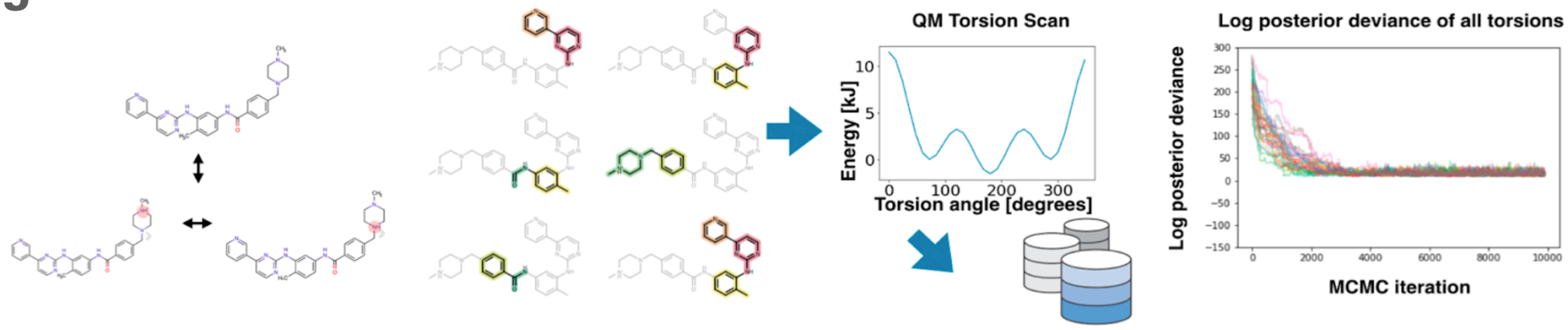
Is the force with us?

Generating chemically relevant data for model fitting

Chaya D. Stern
Chodera lab group meeting
Jun 4, 2019



The Open Forcefield Consortium

INDUSTRY

Boehringer-Ingelheim

Bristol-Myers Squibb

Merck KGaA

Bayer

XtalPi

Roche

Vertex

Qulab

Pfizer

Jeff Wagner

Daniel Smith

COORDINATING INTERMEDIARY

MolSSI

MOLECULAR SOFTWARE SCIENCES INSTITUTE

coordination of funding
while minimizing indirect costs
(7% administrative overhead)

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

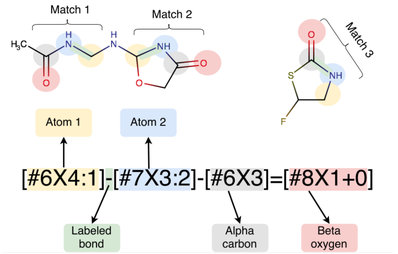
CHRISTOPHER BAYLY
OPENEYE SCIENTIFIC

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

Open Forcefield Initiative objectives

- Develop an open, scalable, extensive toolkit for automatically parameterizing forcefields

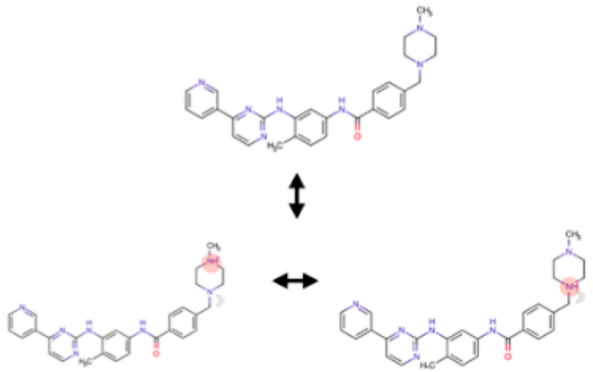
- Generate/curate open datasets necessary for producing high-accuracy small molecule forcefield

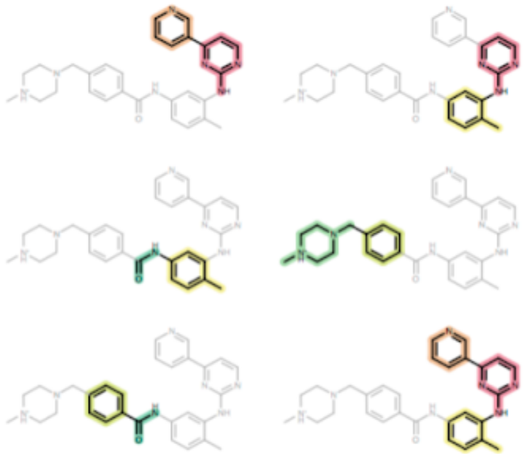


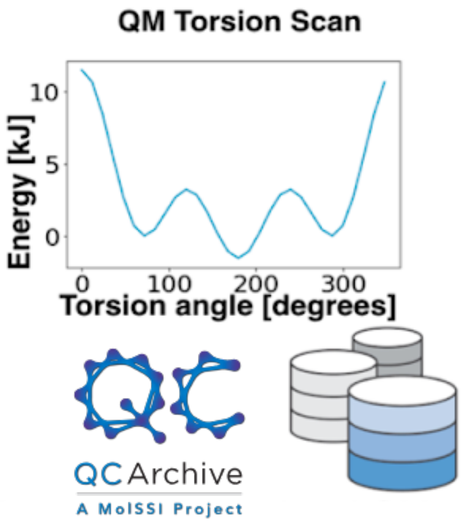
- Generate systematically-improved forcefields

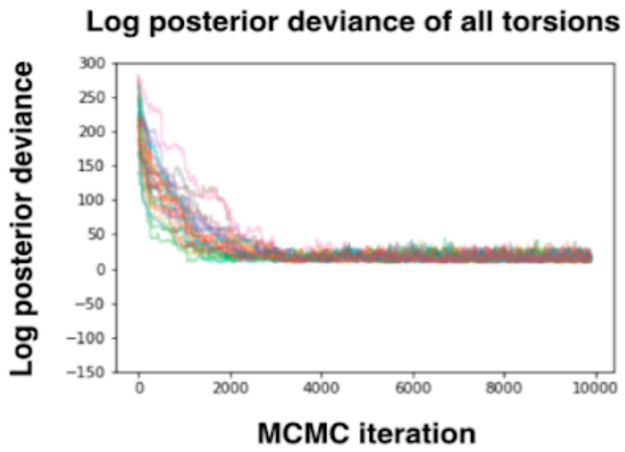
All code, datasets and forcefields: https://github.com/openforcefield

Generating data for torsion parameters



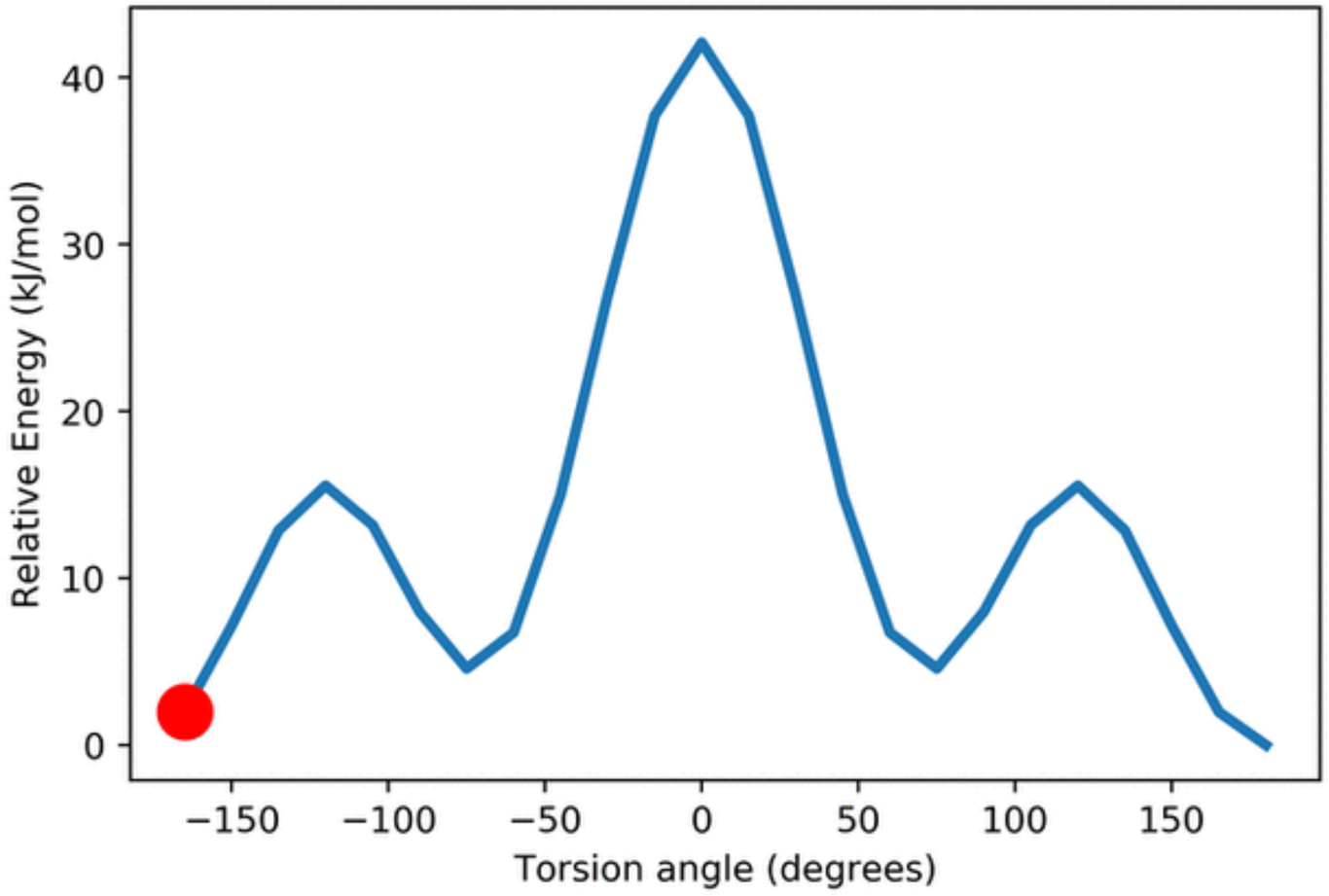


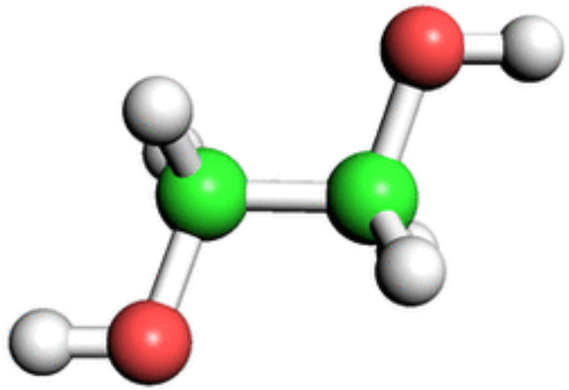




The torsion potential describes the energy of the molecule as it rotates about a central bond

Torsion scans are expensive because of the amount of QM geometry optimizations needed

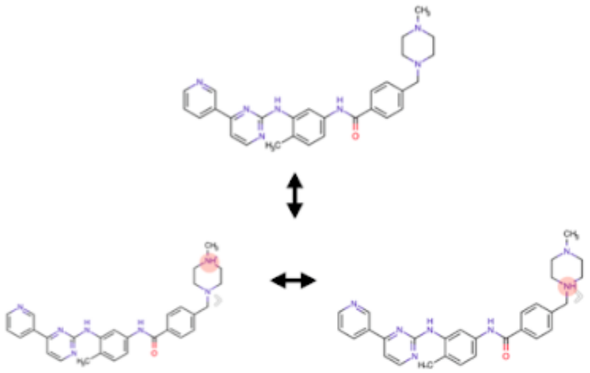




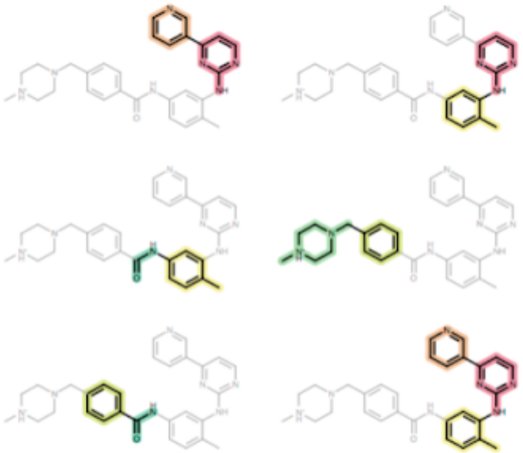
$$\sum_{torsions} k_\chi[1 + cos(n\chi - \delta)]$$

$$\sum_i (V_i^{MM}(k_{torsion} = 0) - V_i^{QM})$$

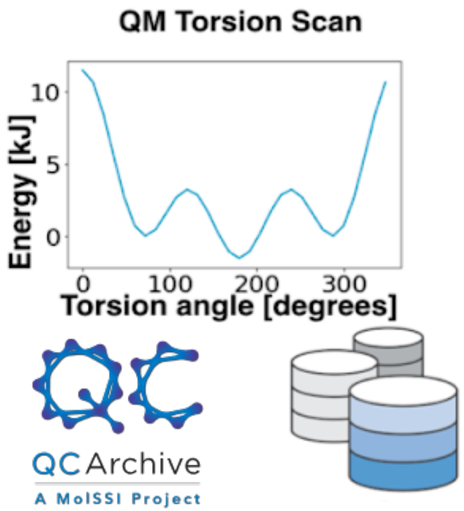
Overview of the torsion fitting pipeline



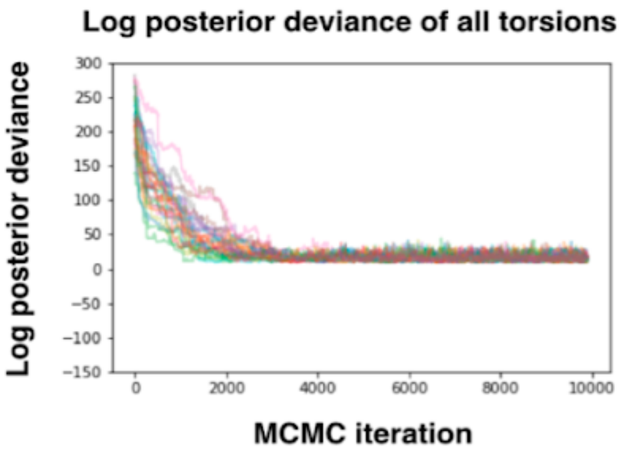
Enumerate ionization states, protonation states and tautomers



Fragment molecules without destroying important chemistry



Run multi-dimensional QM torsion scans on QCArchive and deposit into database



Fit QM torsion profiles using Bayesian inference and MCMC to avoid minima

Related code:

Fragmentation: github.com/openforcefield/fragmenter
Geometry optimization: github.com/leeping/geomeTRIC
Multi dimensional torsion drives: github.com/lpwgroup/torsiondrive
Automated QC parallelization: github.com/MolSSI/QCFractal
QC Database indices: github.com/openforcefield/cmiles
Bayesian torsion fitting: github.com/choderalab/torsionfit

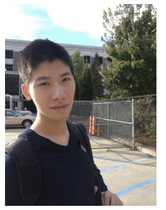
Daniel Smith

Doaa Altarawy

Levi Naden

MolSSI QCArchive Project

Lee-Ping Wang



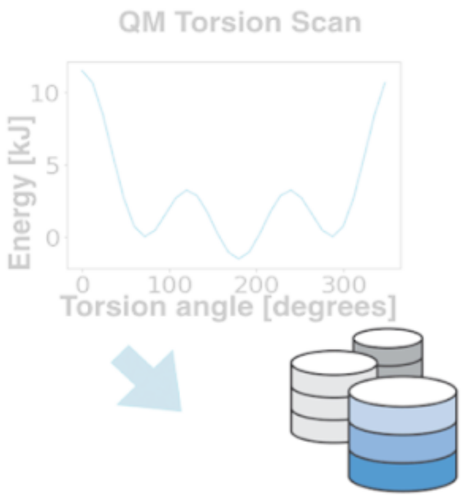
Yudong Qiu

geometric, torsiondrive

Indexing molecules for quantum chemistry database

Enumerate ionization states, protonation states and tautomers

Fragment molecules without destroying important chemistry



Run multi-dimensional QM torsion scans on QCArchive and deposit into database

Fit QM torsion profiles using Bayesian inference and MCMC to avoid minima

Related code:

Fragmentation: github.com/openforcefield/fragmenter

Geometry optimization: github.com/leeping/geomeTRIC

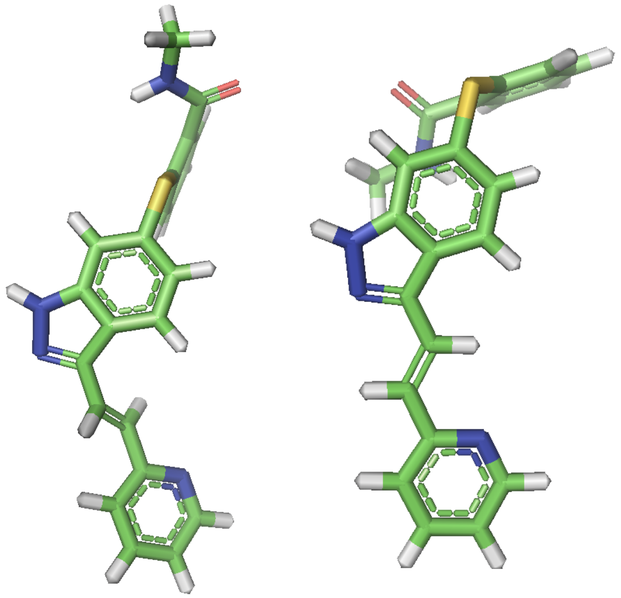
Multi dimensional torsion drives: github.com/lpwgroup/torsiondrive

Automated QC parallelization: github.com/MolSSI/QCFractal

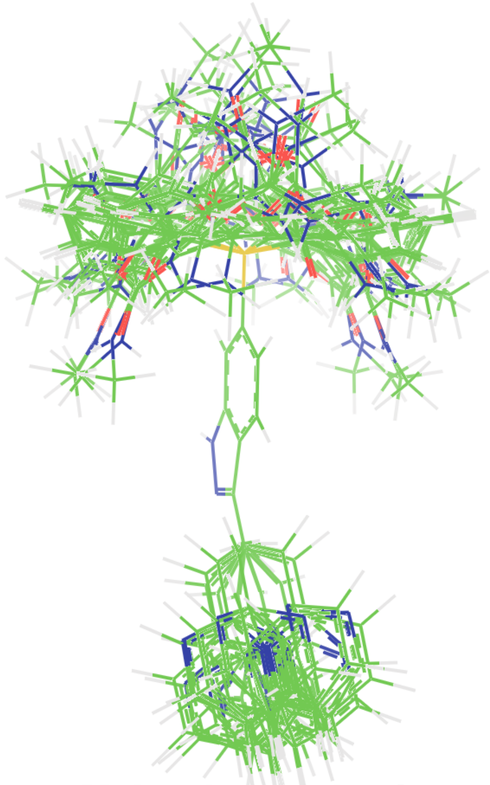
QC Database indices: github.com/openforcefield/cmiles

Bayesian torsion fitting: github.com/choderalab/torsionfit

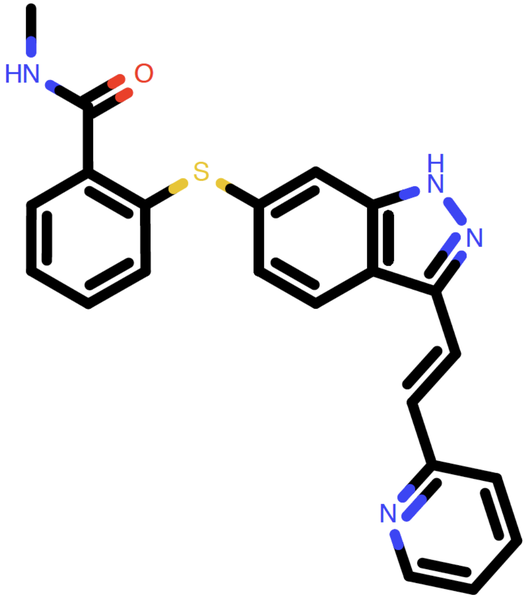
Different communities have conflicting representations of molecules



Quantum chemistry represents molecules by their coordinates.

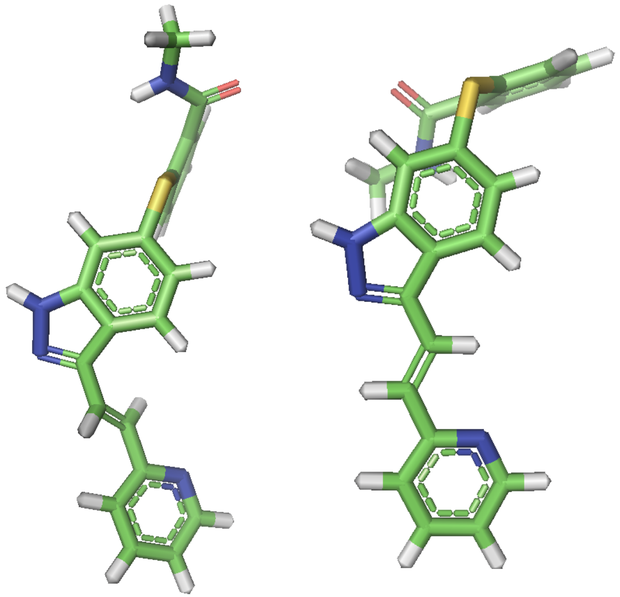


Molecular mechanics represents molecules as conformational distributions

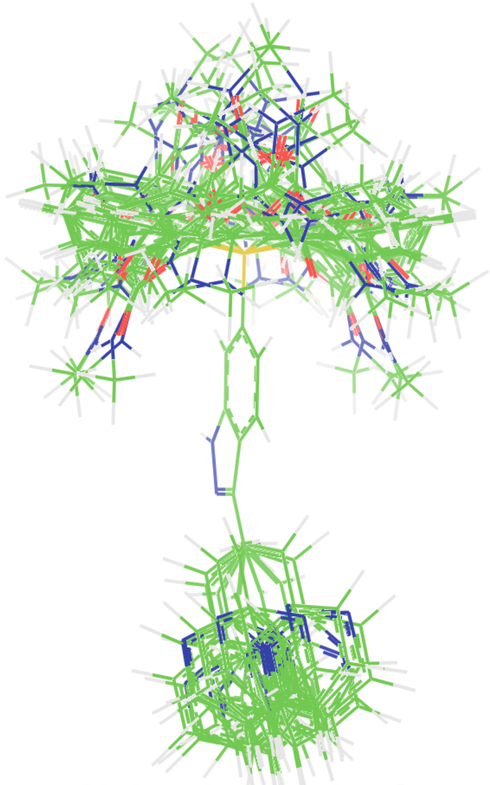


Cheminformatics represent molecules as graphs

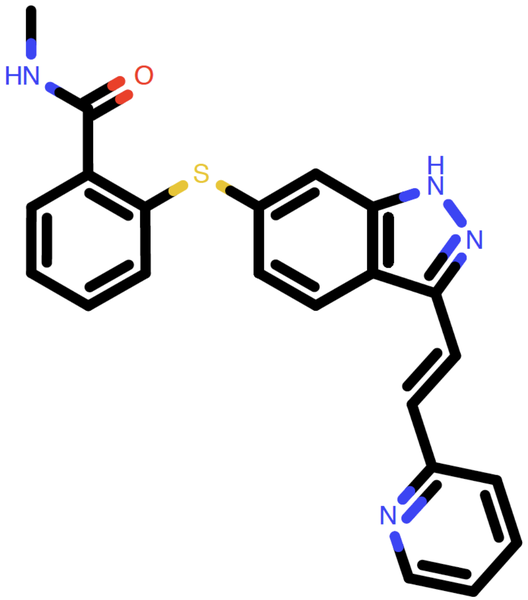
CMILES indices link different representation of molecules



Quantum chemistry represents molecules by their coordinates.



Molecular mechanics represents molecules as conformational distributions



Cheminformatics represent molecules as graphs

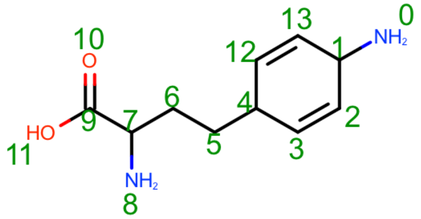
SMILES and InChI are not attached to coordinates so calculations with different geometries can be grouped together

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

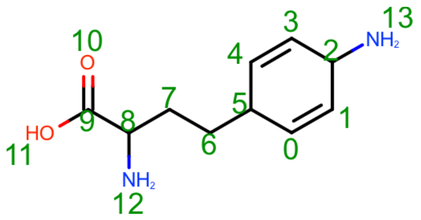
cmiles provides indices that ensure broad usability and sustainability of the database

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

`cmiles` will be distributed as a docker container with pinned toolkit versions



Nodes indices in a graph are arbitrary. This may cause loss of information



SMILES with tags provides a way to recover index order.

```
[H:10][c:1]1[c:2]([c:5]([c:4]([n:7][c:3]1[H:12])[H:13])[C:6](=[O:8])[O:9][H:14])[H:11]
```

https://github.com/openforcefield/cmiles

The Open Forcfield consortium prioritizes software sustainability

cmiles

build passing | codecov 92% | docs passing

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

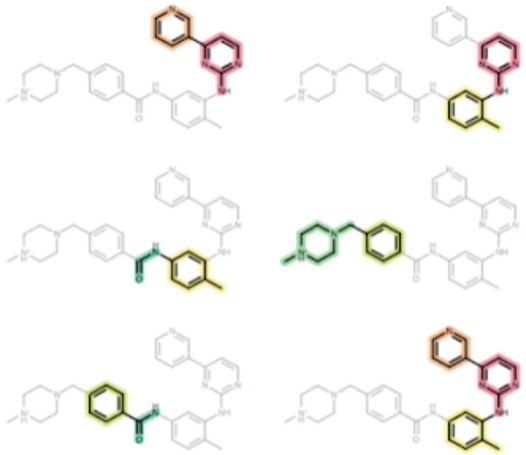
Open Forcfield software scientist ensures smooth handoff of mission critical software.

Jeff Wagner

https://github.com/MolSSI/cookiecutter-cms

Fragmenting molecules for QC torsion scans

Enumerate ionization states, protonation states and tautomers



Fragment molecules without destroying important chemistry

Run multi-dimensional QM torsion scans on QCArchive and deposit into database

Fit QM torsion profiles using Bayesian inference and MCMC to avoid minima

Related code:

Fragmentation: github.com/openforcefield/fragmenter

Geometry optimization: github.com/leeping/geomeTRIC

Multi dimensional torsion drives: github.com/lpwgroup/torsiondrive

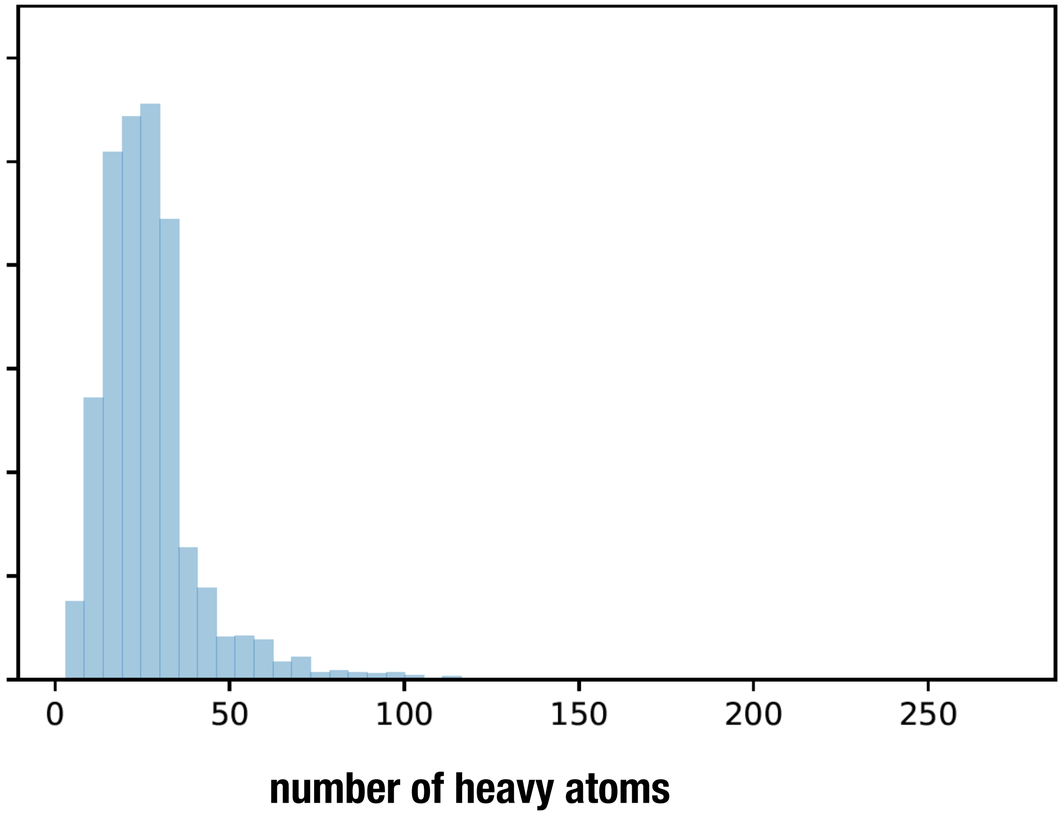
Automated QC parallelization: github.com/MolSSI/QCFractal

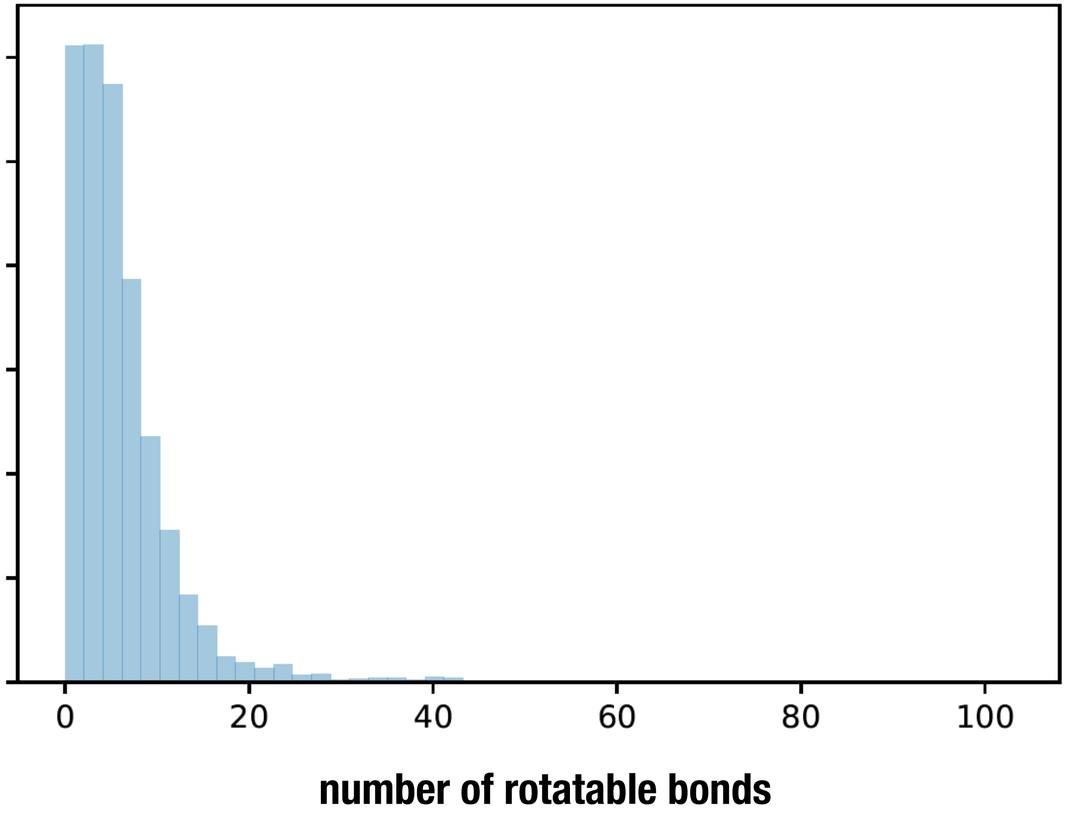
QC Database indices: github.com/openforcefield/cmiles

