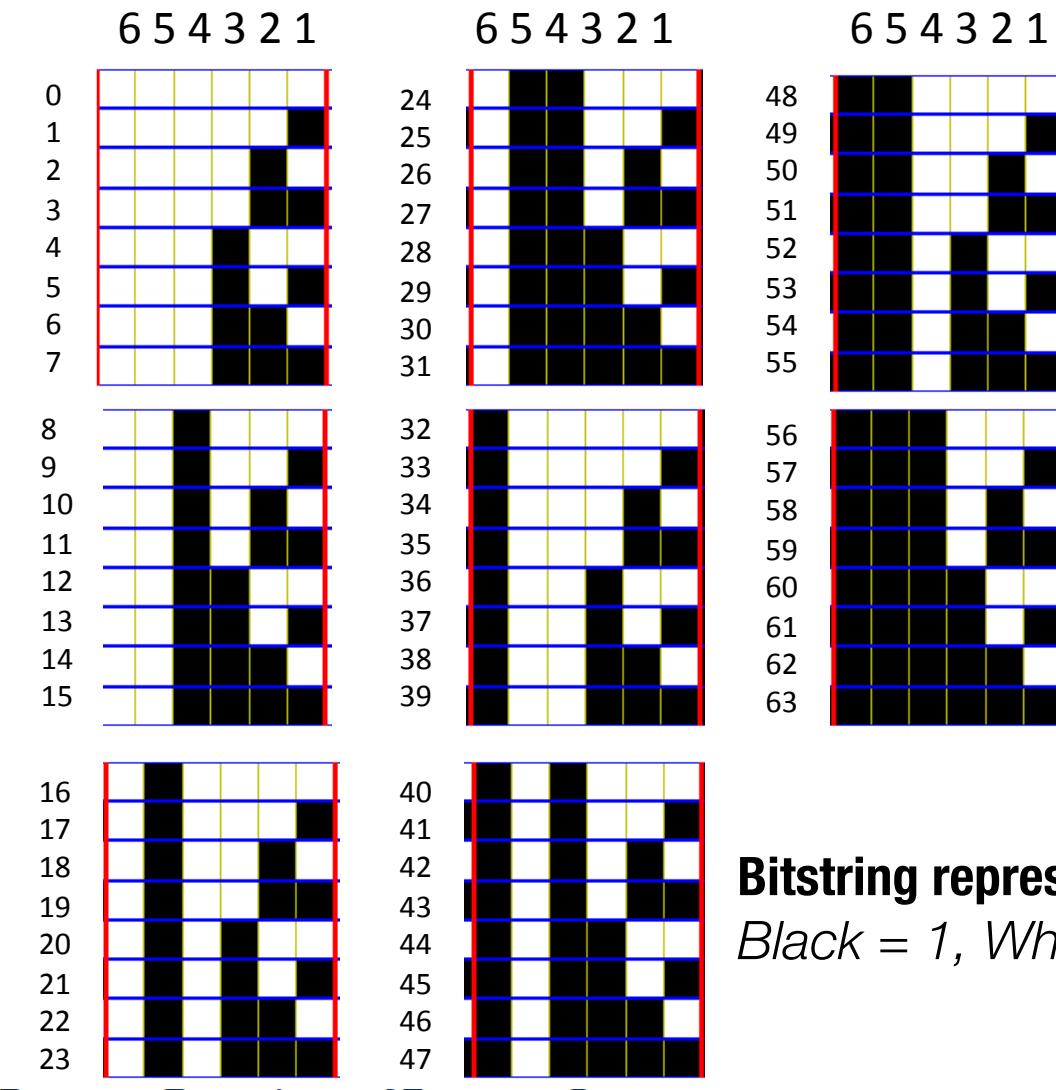
How many **Fourier terms** do we need to reproduce the PES?

Current tools rely on **chemical intuition** to choose periodicity terms.

$$2^6 = 64$$

For each torsion term!

