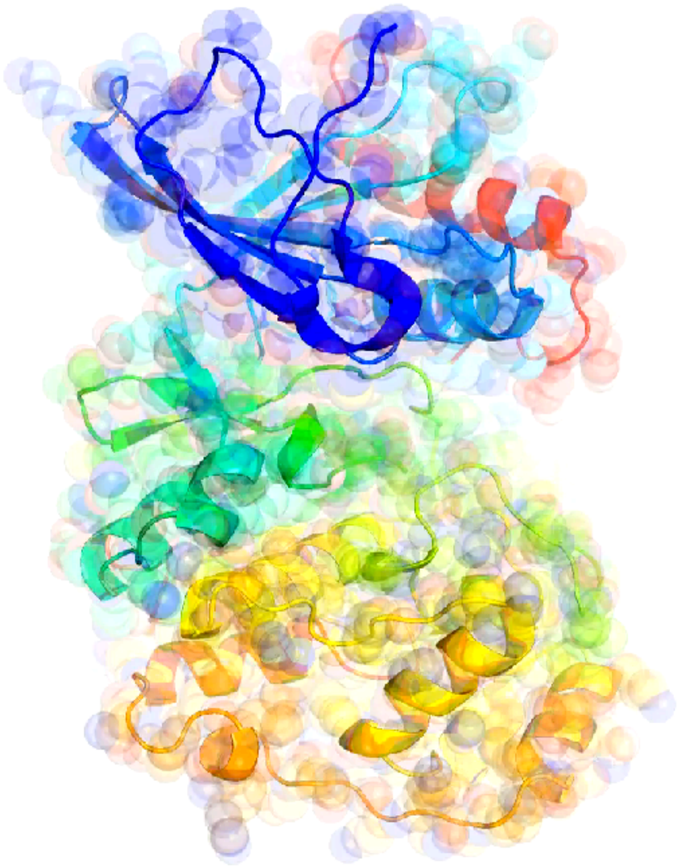
Improving molecular models by generating high-quality quantum chemistry data

Chaya D. Stern, Jeffrey Wagner, Daniel G. A. Smith, Christopher Bayly, Andrea Rizzi, John D. Chodera
RDKit User Group Meeting
Sep 25, 2019

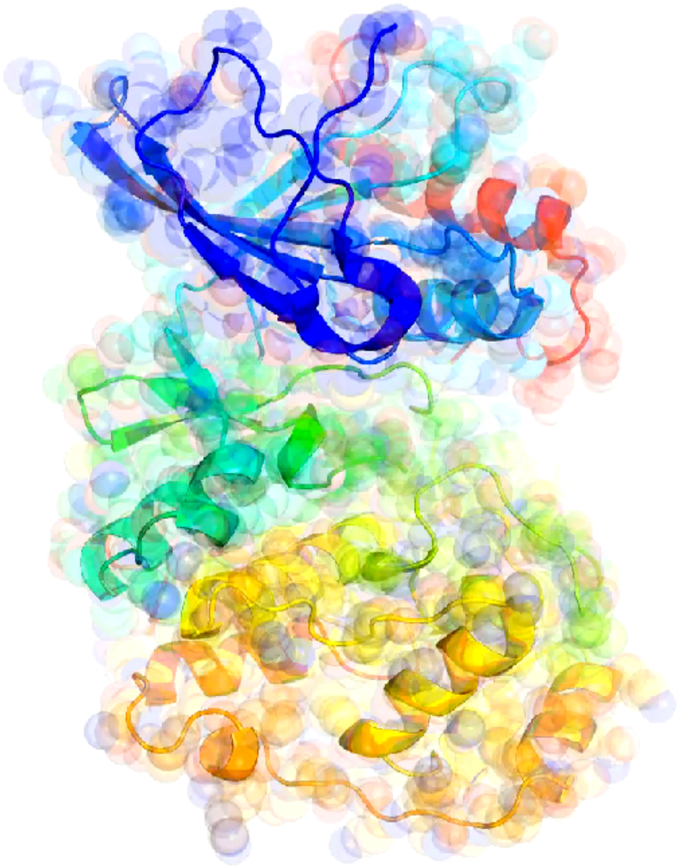
Molecular dynamics simulations provide atomic detail to dynamics that is usually difficult to observe



But more importantly, we can use MD simulations to make quantitative predictions.

Rafal Wiewiora

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



But more importantly, we can use MD simulations to make quantitative predictions.

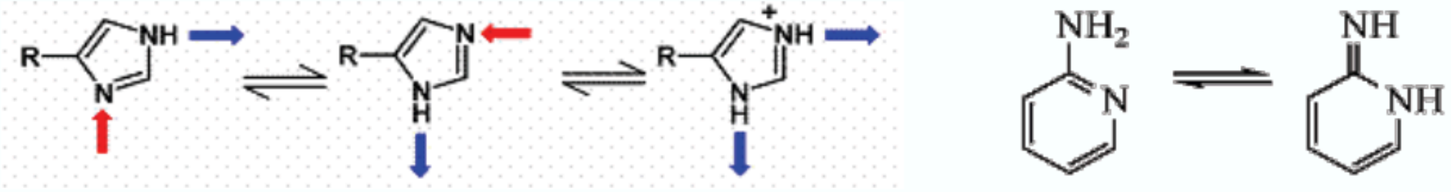
Rafal Wiewiora

Predictions fail for three main reasons.

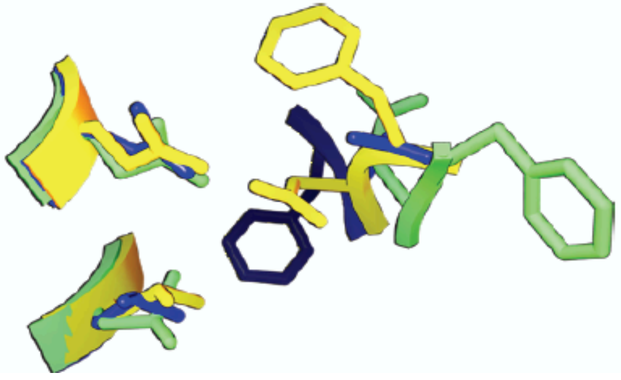
1. The forcefield does a poor job of modeling the physics of our system

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

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



Predictions fail for three main reasons.

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

The Open Forcefield Consortium

INDUSTRY

Jeff Wagner

Daniel Smith

COORDINATING INTERMEDIARY

MOLSSI

MOLECULAR SOFTWARE SCIENCES INSTITUTE

coordination of funding while minimizing indirect costs

ACADEMIC

JOHN CHODERA
SLOAN KETTERING INSTITUTE

MICHAEL GILSON
UNIVERSITY OF CALIFORNIA, SAN DIEGO

DAVID MOBLEY
UNIVERSITY OF CALIFORNIA, IRVINE

MICHAEL SHIRTS
UNIVERSITY OF COLORADO, BOULDER

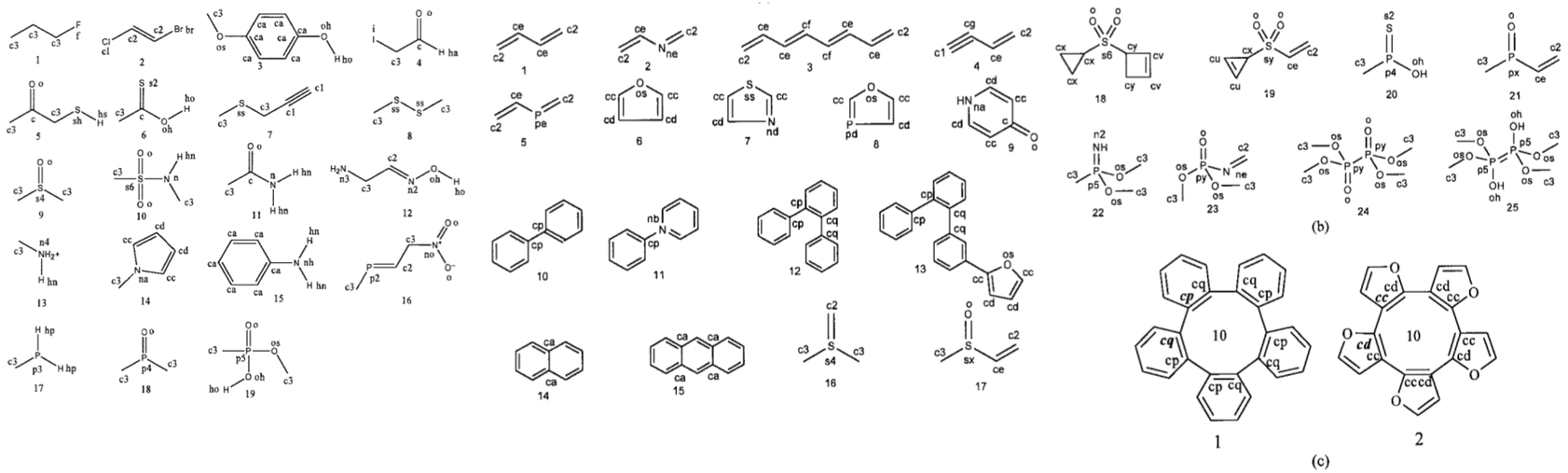
LEE-PING WANG
UNIVERSITY OF CALIFORNIA, DAVIS

CHRISTOPHER BAYLY
OPENEYE SCIENTIFIC

KENNETH KROENLEIN
NIST THERMODYNAMICS RESEARCH CENTER
(NIST is a US federal agency)

As drug discovery explores new parts of chemical space, how can forcefields keep up?

The Generalized Amber Forcefield (GAFF) was parameterized with this chemical universe:



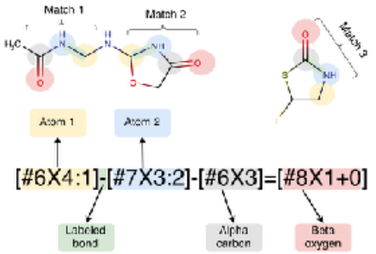
The chemical space of potentially pharmacological active molecules is estimated to be ~ O(10^{63})
Extension of this universe is nontrivial!

Wang J, Wolf RM, Caldwell JW, Kollman PA, and Case DA. J Comput Chem 25:1157, 2004.

Open Forcefield Initiative objectives

- Develop an open, scalable, and interoperable toolkit for automatically parameterizing forcefields

- Generate/curate open datasets necessary for producing high-accuracy small molecule forcefield



- Generate systematically-improved forcefields

All code, datasets and forcefields: https://github.com/openforcefield

The Open Forcefield Toolkit provides a seamless way to parameterize systems

Molecule
(various representations)

Forcefield
(SMIRNOFF format)

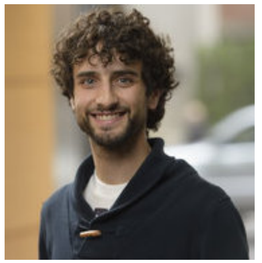
Open Forcefield
Toolkit

Parameterized system
(various representations)

Binder enabled example
https://github.com/openforcefield/openforcefield/tree/master/examples/using_smirnoff_with_amber_protein_forcefield

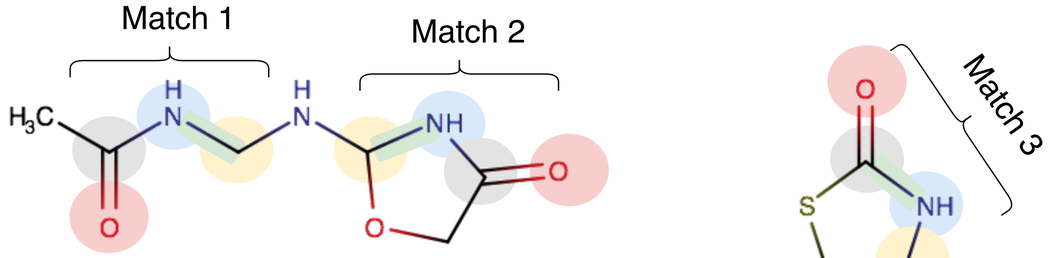
Jeff Wagner

Andrea Rizzi

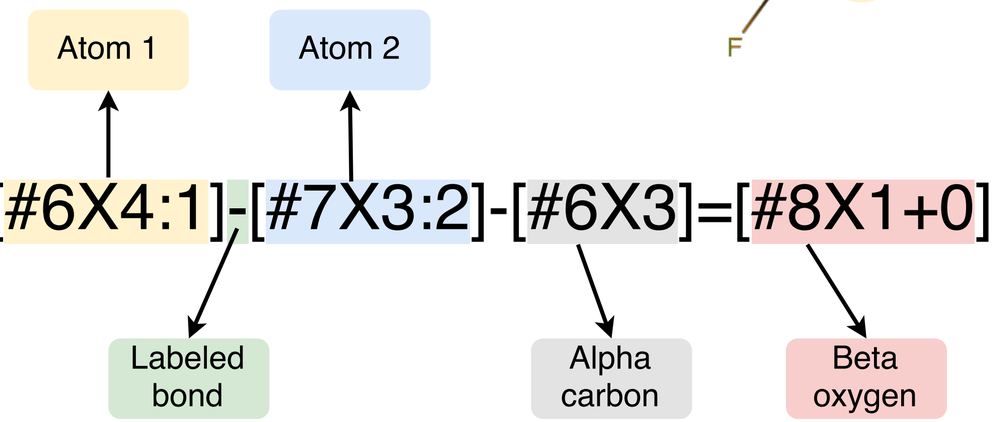


The SMIRKS Native Open Force Field (SMIRNOFF) format avoids the complexity of atom typing

RDKit enables OFFToolkit to be fully Open Sourced



MATCH BONDS DIRECTLY:



Caitlin Bannan

David Mobley

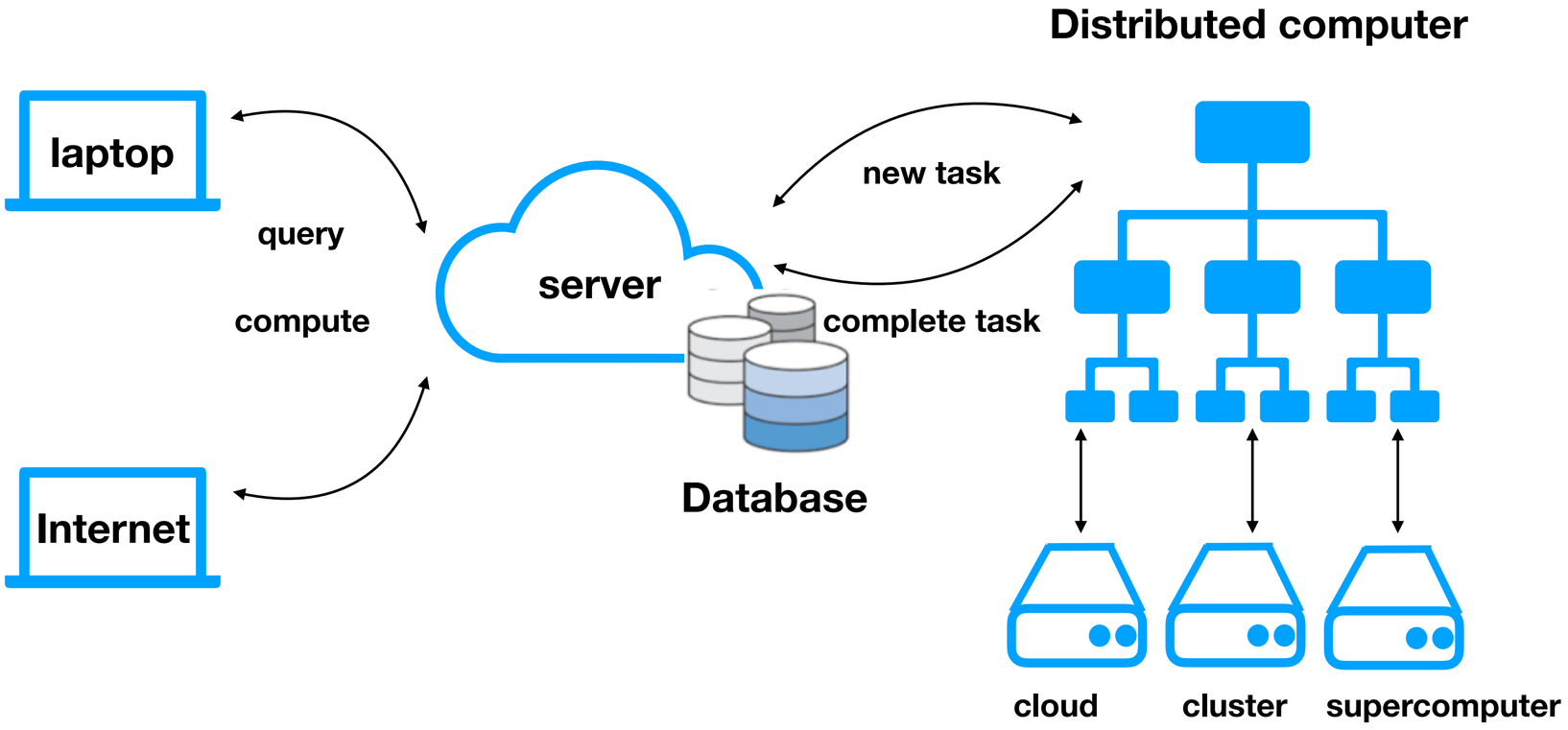
Christopher Bayly

Mobley, D., et. al (2018). JCTC, 14(11), 6076–6092. http://doi.org/10.1021/acs.jctc.8b00640

Use of industry-standard SMARTS/SMIRKS chemical perception greatly simplifies tooling for parameter assignment while solving issues with extensibility and flexibility

https://github.com/open-forcefield-group/openforcefield/blob/master/The-SMIRNOFF-force-field-format.md

QCArchive enables automatic generation and sharing of high-quality quantum chemistry data




Daniel Smith

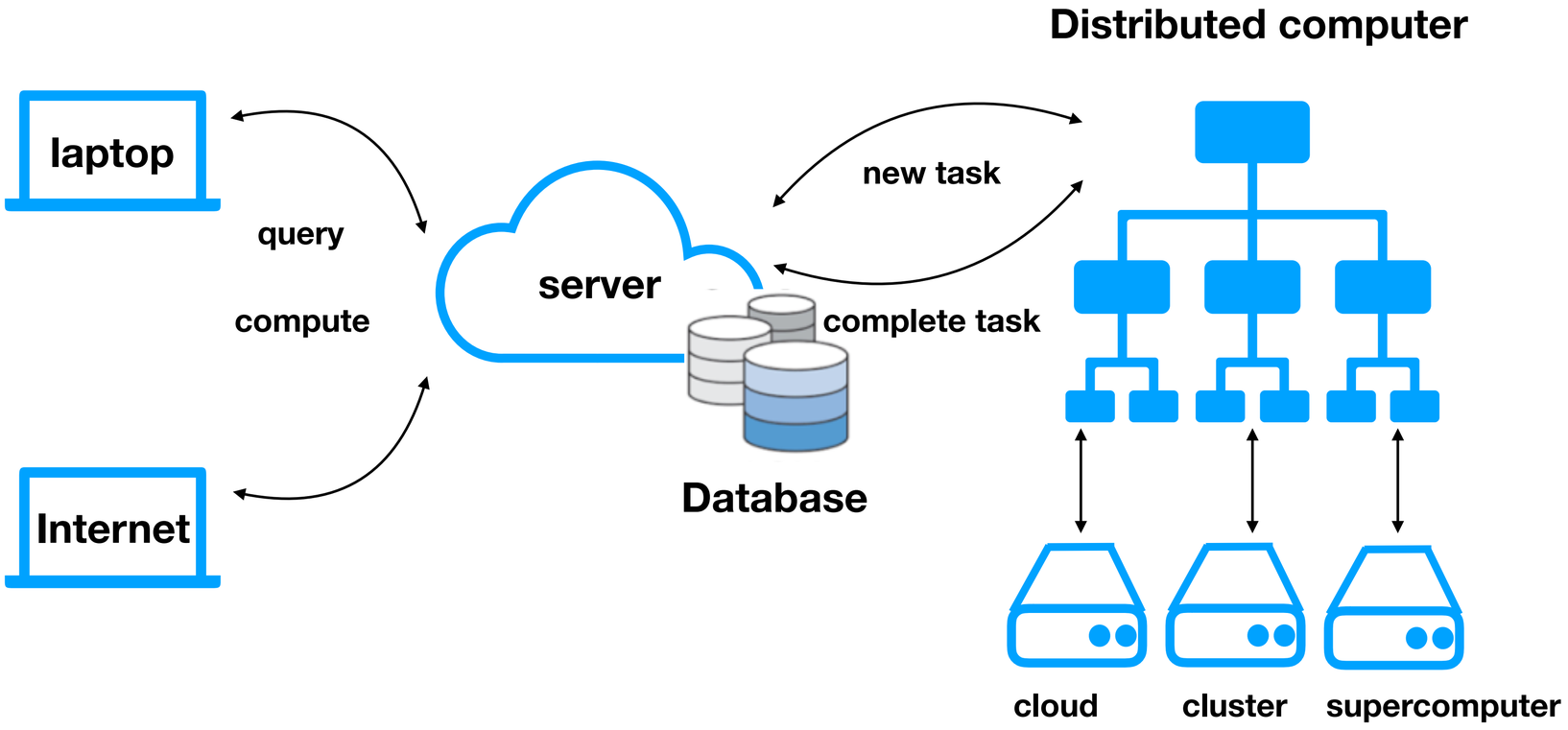

Doaa Altarawy


Levi Naden

qcarchive.molssi.org

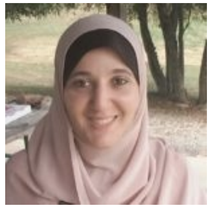
MolSSI QCArchive Project

QCArchive enables automatic generation and sharing of high-quality quantum chemistry data



QCArchive
A MolSSI Project


Daniel Smith

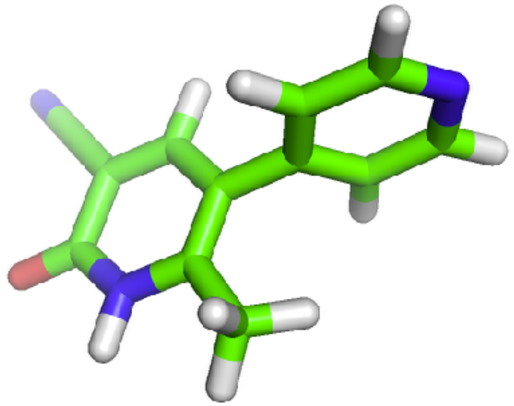

Doaa Altarawy

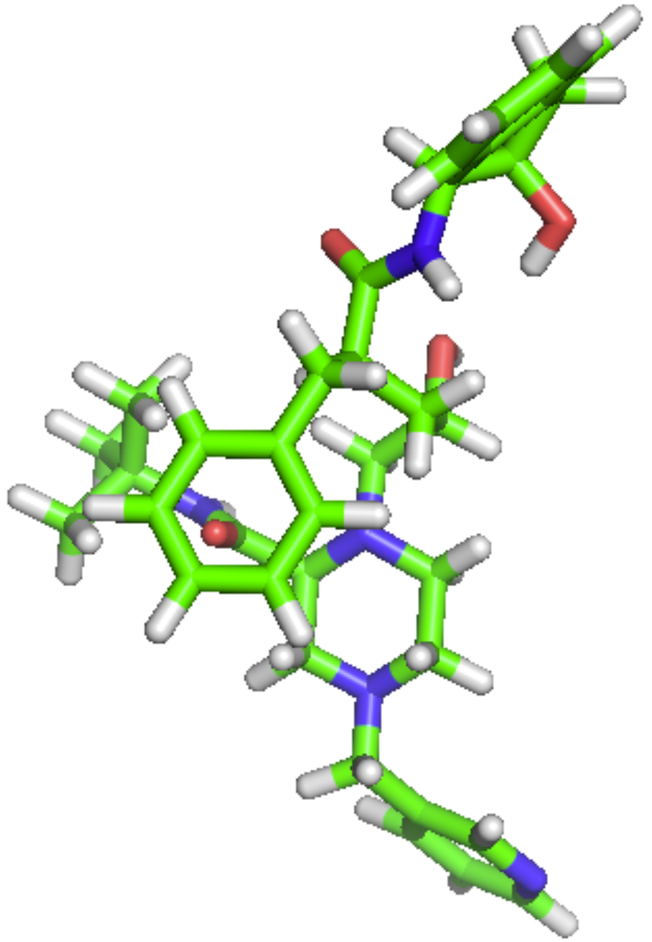

Levi Naden

qcarchive.molssi.org

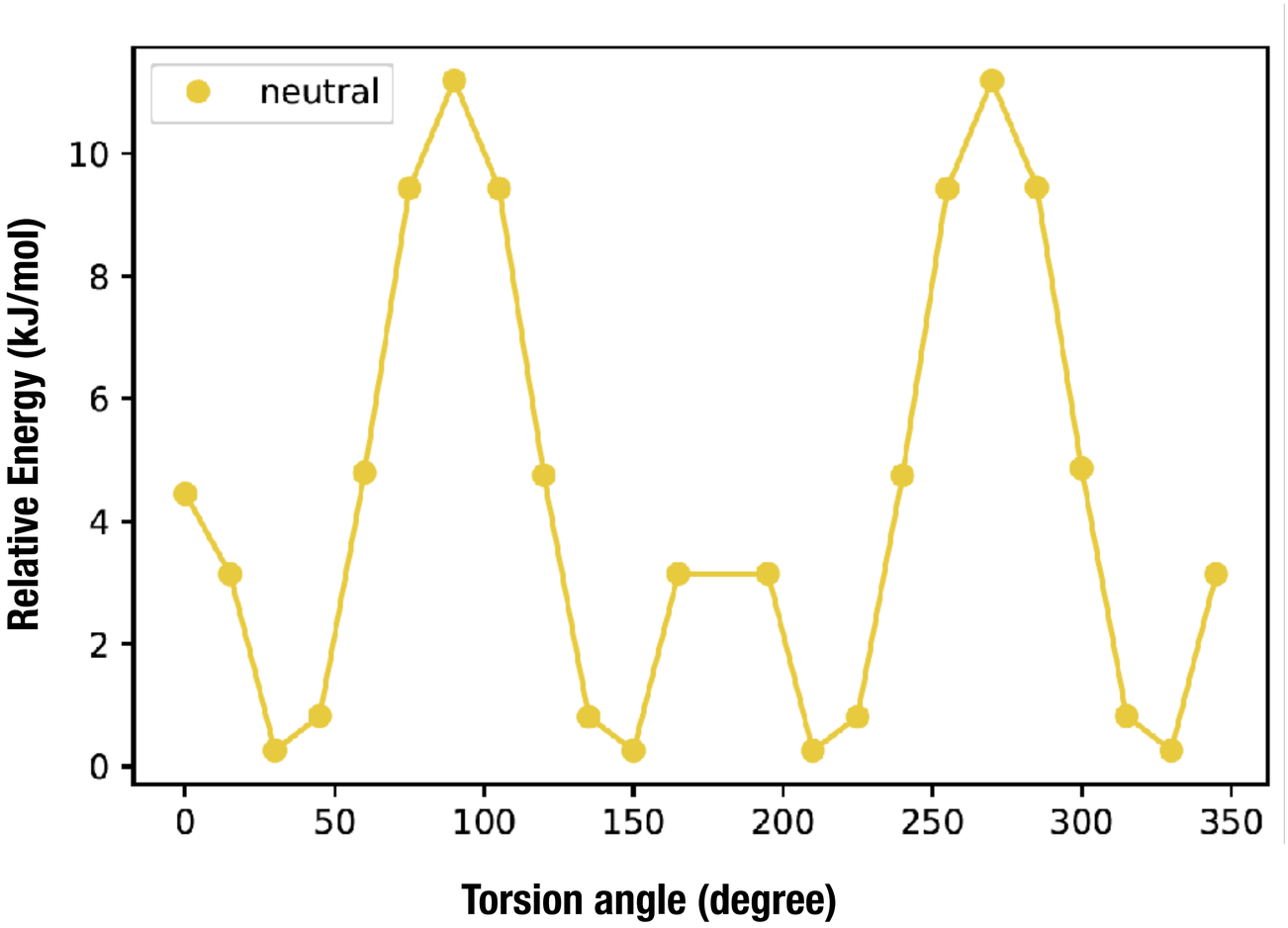
MolSSI QCArchive Project

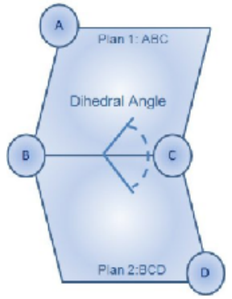
Fragmenting molecules is necessary to avoid high computational cost of generating QC data