Bayesian torsion fitting: github.com/choderalab/torsionfit

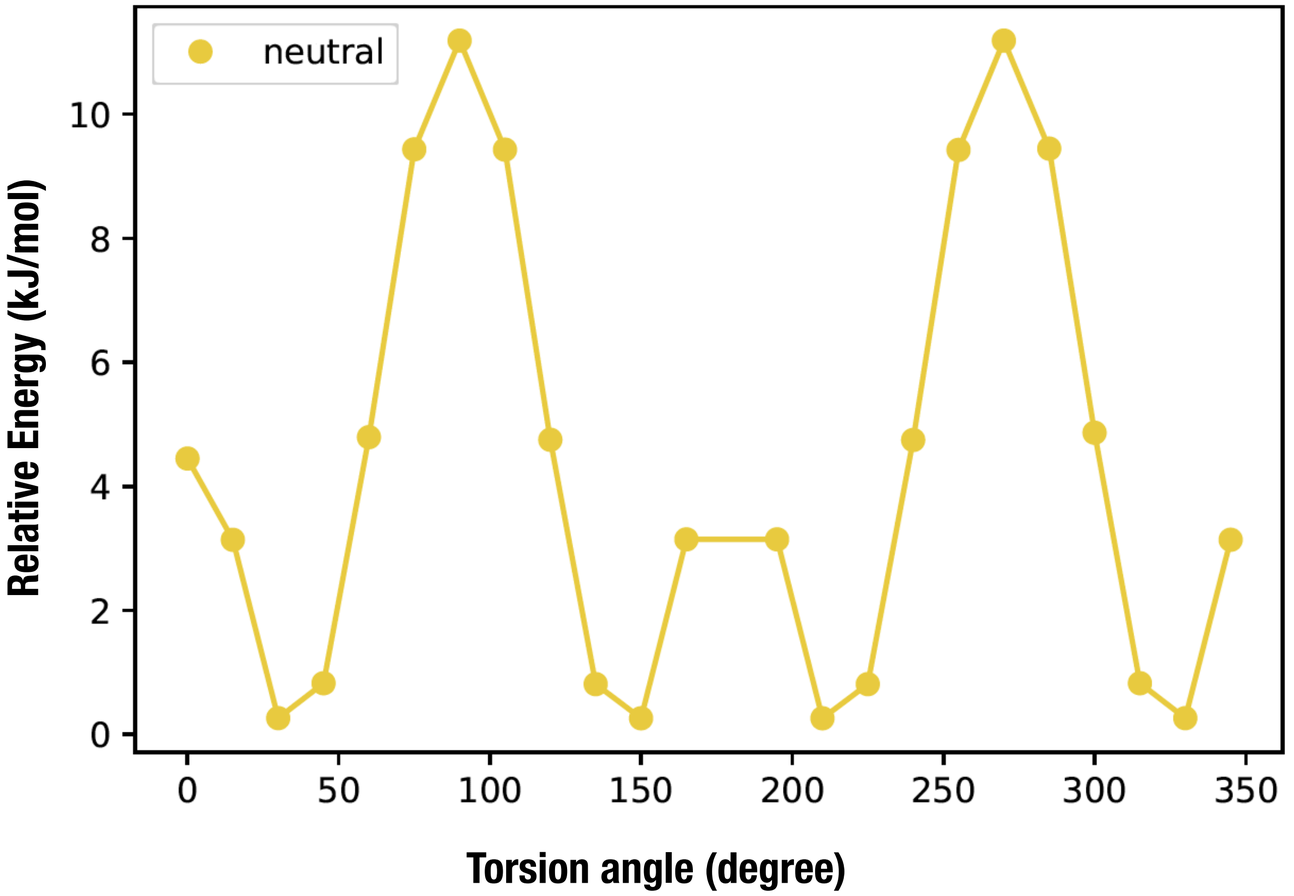
Fragmenting molecules is necessary to avoid high computational cost of generating QC data and avoid intramolecular interactions

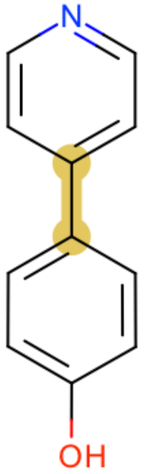
Distribution of molecular size and rotatable bonds of small molecules in drug bank



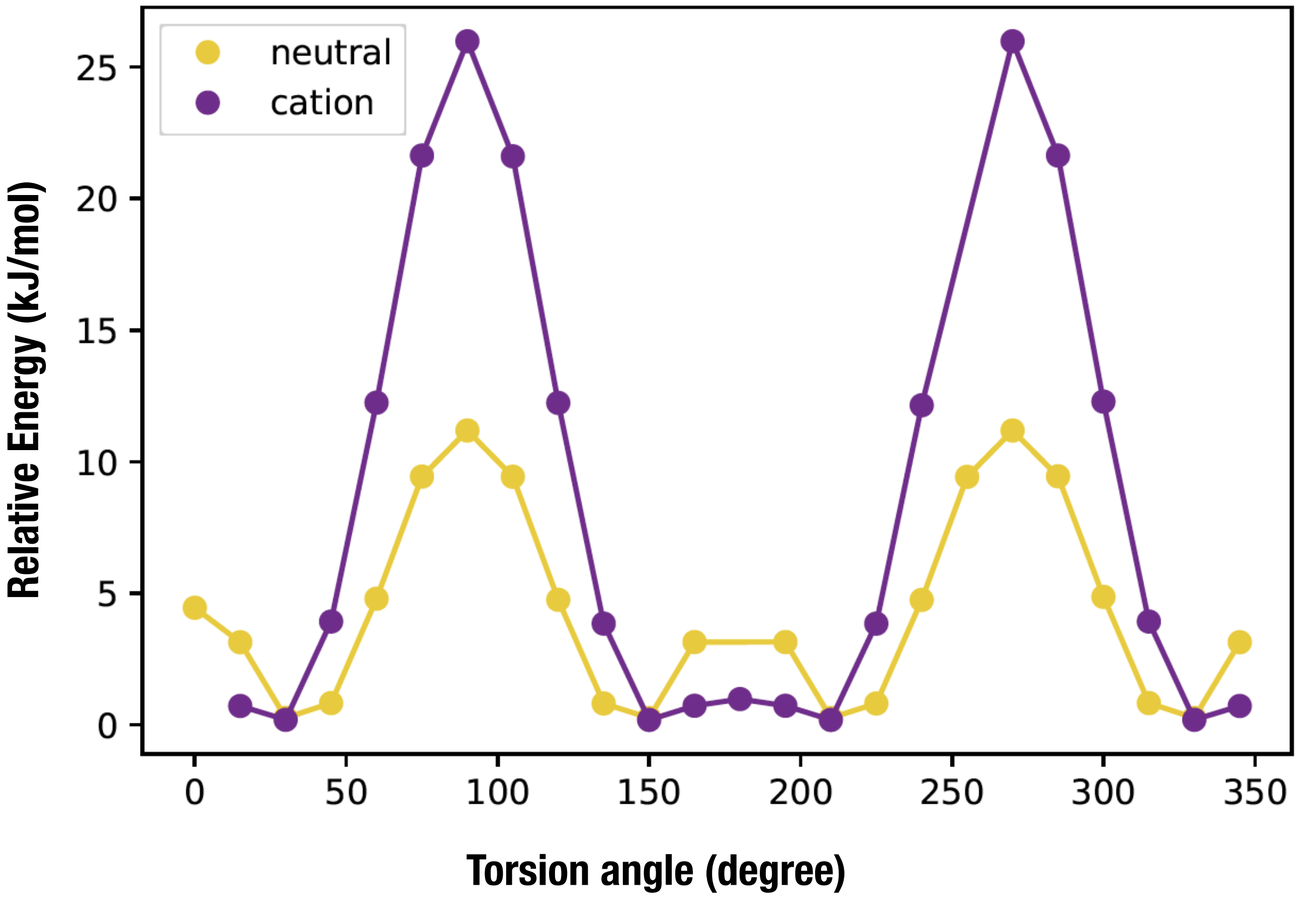


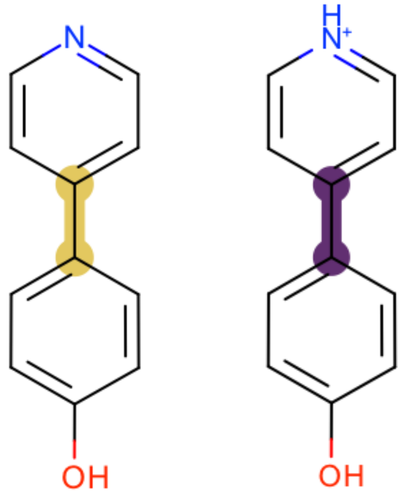
What are the pitfalls when fragmenting molecules for torsion scans?



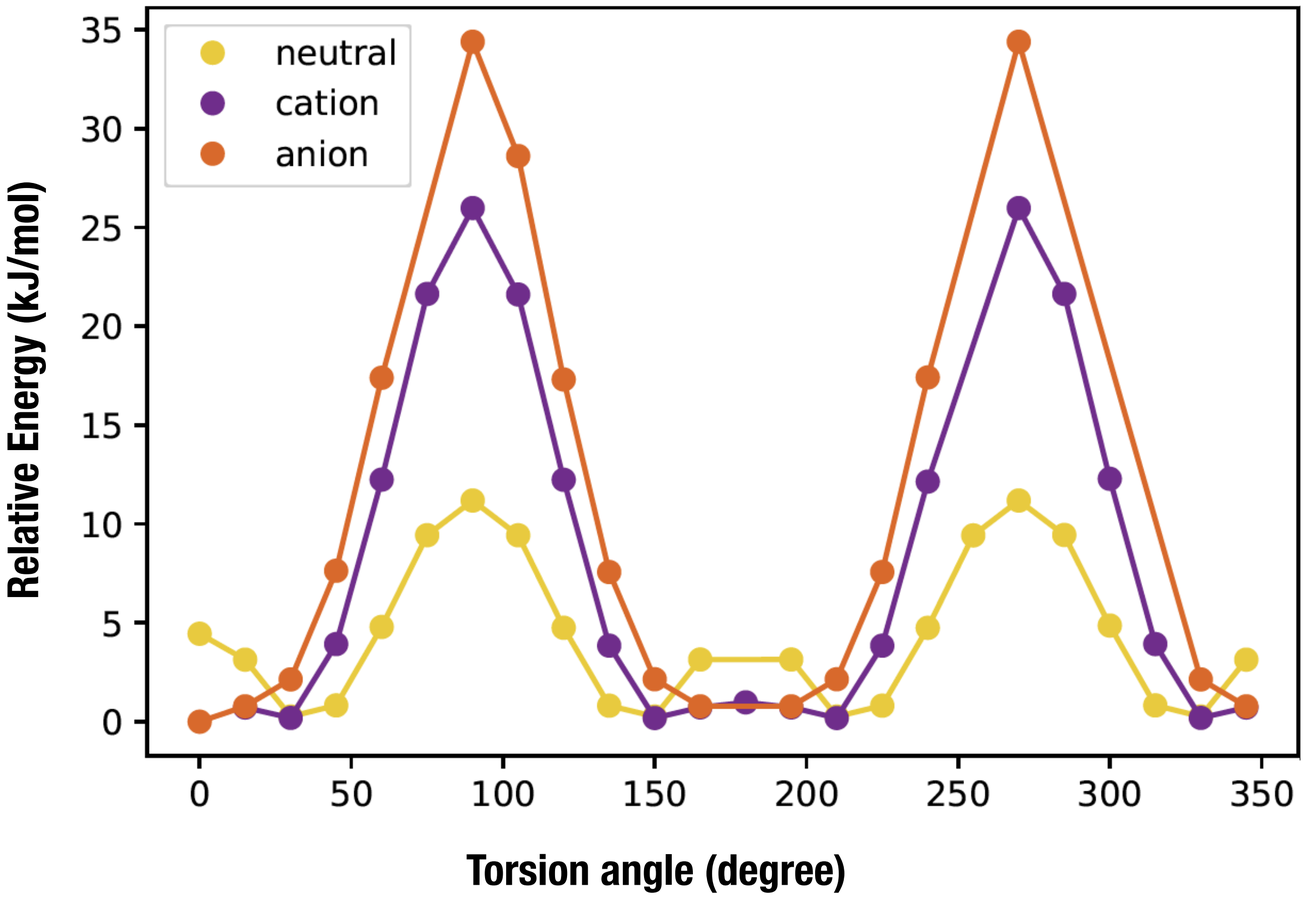


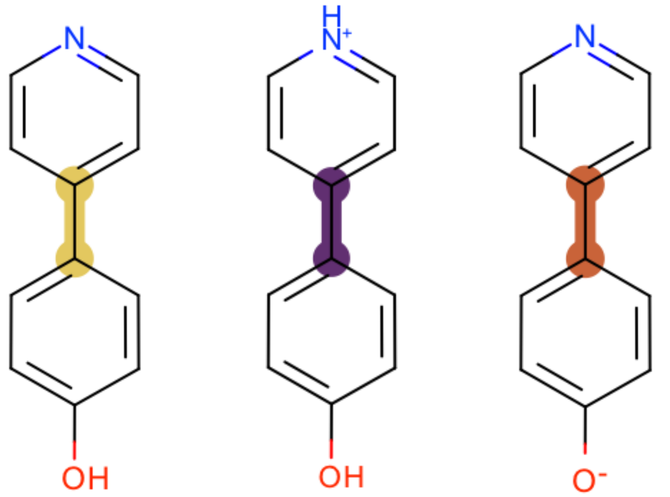
What are the pitfalls when fragmenting molecules for torsion scans?



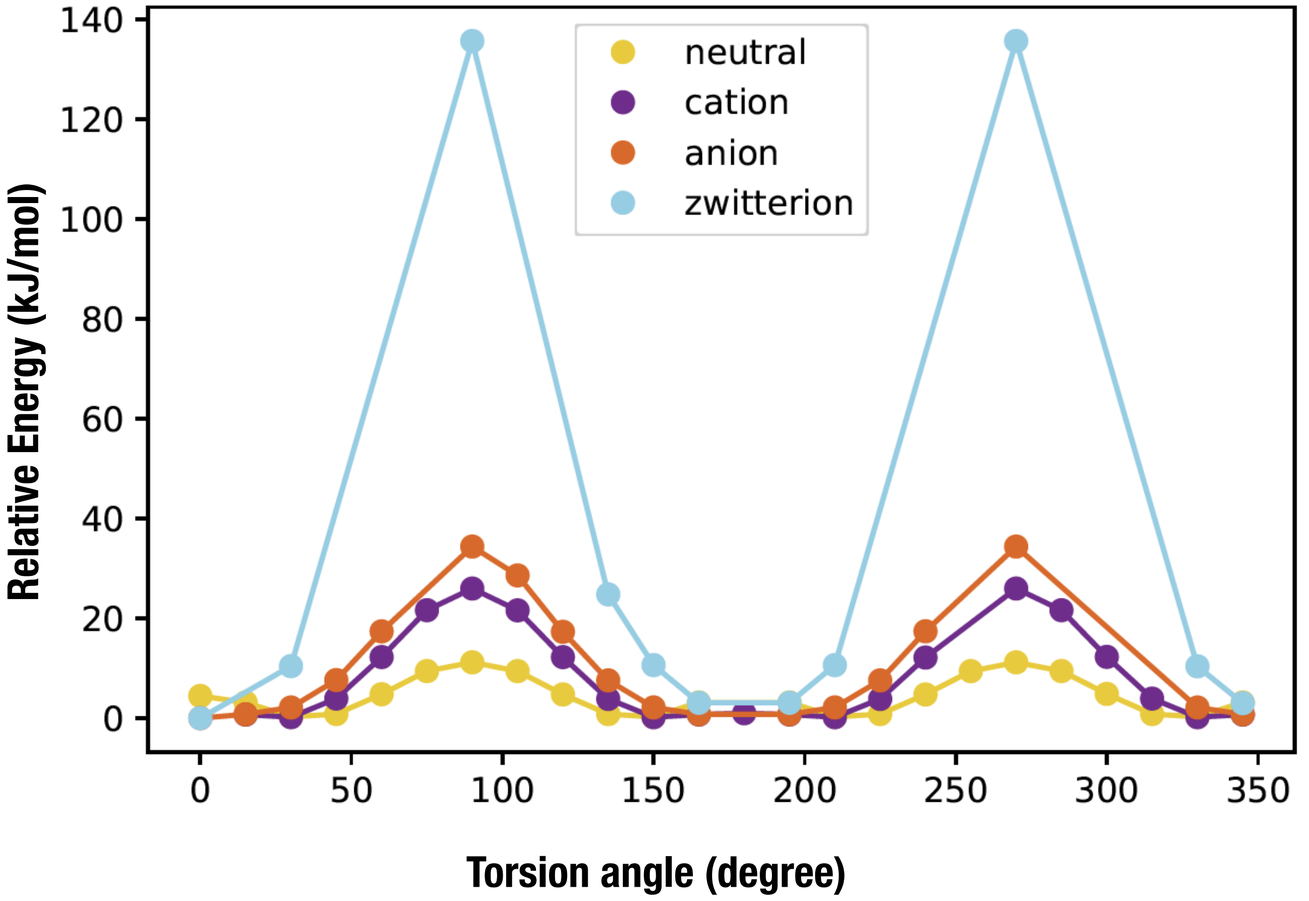


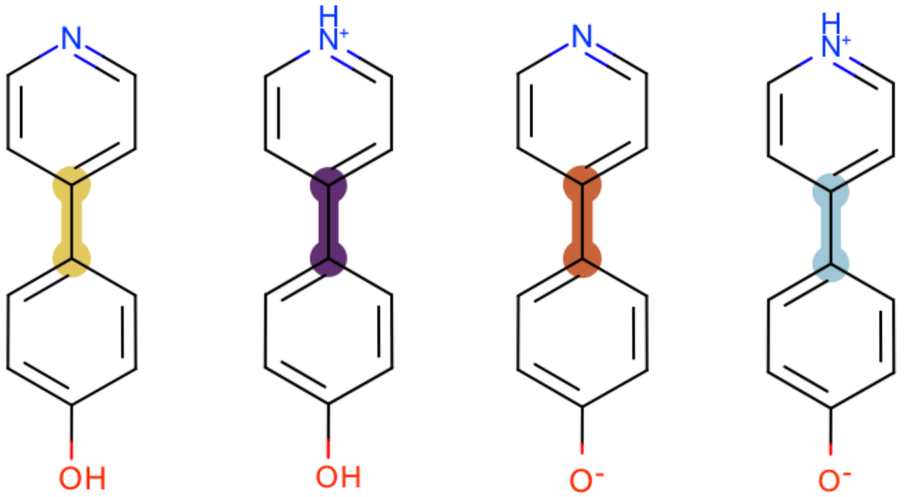
What are the pitfalls when fragmenting molecules for torsion scans?



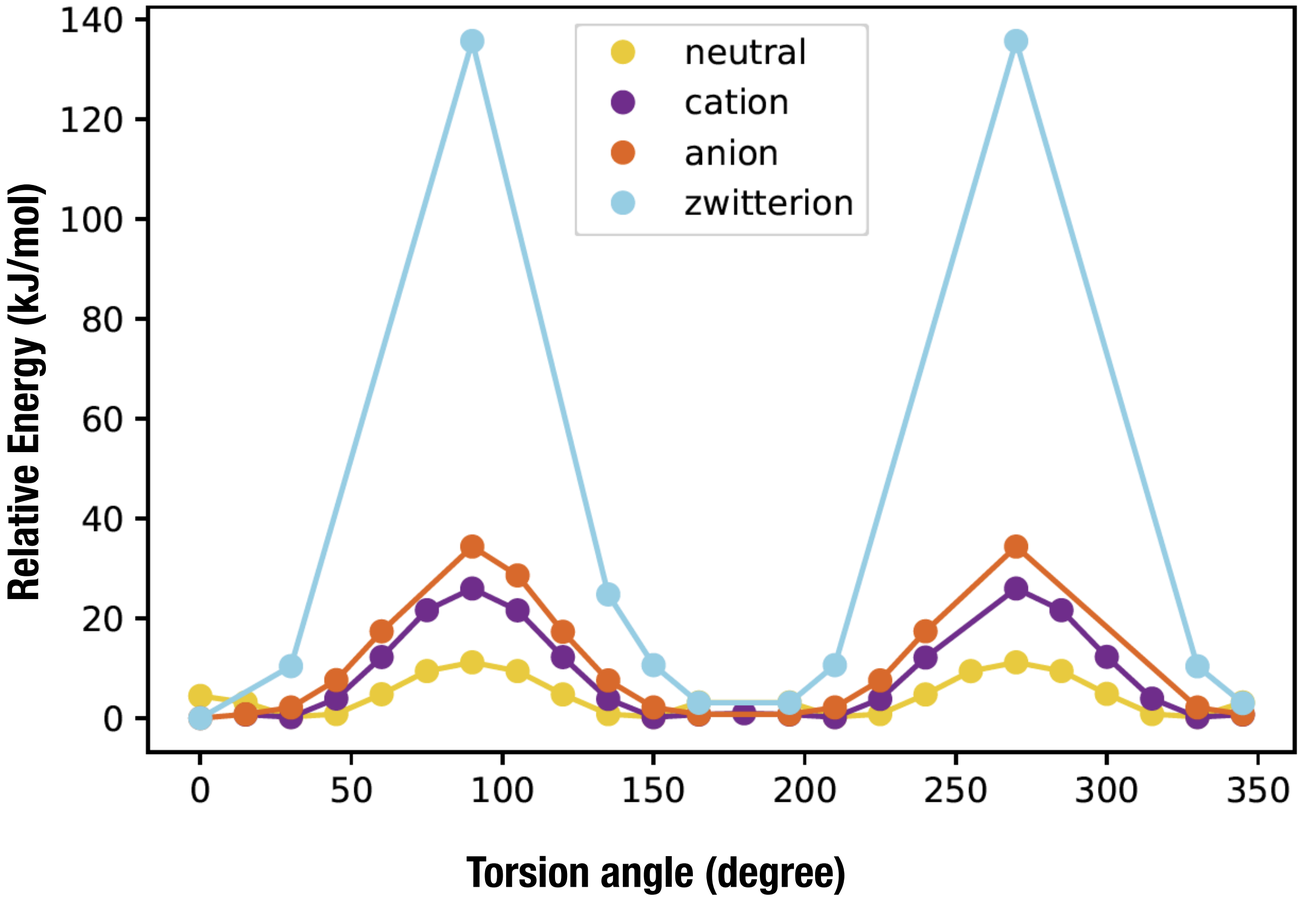


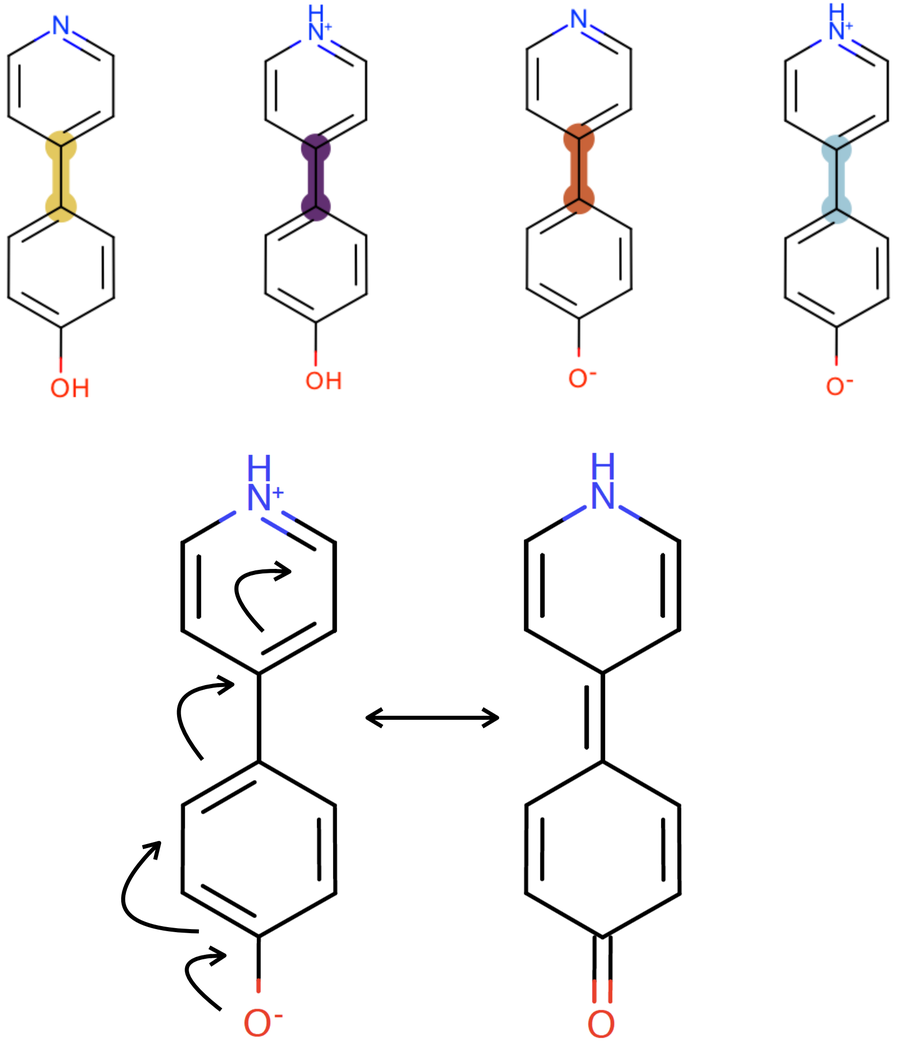
What are the pitfalls when fragmenting molecules for torsion scans?





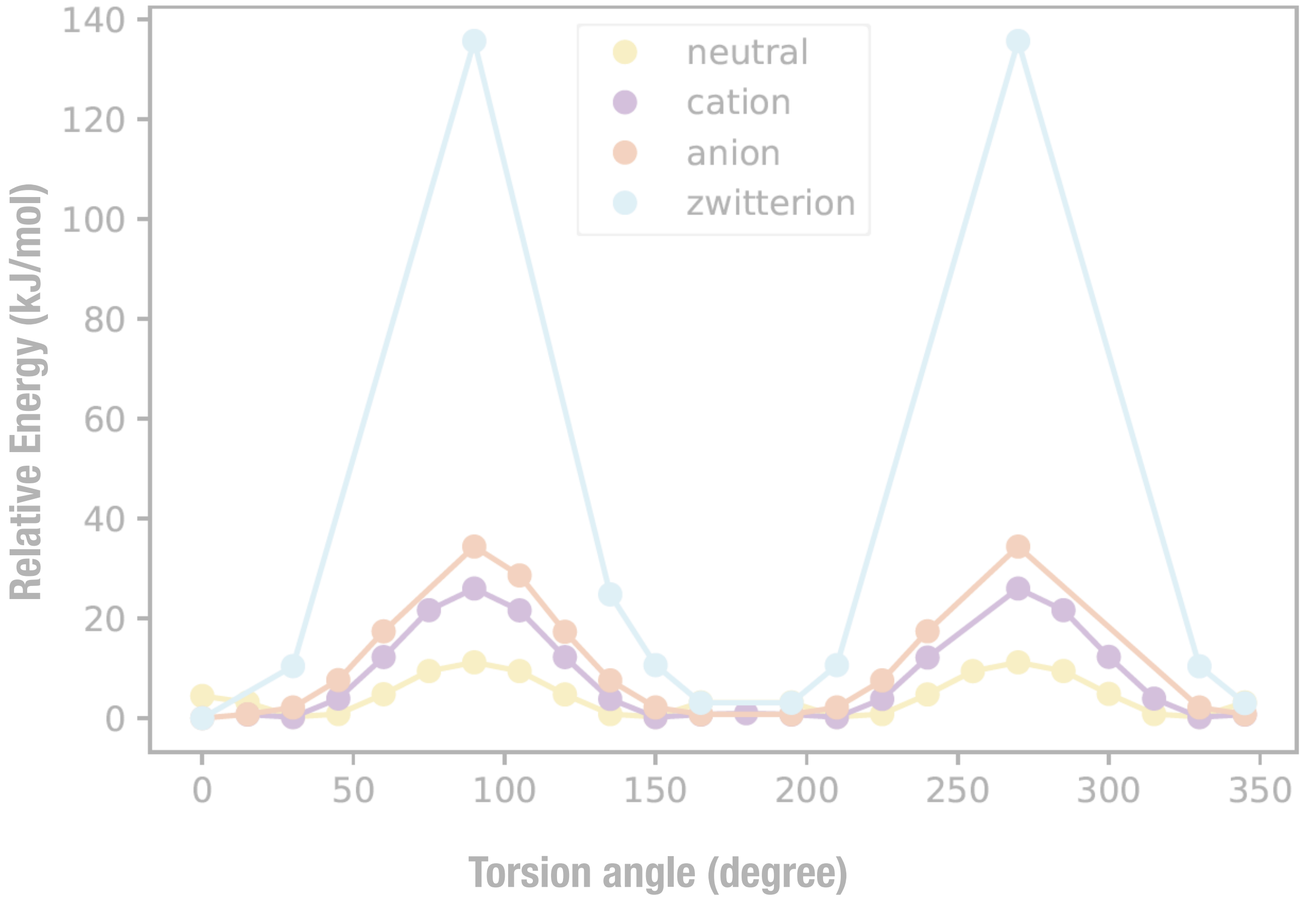
Chemical environment impacts the torsion profile of small molecules

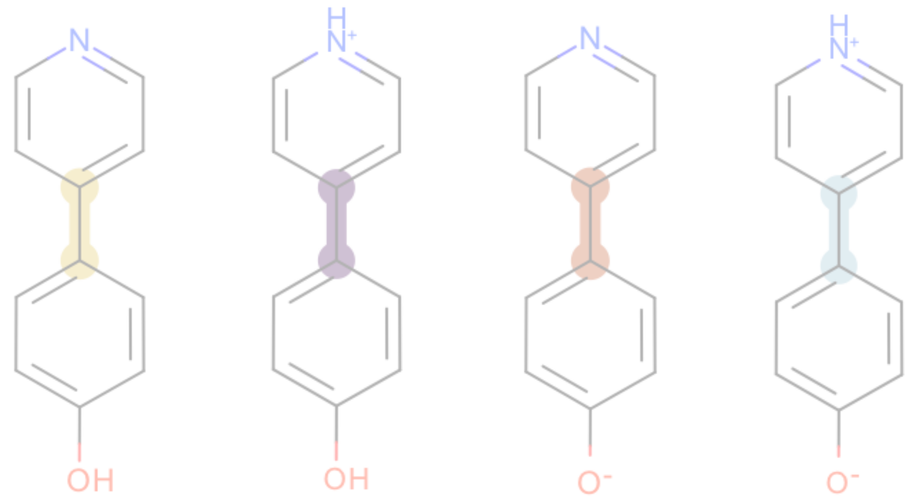




The Wiberg Bond Order is a measure of electron population overlap between two atoms