Algorithm for sampling over models in Pymc



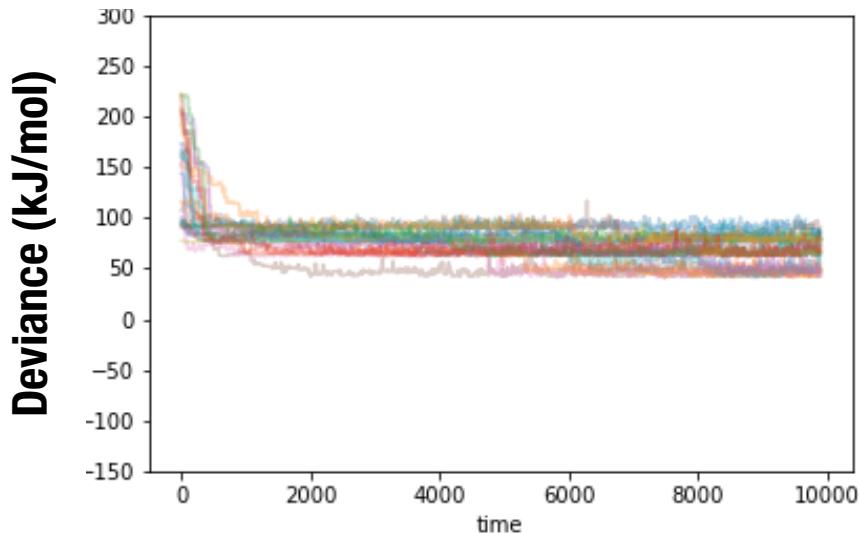
Bitstring representation of integers {0, 63}
Black = 1, White = 0

```
bitstring = pymc.DiscreteUniform(name, lower=0, upper=63, value=multiplicity_bitstrings[torsion_name])
```

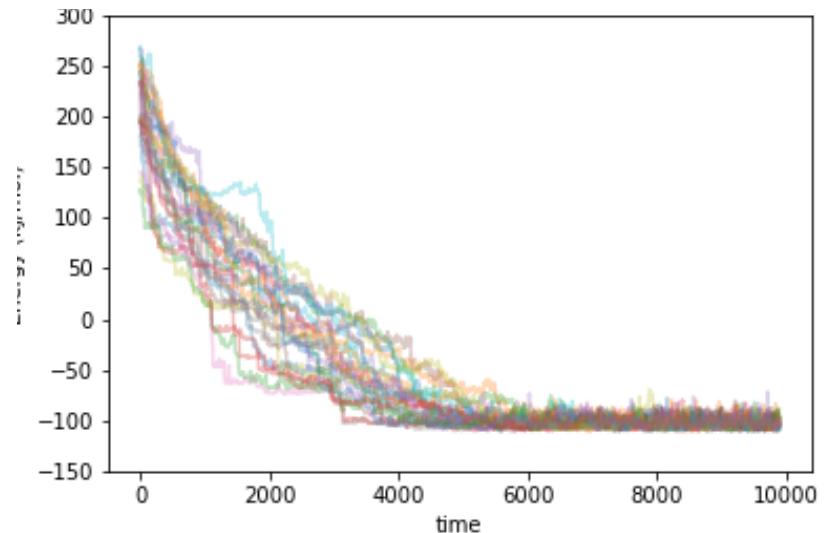
RJMC leads to poor mixing

$$-2 \log[p(D|\theta)]$$

Reversible jump



Model 63 (includes all terms)



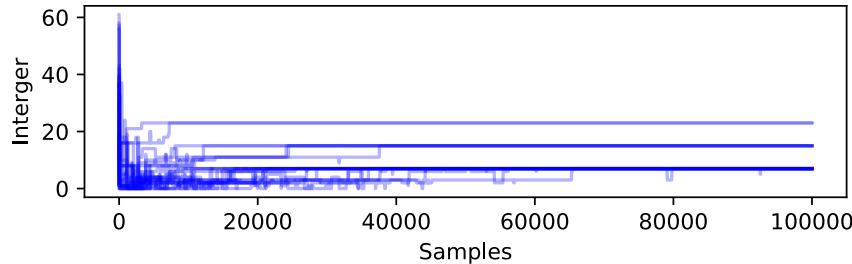
Why does this implementation lead to poor mixing?

```
for n in range(len(param.dihedral_types[t])):
    m = int(param.dihedral_types[t][n].per)
    multiplicity_bitmask = 2 ** (m - 1) # multiplicity bitmask
    if (multiplicity_bitstring & multiplicity_bitmask) or not rj:
        sample = model.pymc_parameters[k].value
    else:
        # This torsion periodicity is disabled.
        logger().debug('Turning off {}'.format(m))
        param.dihedral_types[t][n].phi_k = 0
        param.dihedral_types[reverse_t][n].phi_k = 0
```