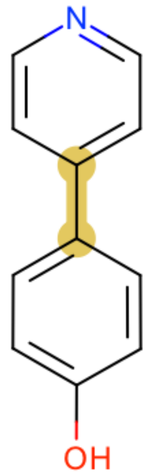




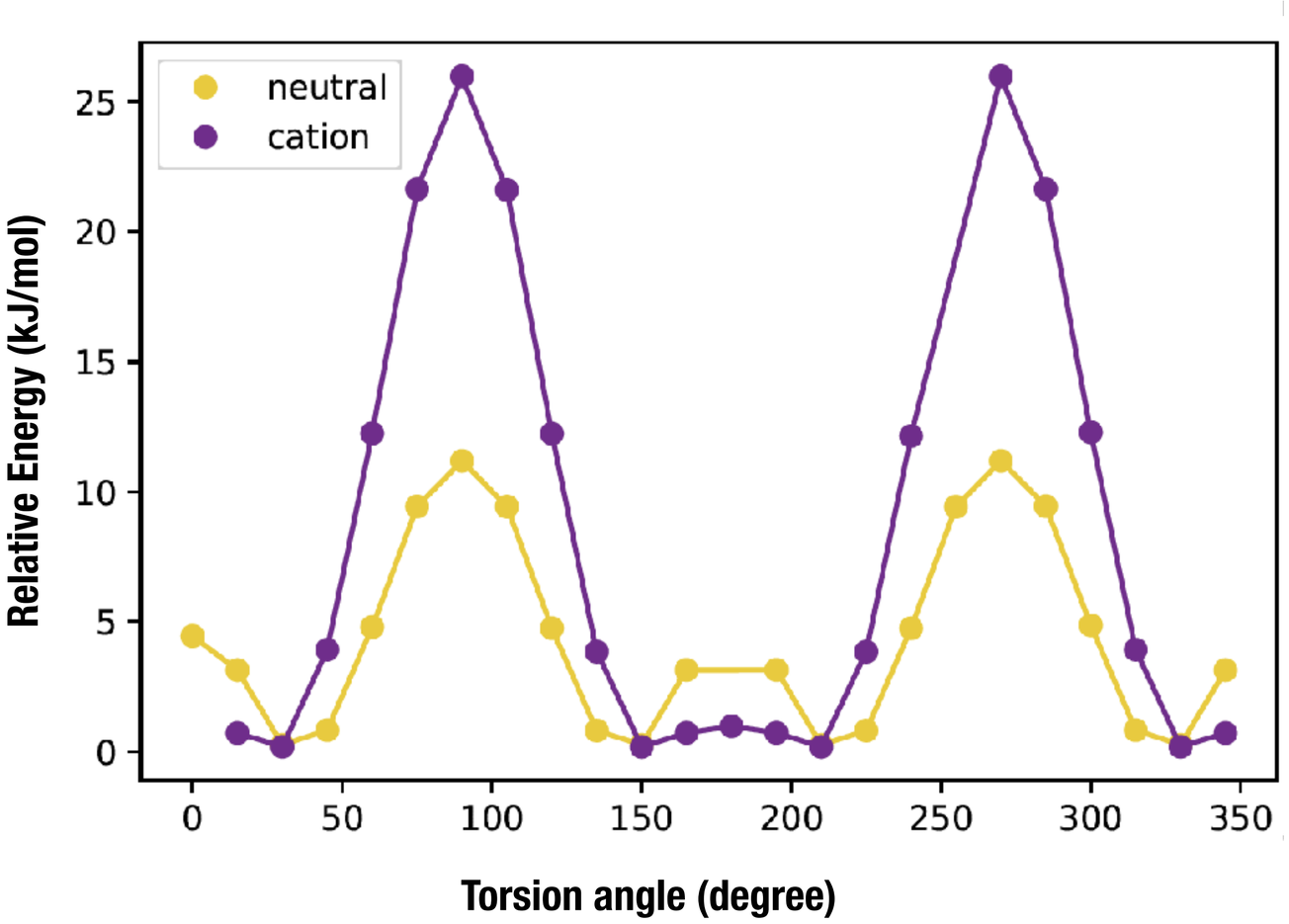
The chemical environment of bonds is sensitive to small remote substituent changes

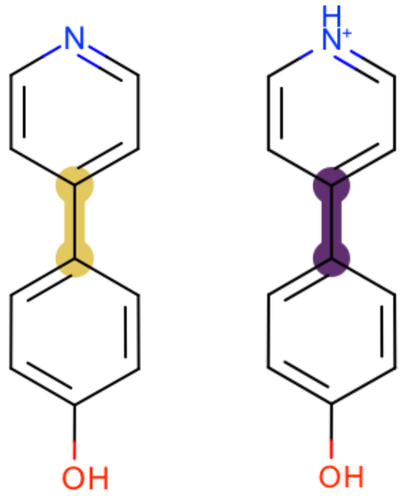




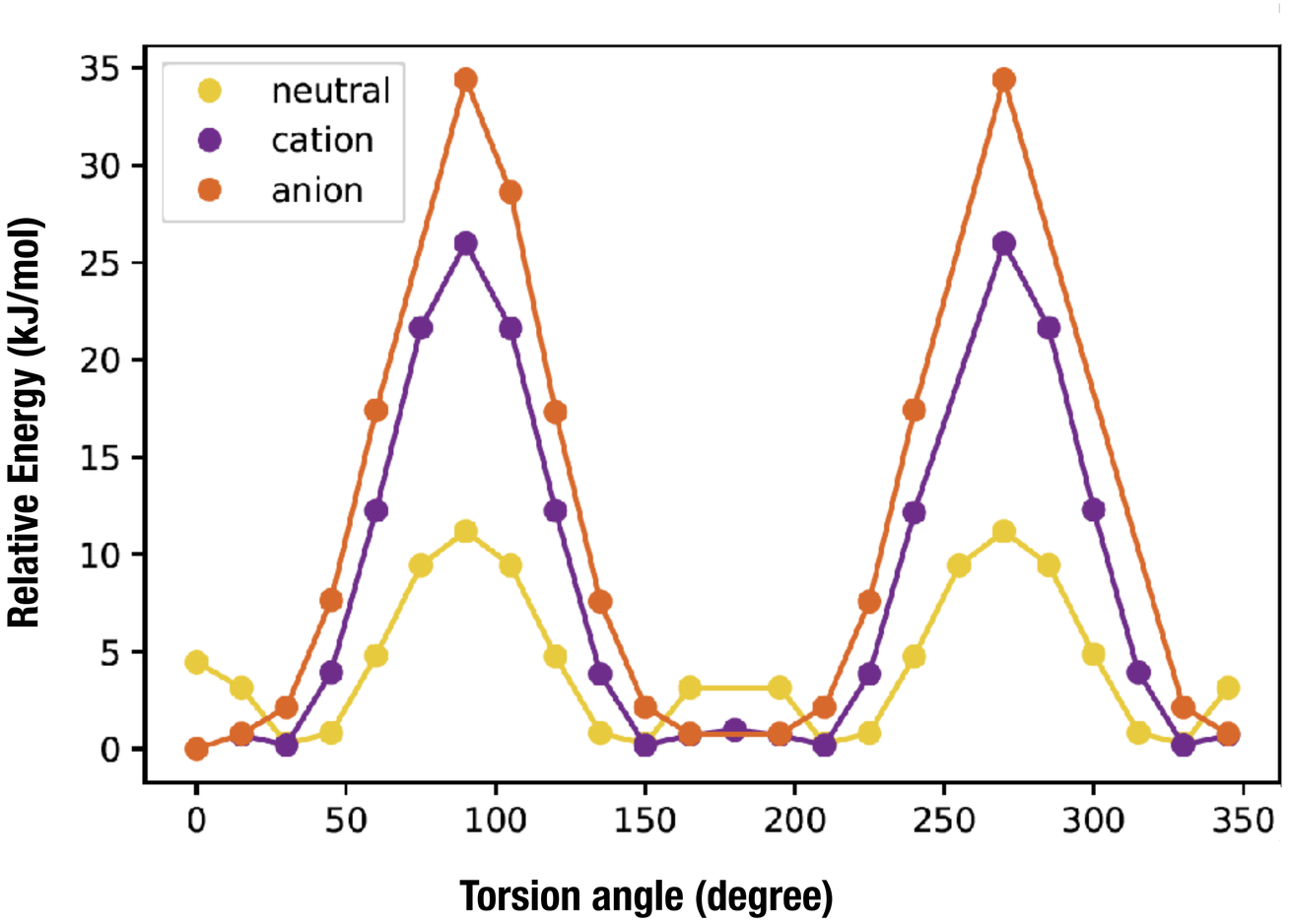


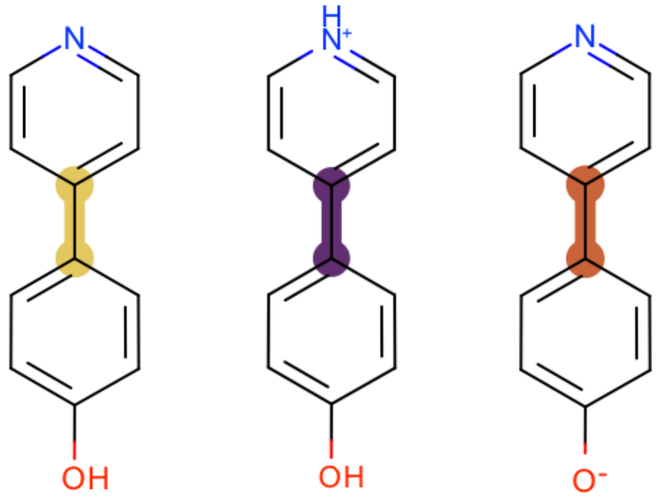
The chemical environment of bonds is sensitive to small remote substituent changes



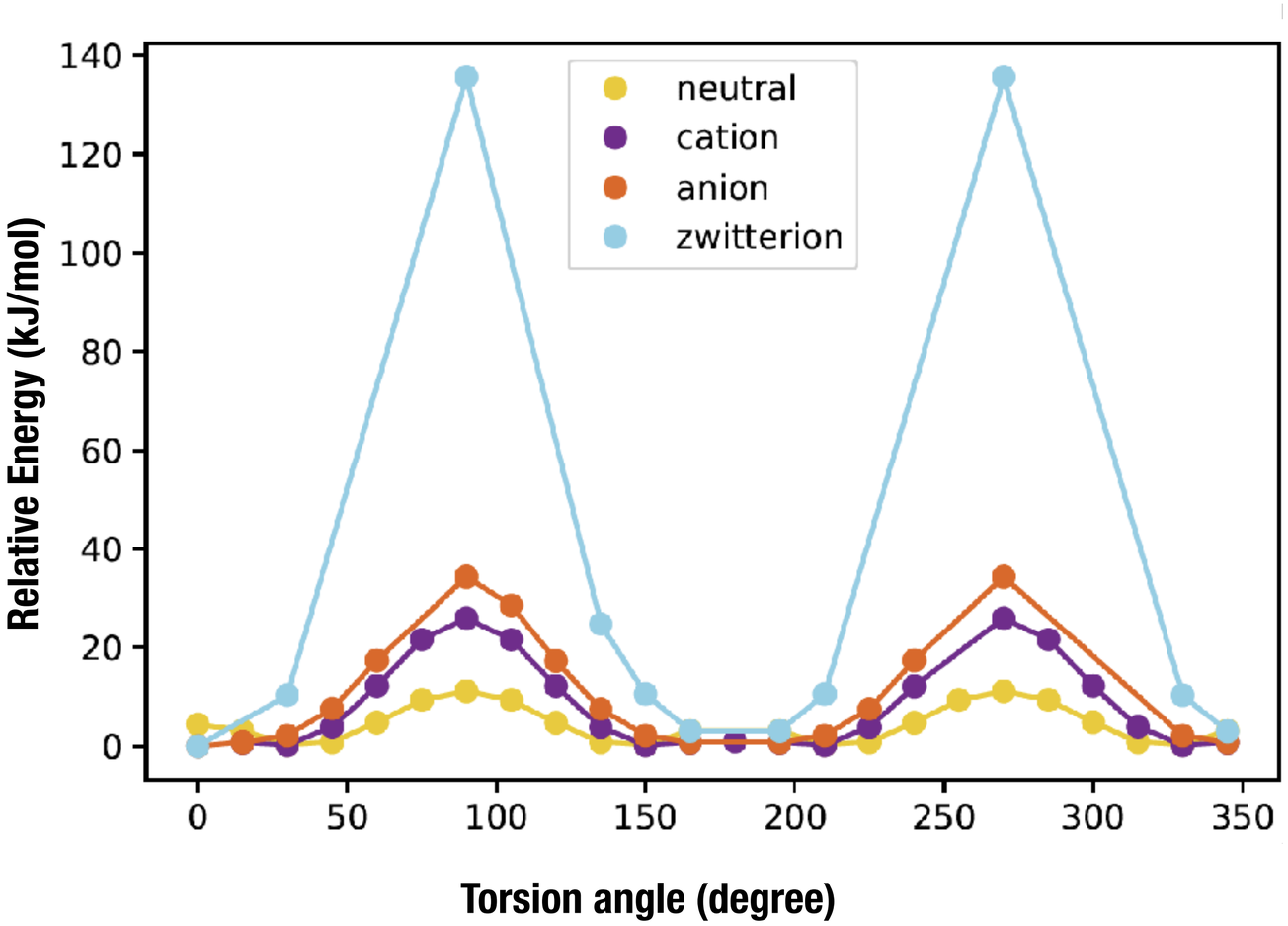


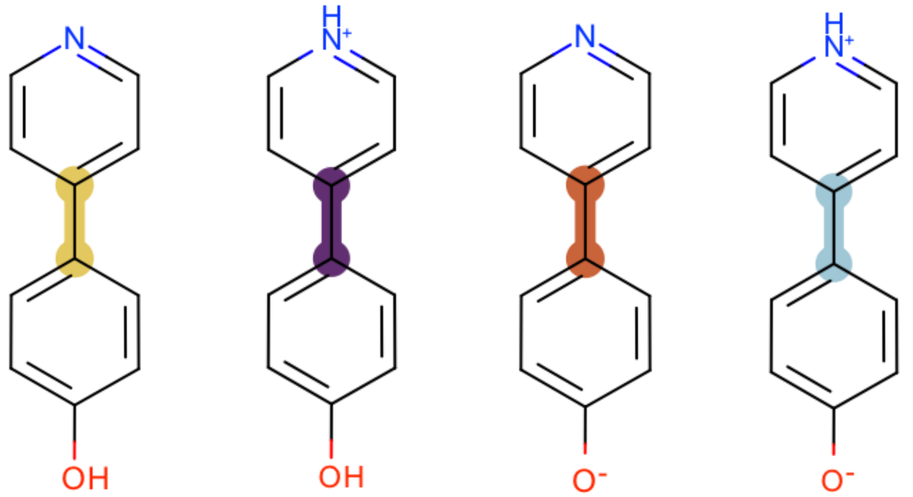
The chemical environment of bonds is sensitive to small remote substituent changes