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

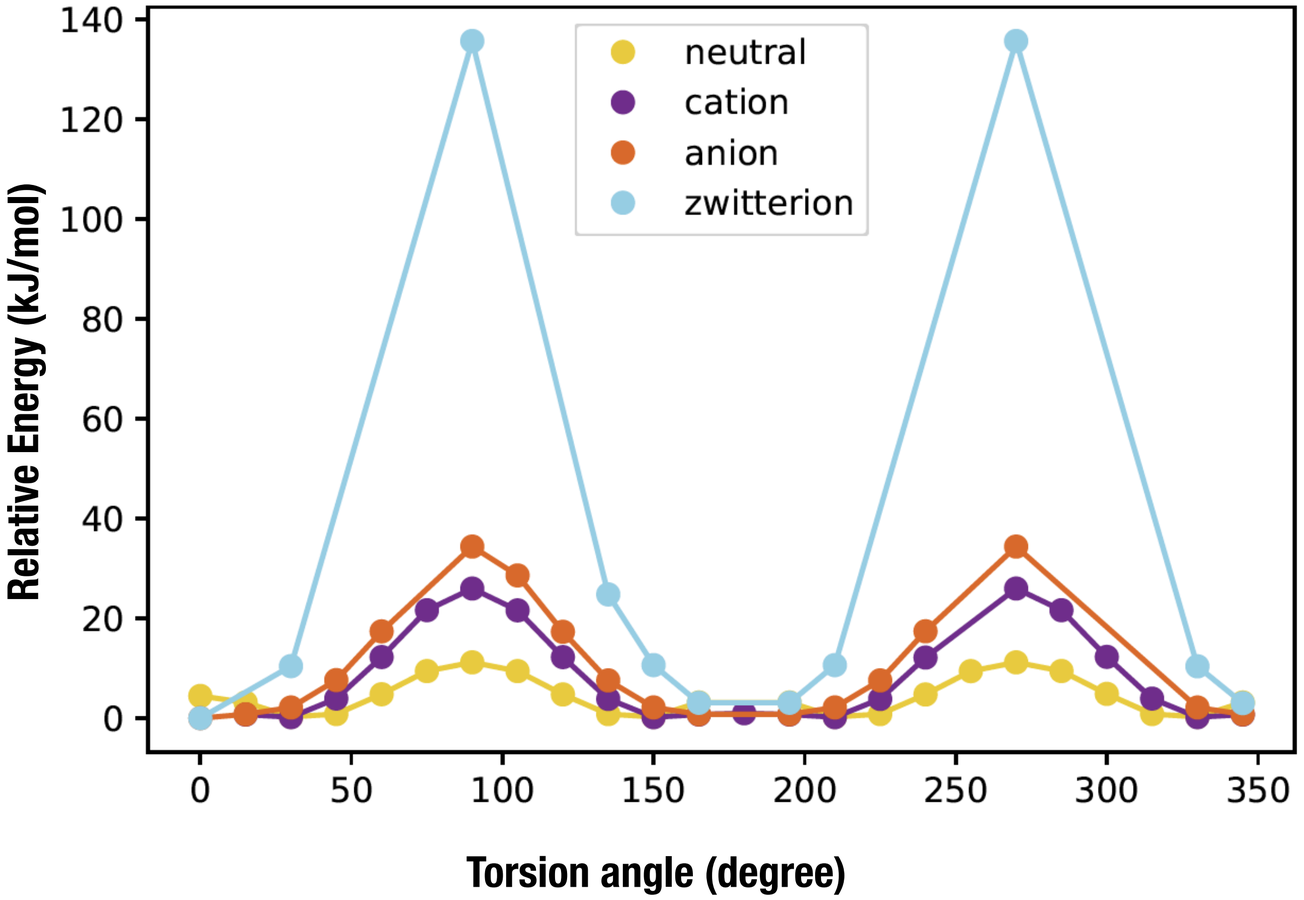


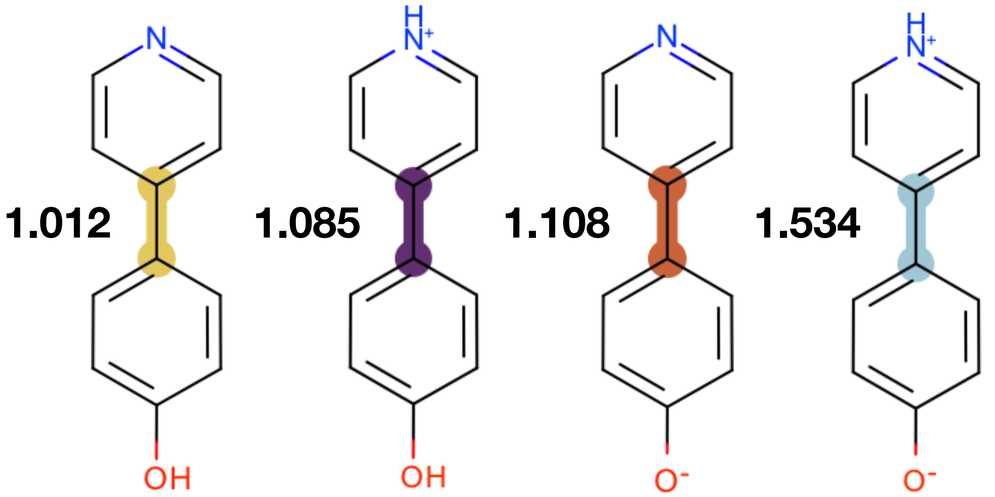


$$D_{\mu\nu} = 2 \sum_{i}^{\text{occ.}} C_{\mu i} C_{\nu i}^{*},$$

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

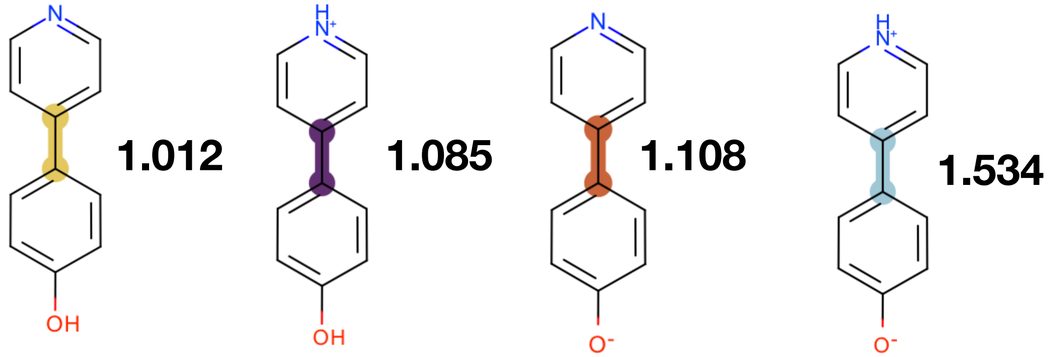
$$W_{AB} = \sum_{\mu \in A} \sum_{\nu \in B} |D_{\mu\nu}|^2.$$

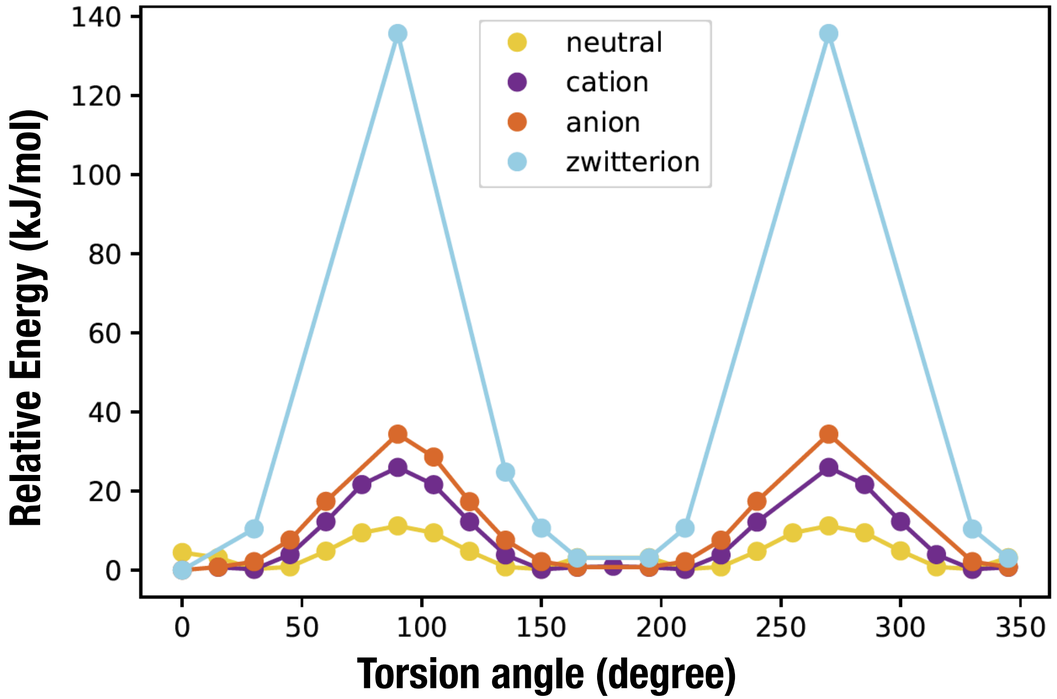


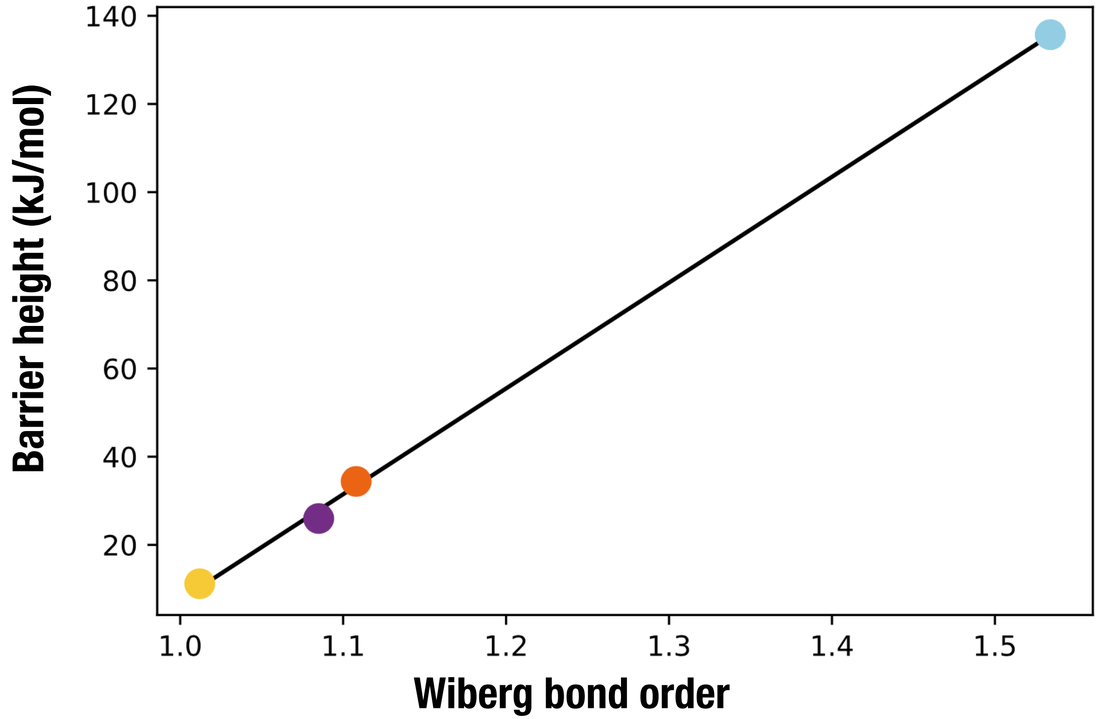


$$D_{\mu\nu} = 2 \sum_{i}^{\text{occ.}} C_{\mu i} C_{\nu i}^{*},$$

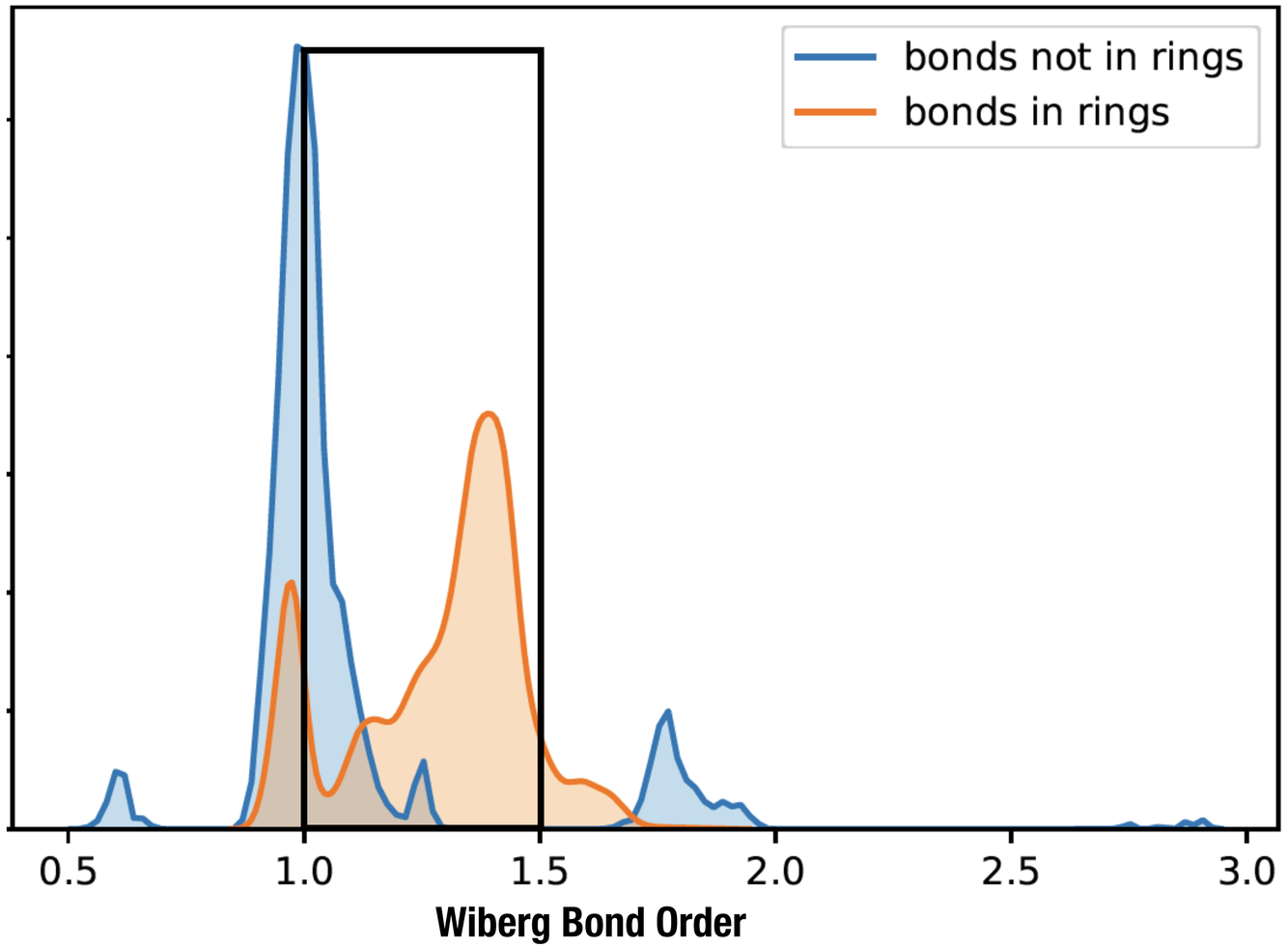
Torsion potential barrier heights are linear with Wiberg Bond Orders



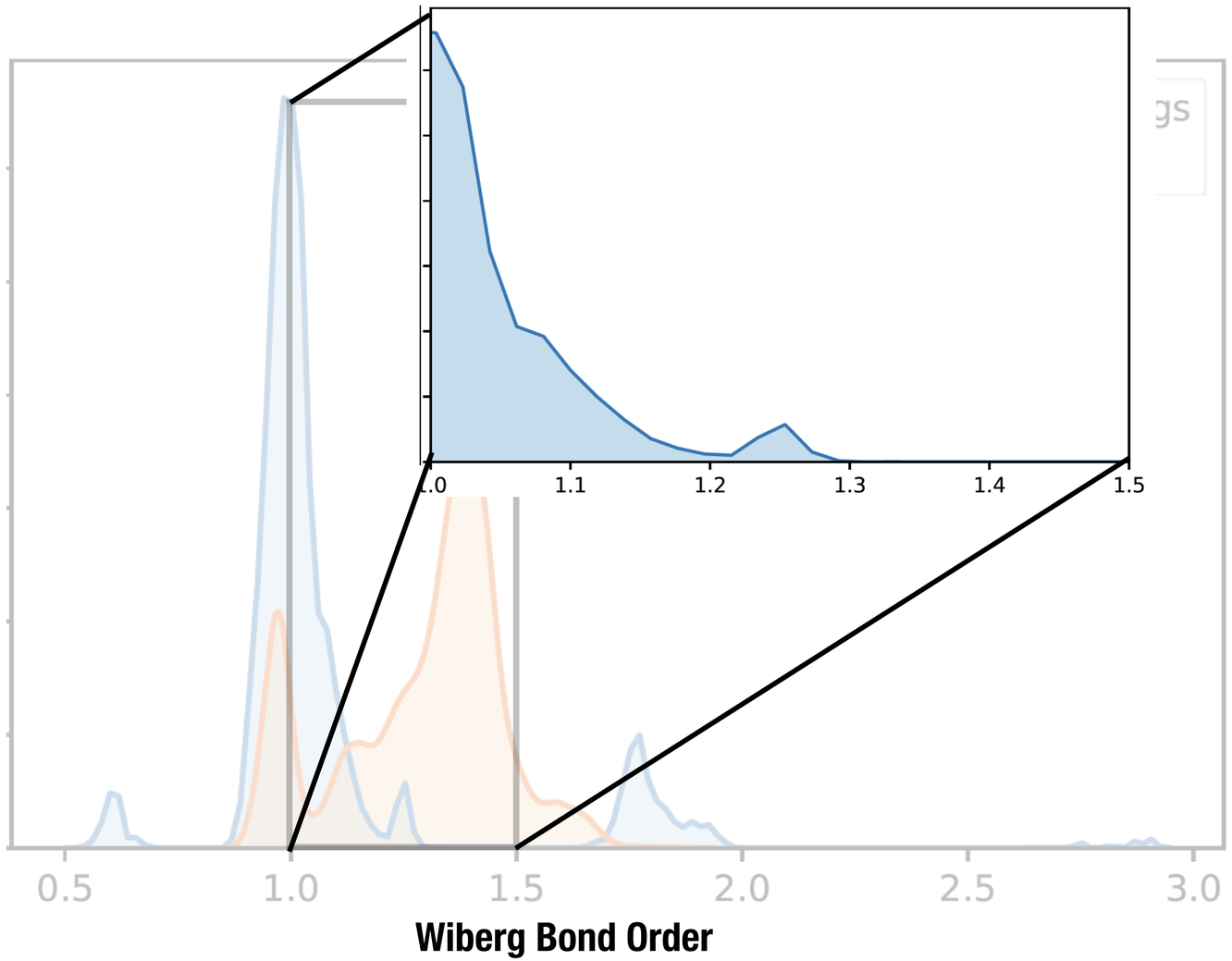




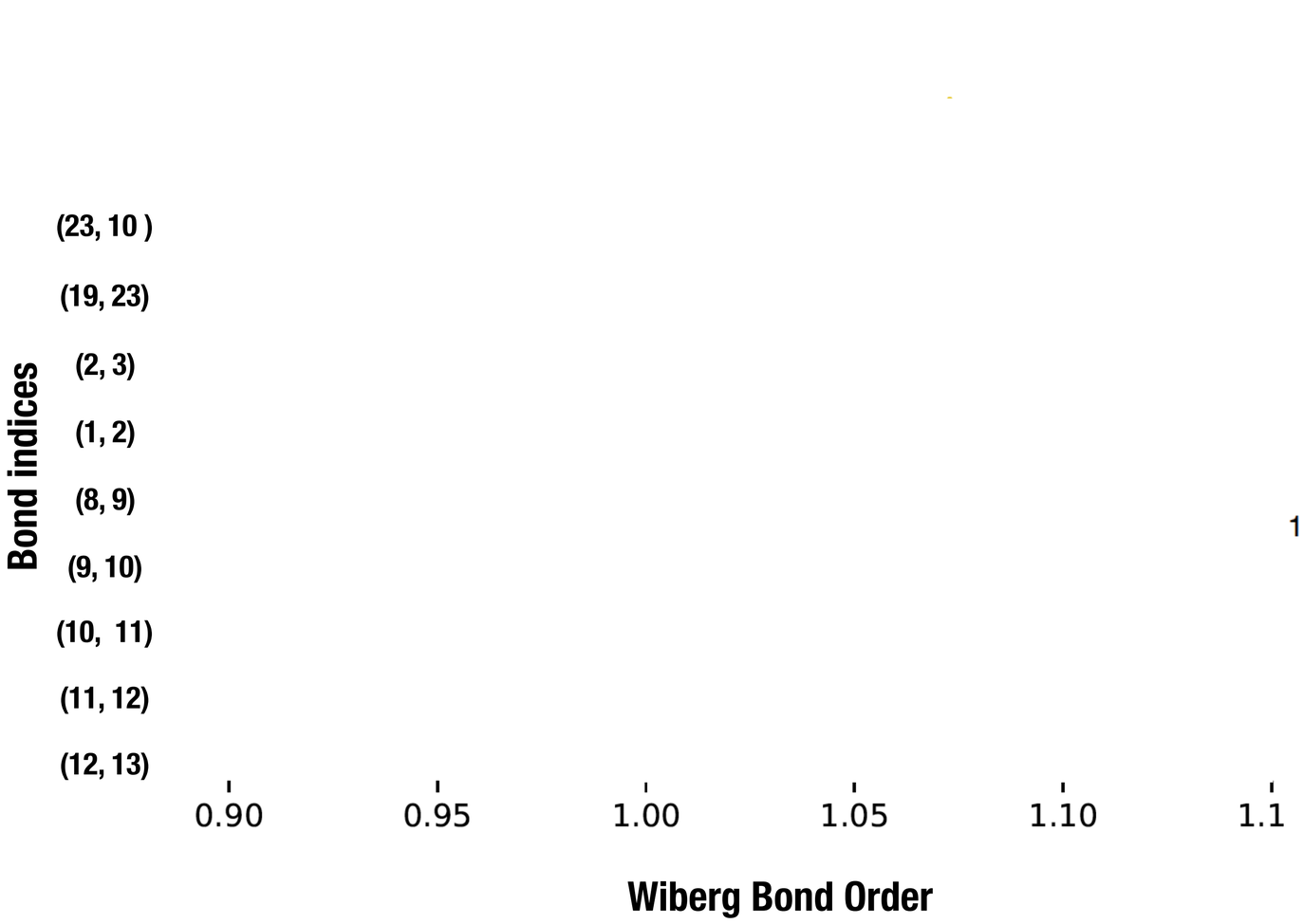
Distribution of WBO of drug like molecules shows significant density of conjugated bonds



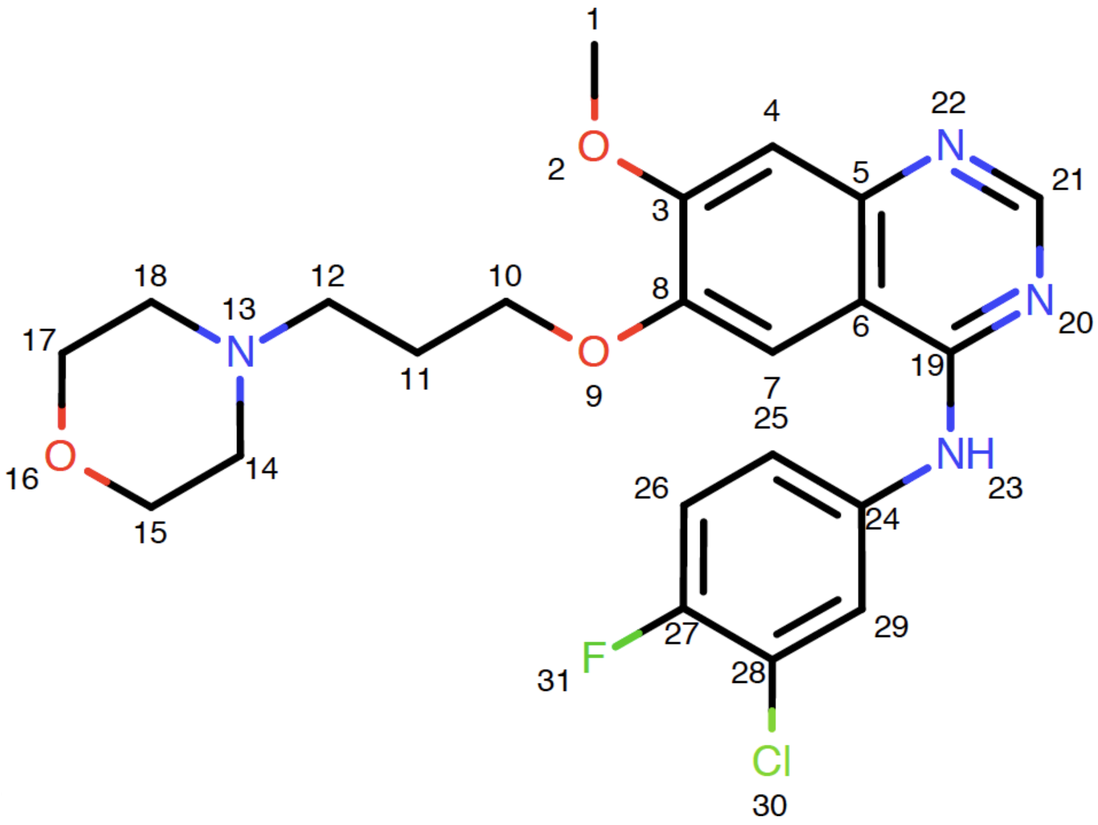
Distribution of WBO of drug like molecules shows significant density of conjugated bonds



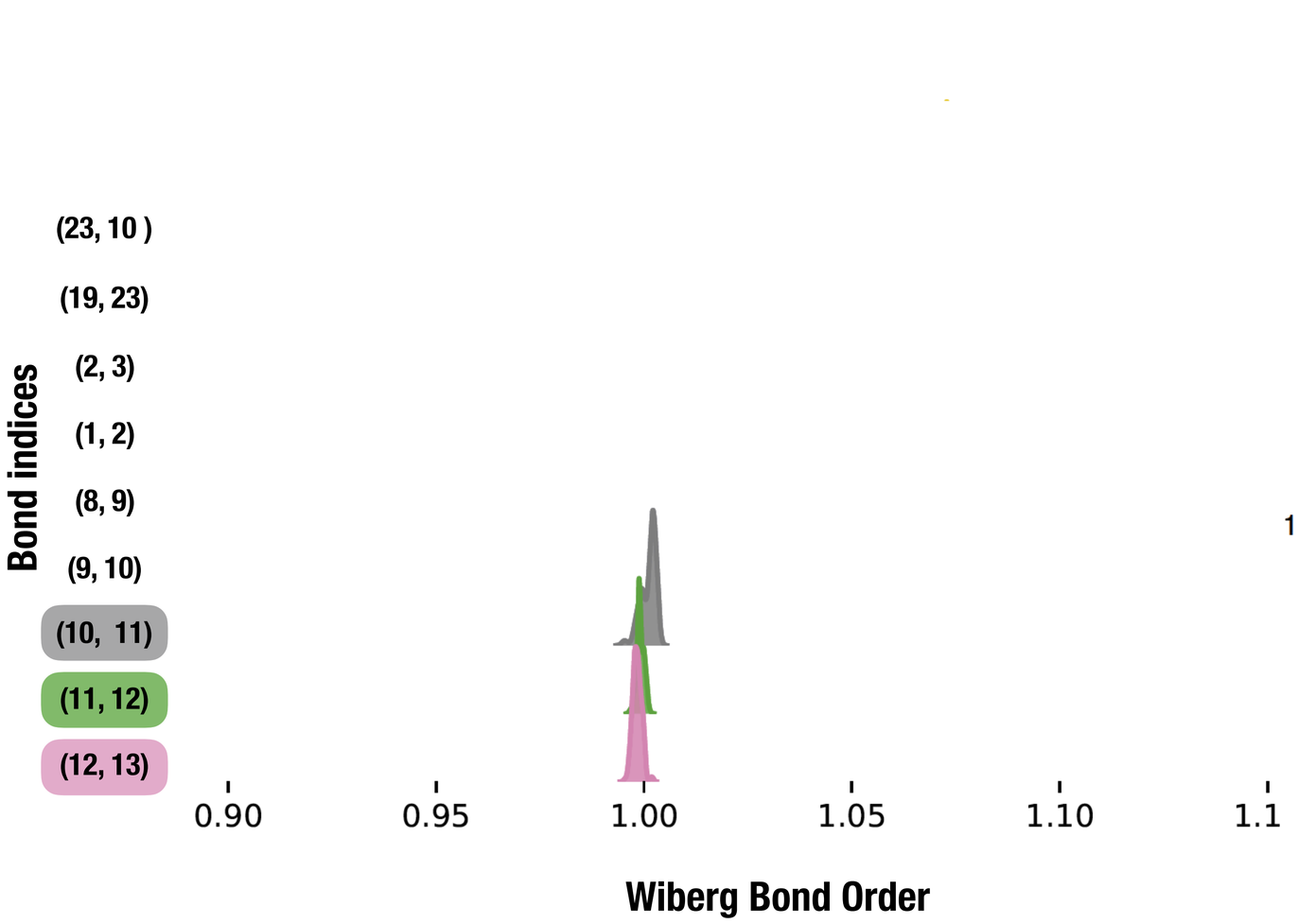
Variance of Wiberg Bond Order with respect to conformations are higher for conjugated bonds

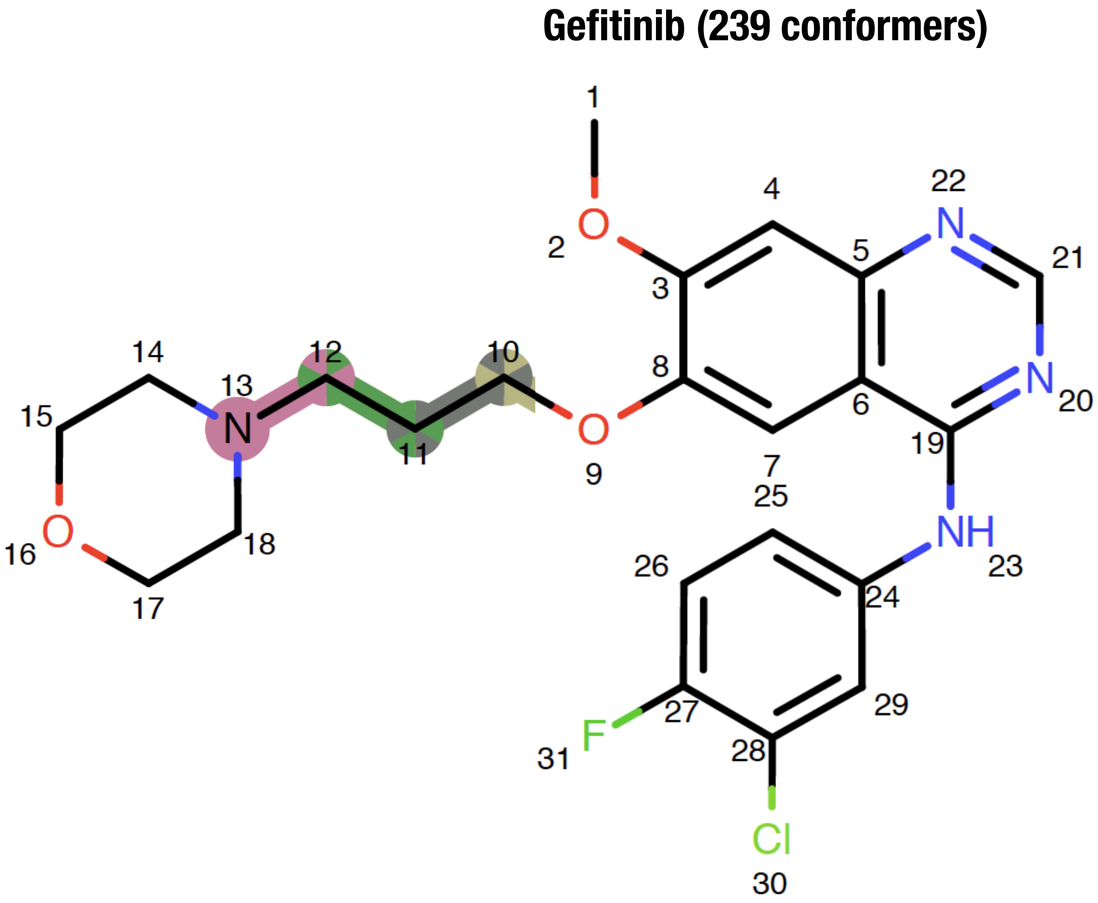


Gefitinib (239 conformers)

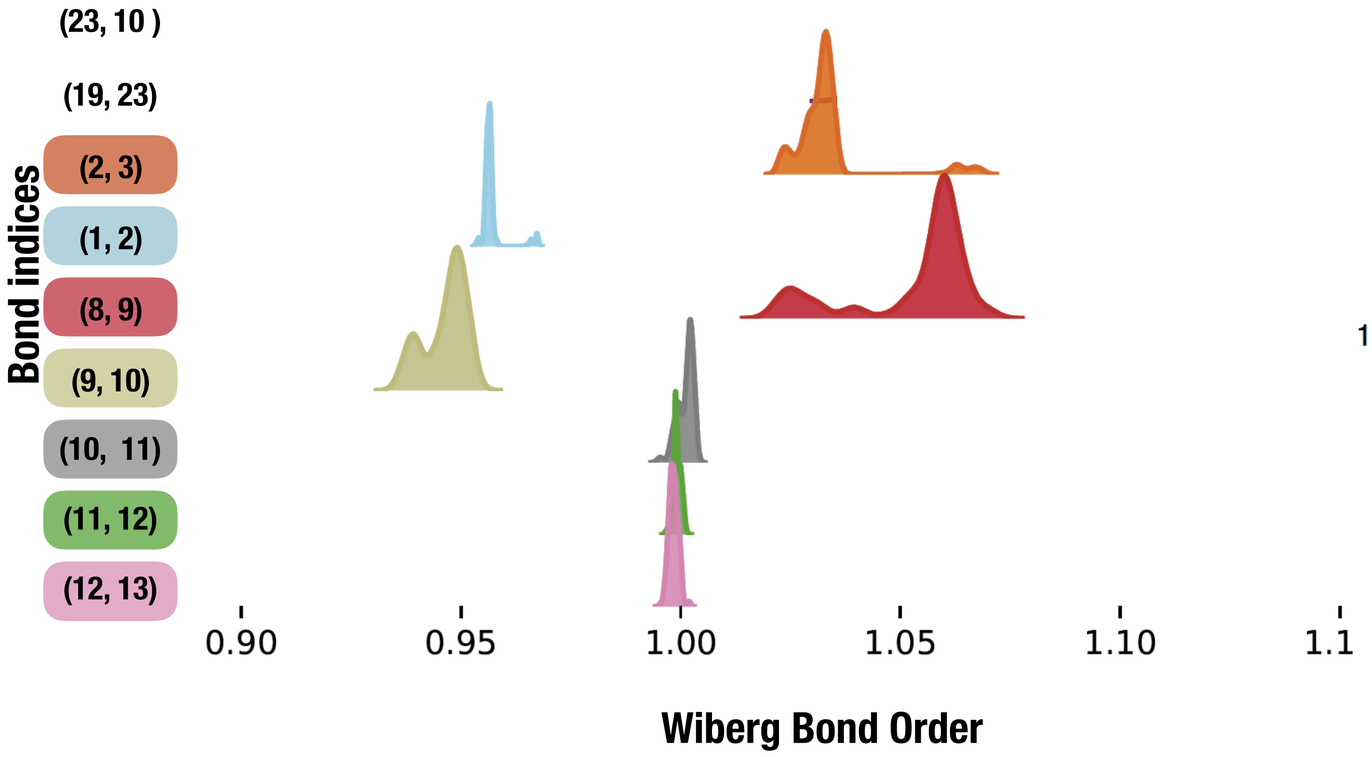


Variance of Wiberg Bond Order with respect to conformations are higher for conjugated bonds