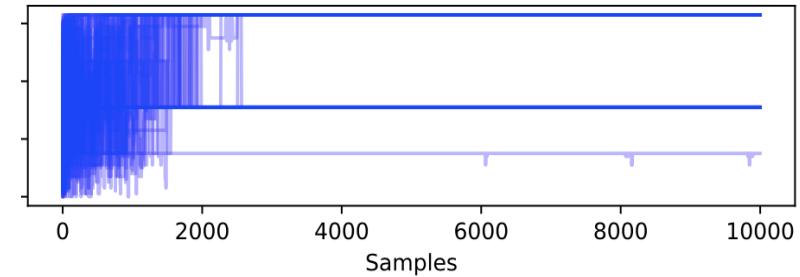
A biased prior was used to improve mixing between models.

Traces and histograms of multiplicity bit string

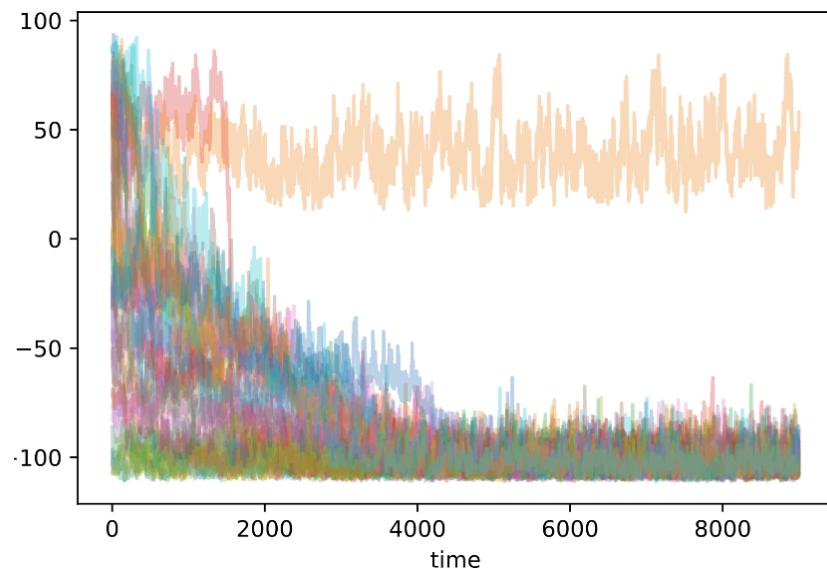
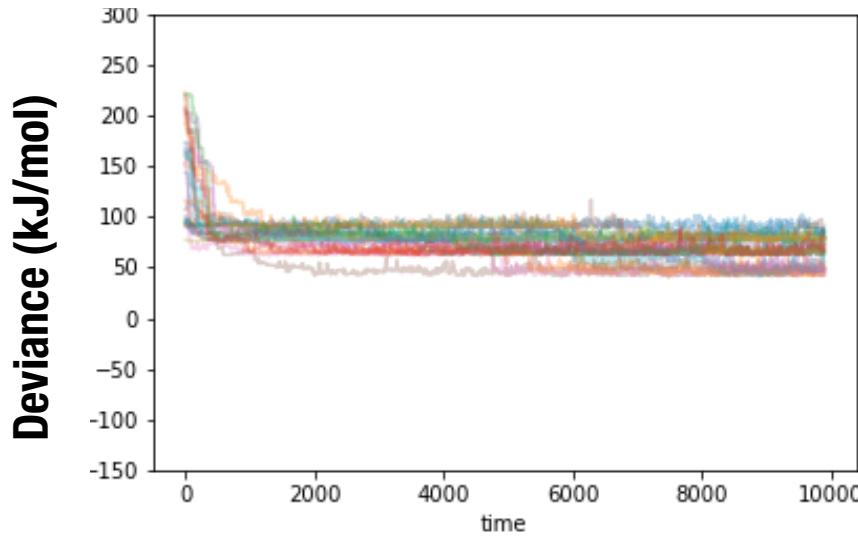
Uninformative prior