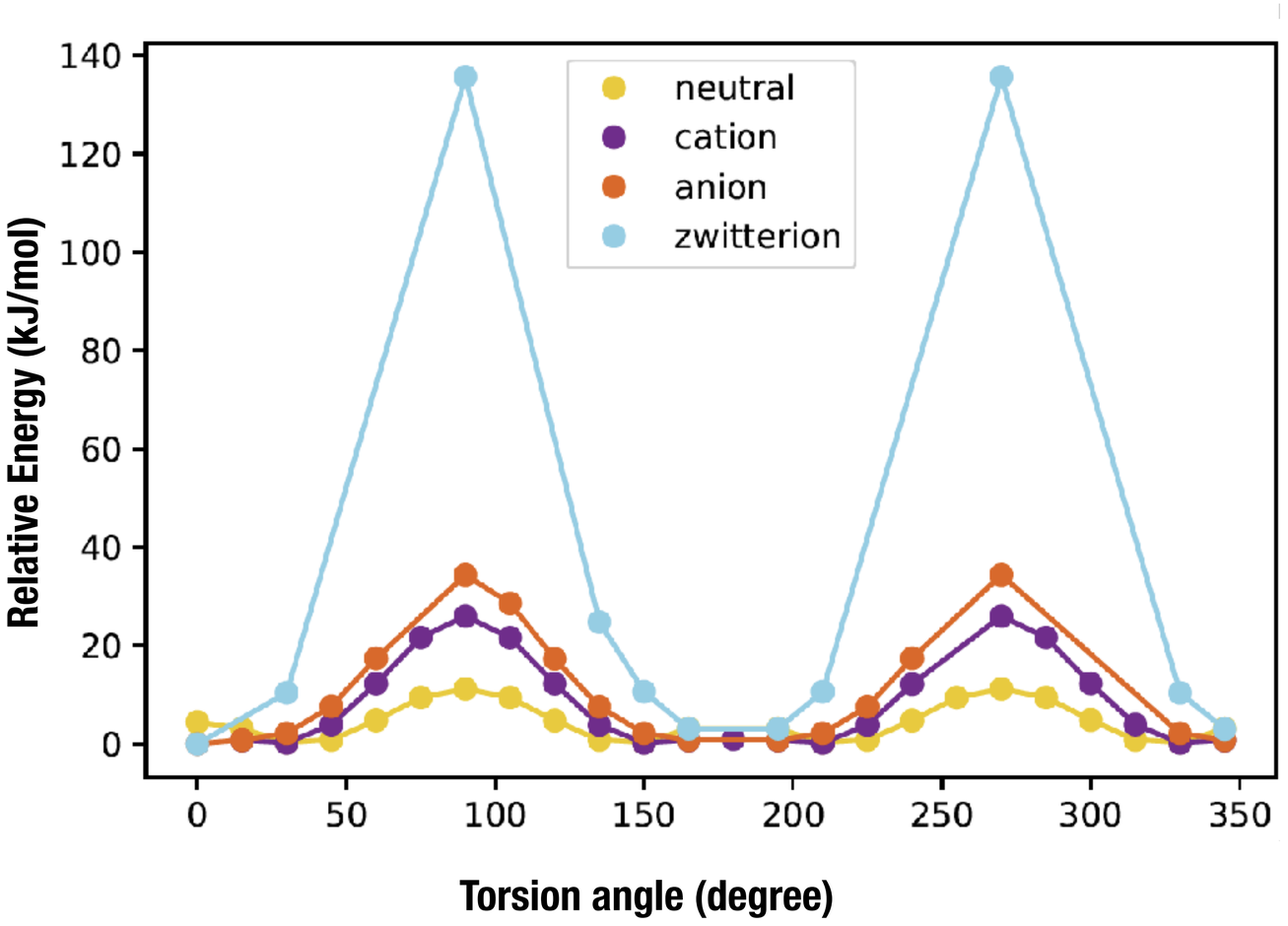


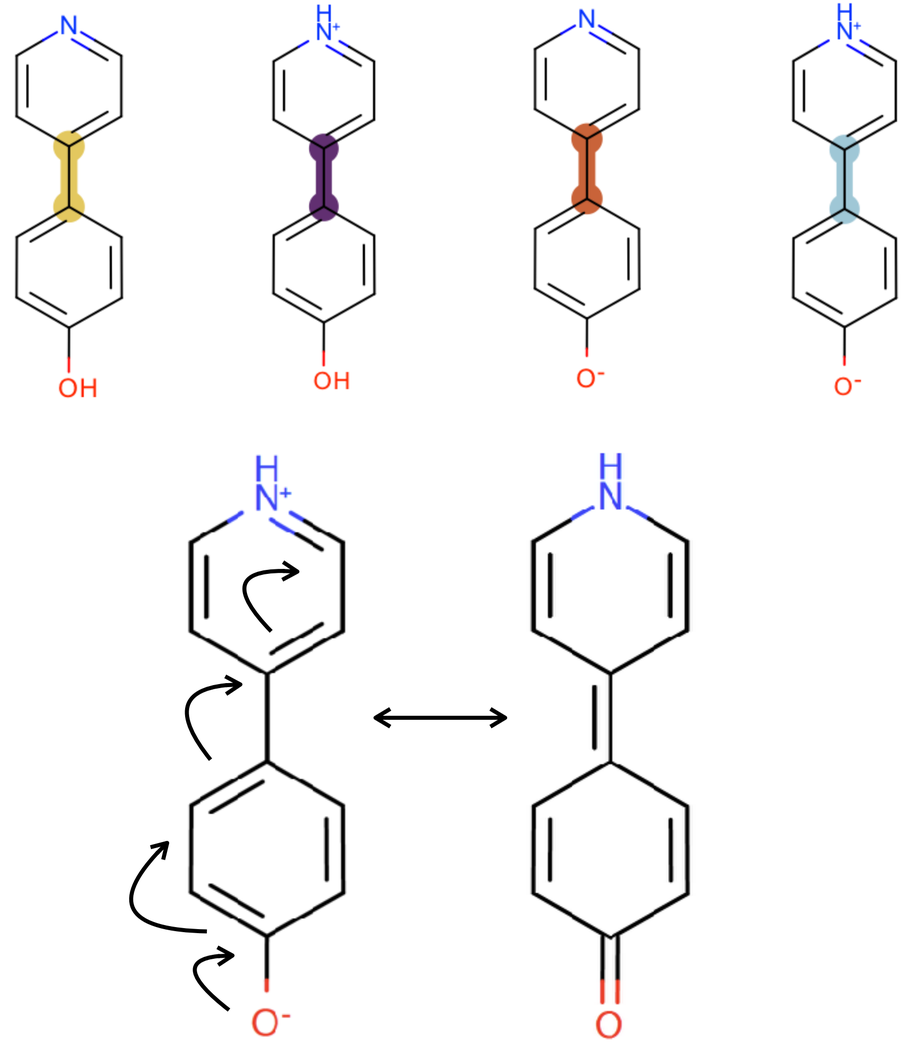
Considering chemical environment is crucial when fragmenting bonds





Chemical environment impacts the torsion profile of small molecules



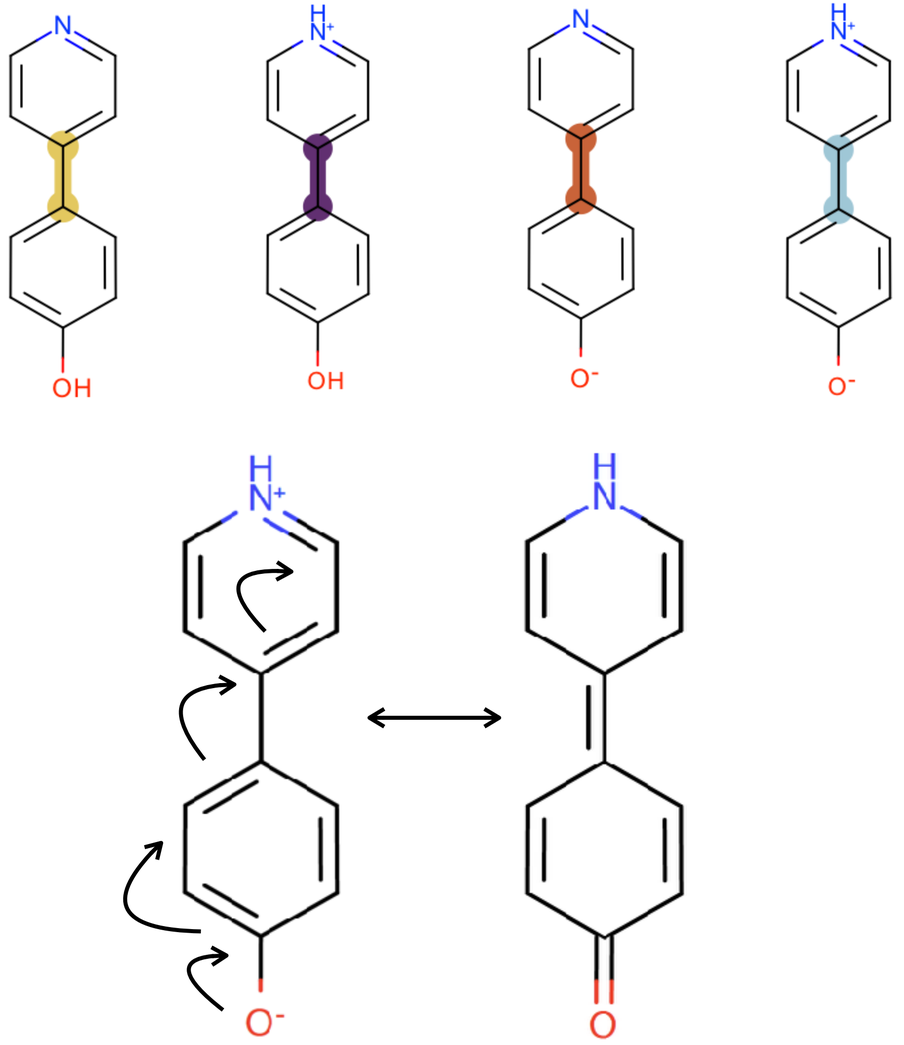


Chemical environment impacts the torsion profile of small molecules

This issue poses several problems when automating molecule fragmentation:

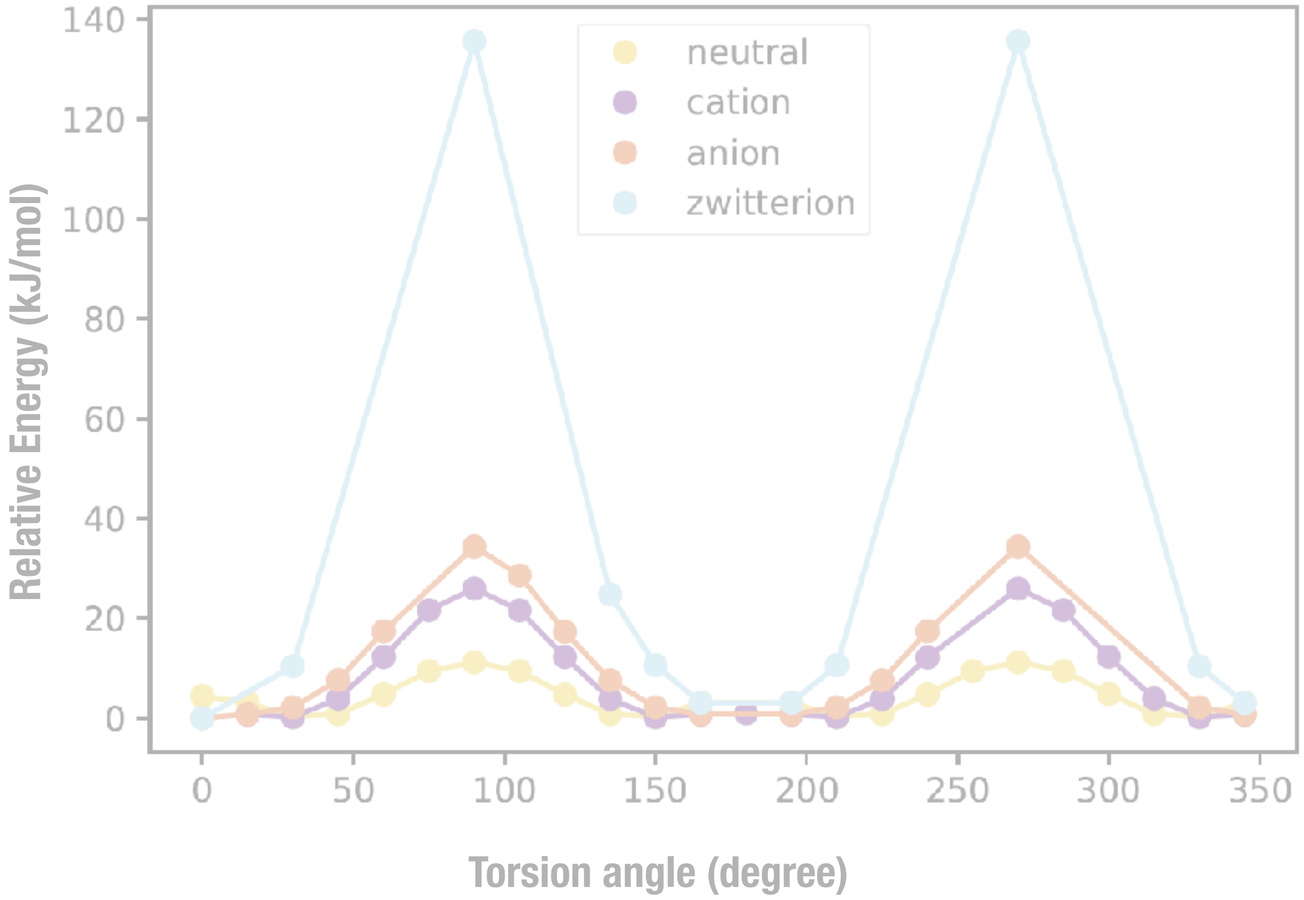
1. Most cheminformatics tools label the central bond as rotatable