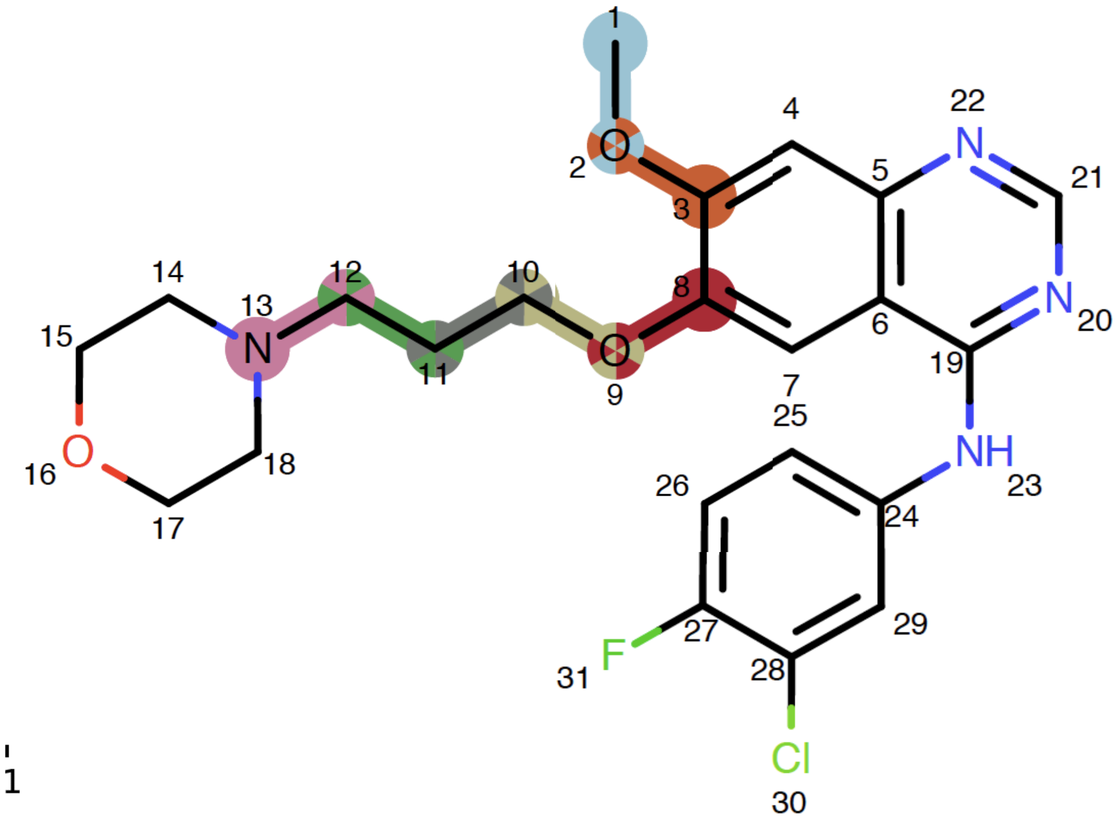




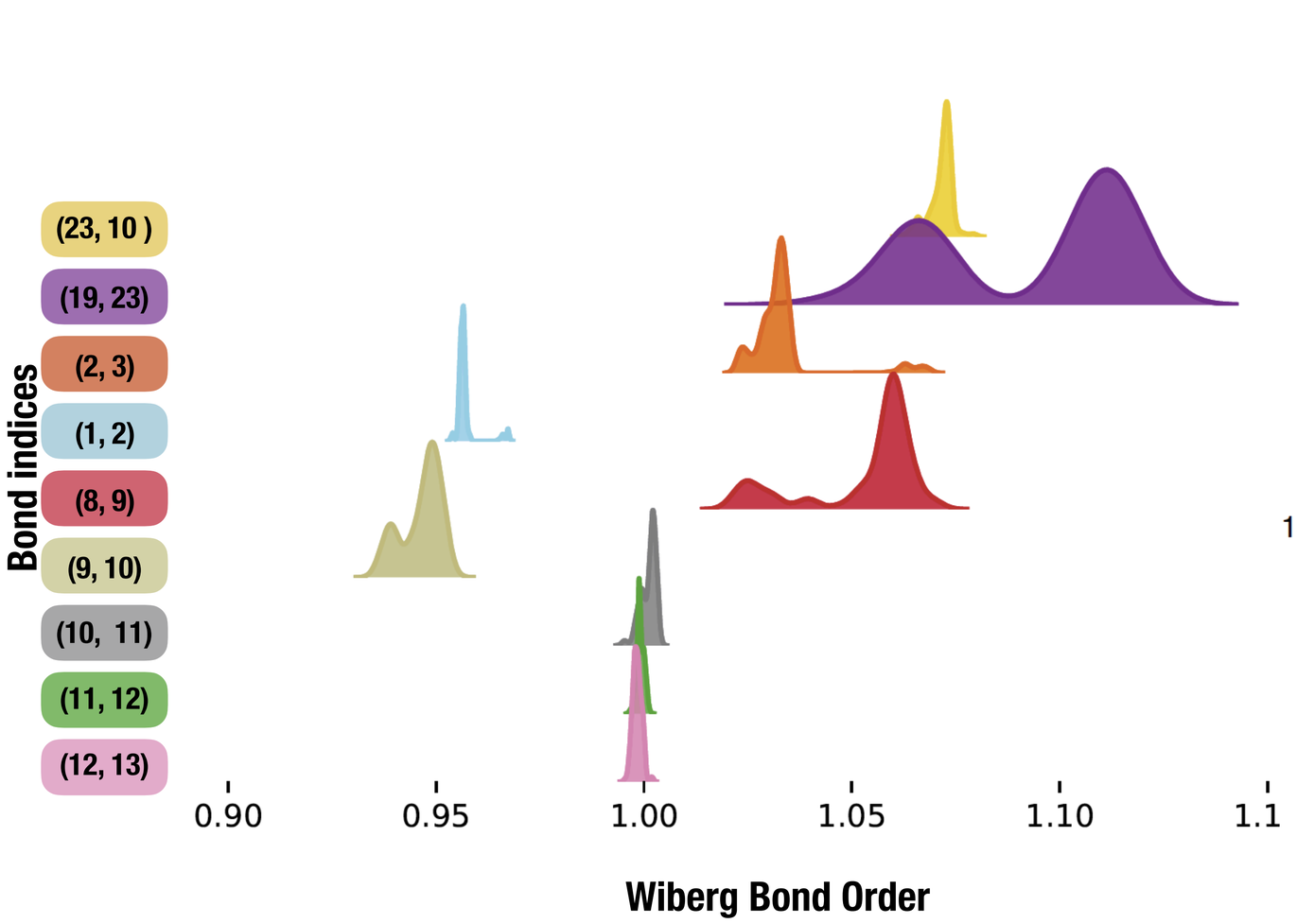
Variance of Wiberg Bond Order with respect to conformations are higher for conjugated bonds

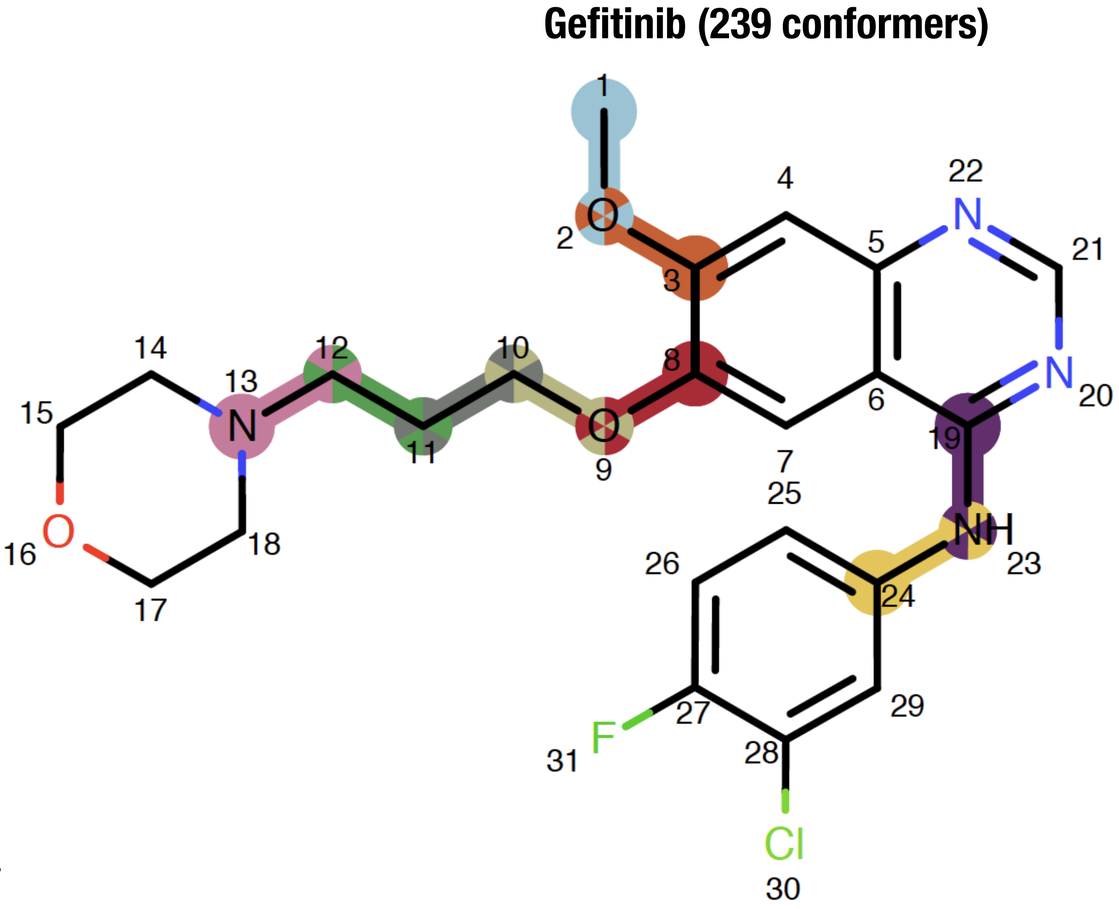


Gefitinib (239 conformers)

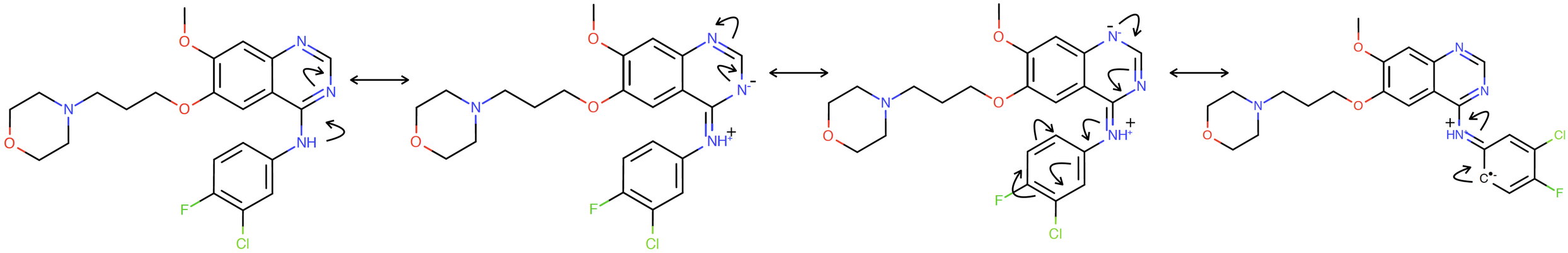


Variance of Wiberg Bond Order with respect to conformations are higher for conjugated bonds





Wiberg Bond Order variance aligns with the stability of resonance structures



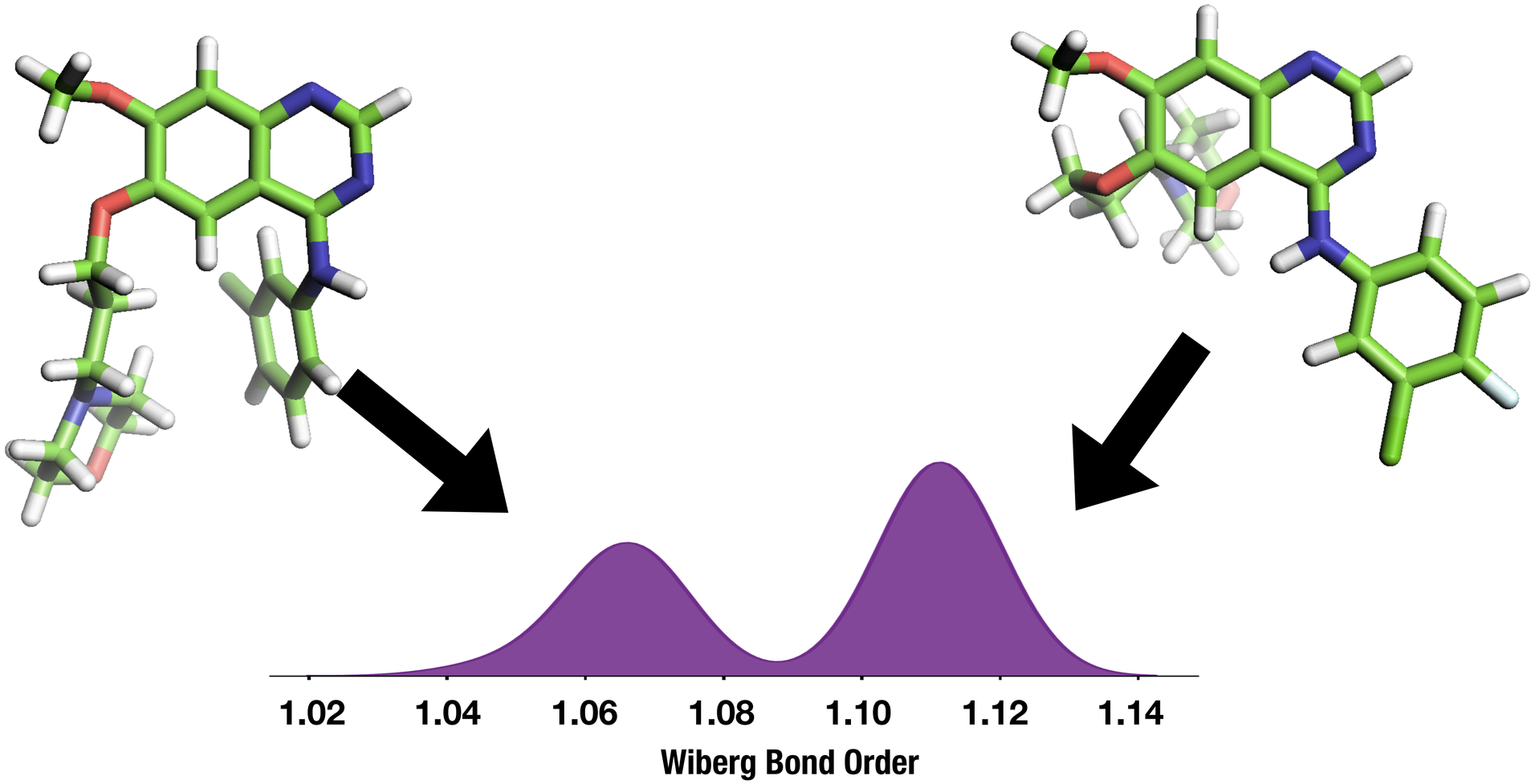
More stable
negative charge is on a nitrogen

More stable
negative charge is on a nitrogen

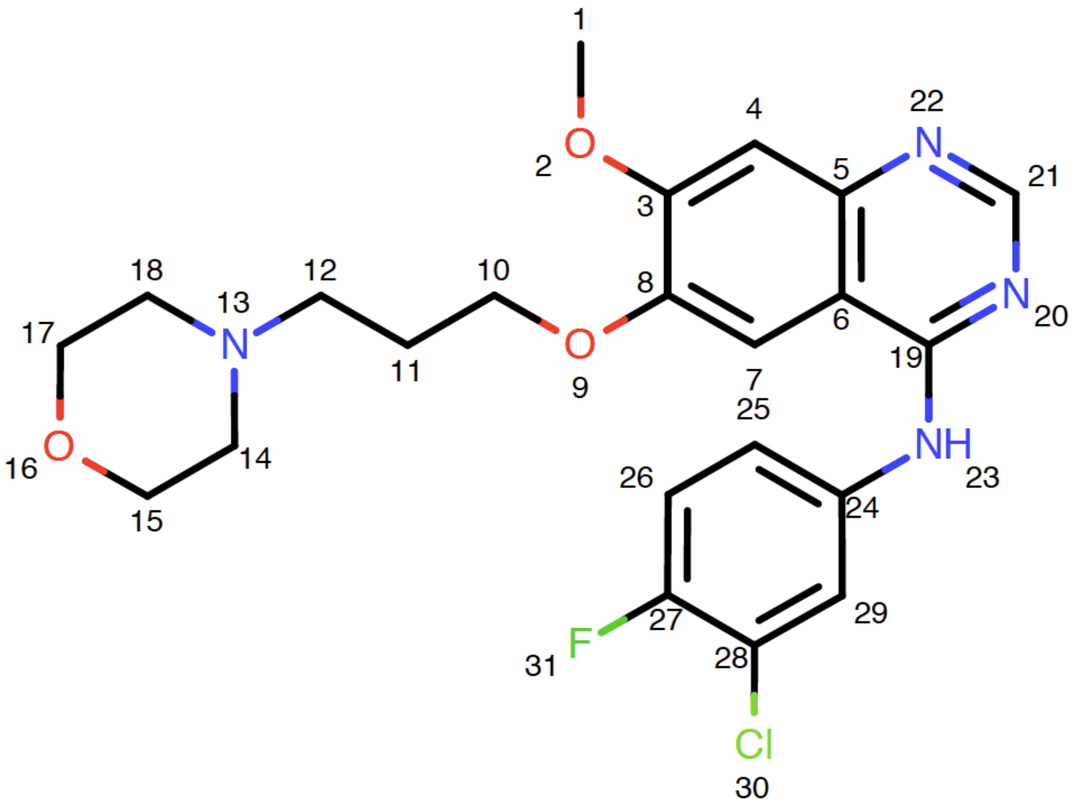
Less stable
negative charge is on a carbon

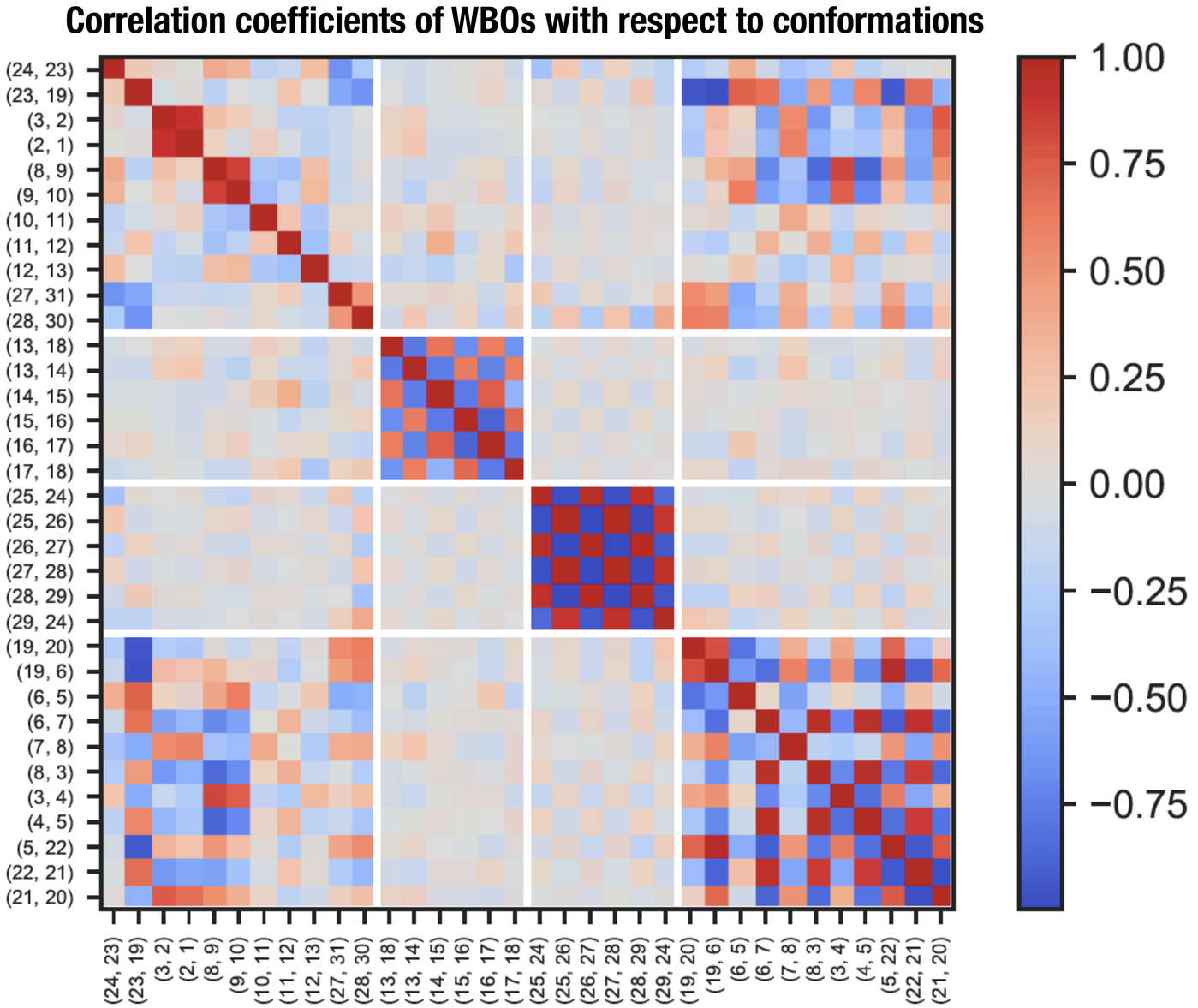
The bond between the secondary amine and quinazoline is more conjugated than the bond between the amine and the chloro fluoro benzene

Conjugated system is planar in the higher mode of the Wiberg Bond Order distribution

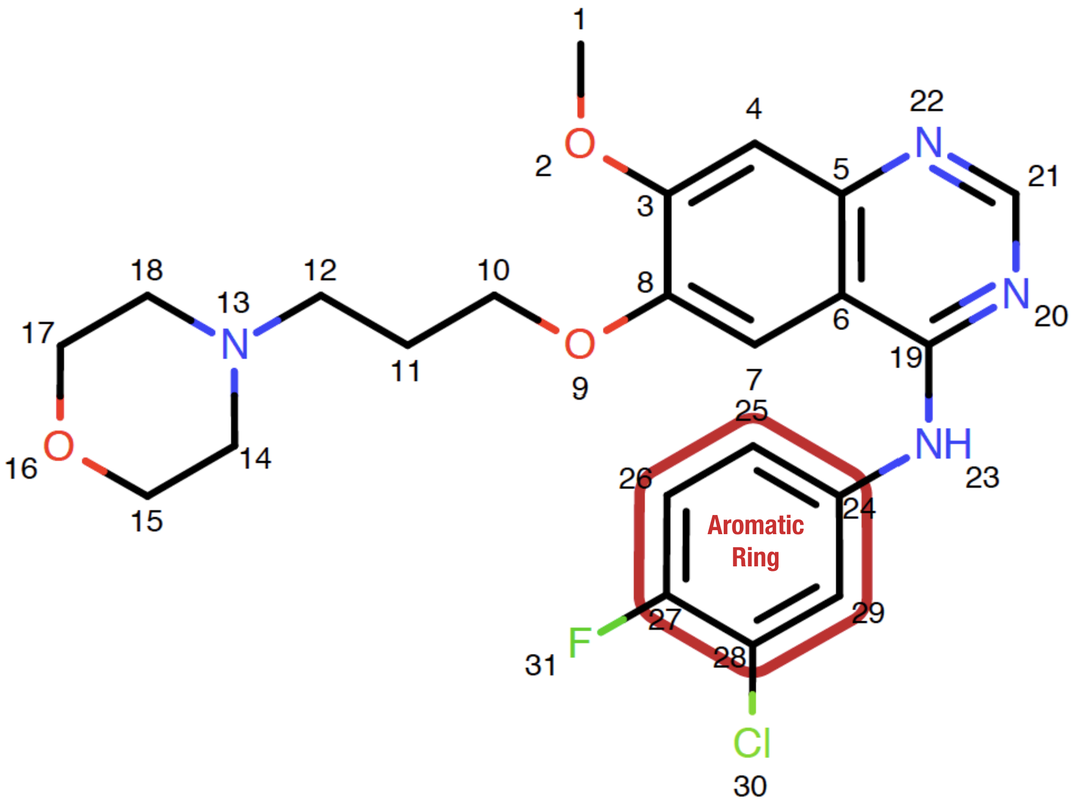


Wiberg Bond Orders of bonds in conjugated systems are correlated with each other

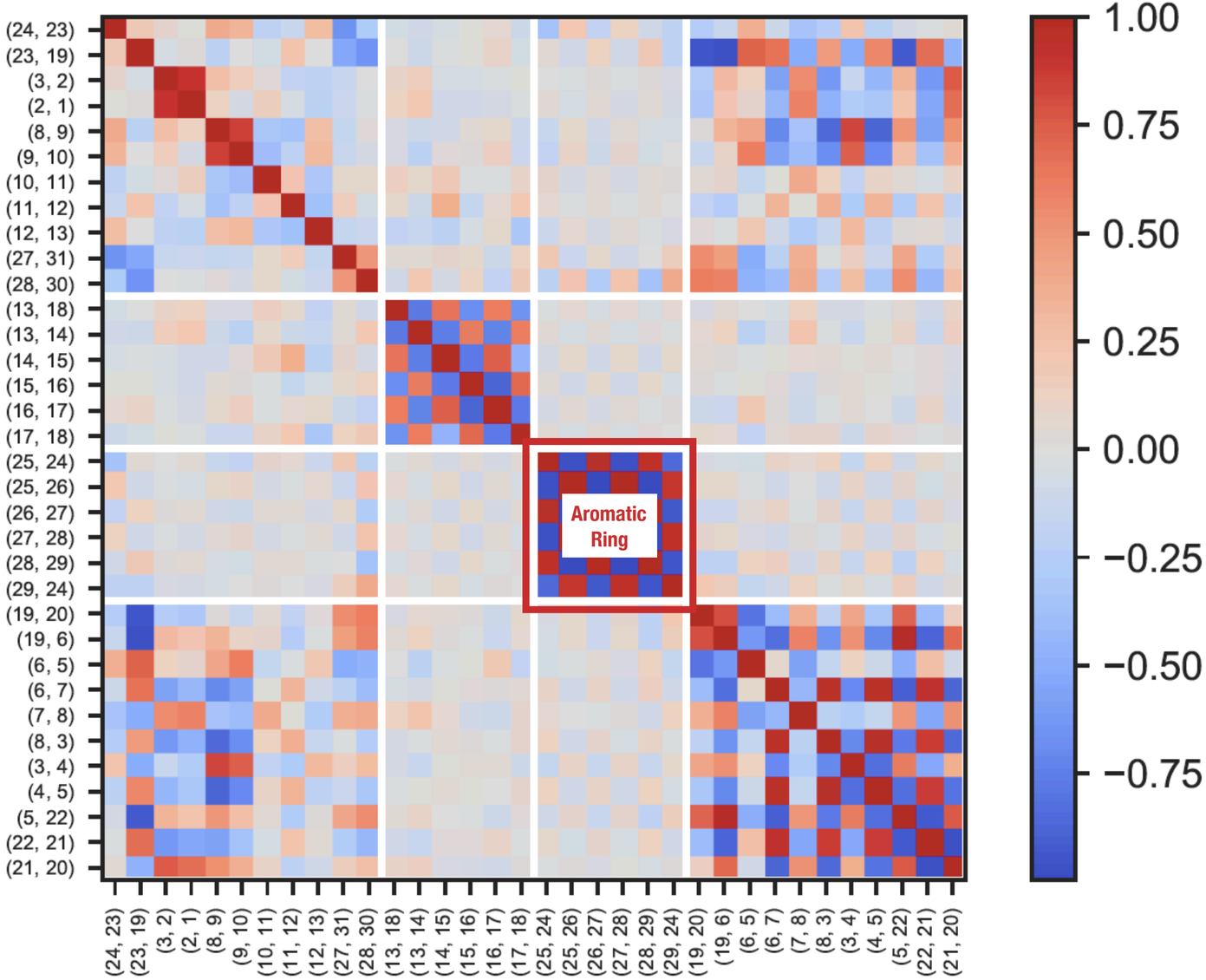




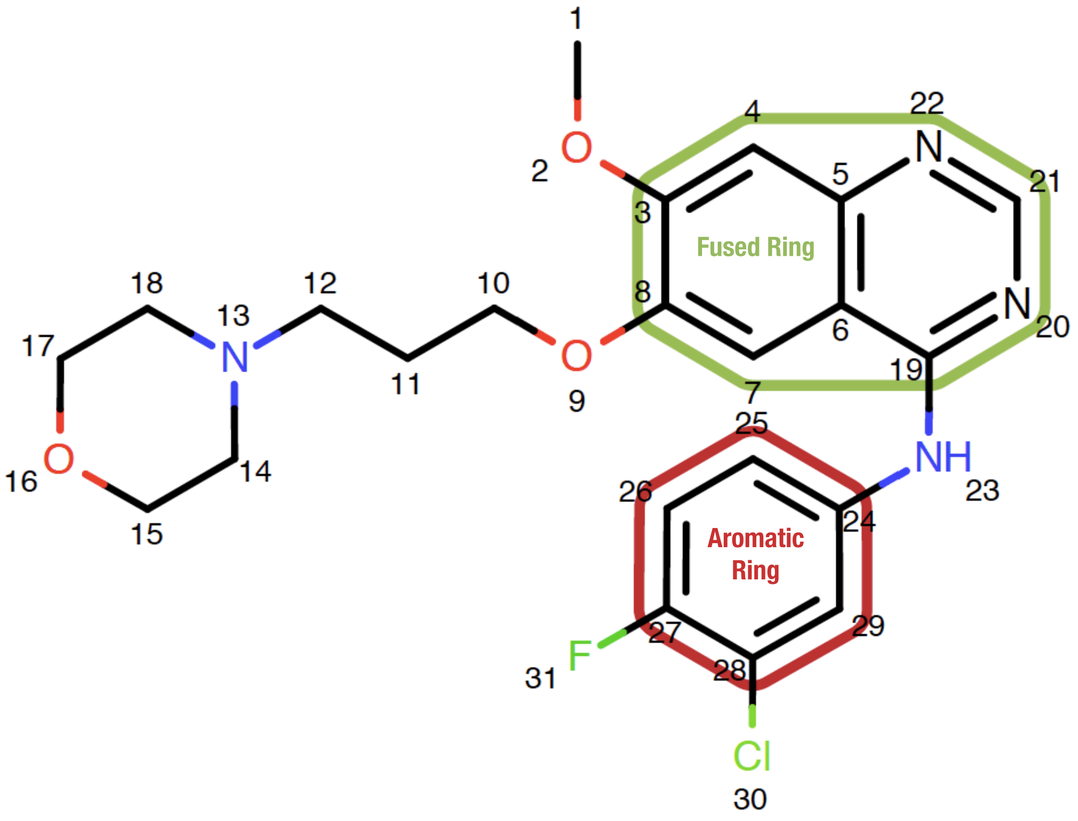
Wiberg Bond Orders of bonds in conjugated systems are correlated with each other