Informative (biased) prior

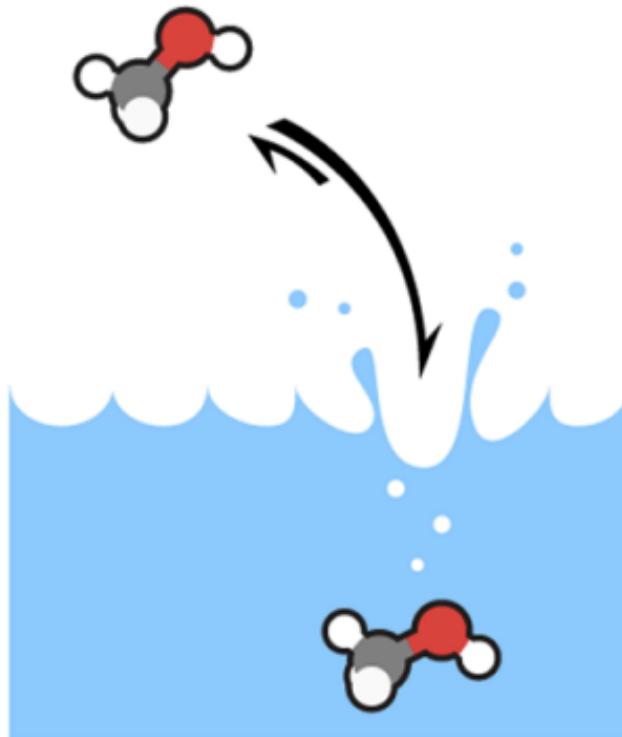


Deviance (kJ/mol)



Propagating the **error** to a computed property using the posterior.

Hydration Free Energy.