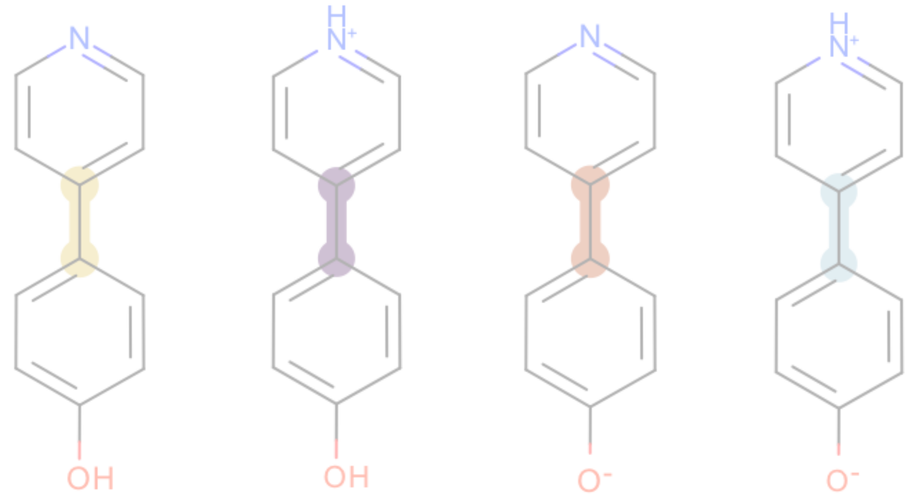
2. How do we ensure that we do not naively destroy the chemical environment by fragmenting a remote substituent?



The AM1 Wiberg Bond Order is a cheap measure of electron population overlap between two atoms

$$W_{AB} = \sum_{\mu \in A} \sum_{\nu \in B} |D_{\mu\nu}|^2.$$

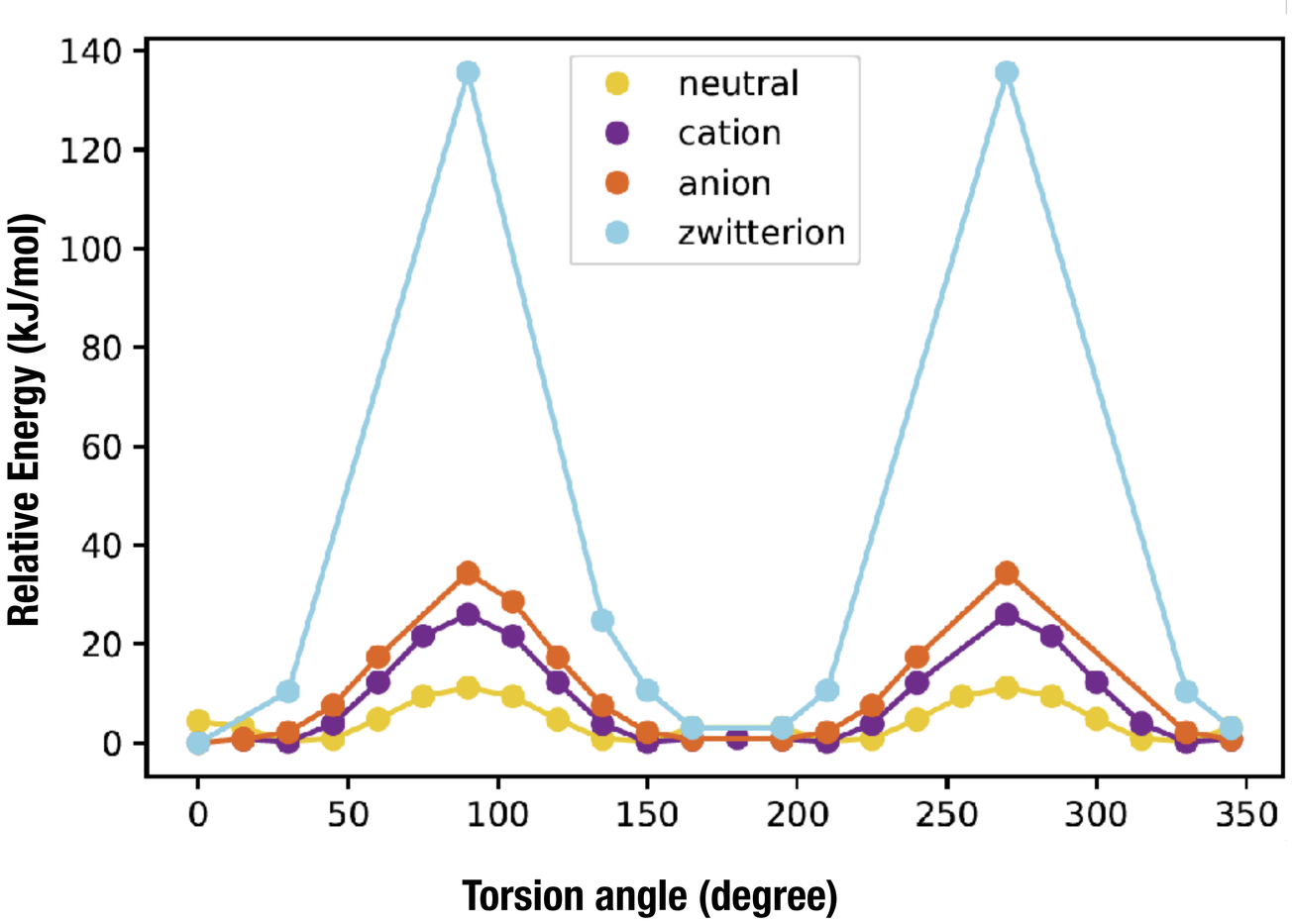


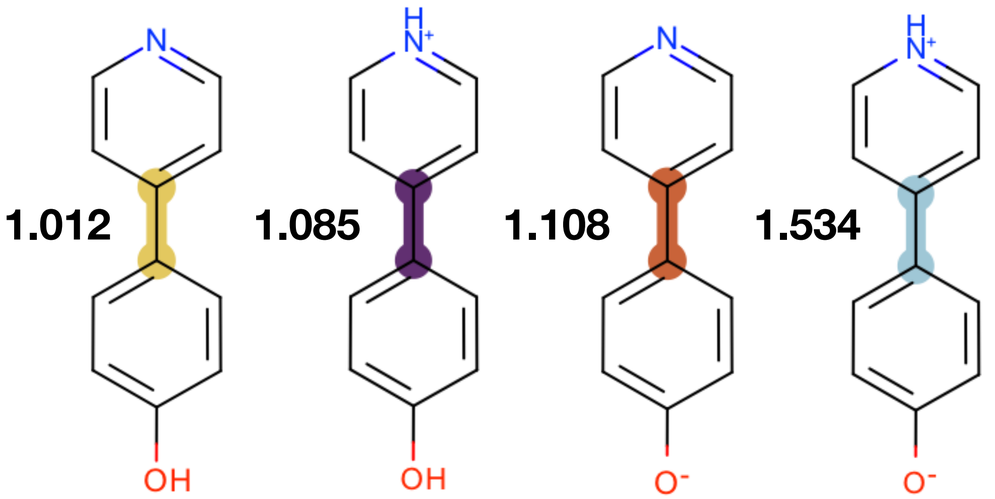


$$D_{\mu\nu} = 2 \sum_{i}^{\text{occ.}} C_{\mu i} C_{\nu i}^{*},$$

The Wiberg Bond Order is a cheap measure of electron population overlap between two atoms

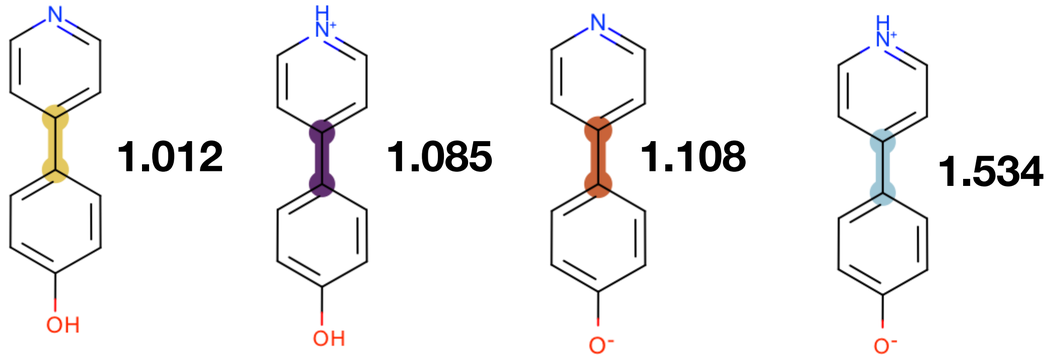
$$W_{AB} = \sum_{\mu \in A} \sum_{\nu \in B} |D_{\mu\nu}|^2.$$