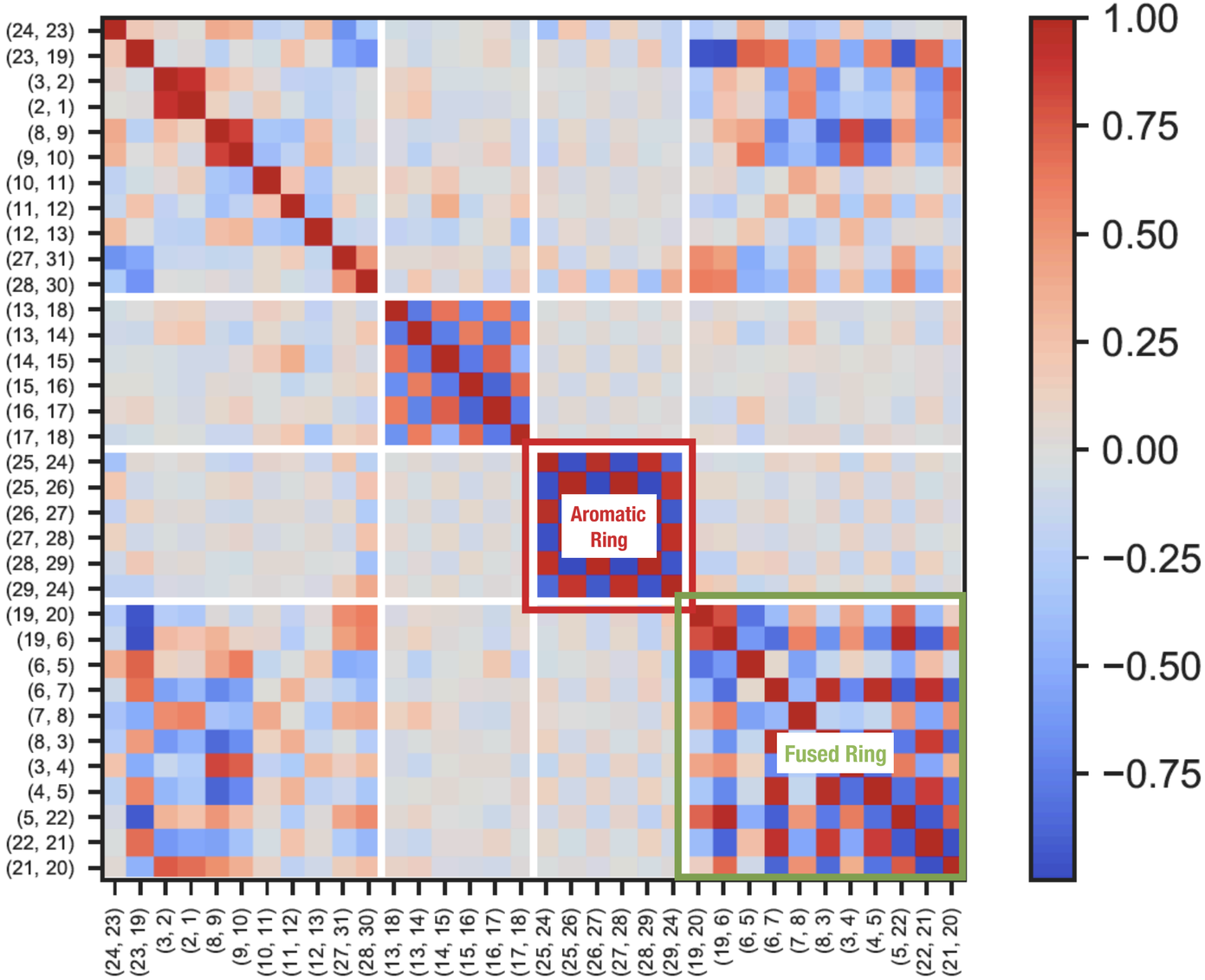
Correlation coefficients of WBOs with respect to conformations



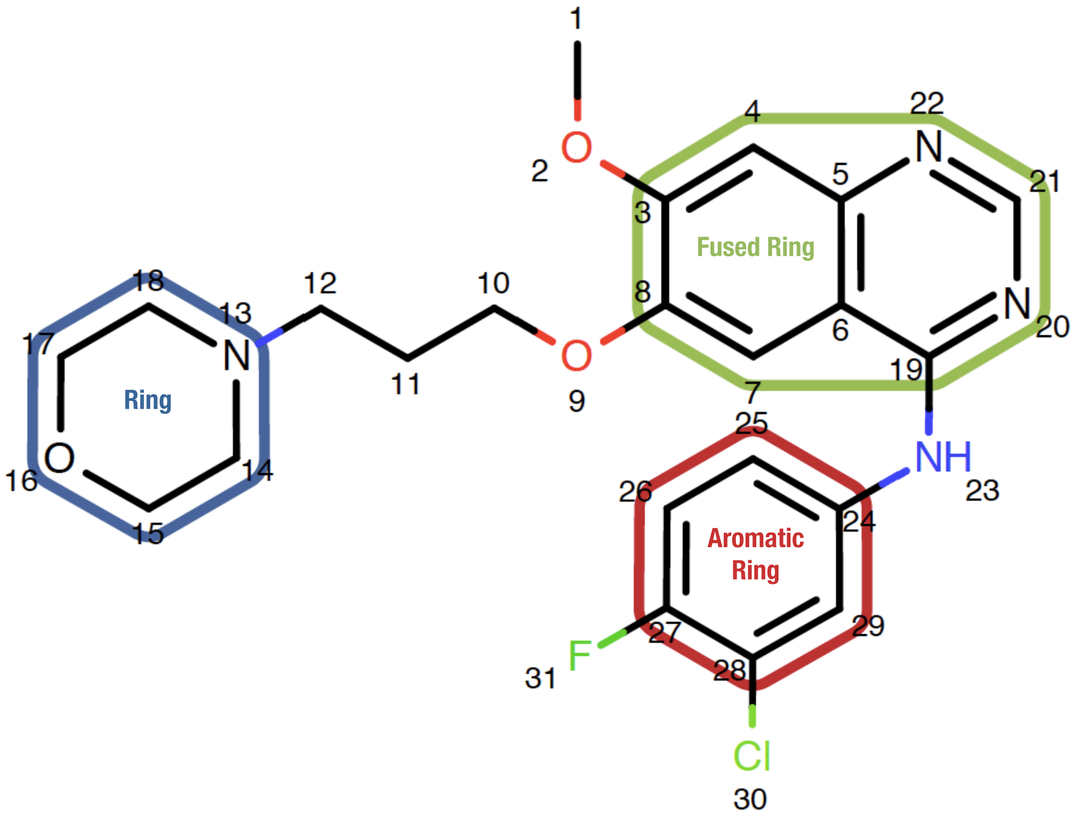
Wiberg Bond Orders of bonds in conjugated systems are correlated with each other

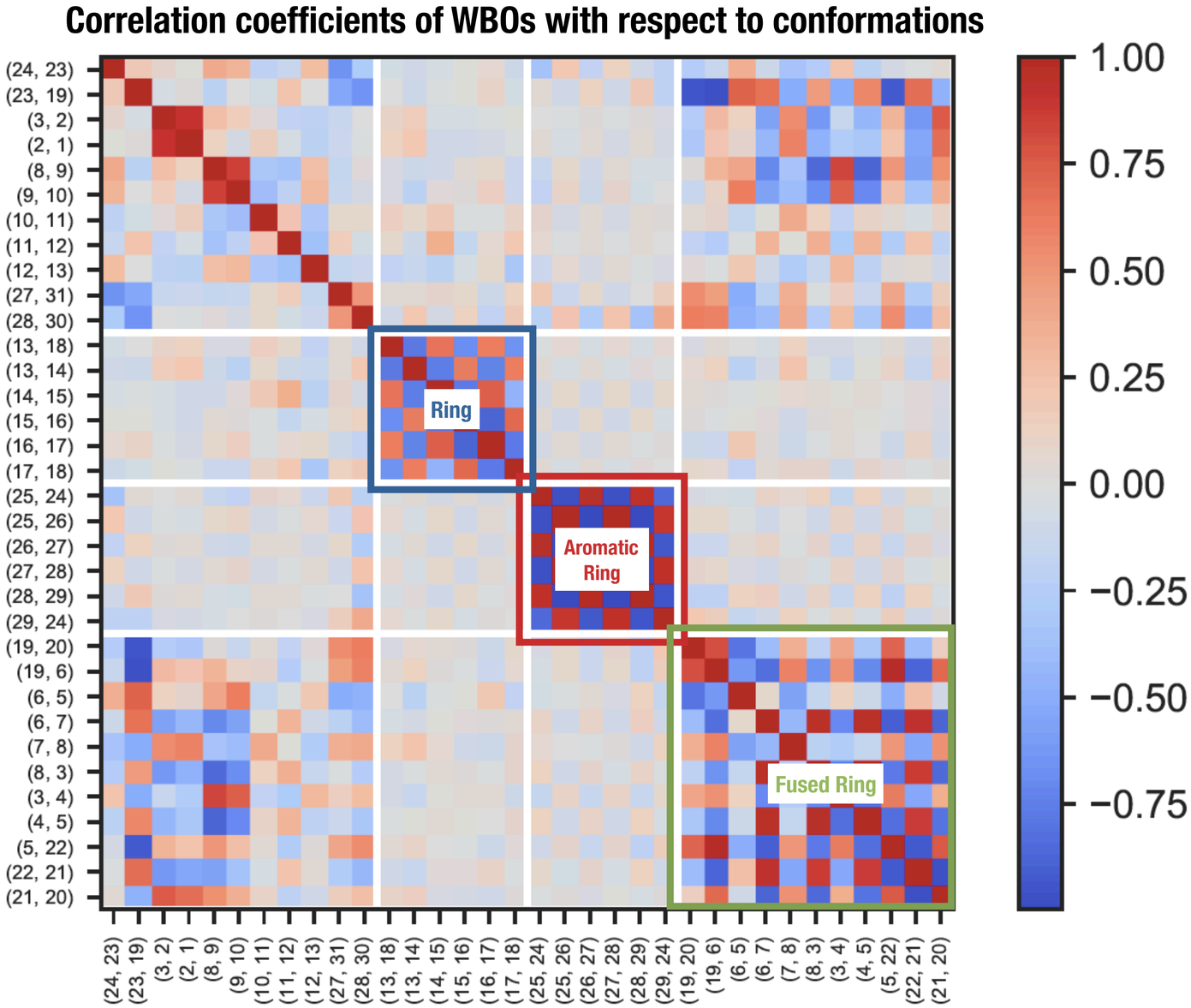


Correlation coefficients of WBOs with respect to conformations

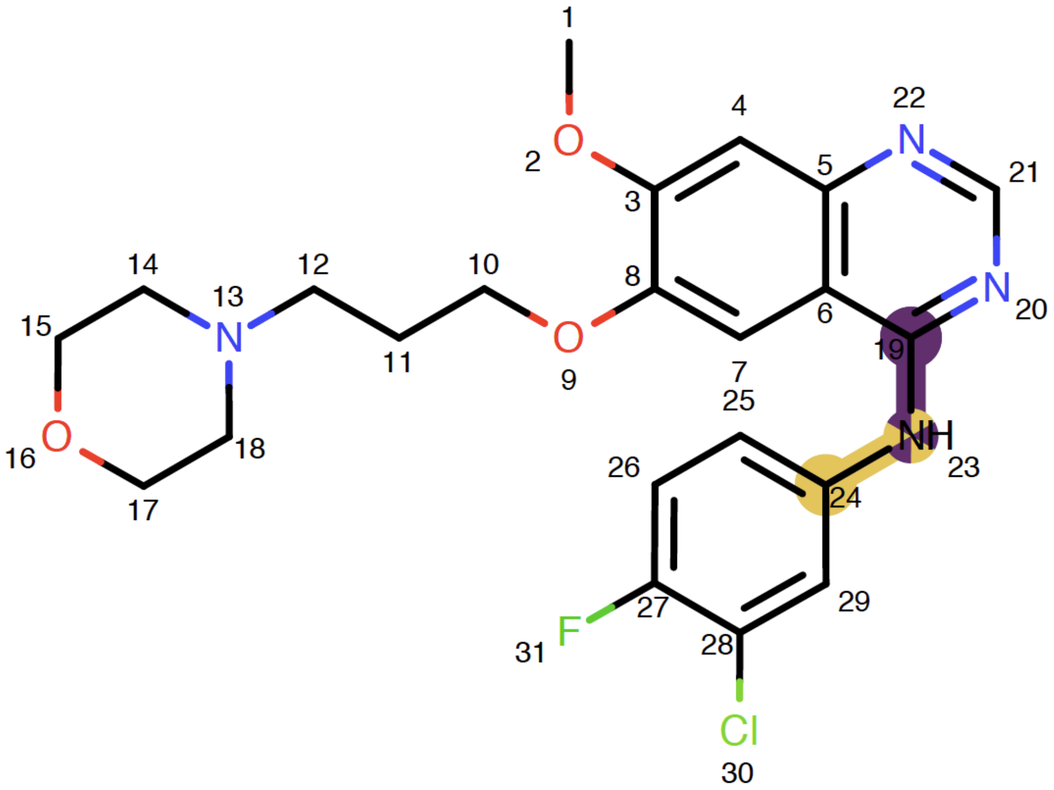


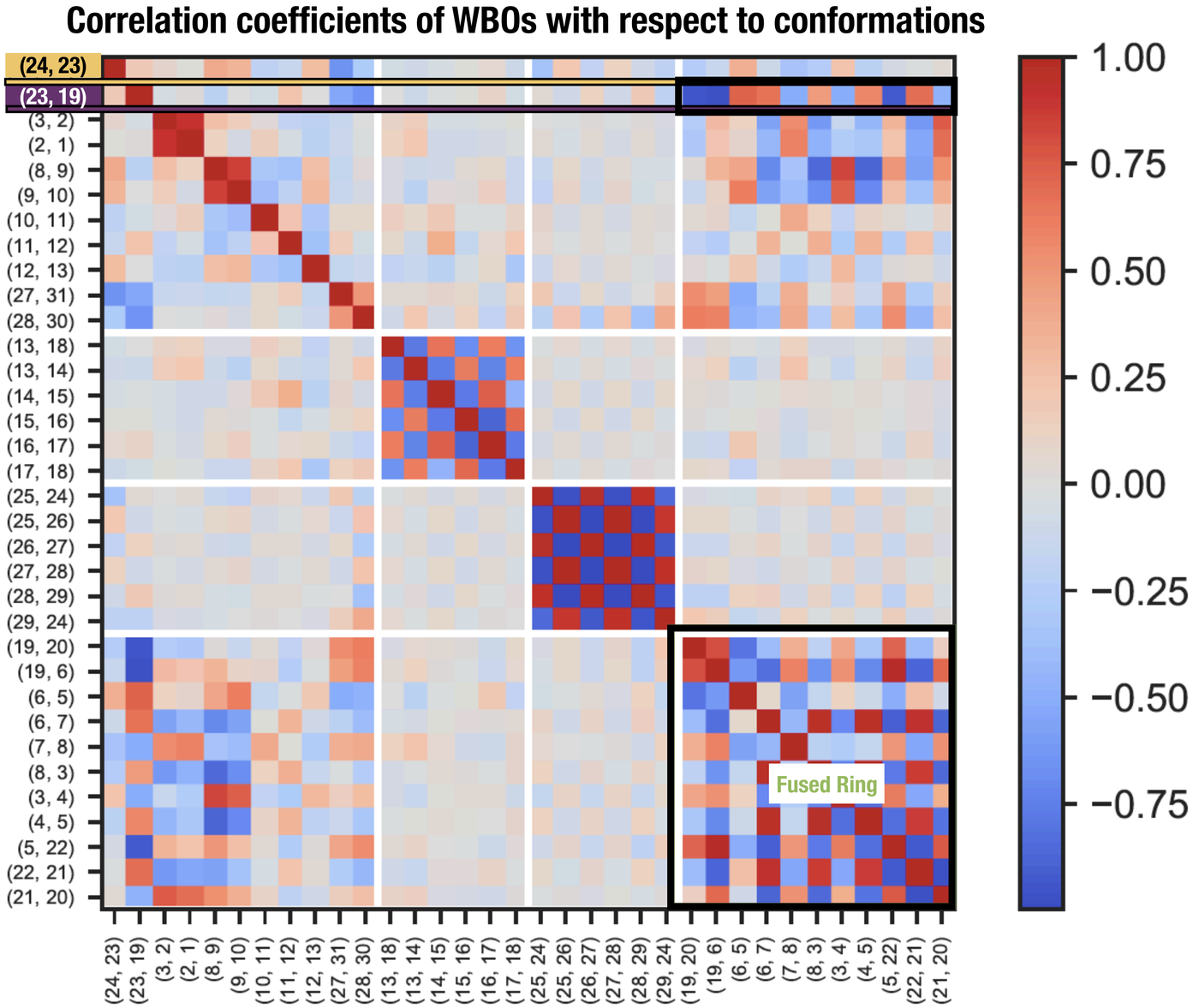
Wiberg Bond Orders of bonds in conjugated systems are correlated with each other



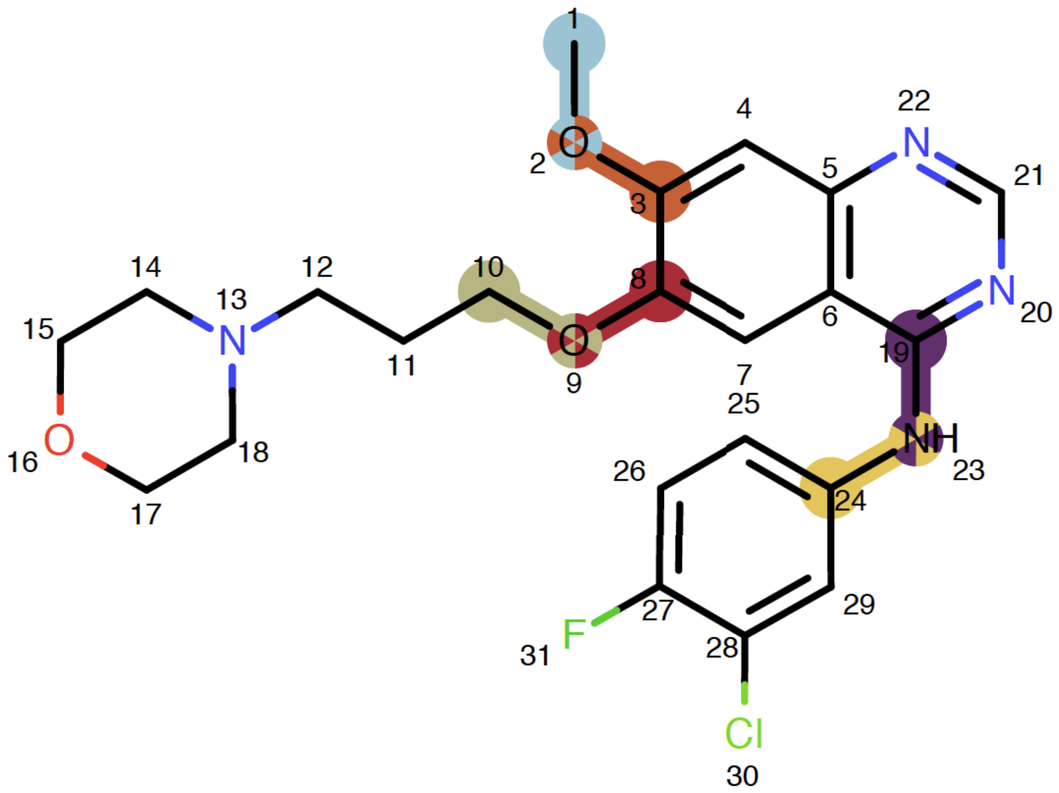


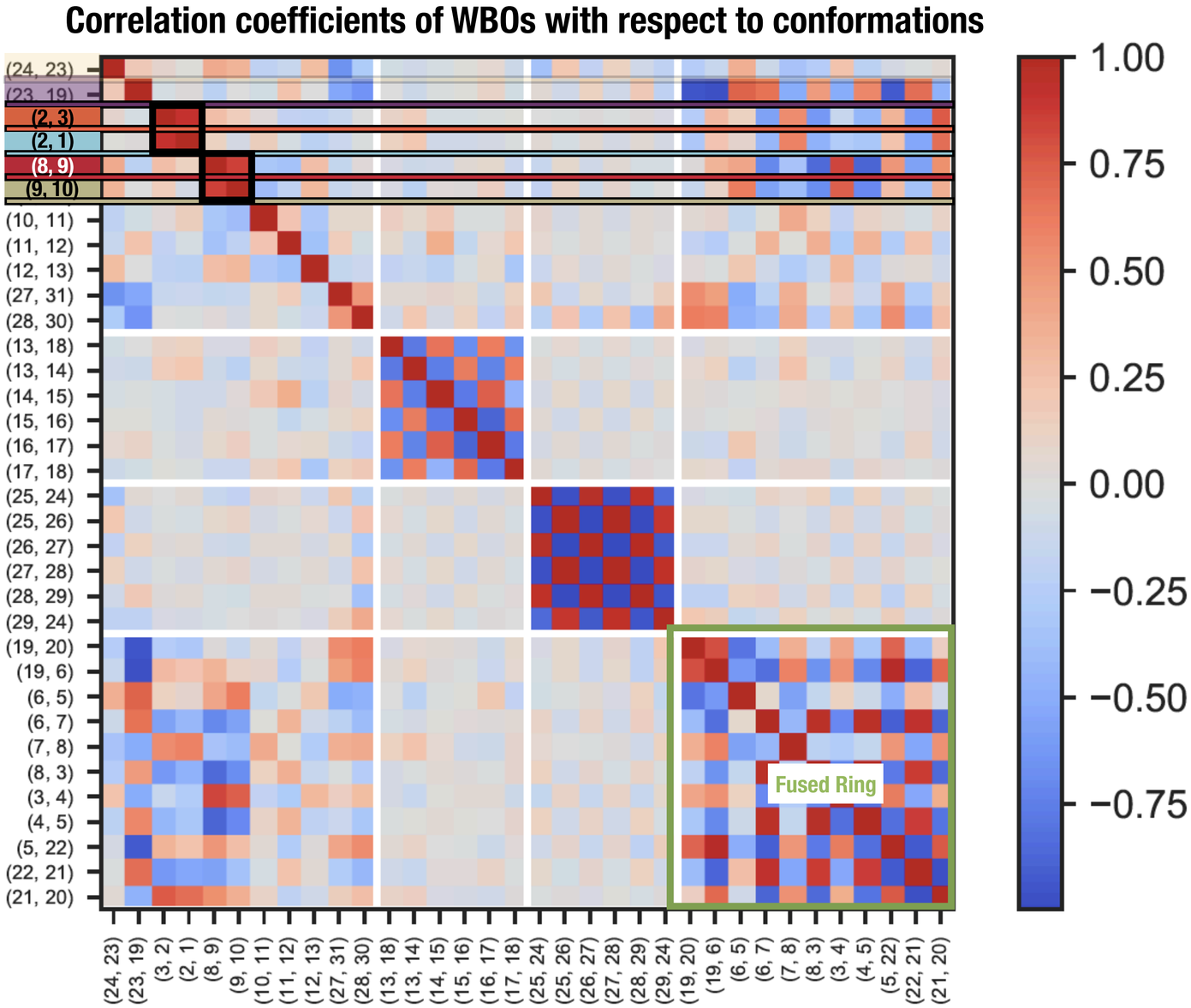
Wiberg Bond Orders of bonds in conjugated systems are correlated with each other



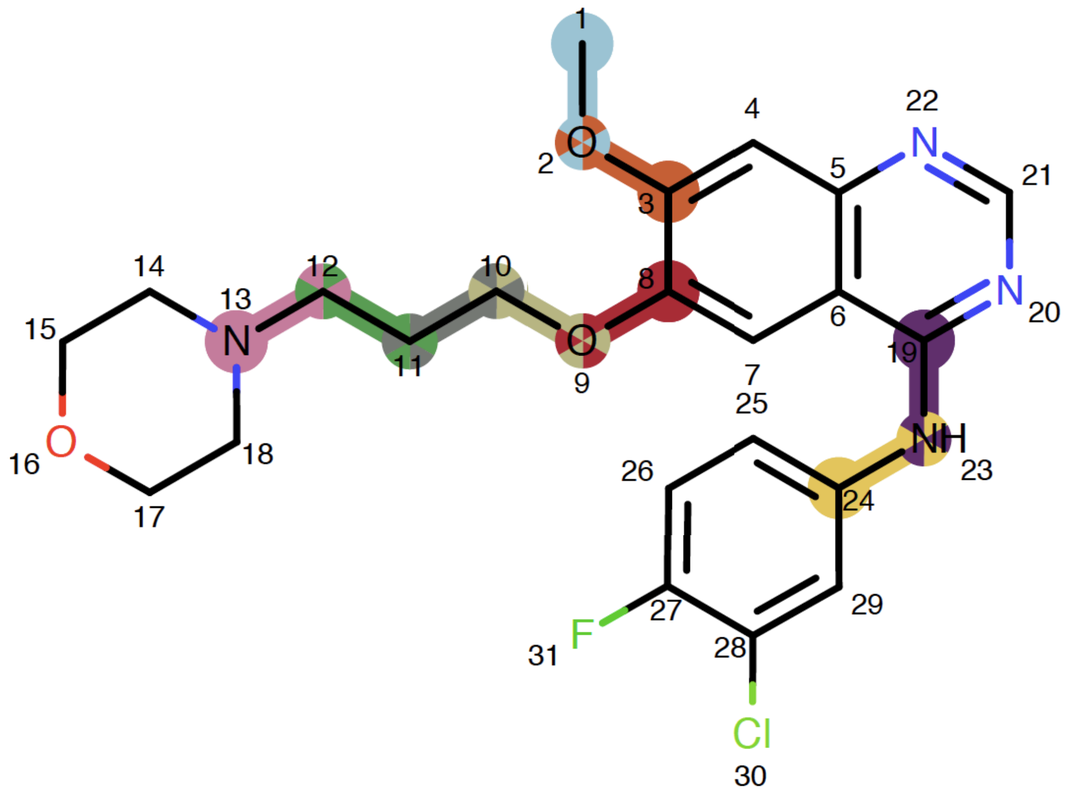


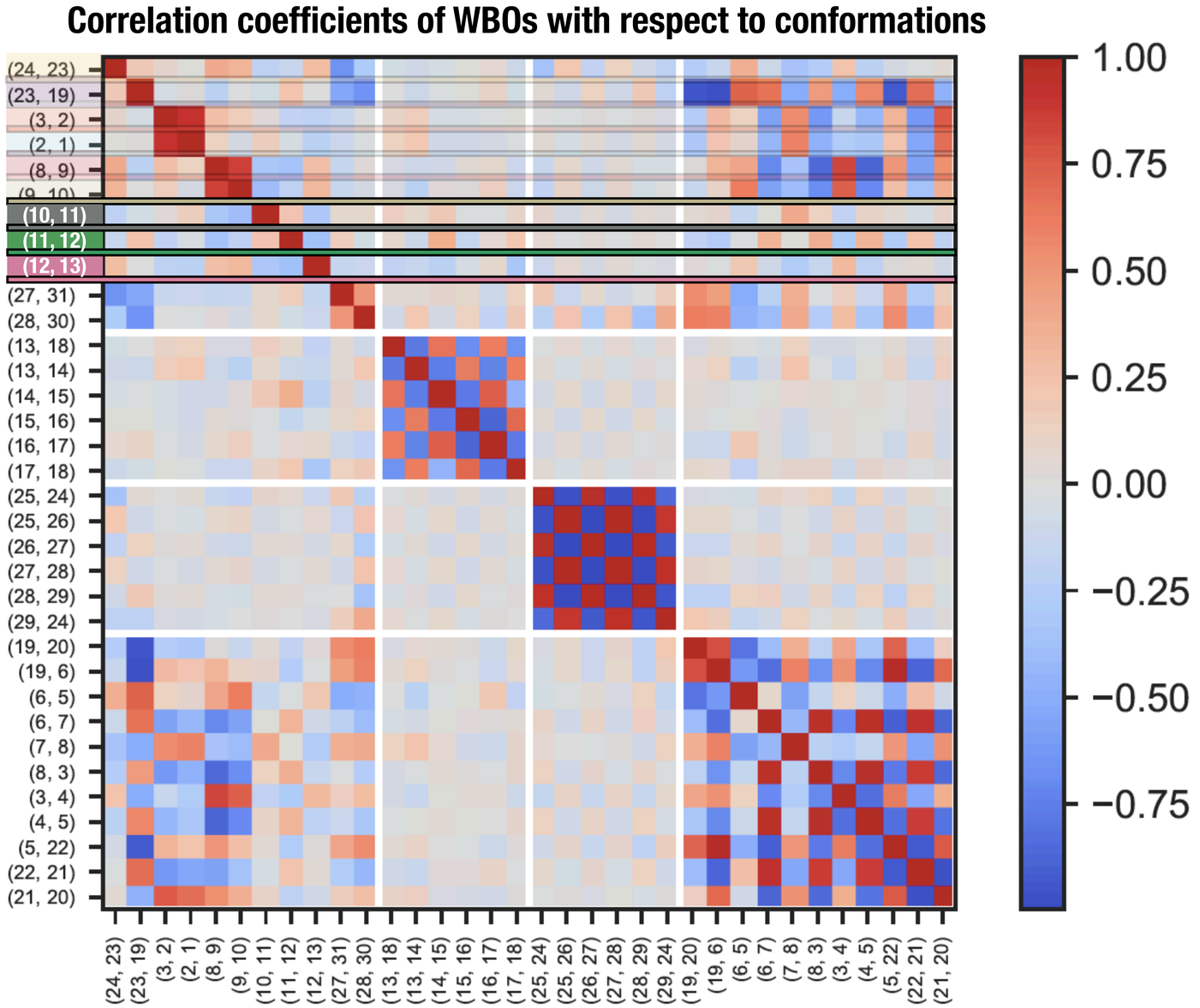
Wiberg Bond Orders of bonds in conjugated systems are correlated with each other



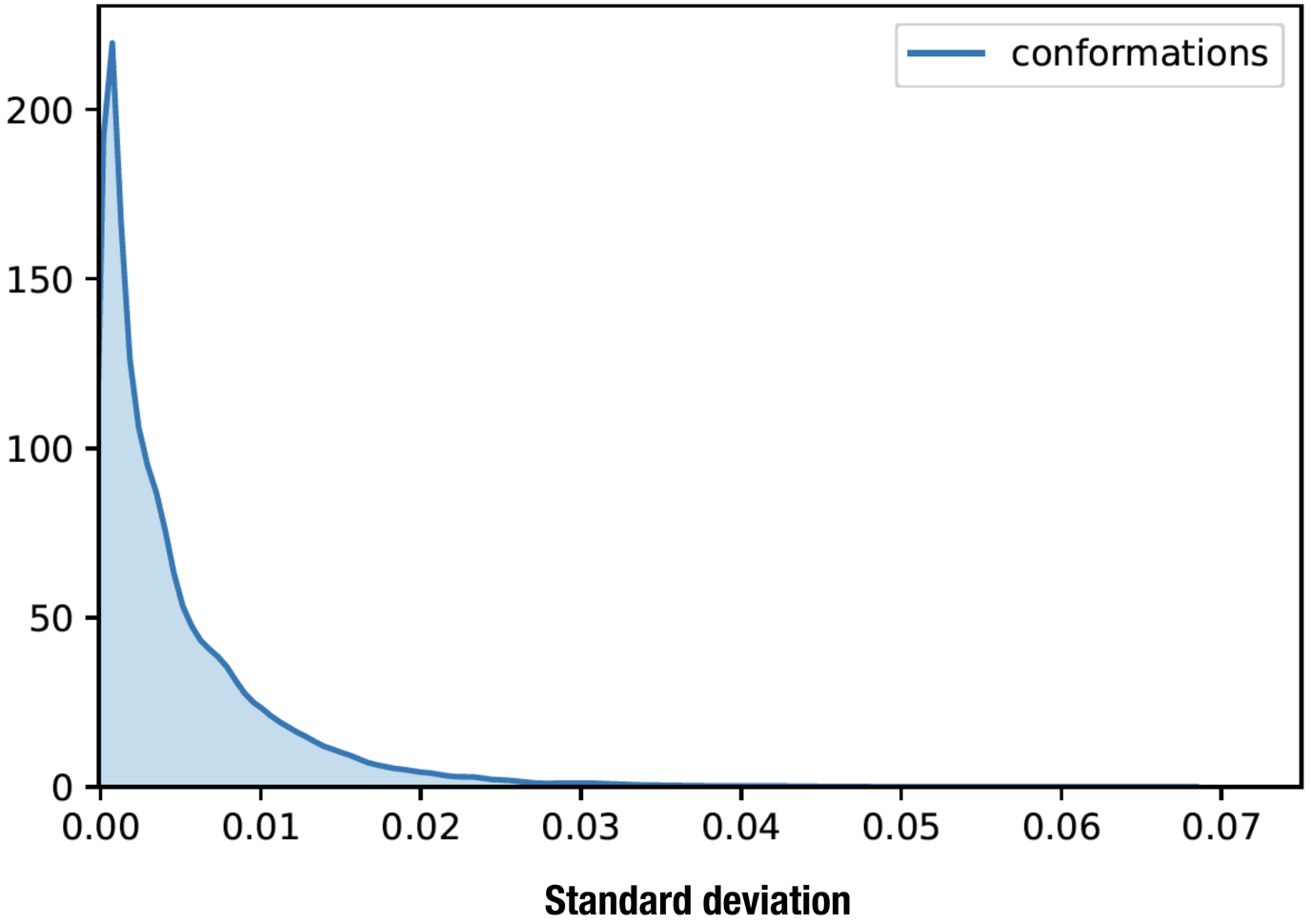


Wiberg Bond Orders of bonds in conjugated systems are correlated with each other



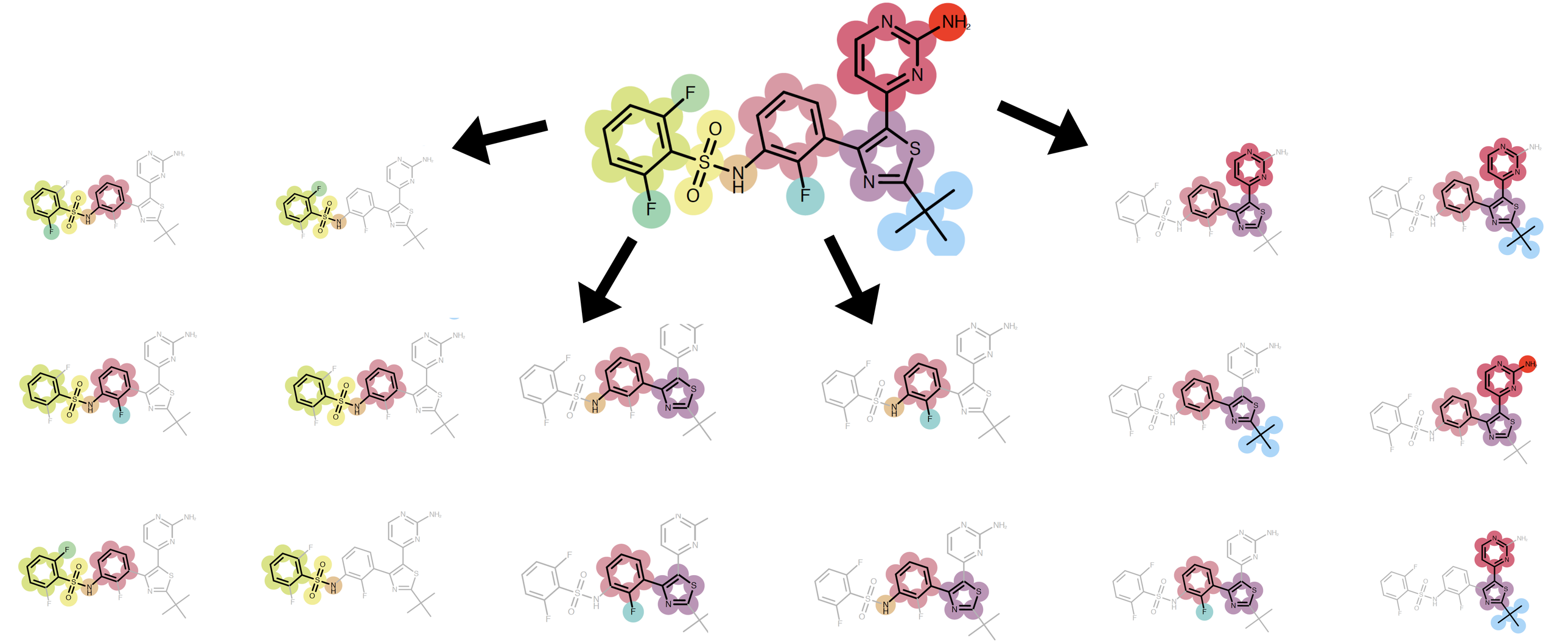


If the Wiberg Bond Order is dependent on conformation, can it be a meaningful indicator?