$$\Delta G_{hyd} = -k_B T \ln \left[\frac{Z_s}{Z_v} \right]$$

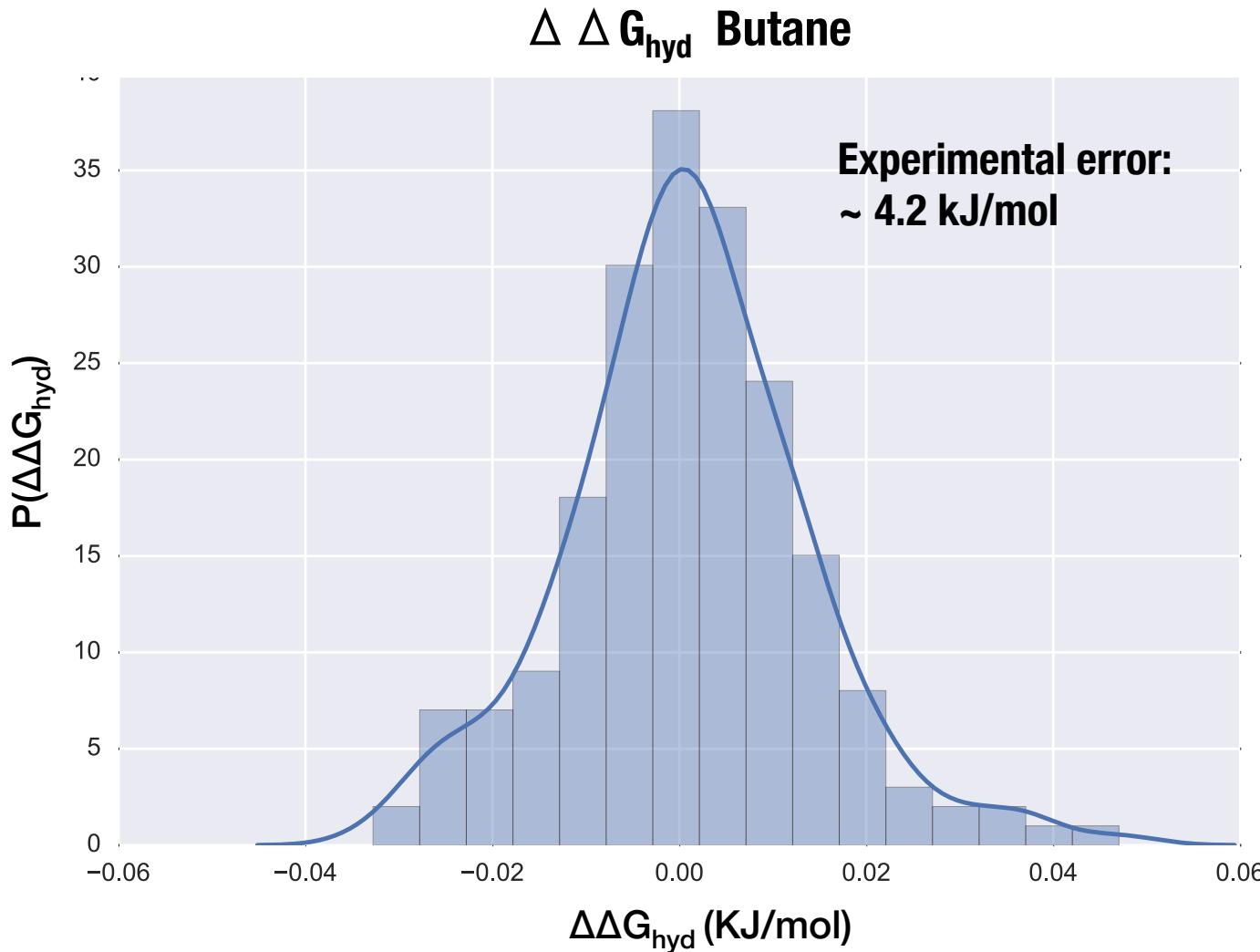
k_B : Boltzmann constant

Z_s : Normalizing constant of molecule in solvent

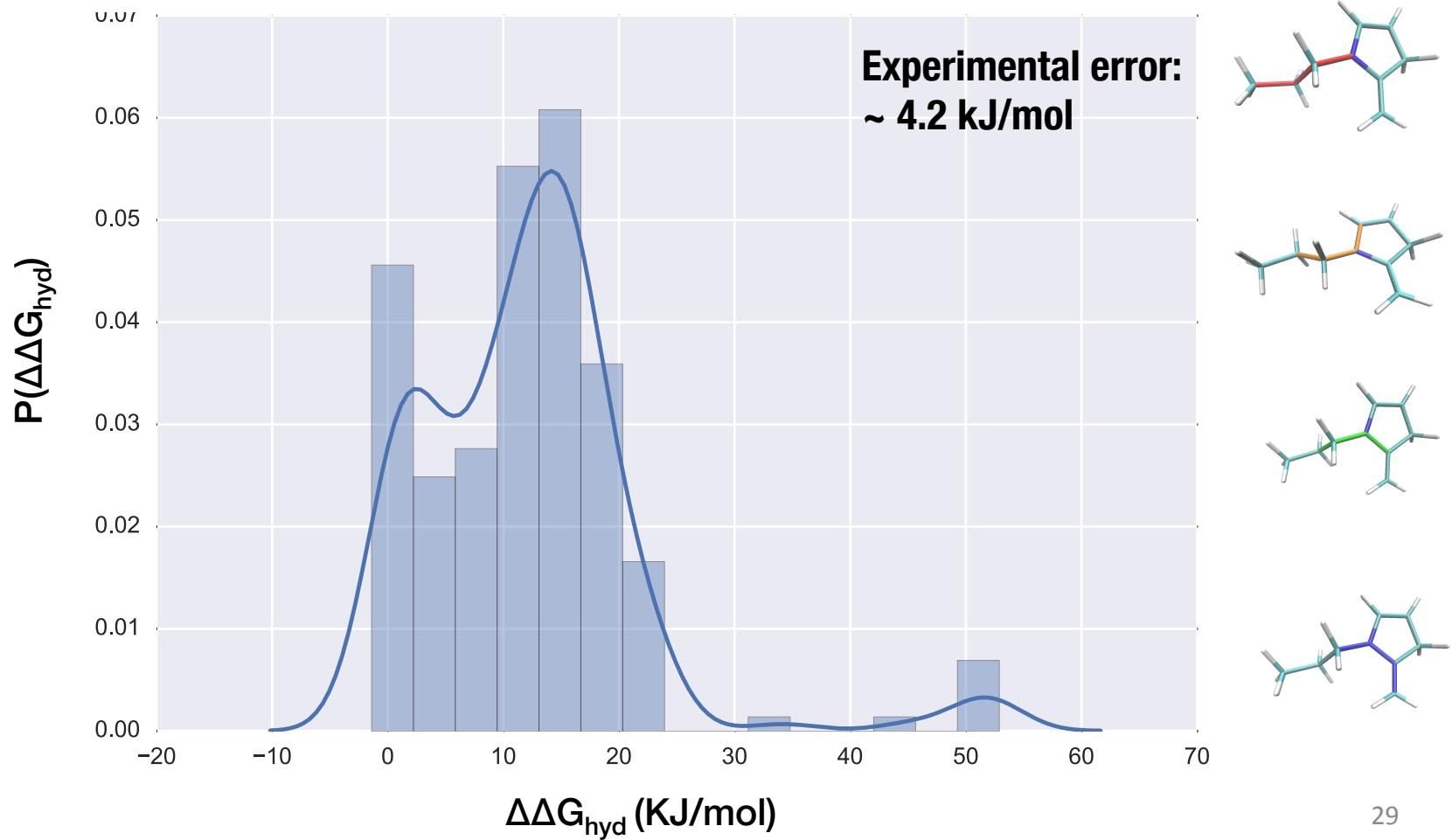
Z_v : Normalizing constant of molecule in vacuum