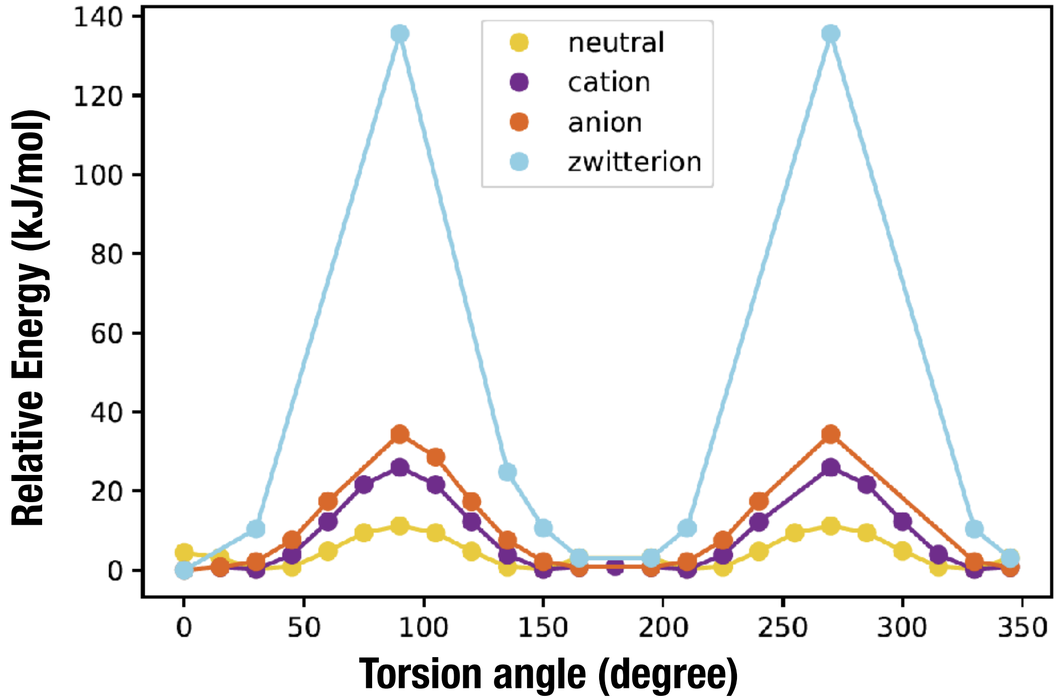


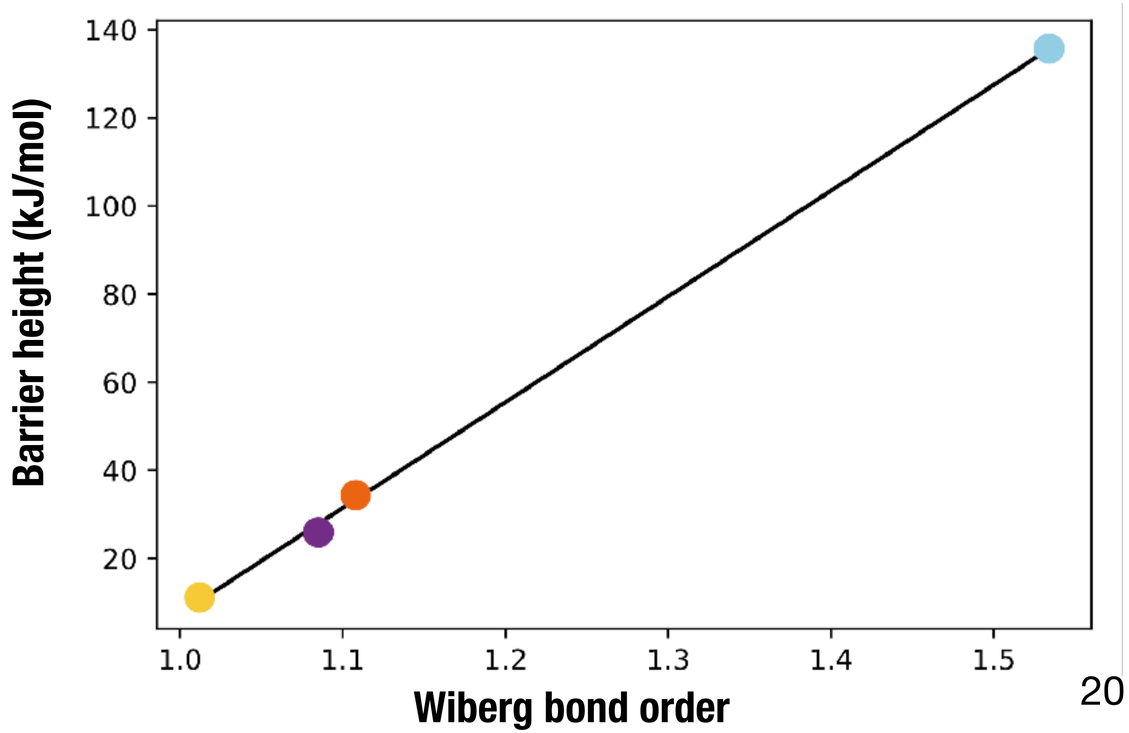


$$D_{\mu\nu} = 2 \sum_{i}^{\text{occ.}} C_{\mu i} C_{\nu i}^{*},$$

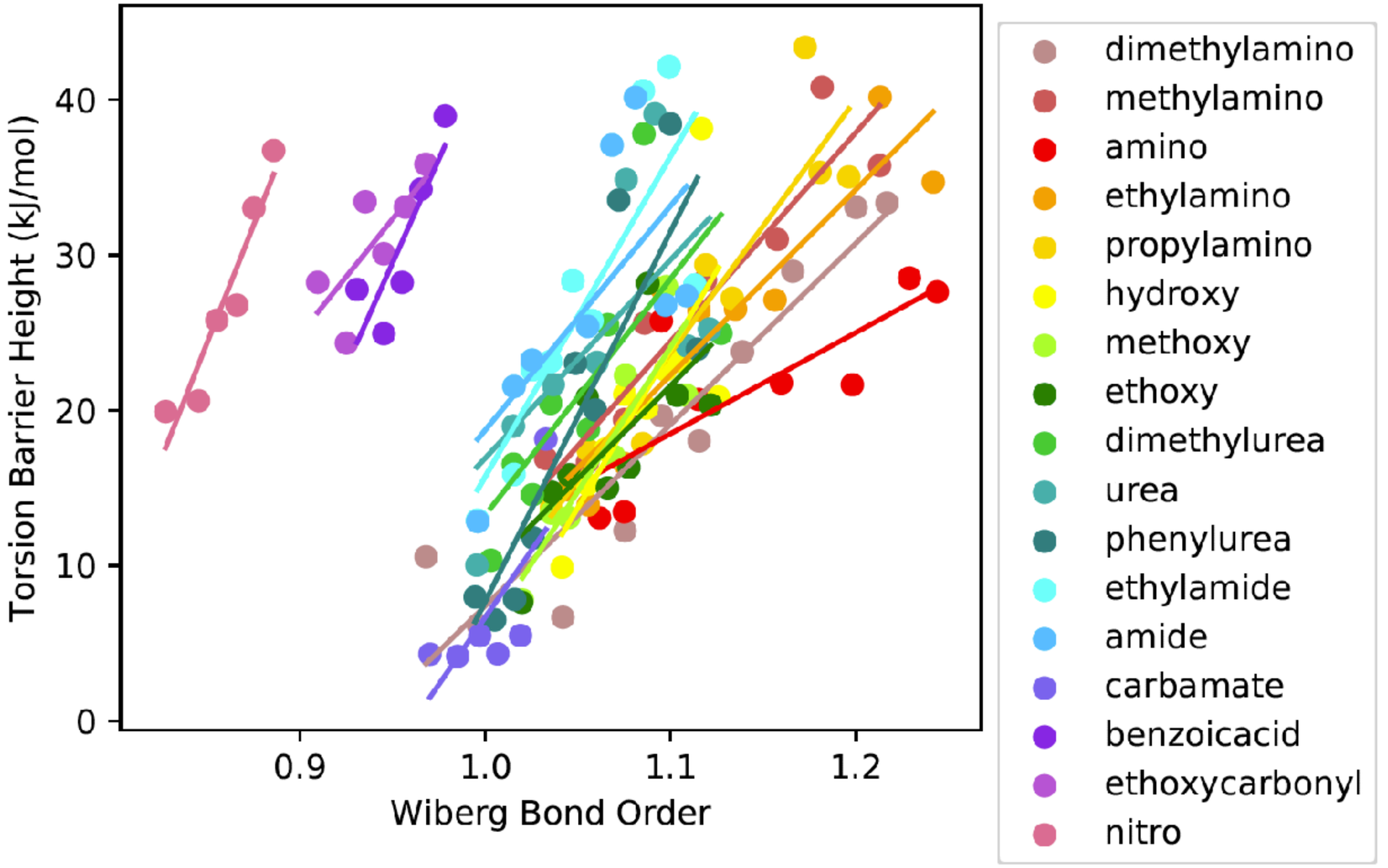
Torsion potential barrier heights are linear with Wiberg Bond Orders







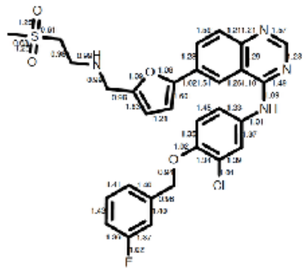
Torsion barrier height increases with increasing Wiberg Bond Order for other functional groups



QC Archive
A MolSSI Project

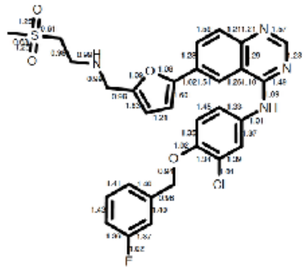
Intelligent fragmentation can reduce the misrepresentation of torsions in QM database

Calculate Wiberg Bond Order from AM1 calculation

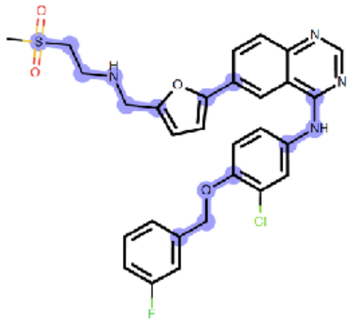


Intelligent fragmentation can reduce the misrepresentation of torsions in QM database

Calculate Wiberg Bond Order from AM1 calculation

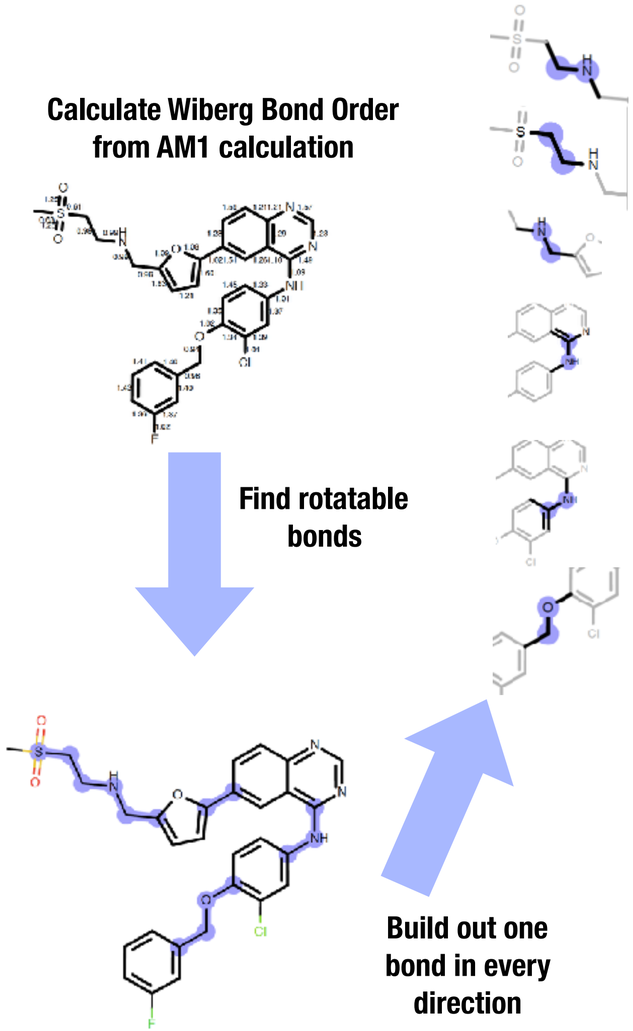


Find rotatable bonds



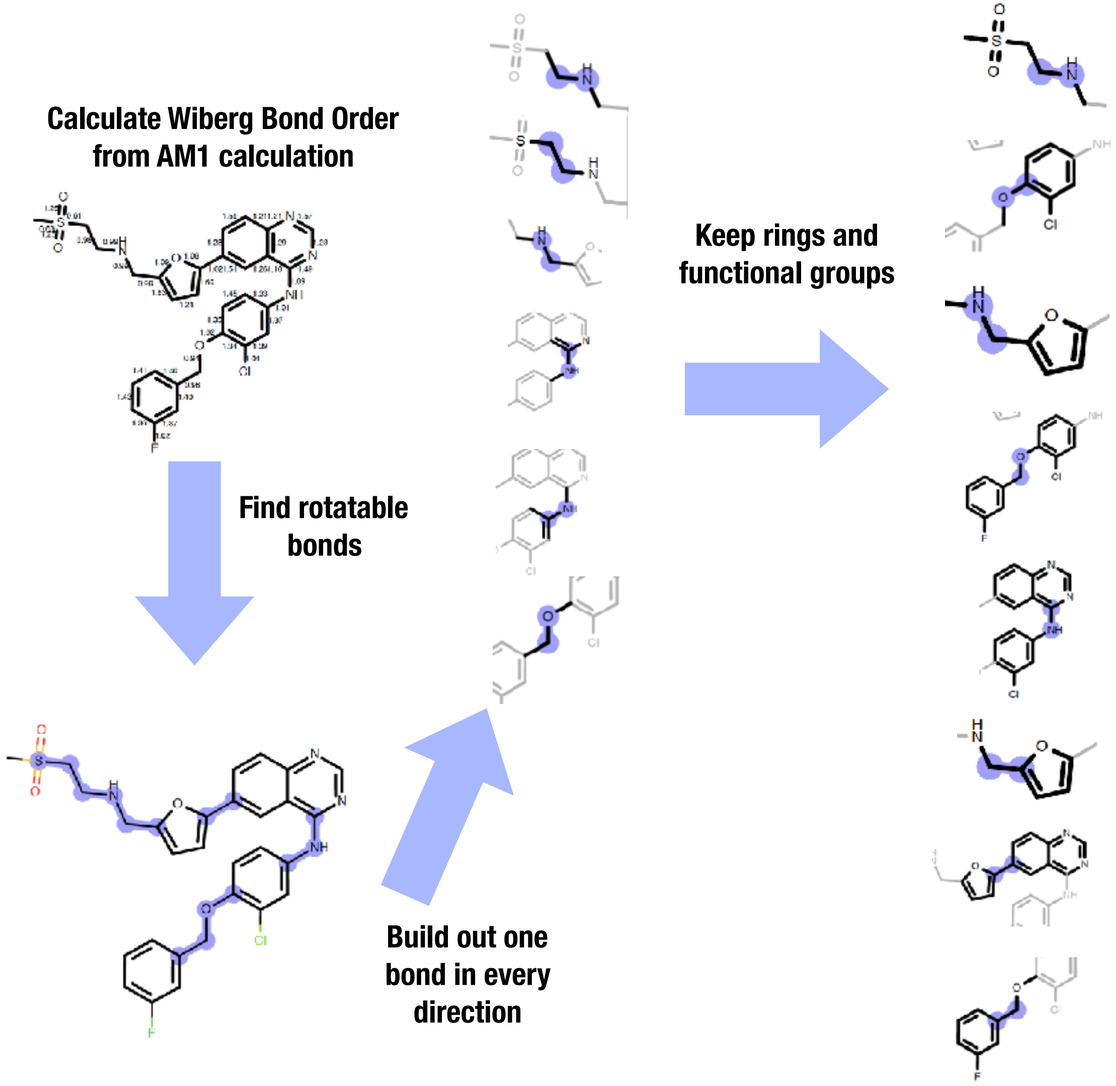
Xinjun Hou
Vishnu Sresth

Intelligent fragmentation can reduce the misrepresentation of torsions in QM database



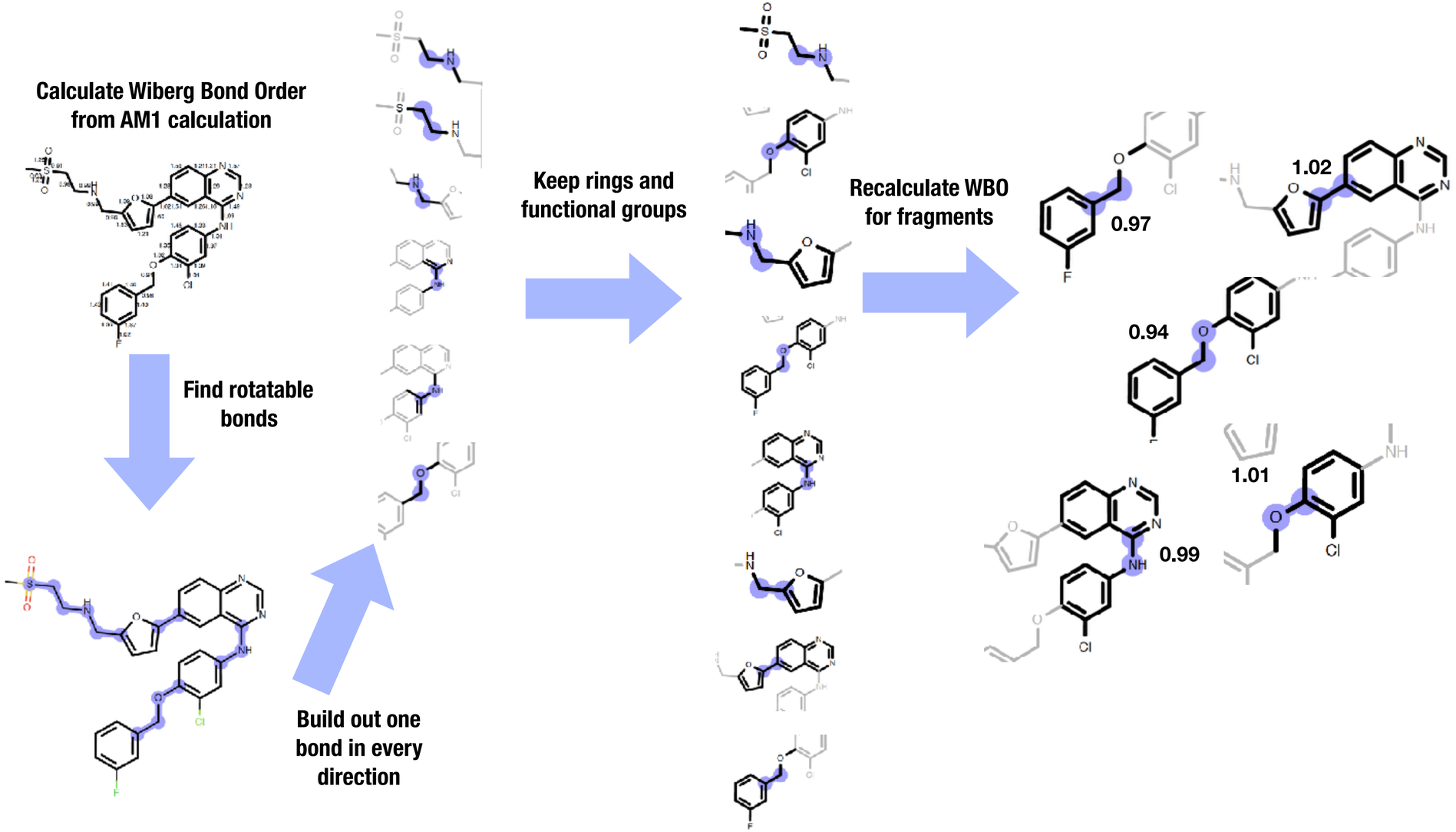
Xinjun Hou
Vishnu Sresth

Intelligent fragmentation can reduce the misrepresentation of torsions in QM database