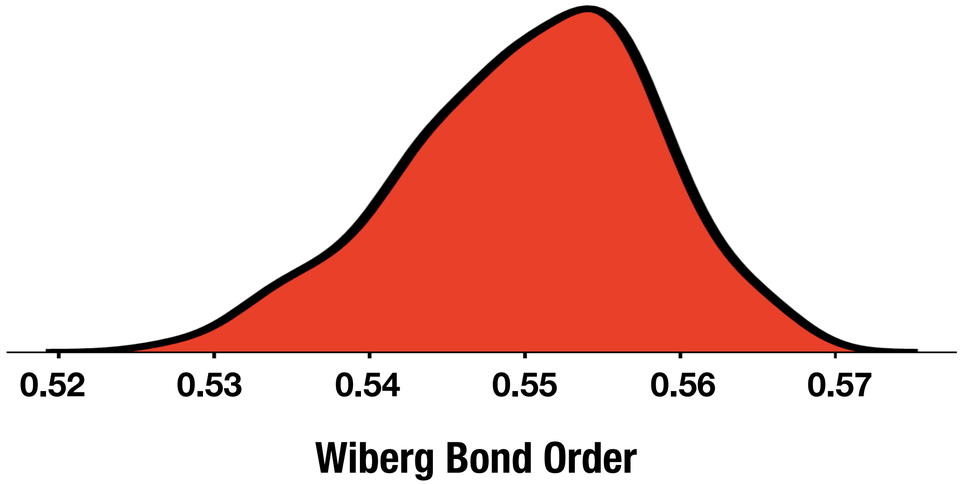


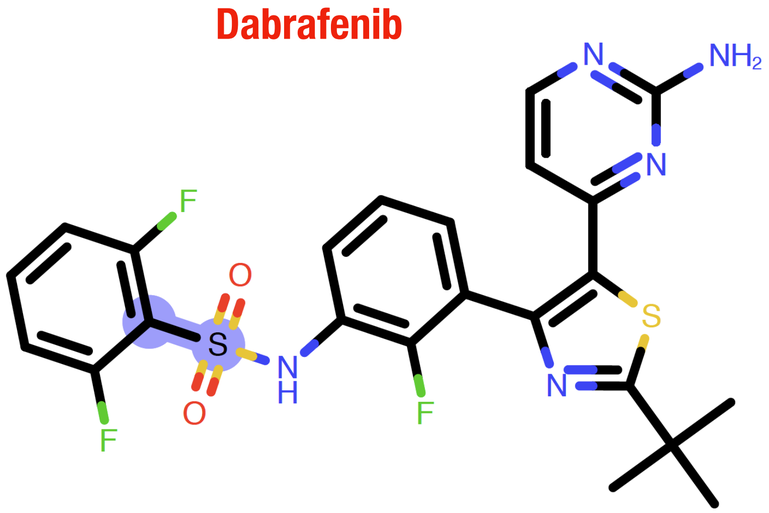
Standard deviations of Wiberg bond orders on the same bond in different conformations (~2,000 molecules)

Generate all possible fragments of kinase inhibitors to see how WBO of bonds change in different chemical environments

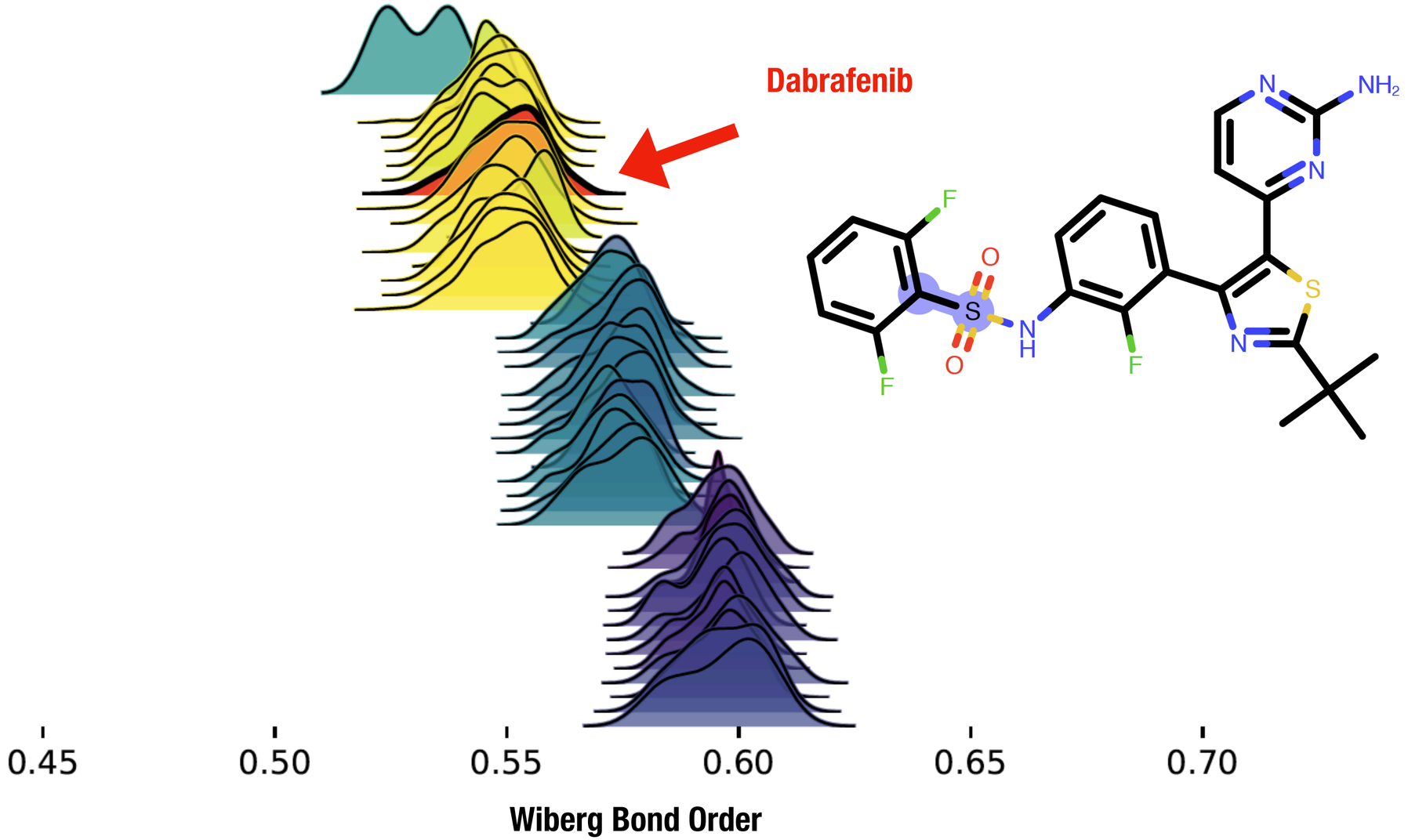


Distribution of Wiberg Bond Order for the bond in its parent chemical environment

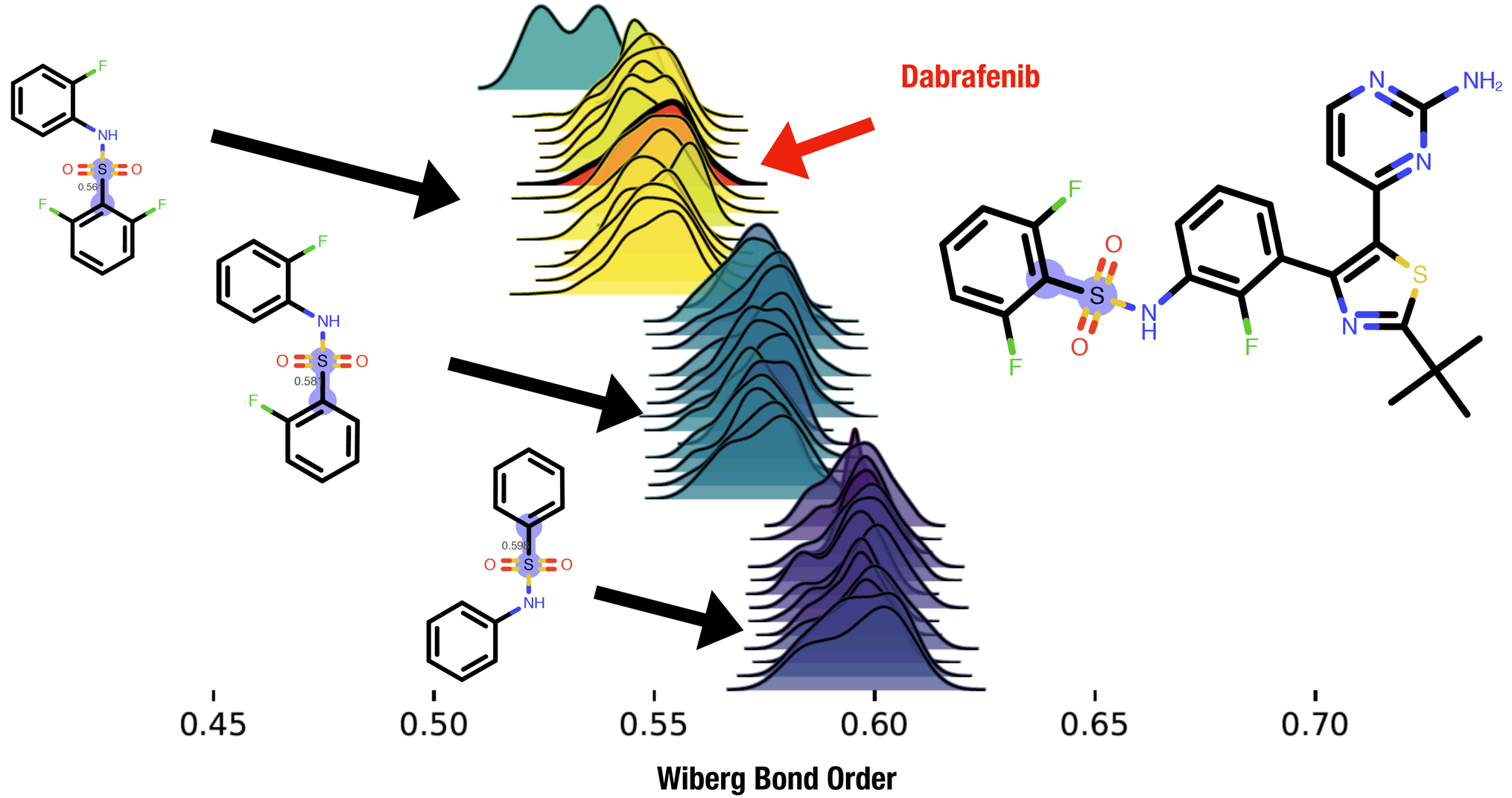




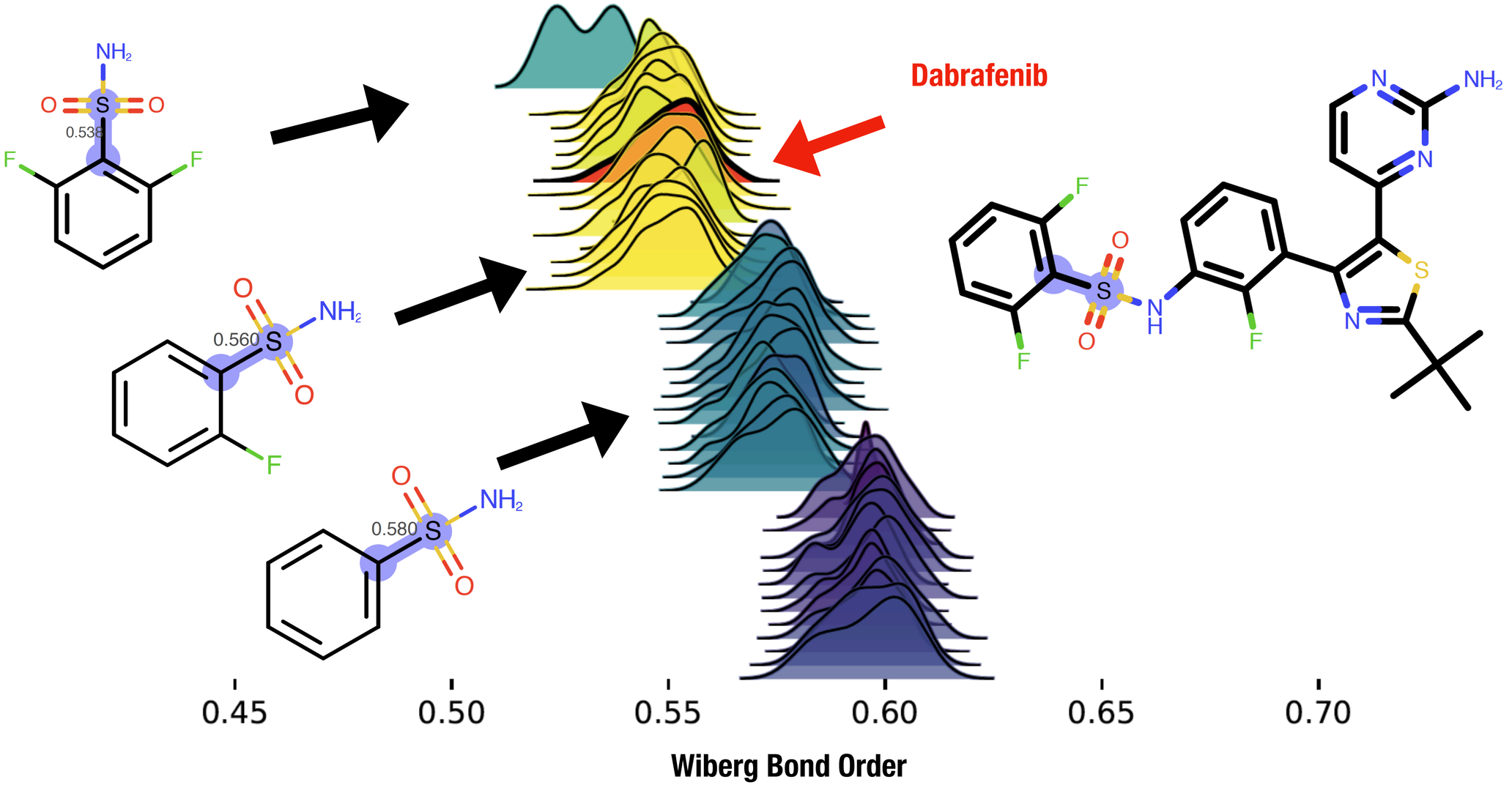
The distribution of the WBO of the same bond in different fragments fall into distinct bins



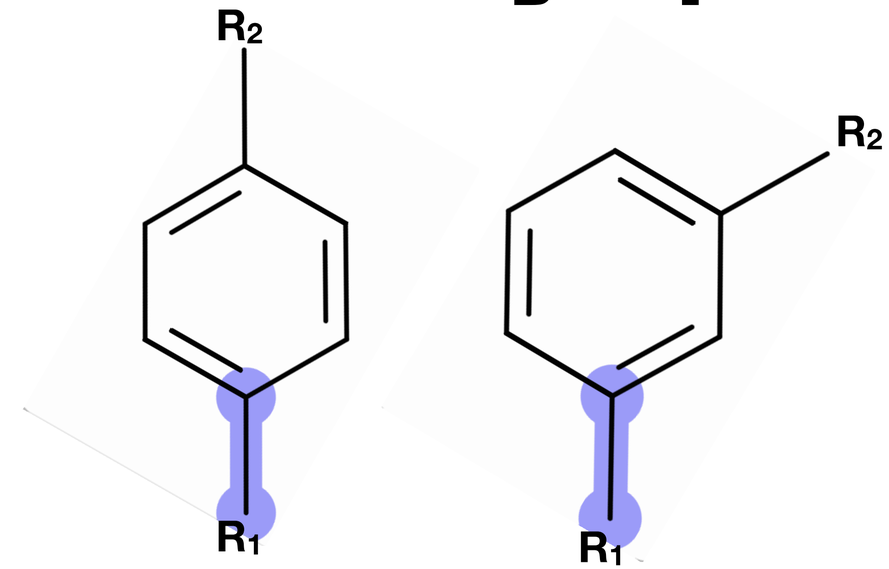
Removal of fluorine shifts the Wiberg bond order of the bond

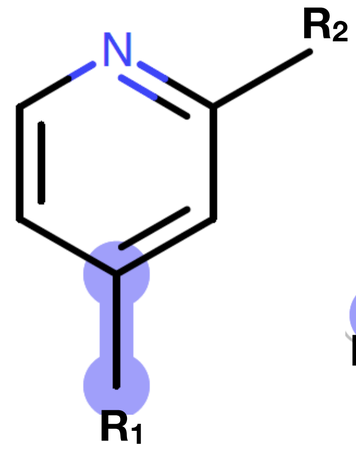


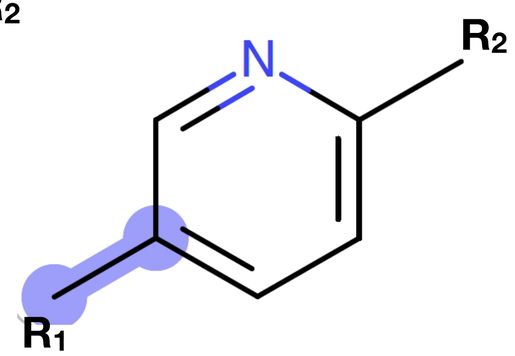
Removal of the ring shifts the Wiberg Bond Order in the opposite direction



Isolating resonance effect from steric effects show how sensitive some functional groups are to chemical changes several bonds away

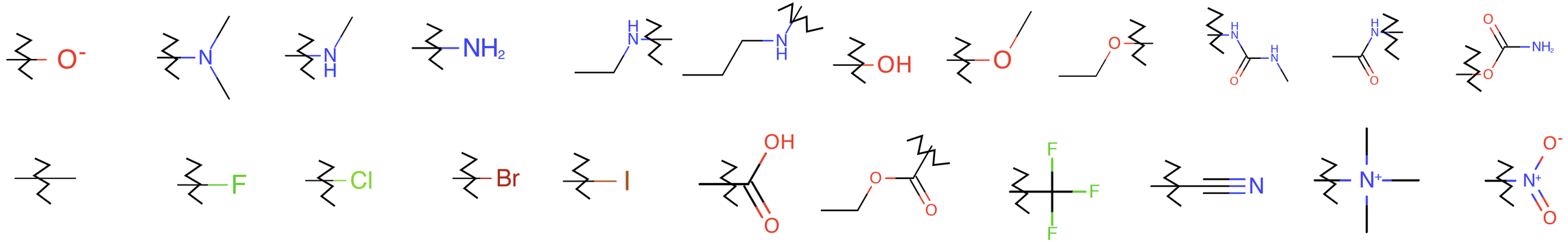






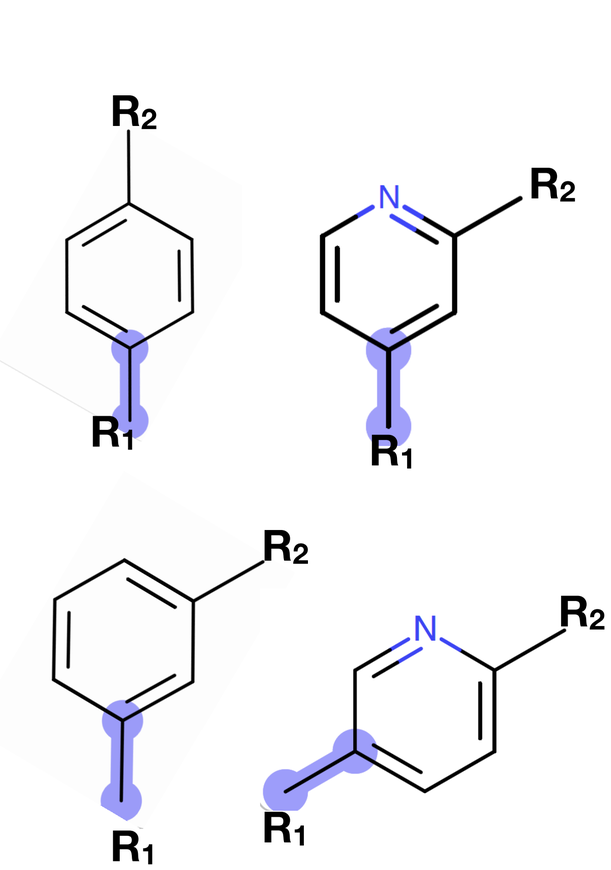
Electron donating

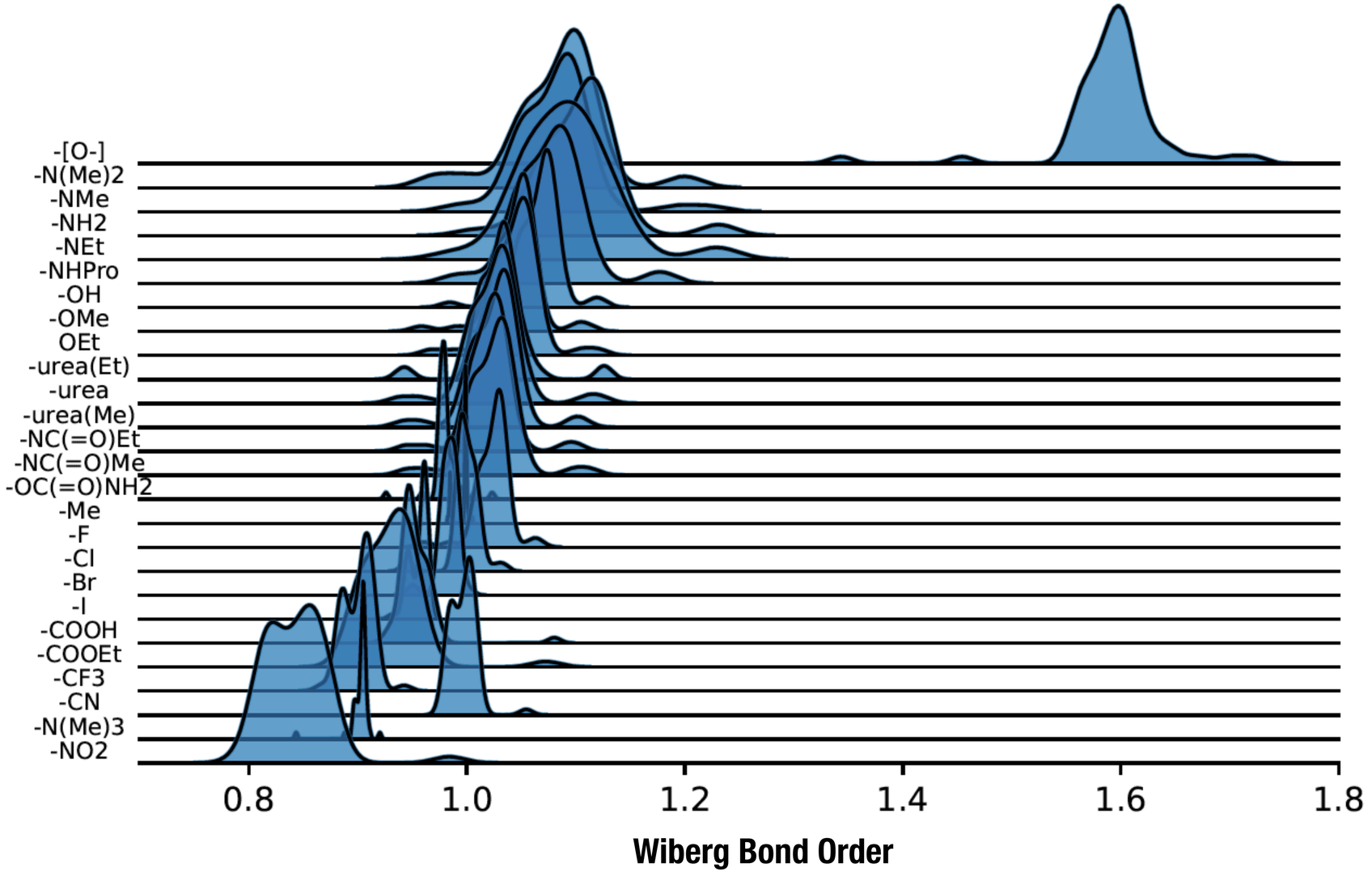
R_1, R_2 :



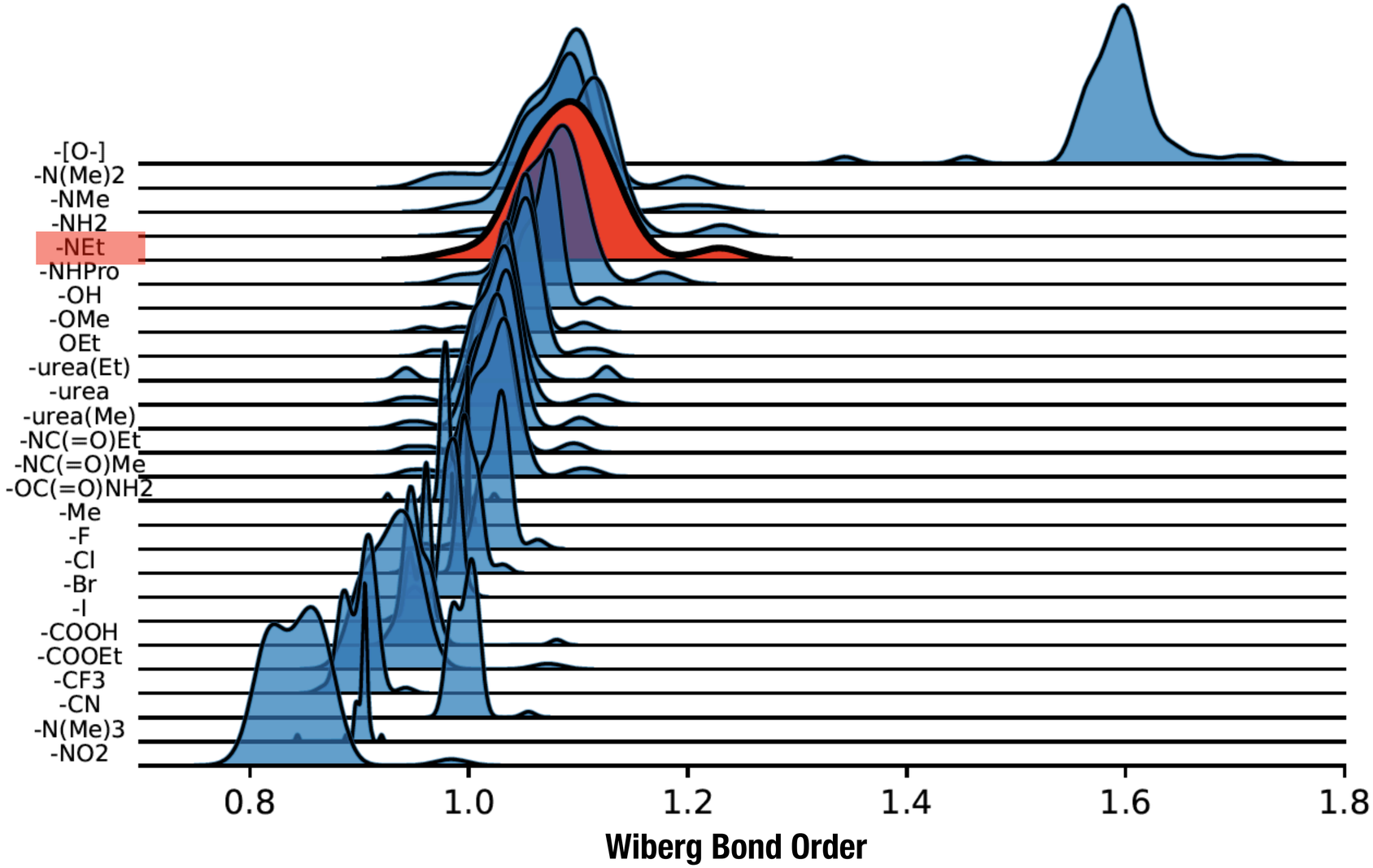
Electron withdrawing

Changes of substituents far from the central bond changes the Wiberg Bond order of that bond

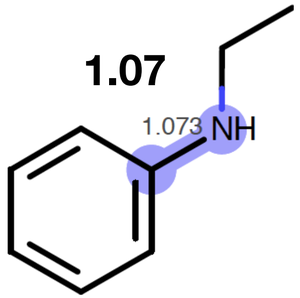


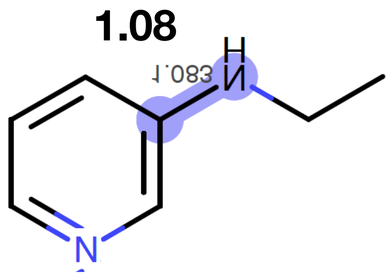


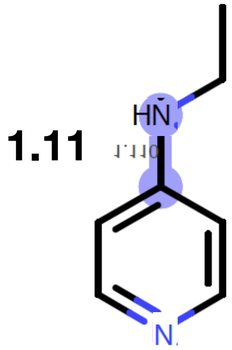
Isolating resonance effect from steric effects show how sensitive some functional groups are to chemical changes several bonds away