Fennell, et al. 2014.

For butane, error in $\Delta \Delta G$ is very small relative to experimental error.

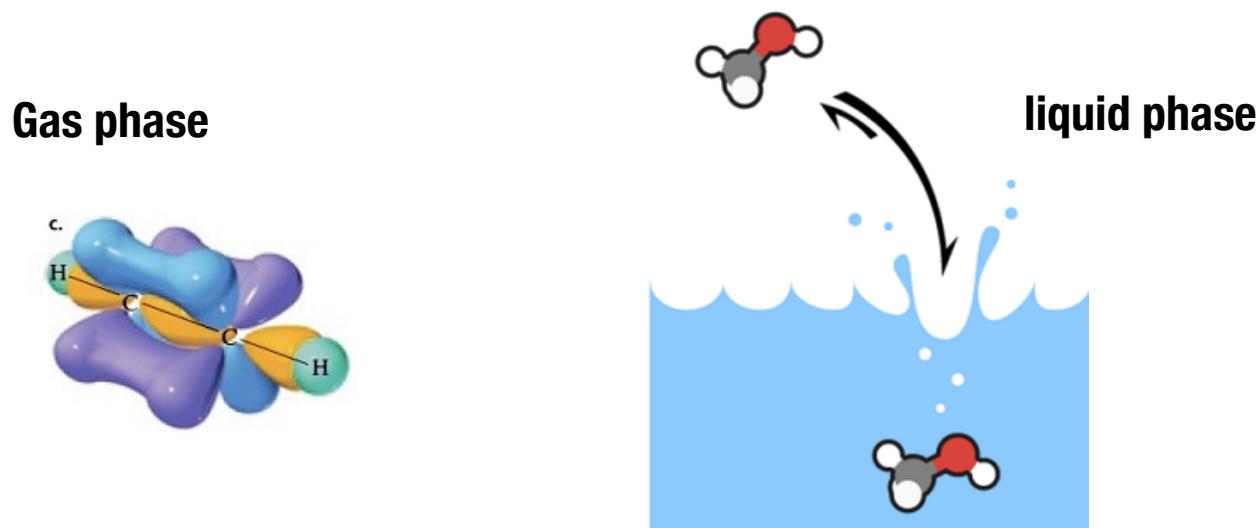


The error for more complex molecules is significant relative to experimental error.



Why are the errors so large?

- **Do we have enough data?**
 - Is QM data enough?
 - Are we using the right kind of data?
 - Is the functional form adequate (model misspecification) ?



The Open Forcefield Group



CHRISTOPHER BAYLY
OPENEYE SCIENTIFIC



JOHN CHODERA
SLOAN KETTERING INSTITUTE



MICHAEL GILSON
UNIVERSITY OF CALIFORNIA, SAN DIEGO



KENNETH KROENLEIN
NIST THERMODYNAMICS RESEARCH CENTER



DAVID MOBLEY
UNIVERSITY OF CALIFORNIA, IRVINE



MICHAEL SHIRTS
UNIVERSITY OF COLORADO, BOULDER

<http://openforcefield.org>

Acknowledgment



**John Chodera
Gregory Ross**

Funding

Tri-Institutional PhD Program
Chemical Biology