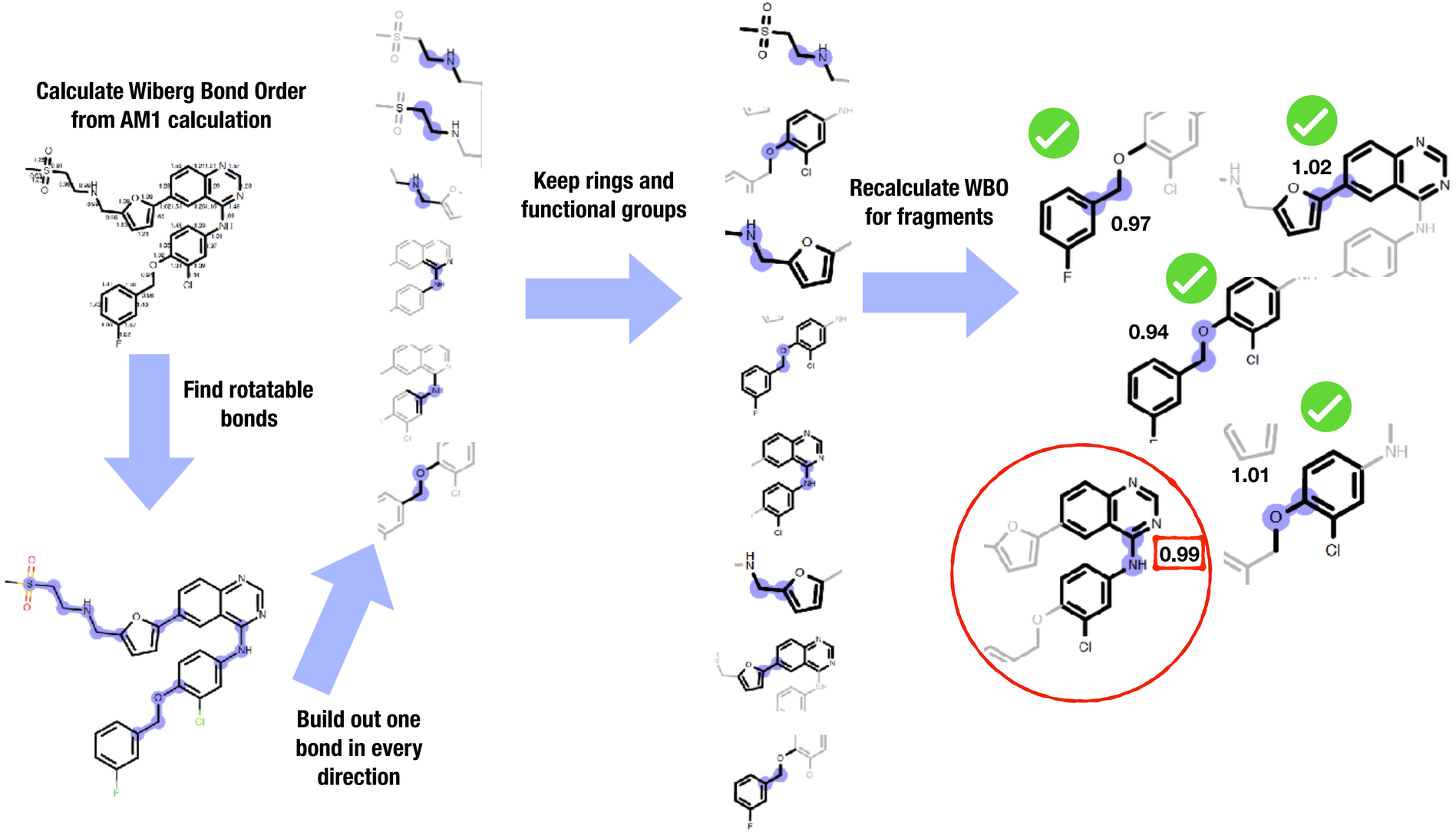
Xinjun Hou
Vishnu Sresth

Intelligent fragmentation can reduce the misrepresentation of torsions in QM database



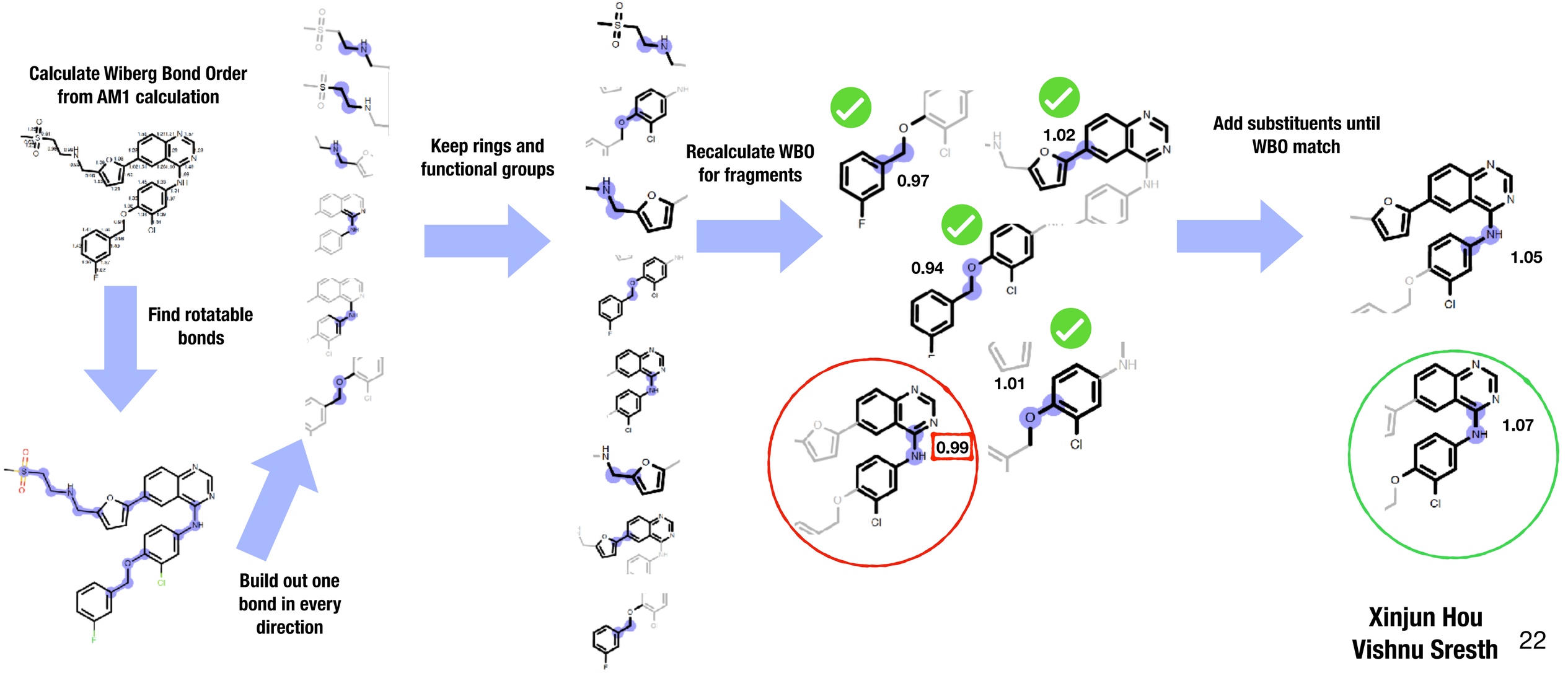
Xinjun Hou
Vishnu Sresth

Intelligent fragmentation can reduce the misrepresentation of torsions in QM database



Xinjun Hou
Vishnu Sresth

Intelligent fragmentation can reduce the misrepresentation of torsions in QM database



Xinjun Hou
Vishnu Sresth

OFF is pursuing several routes to integrate open source AM1 WBO calculations.

1. **AmberTools SQM**
   Dustin Tracy from the Roitberg lab added WBO to SQM.
   Needs testing and integration into an AmberTool conda package so it can be integrated with OFFToolKit

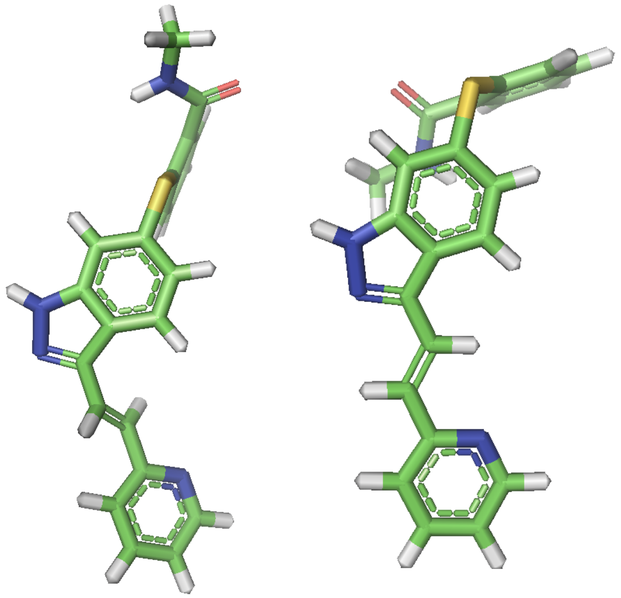
2. **Graph Nets trained on AM1 WBO**
   Can potentially be a lot cheaper than AM1

Yuanqing Wang

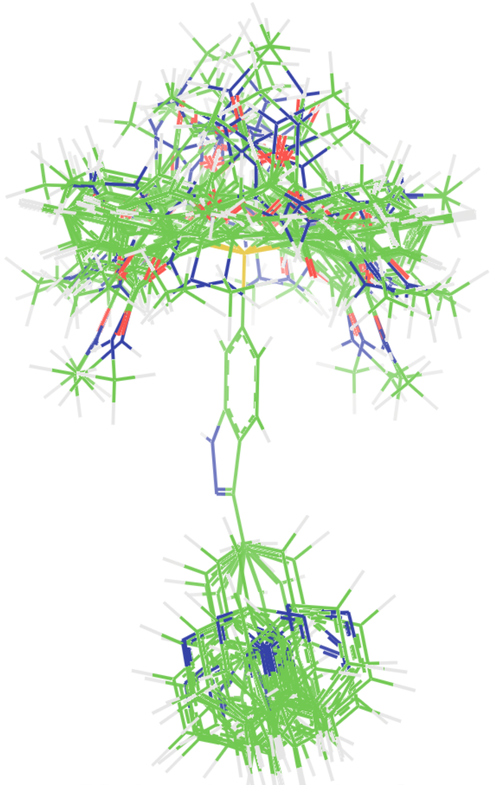
Adrian Roitberg

Dustin Tracy

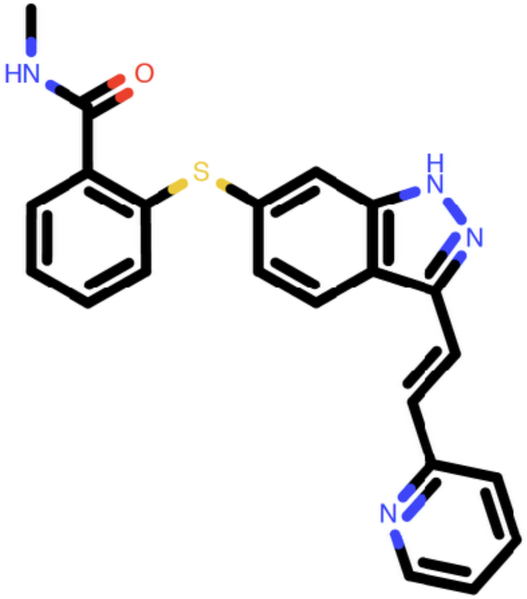
Indexing molecules for QC database



Quantum chemistry represents molecules by their coordinates.