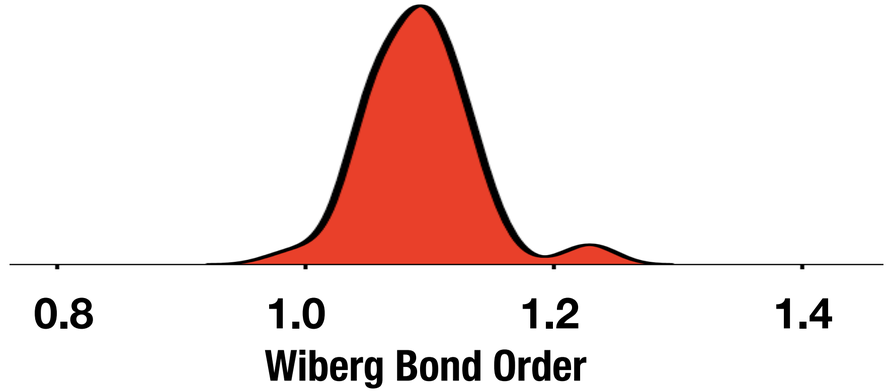


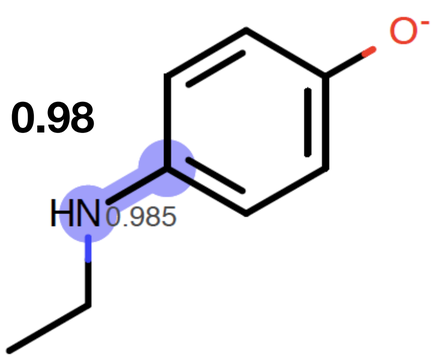
Electron withdrawing groups para to R_1 increase the Wiberg Bond Order for electron donating R_1

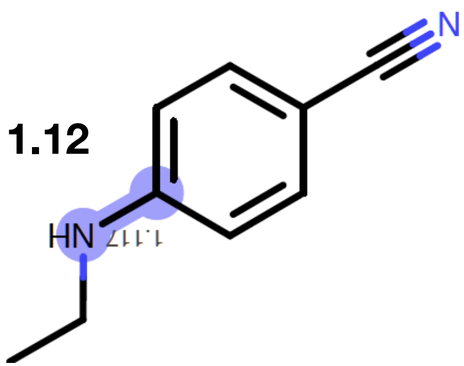


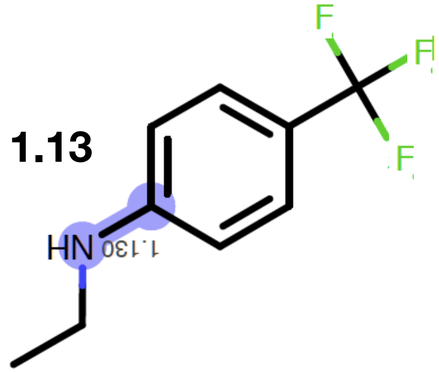


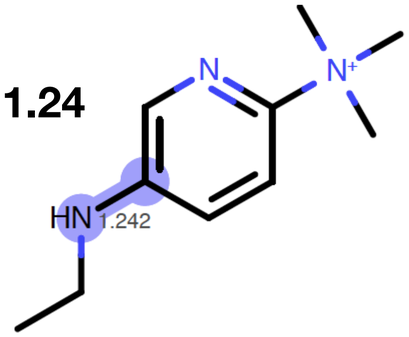






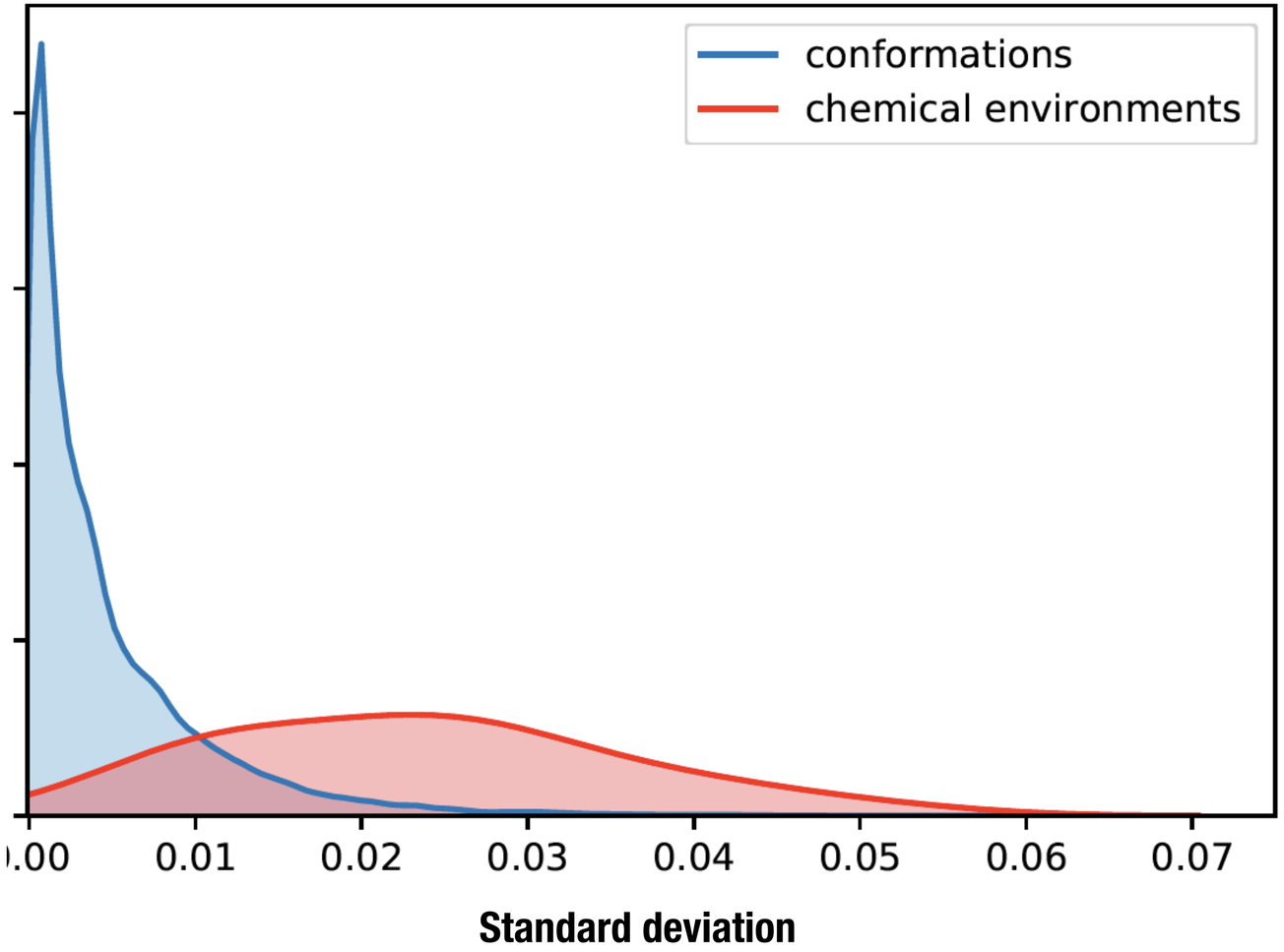






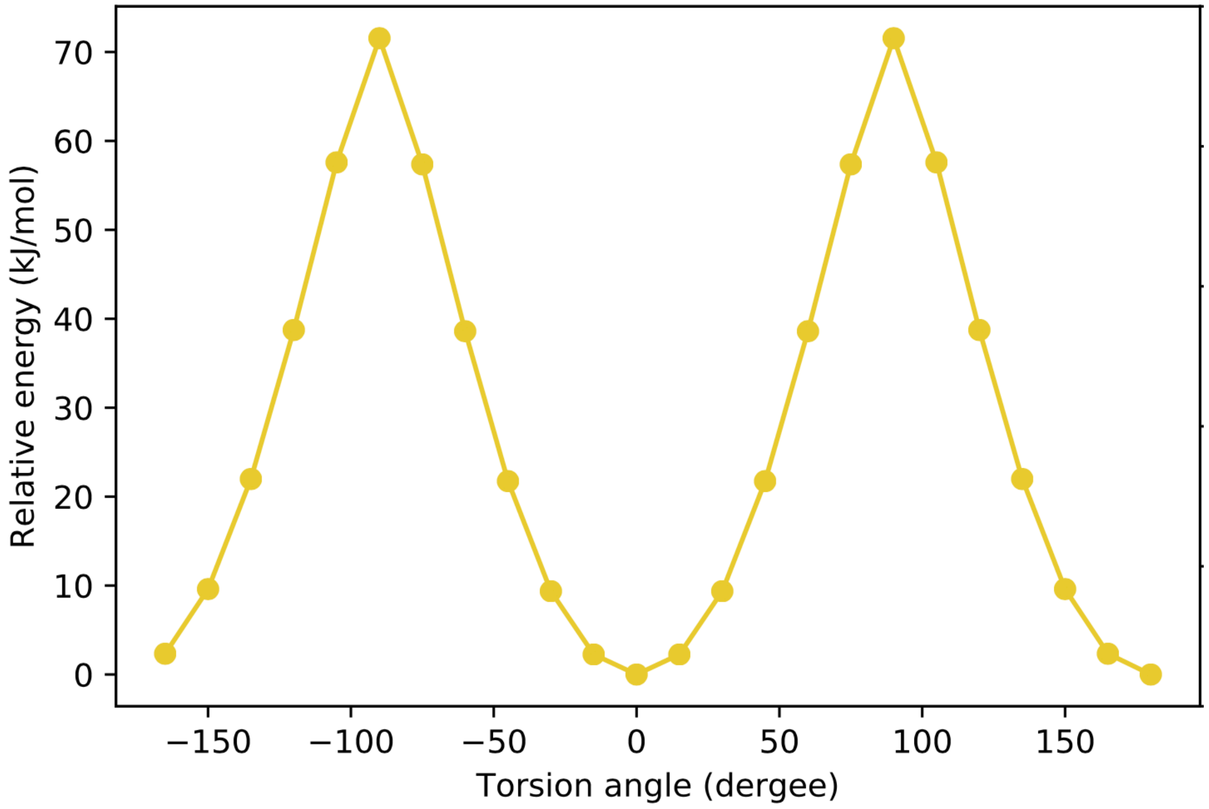
Opposite trend for electron withdrawing R_1 groups

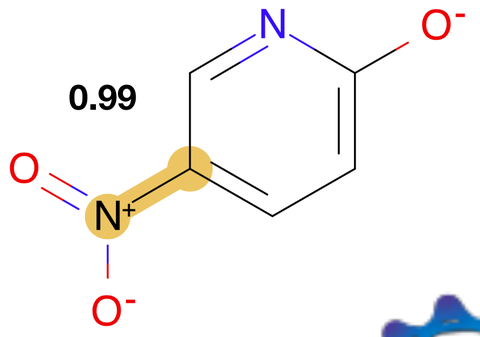
Changes in WBO due to conformation are smaller than changes in WBO due to different chemical environments



Standard deviations of Wiberg Bond Orders on the same bond in different chemical environments (~2000 molecules)

The trend of increasing torsion barrier height with increasing WBO is there even for lower WBOs



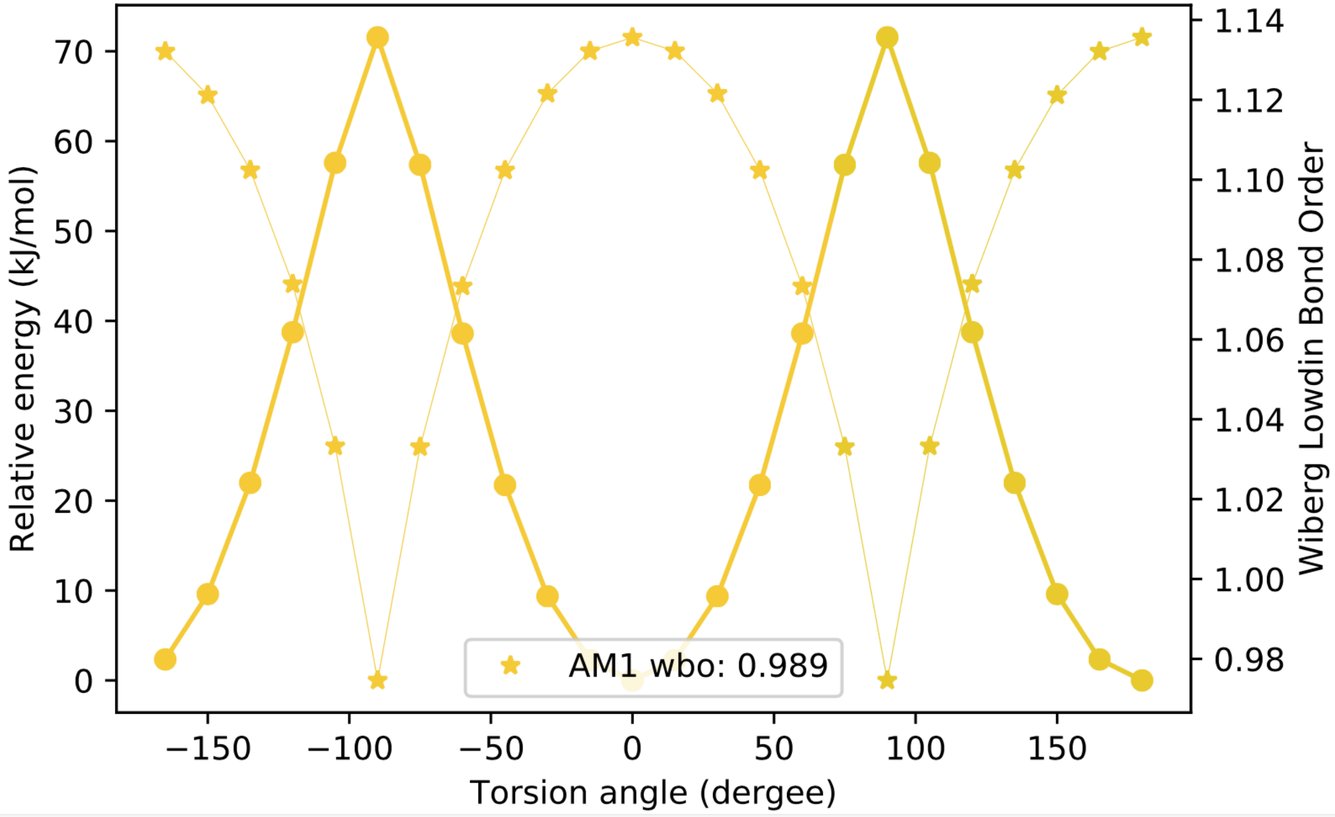


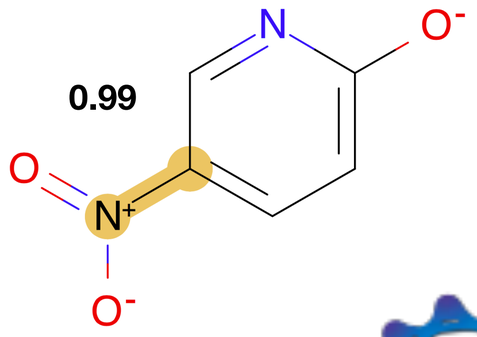
QC Archive
A MolSSI Project

The data is available on QCArchive

http://docs.qcarchive.molssi.org/en/latest/basic_examples/torsiondrive_datasets.html

Wiberg bond order is anti correlated with the torsion scan



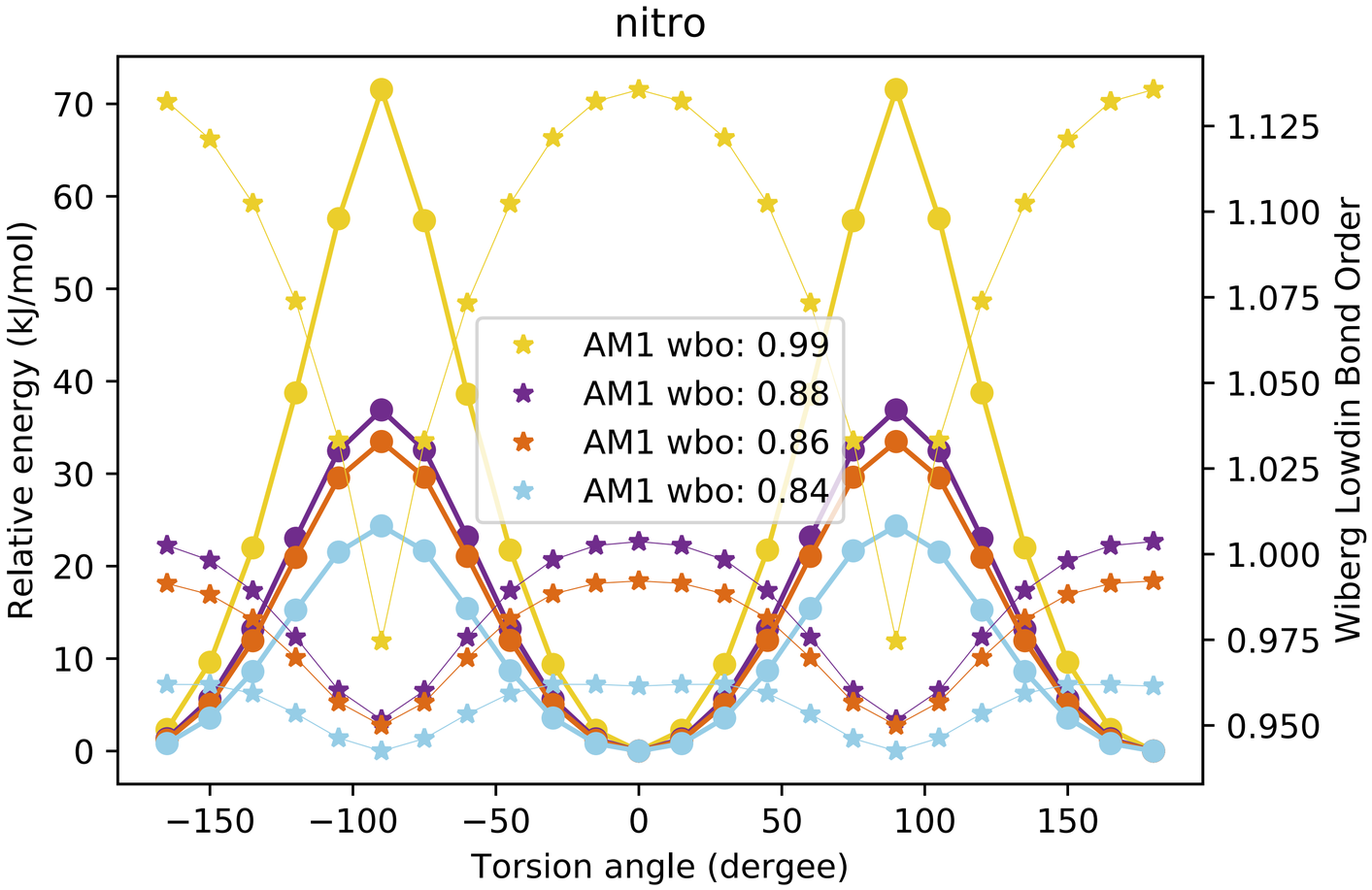


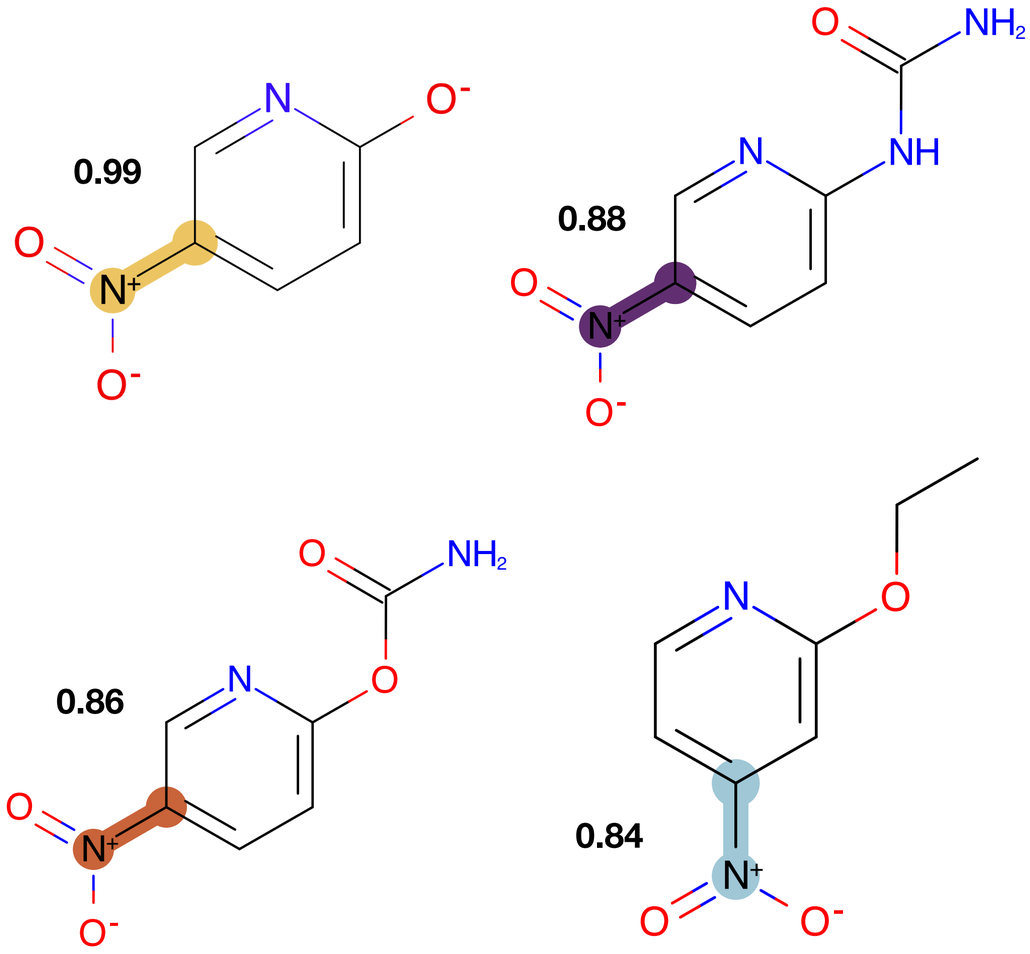
QCArchive
A MolSSI Project

The data is available on QCArchive

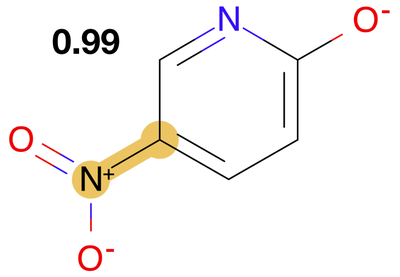
http://docs.qcarchive.molssi.org/en/latest/basic_examples/torsiondrive_datasets.html

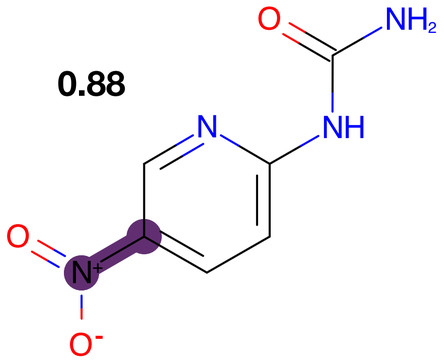
The trend of increasing torsion barrier height with increasing WBO is there even for lower WBOs

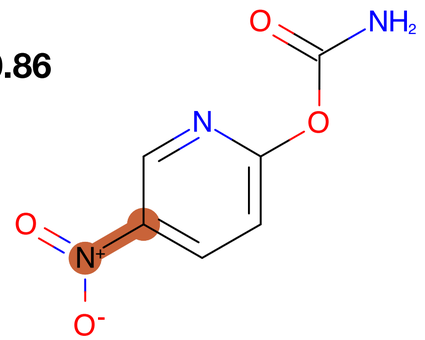


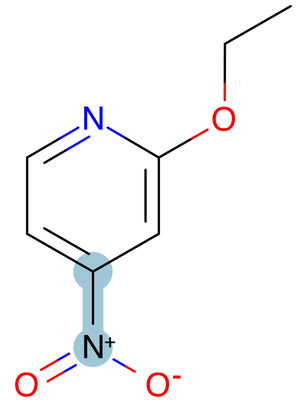


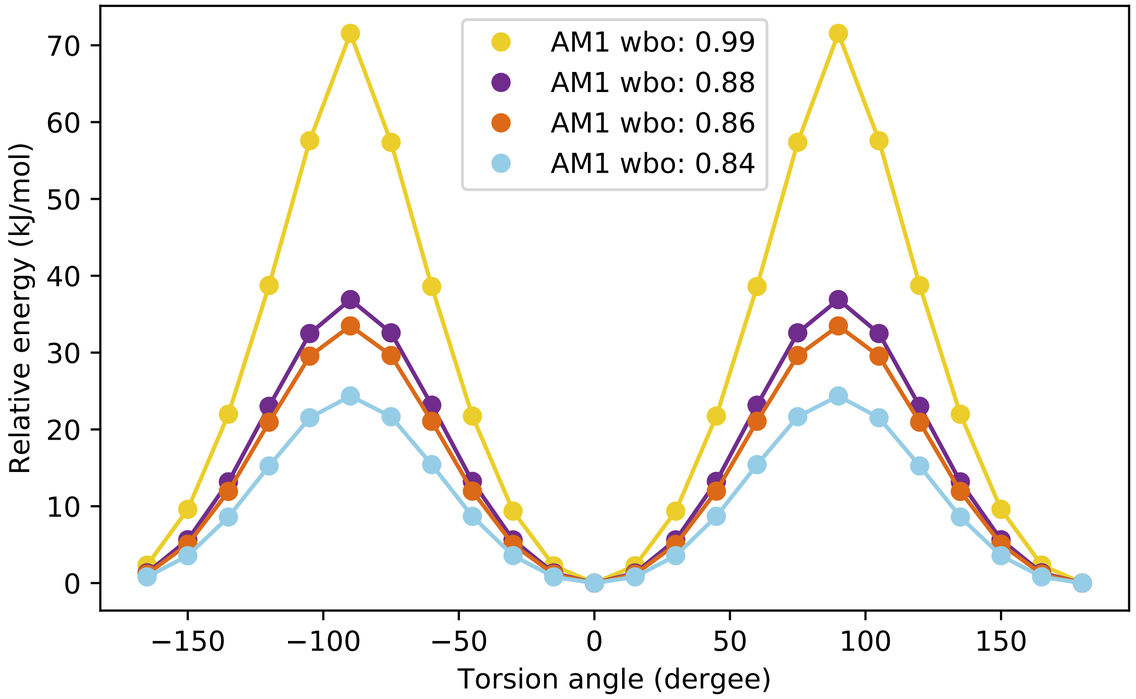
Torsion scans on combinatorial phenyl set reveal correlation of Wiberg Bond Order with torsion barrier height

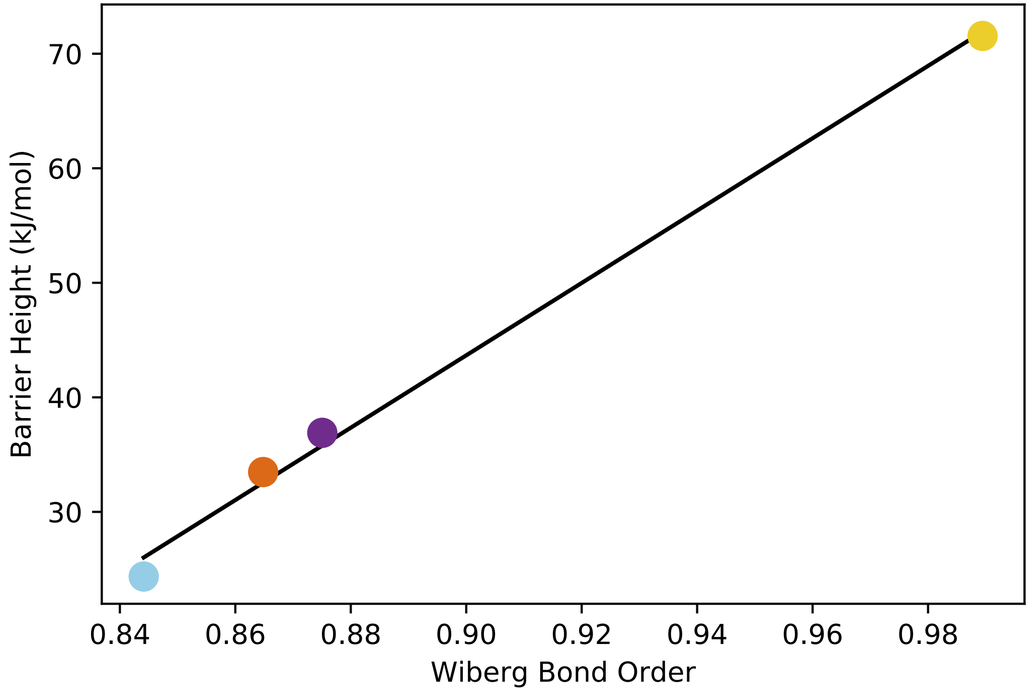




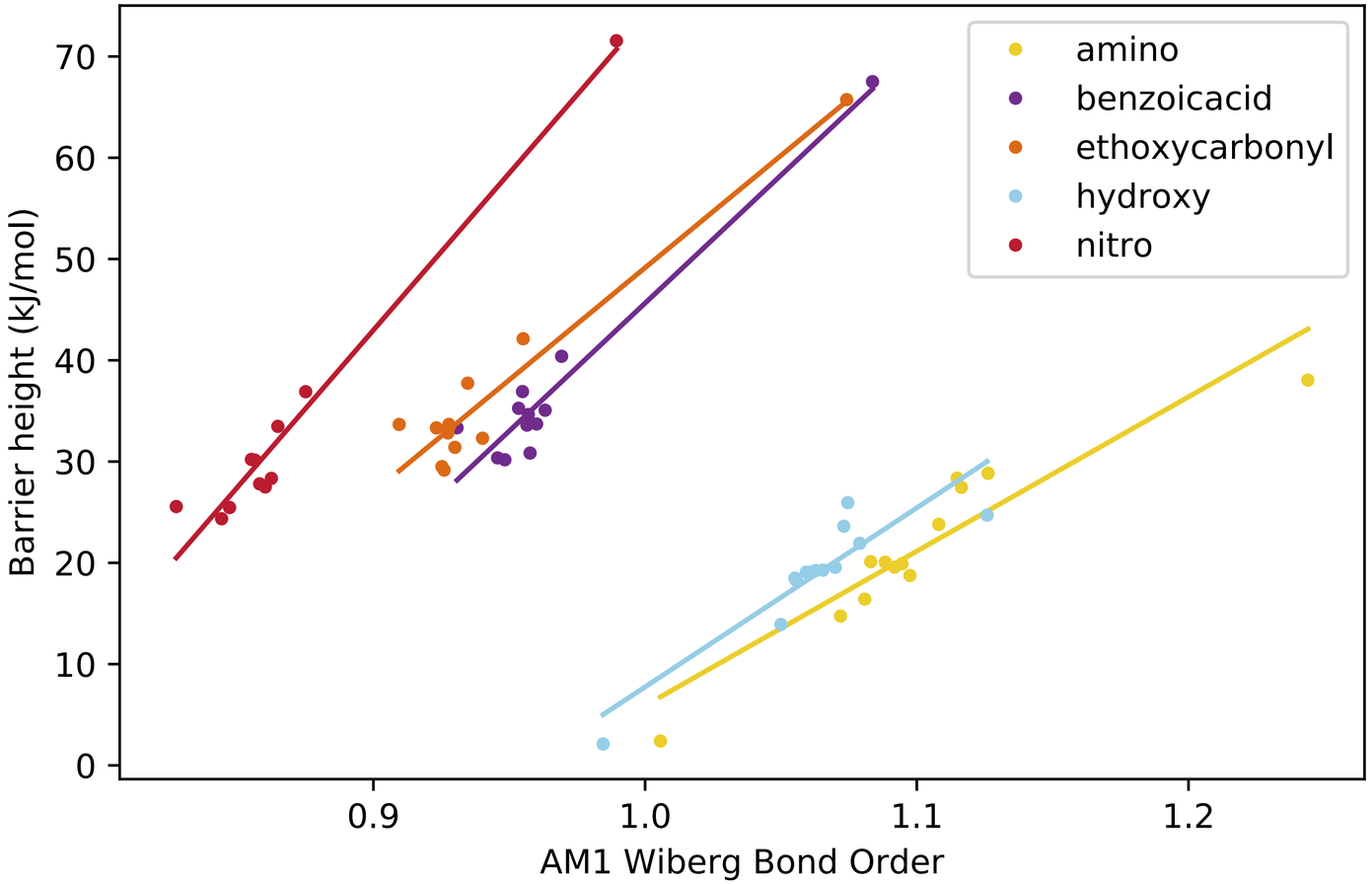








Torsion barrier height increases with increasing Wiberg Bond Order for other functional groups

