Efficient Bayesian Sampling of Molecular Conformers

Understanding Torsional Correlations & Entropy

Chemistry

Prof. Geoffrey Hutchison

Departments of Chemistry & Chemical Engineering
University of Pittsburgh

Emerging Tech in Computational Chemistry

March 23, 2020

<http://hutchison.chem.pitt.edu>

Computationally-Driven Materials Discovery

How do we rapidly find organic materials with interesting properties?

To Understand Properties, Need Conformers

Many properties depend on molecular conformation:

P3HT

PCPDTBT

Side Note: Energy Accuracy for Conformers

Many conformers only have small ΔE between them...

How reliable are relative energy rankings?

- Want to compare energies
force fields, semiempirical,
ML, density functional, ..
- Recent preprint:
<https://bit.ly/ConfMark>
- Previous benchmark:
Int. J Quant. Chem. (2018) 118, e25512
<https://doi.org/10.1002/qua.25512>

Poster: Dakota Folmsbee
"Assessing conformer energies"

How Good is Conformer Sampling?

Current Benchmarks

- Pick a ligand from PDB
- Generate X conformers (e.g., 50, 100, 100,000)
- Compare generated geometries to PDB
- Compute lowest RMSD for best computed geometry

"If I throw enough darts, I should hit the crystal structure, right?"

How Good is Conformer Sampling?

Current Techniques

- Mostly stochastic - generate random / diverse geometries
- Distance geometry (RDKit)
- Torsion driving (Omega, Confab, others)
- Some use of low-mode MD (still need basin-hopping)

Few methods learn (sample) from the potential energy surface

Conventional Conformer Sampling

Most torsion driving approaches:

- Assume independent free rotation of dihedrals
- Usually rely on expected average dihedral angles
- Fixed frequencies for dihedral angles
- Add extra rules for unusual cases...

**Most molecules in crystal databases
have dihedrals outside $\pm 10^\circ$ range
of common angles!**

Who Was Bayes?

Thomas Bayes (?)
~1701-1761

(Image: Wikipedia)

Became really interested in probability and "inverse probability"

If I make X observations, what can I infer about the distribution?

Bayes Theorem

1. Start with initial assumption of probability - the "prior distribution"
 $p(\text{blue})$
2. Make observations
3. Update probability distribution
 $p(\text{blue} | X) = p(\text{blue})$ given X happened

Bayesian Sampling and Stat Mech

- Boltzmann statistics connect **Energies** and **Probabilities**
- We don't always know the potential energy surface *a priori*
- We **do** have prior expectations (e.g. 60° increments are most likely)
- Sample from the posterior (generate conformer, update probabilities)

*True probabilities
(CCSDT energies)*

*Current Estimate
(with uncertainty)*

**Bayesian sampling works
with any energy function**

Does Bayesian Sampling Work?

- **First try** - compared BOA to uniform random & exhaustive search (Confab)
- Fixed 'budget' of 100 samples to **exhaustive (10^4 - 10^6)**
- Still frequently finds lower E conformer - or very close
100x to 10,000x fewer samples
- Clearly still problems with 5+ bonds .. *how do we do better?*

Improving on Bayesian Sampling...

- **First try:** BOA
Can work with **any** energy function
Outperforms exhaustive search
.. on small molecules
- Assumed that all dihedral angles are independent free rotors...
- What if we look for correlated dihedral angles?
- If we can improve the "initial guess" of probable dihedrals...

BOKEI - Using Correlated Dihedrals

- Uses knowledge of correlated torsions avoids high E areas, suggests low E
- Correlated non-ring dihedrals
 - h-bonds
 - π - π stacking
 - steric clashes
- PCCP (2020) 22, 5211-5219
<https://doi.org/10.1039/C9CP06688H>

E. bokei cactus

BOKEI - Using Correlated Dihedrals

- Uses knowledge of correlated torsions avoids high E areas, suggests low E
- Correlated non-ring dihedrals
 - h-bonds
 - π - π stacking
 - steric clashes
- PCCP (2020) 22, 5211-5219
<https://doi.org/10.1039/C9CP06688H>

E. bokei cactus

Poster: Leung Sing Chan (Oxford)
"BOKEI - Bayesian Optimization Using Knowledge..."

BOKEI - Using Correlated Dihedrals

- Uses knowledge of correlated torsions
avoids high E areas, suggests low E

Big advantage
in early iterations

"knows more"
about PES

PCCP (2020) 22, 5211-5219

<https://doi.org/10.1039/C9CP06688H>

BOKEI - Using Correlated Dihedrals

BOKEI can use any energy method (e.g. MMFF94 optimizations and GFN2)

Beats both GA (up to ~12 bonds) and BOA - often by 2-5 kcal/mol

Need better priors or more evaluations to beat GA on large molecules

PCCP (**2020**) 22, 5211-5219

<https://doi.org/10.1039/C9CP06688H>

BOKEI - Correlated Dihedral Patterns

Combine common torsion SMARTS patterns A+B

Currently need many matches to generate probability densities

**Crystallographic
Open Database**

MMFF94

GFN2

What are the correlated dihedrals?

BOKEI includes patterns avoiding steric clashes (high E regions)
.. but also favorable interactions

Intramolecular h-bonds

π - π stacking

PCCP (2020) 22, 5211-5219

<https://doi.org/10.1039/C9CP06688H>

What's Next?

- BOKEI encodes some 2D neighboring correlations (match ~15% of molecules)
- Other patterns likely
- Longer-range correlations known with biomolecules
- Rings clearly have correlated dihedrals

N-3 degrees of freedom

What's Next?

- BOKEI encodes some 2D neighboring correlations (match ~15% of molecules)
- Other patterns likely
- Longer-range correlations known with biomolecules
- Rings clearly have correlated dihedrals

N-3 degrees of freedom

BOKEI can continually improve through data

Are Correlated Dihedrals Common?

BOKEI patterns ~15% in large molecule databases
(not uncommon, but not huge either...)

Are correlated dihedrals more common?

**Many
high energy
conformers**

**Few conformers
under
10 kcal/mol**

Are Correlated Dihedrals Common?

Even "simple" n-alkane chains show correlated dihedrals

.. **few** low-energy conformers

.. most conformers are high E due to steric clashes

Are Correlated Dihedrals Common?

Preprint coming soon... Sampled ~150k molecules ~15M conformers

Number of low-energy conformers increases slowly

.. "Rotatable bonds" are **not** independent free rotors

Take-Home Messages

Bayesian sampling works for conformers

- Can use any energy function
 - Learning dihedral correlations helps immensely
 - Better "prior" knowledge = better searches
- Correlated dihedrals are common
 - Work to come: ring motion
 - Work to come: accurate conformer entropies

Any Questions?