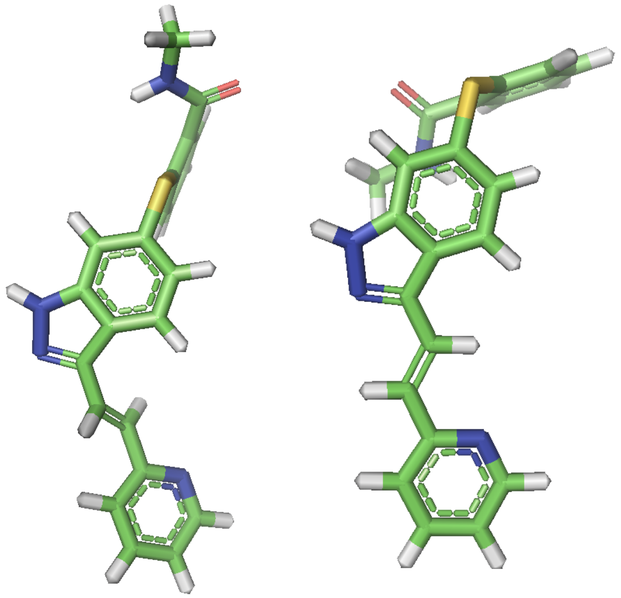


Molecular mechanics represents molecules as conformational distributions

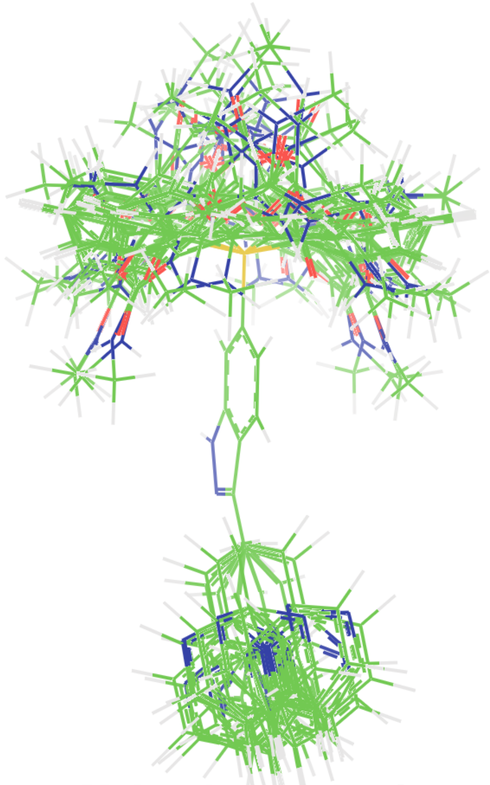


Cheminformatics represent molecules as graphs

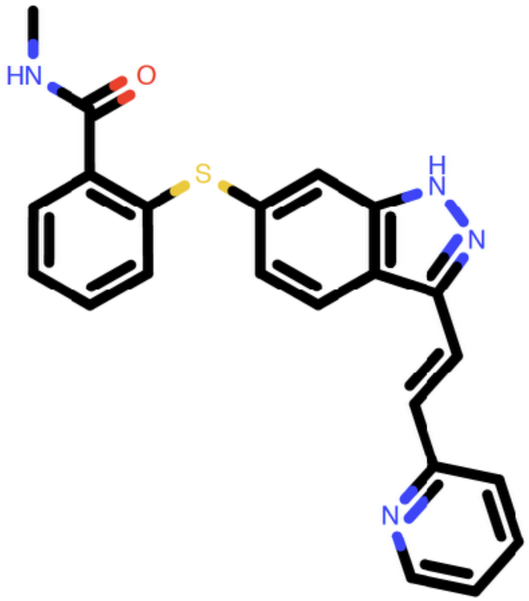
CMILES indices link different representation of molecules



Quantum chemistry represents molecules by their coordinates.



Molecular mechanics represents molecules as conformational distributions



Cheminformatics represent molecules as graphs

SMILES and InChI are not attached to coordinates so calculations with different geometries can be grouped together

SMILES: Simplified Molecular Input Line Entry Specification
InChI: The IUPAC International Chemical Identifier

cmiles provides indices that ensure broad usability and sustainability of the database

cmiles

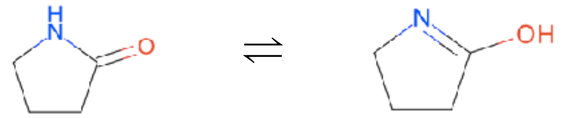
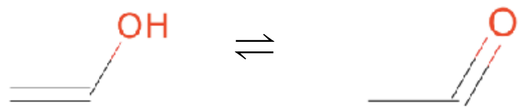
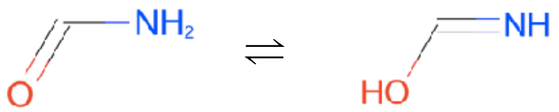
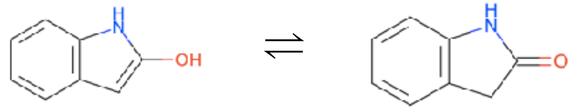
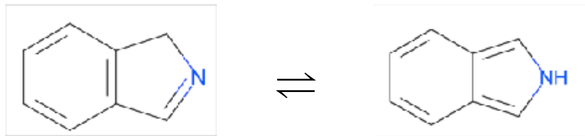
build passing | codecov 92% | docs passing

SMILES must be canonical to avoid redundancy and search failures
Canonical SMILES are only canonical with respect to toolkit and toolkit version

cmiles is set up to test if updates in dependent cheminformatics toolkits changed the canonicalization algorithm.

~ 0.5 % of SMILES tested changed with RDKit updates - mostly stereochemistry

Standardizing molecules for QCArchive

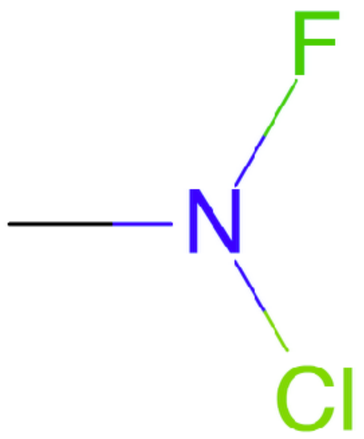
	Tautomer	InChI	RDKit	OpenEye
lactam - lactim		✓	✓	✓
keto - enol		✗	✓	✓
amide - imidic acid		✓	✗	✓
indole-2-ol - indole-2-one		✗	✓	✗
2H-isoindole - 1H-isoindole		✗	✗	✗

Sayle, R. A. (2010). So you think you understand tautomerism? JCAMD 24(6–7), 485–496. http://doi.org/10.1007/s10822-010-9329-5

OFF Toolkit requires fully defined stereochemistry to avoid incorrect partial charges and/or partial bond orders

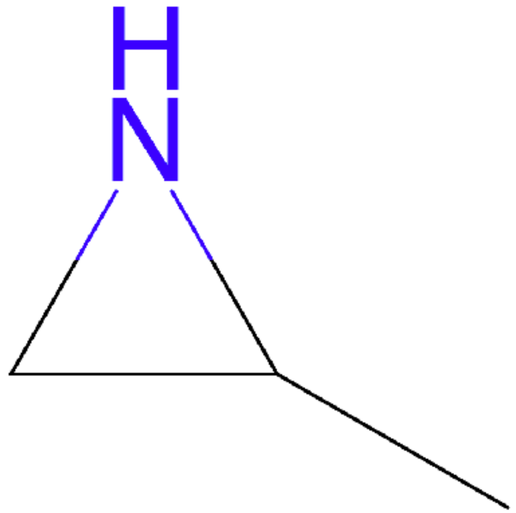
Why is this an issue?
Stereochemical differences can lead to different partial charges or partial bond orders and potentially different parameter assignments

pyramidal nitrogen



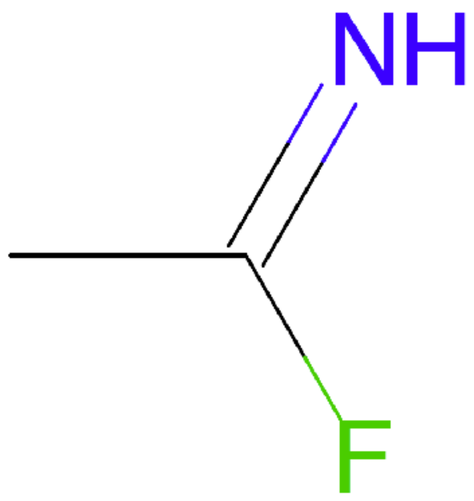
Conversion timescale: 10^3 - 10^5 per sec

N in three remembered ring



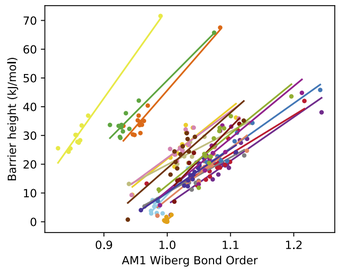
OE does not consider this stereogenic

primary imine



RDKit flags as potential stereogenic if hydrogen is explicit

Summary



1. WBO is a powerful, cheap measure that provides robust signal of a bond's chemical environment and can be used to automate fragmentation and potentially finding rotatable bonds.

QCArchive
A MolSSI Project

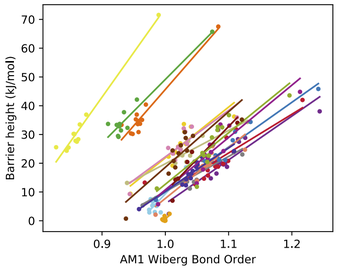
2. Automatic submission to QCArchive and massively parallel computation allows multidimensional torsion scan

3. The QCArchive OFF datasets are indexed with multiple canonical identifiers with their provenance

Future Directions

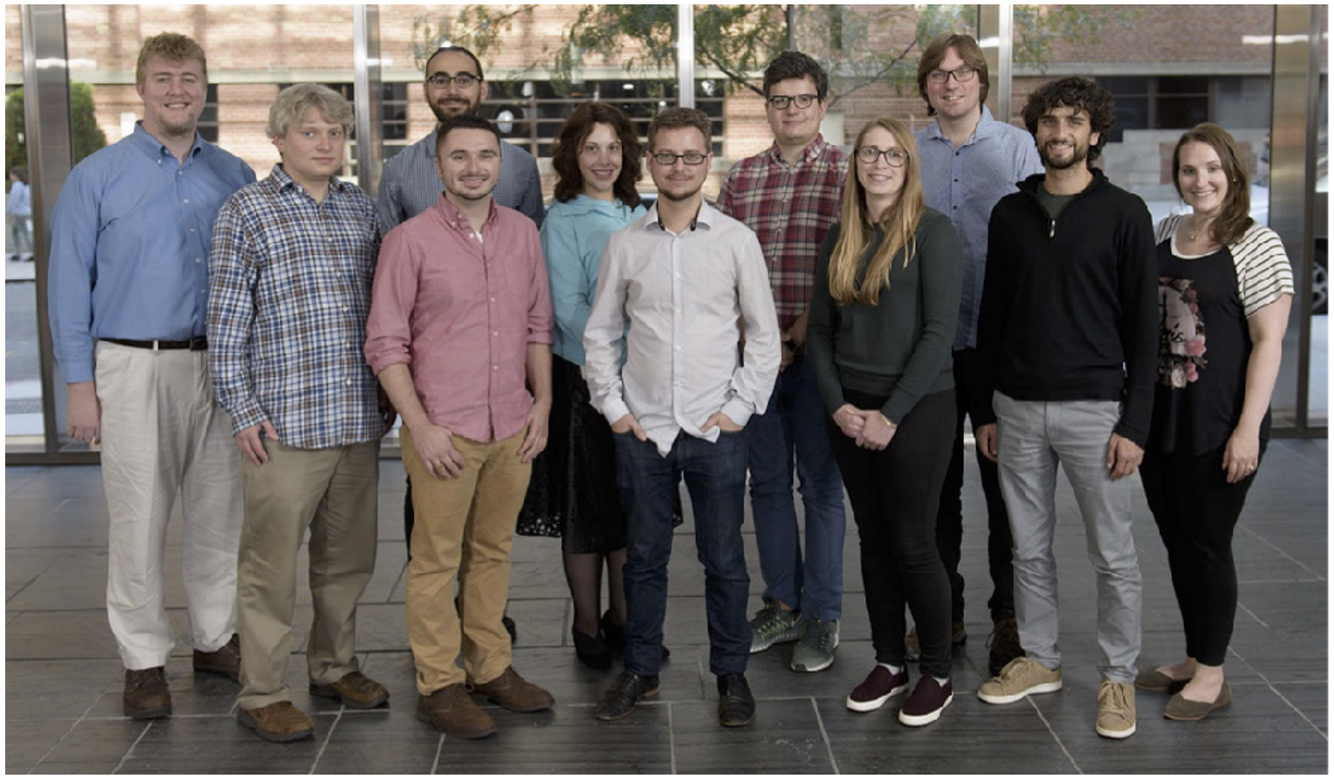
- Integrate AM1 WBO and `fragmenter` with the rest of the OFF software stack

- Optimize fragmentation scheme via learning patterns not to fragment from the validation set



- Generalize the torsion barrier heights vs WBO relationship and interpolate torsion force constants

Acknowledgment



John Chodera
Josh Fass
Mehtap Işık
Andrea Rizzi
Bas Rustenberg
Marcus Wieder
Simon Boothroyd
Rafal Wiewiora
Melissa Boby

OpenEye
Christopher Bayly

#Torsions channel
Alberto Gobbi
Adrian Roitberg
Lee-Ping Wang
Yudong Qiu

David Mobley
Caitlin Bannan

MolSSI
Daniel Smith
Doaa Altarawy
Levi Naden

QC Archive
A MolSSI Project
https://qcarchive.molssi.org/

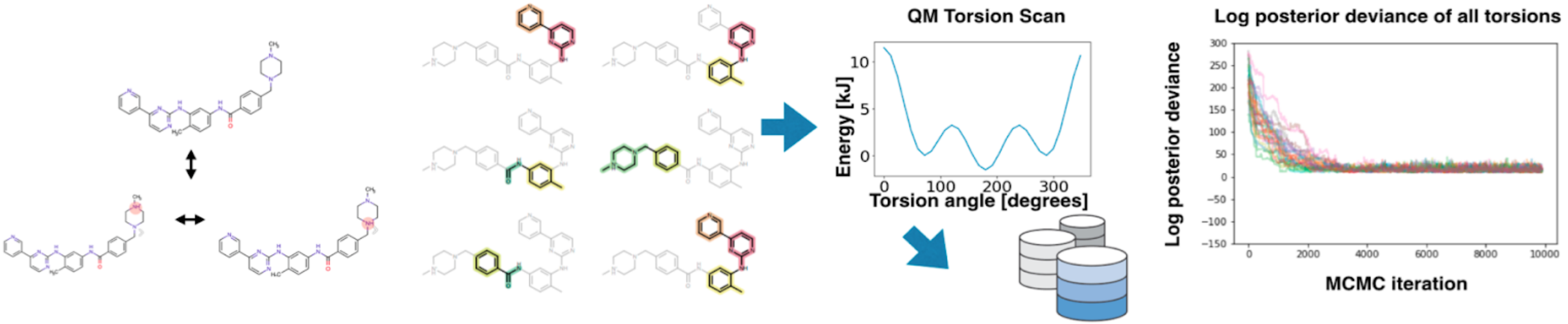
Funding

Tri-Institutional PhD Program
Chemical Biology

NSF

MolSSI

Questions


QM Torsion Scan
Energy [kJ]
10
5
0
0
100
200
300
Torsion angle [degrees]
Log posterior deviance of all torsions
Log posterior deviance
300
250
200
150
100
50
0
−50
−100
−150
0
2000
4000
6000
8000
10000
MCMC iteration