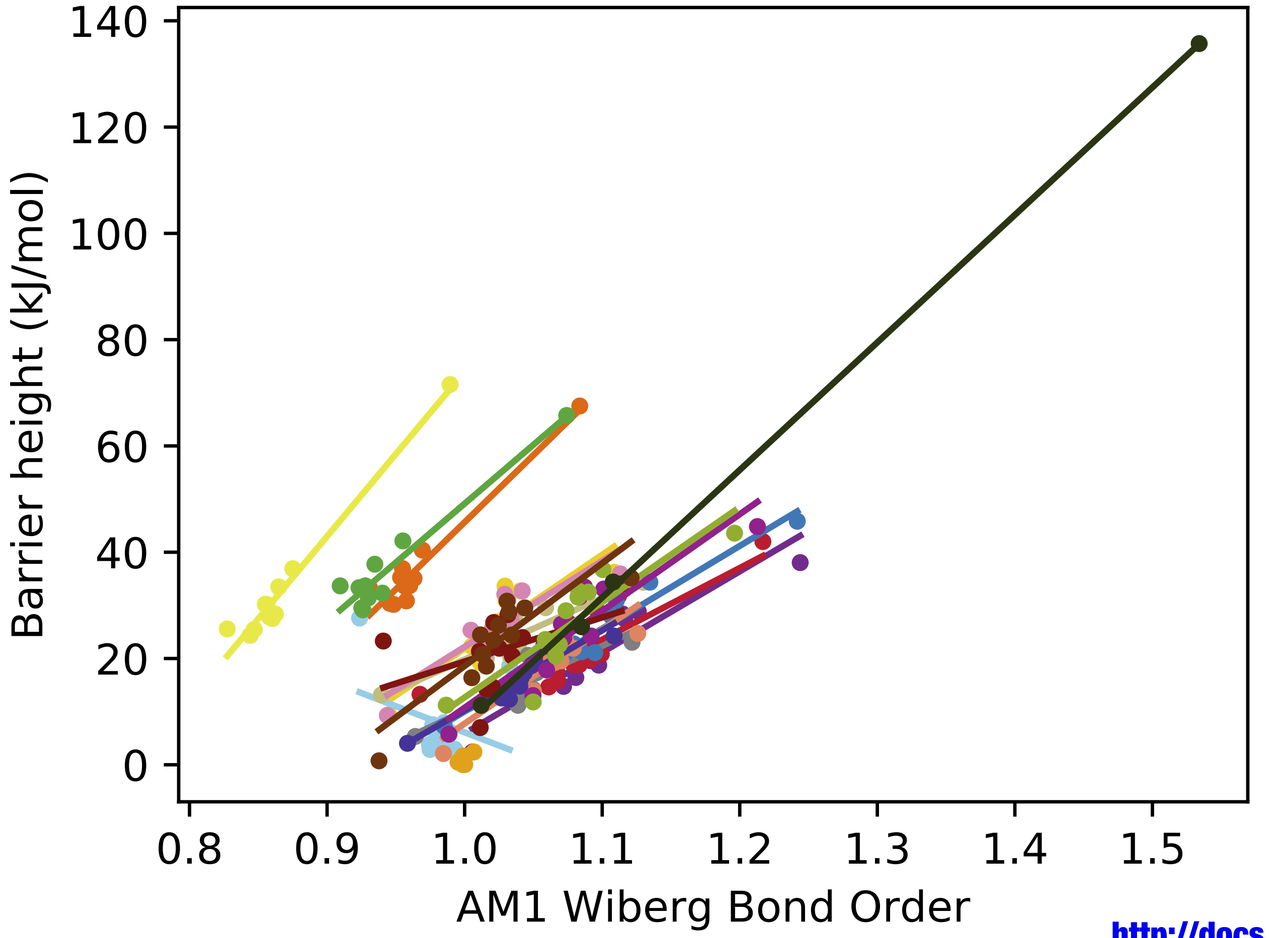


QCArchive
A MolSSI Project

The data is available on QCArchive

http://docs.qcarchive.molssi.org/en/latest/basic_examples/torsiondrive_datasets.html

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



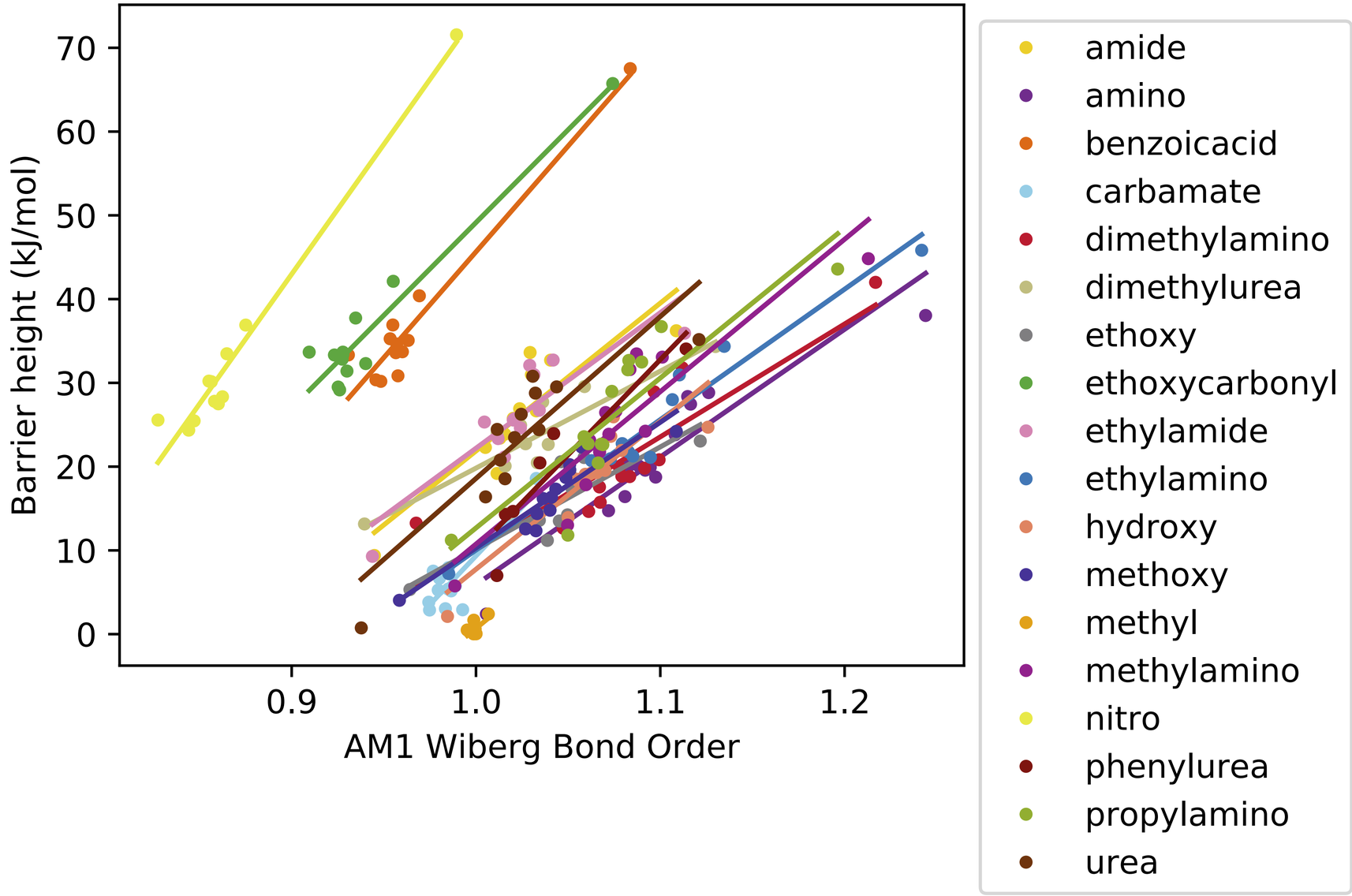
QCArchive
A MolSSI Project

The data is available on QCArchive

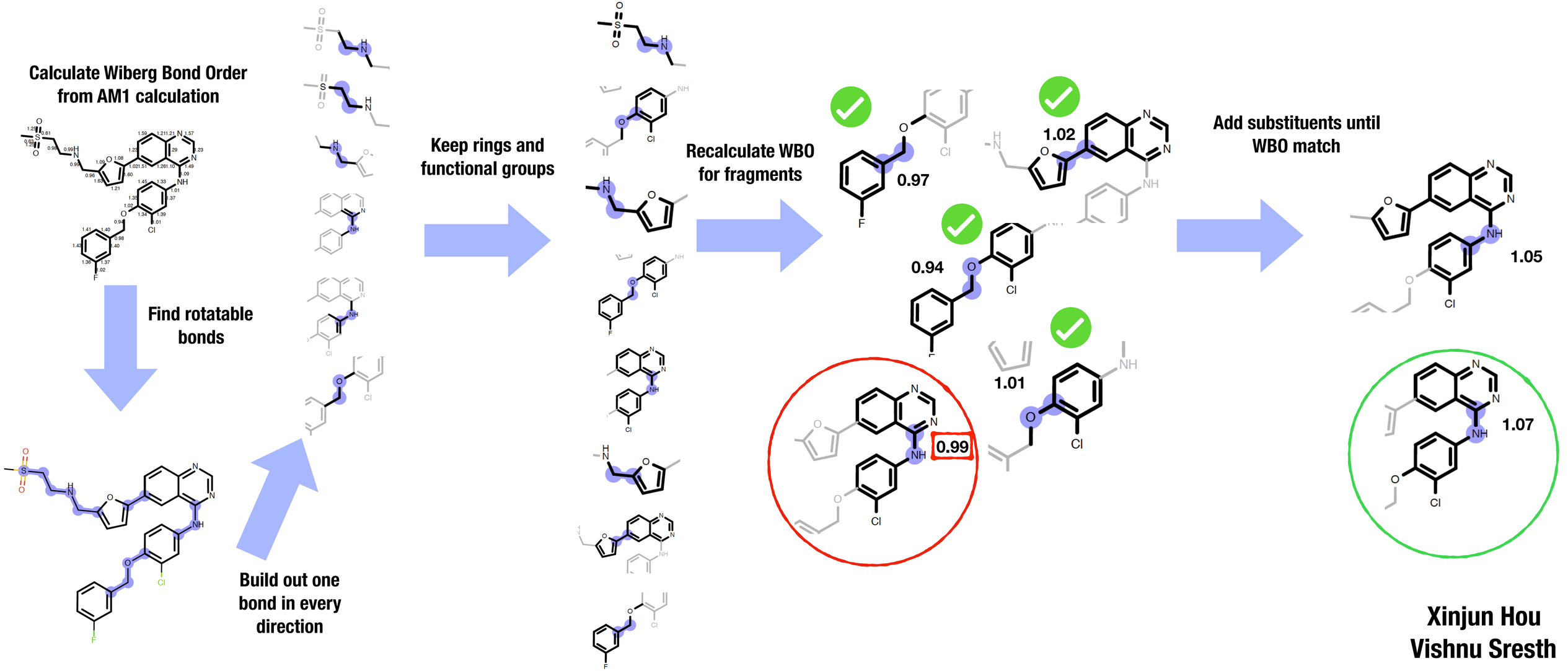
http://docs.qcarchive.molssi.org/en/latest/basic_examples/torsiondrive_datasets.html

There are two trends for outliers

1. **Trivalent nitrogen is pyramidal instead of planar (mostly with phenoxide)**
2. **Neighboring dihedrals are in different conformations.**

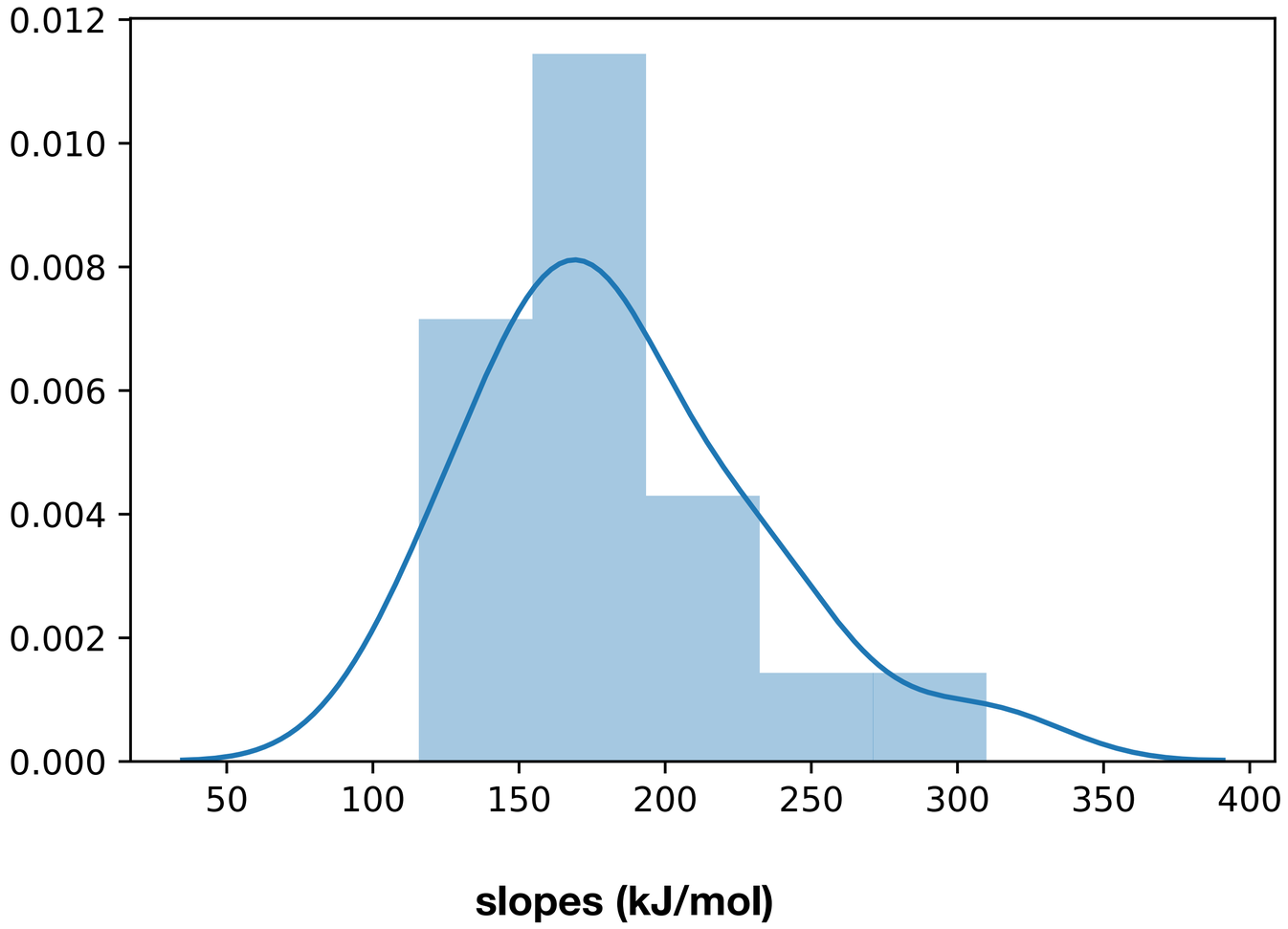


Intelligent fragmentation can reduce the misrepresentation of torsions in QM database



Xinjun Hou
Vishnu Sresth

What should the threshold be?



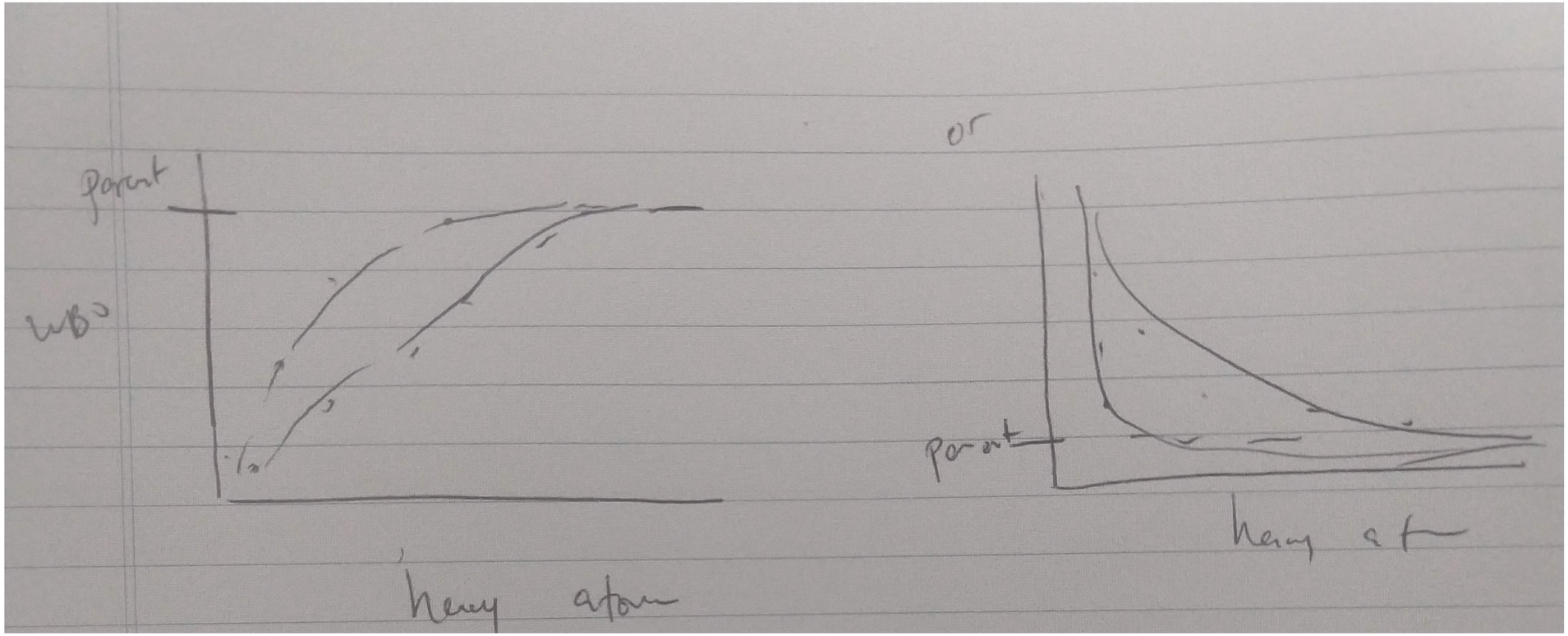
mean slope: 188 kJ/mol
median slope: 179 kJ/mol

Change of 0.1 in WBO is roughly a change of 7kT

Many paths exist to grow out a fragment

Heuristics:

1. **Rank by Wiberg bond order.**
2. **Shortest path**

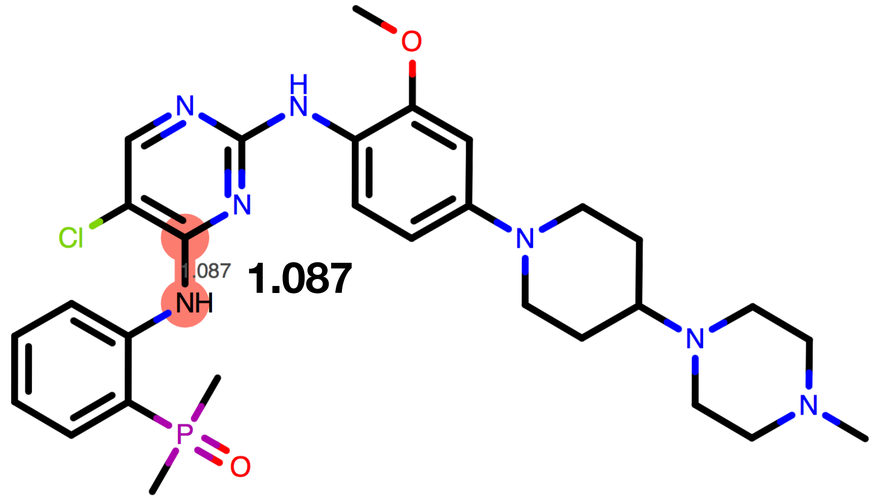


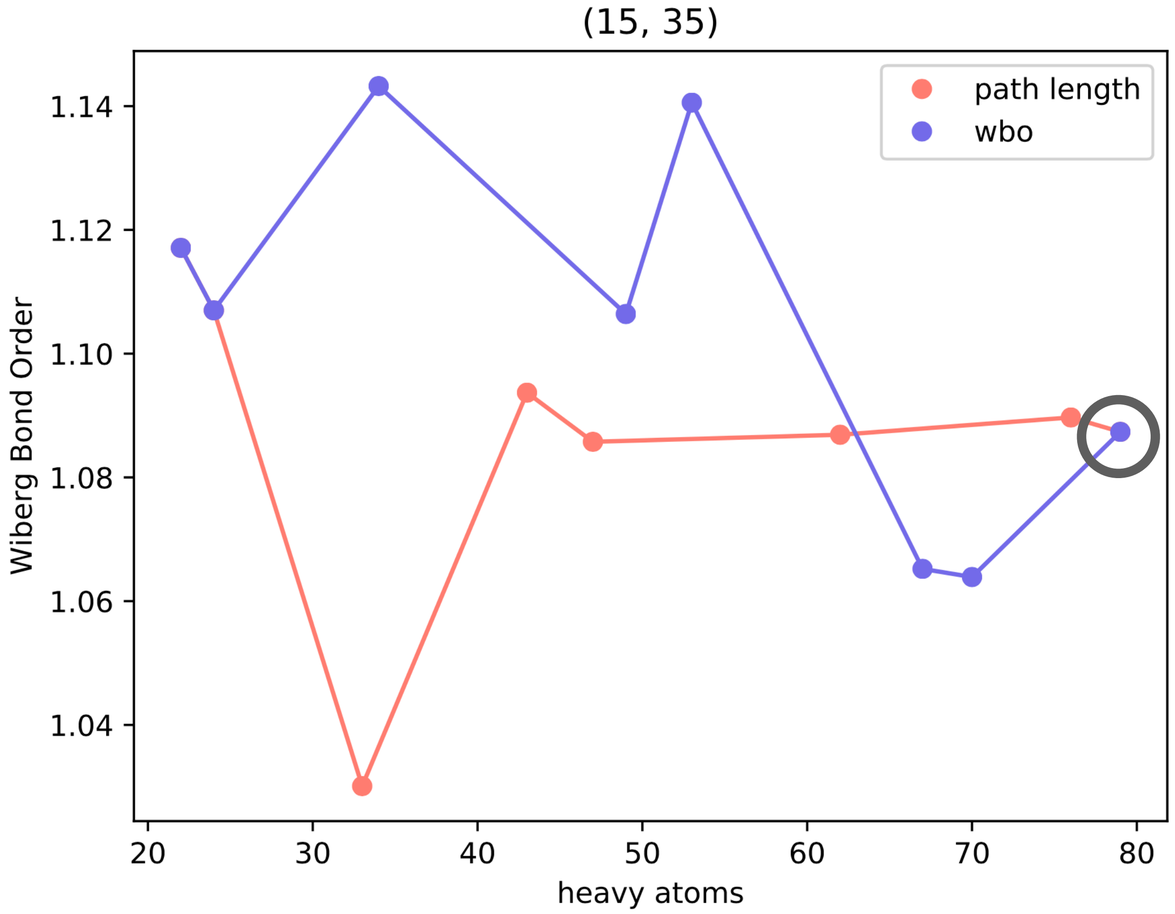
What I expected to see (stability of fragments)

Many paths exist to grow out a fragment

Heuristics:

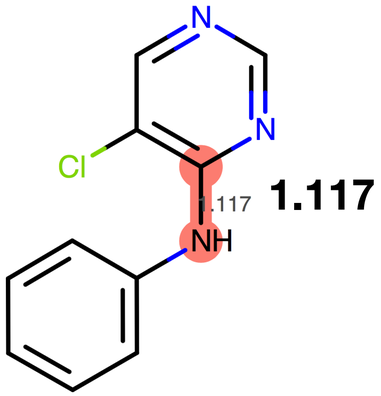
1. **Rank by Wiberg bond order.**
2. **Shortest path**

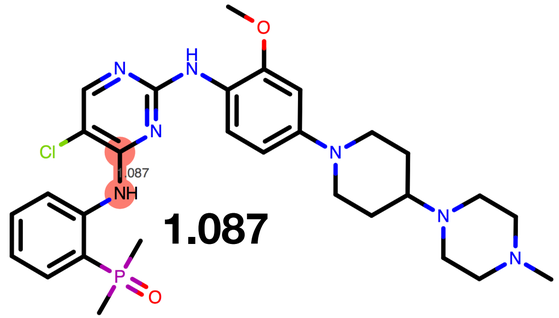


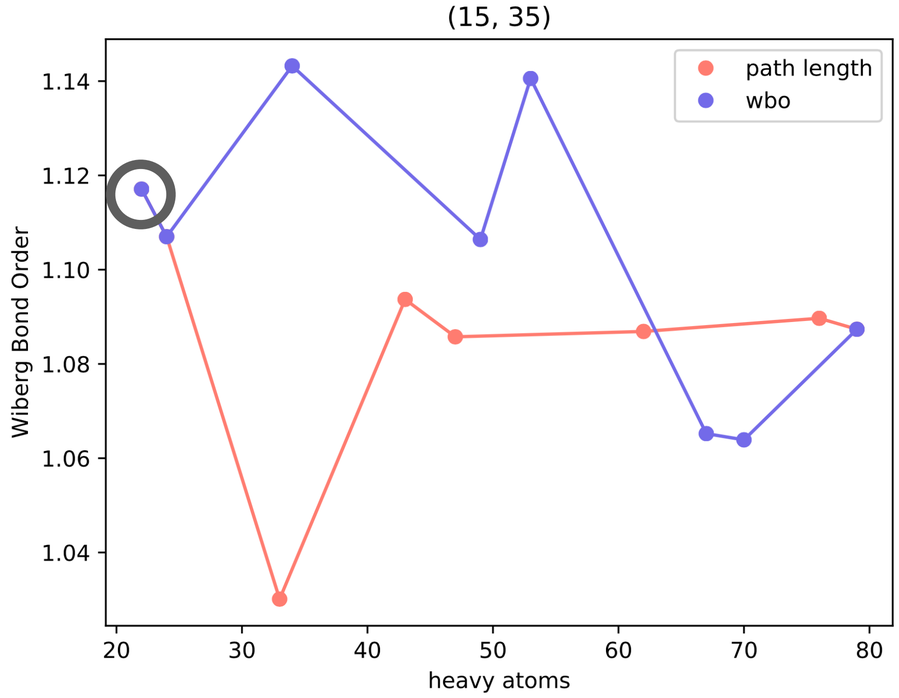


What the data actually looks like

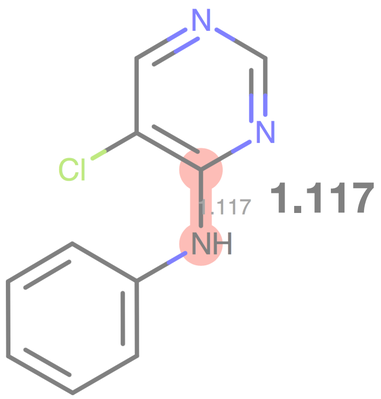
Many paths exist to grow out a fragment



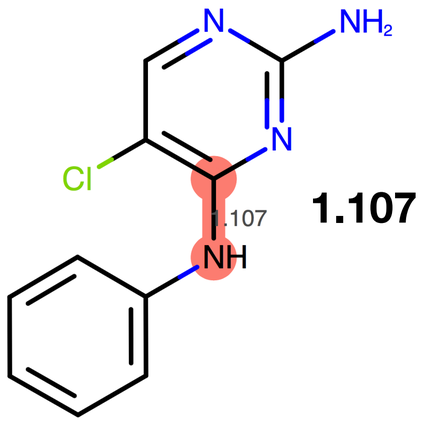




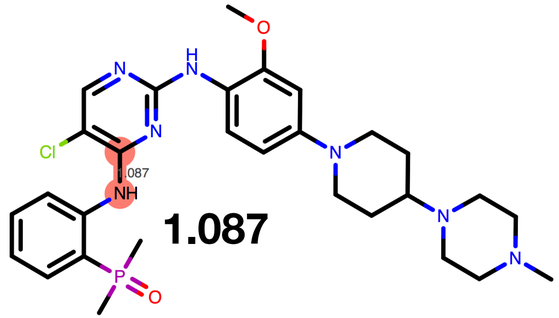
Many paths exist to grow out a fragment



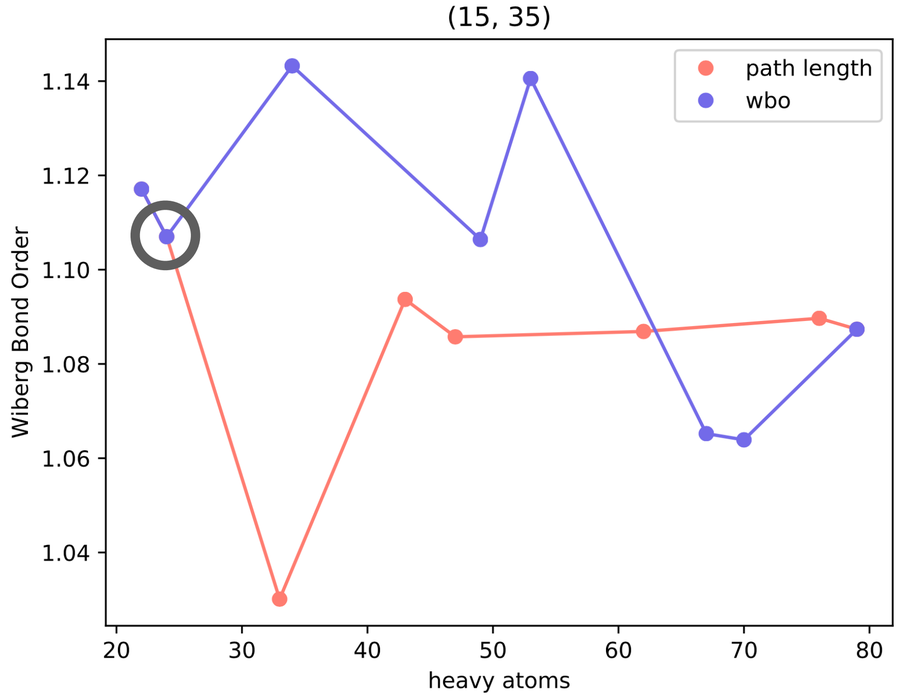
1.117



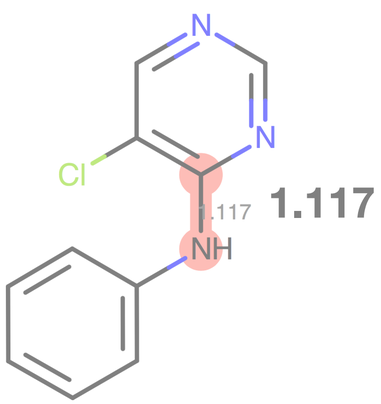
1.107



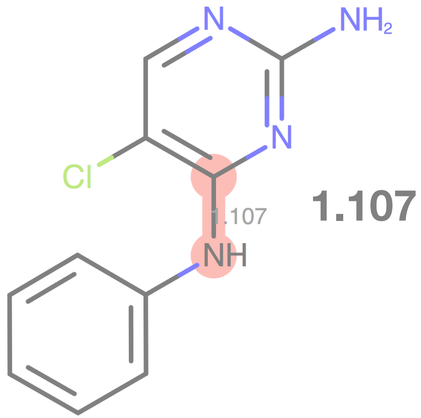
1.087



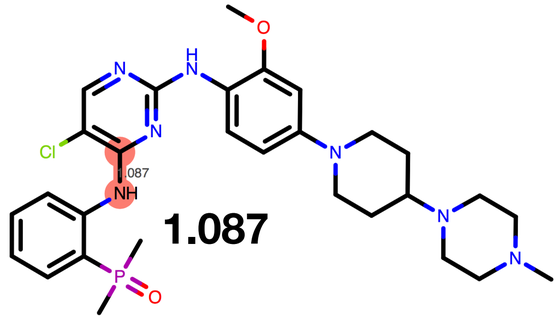
Many paths exist to grow out a fragment



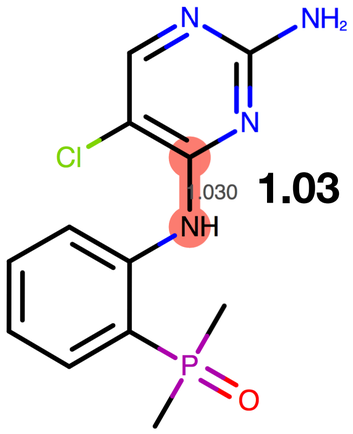
1.117



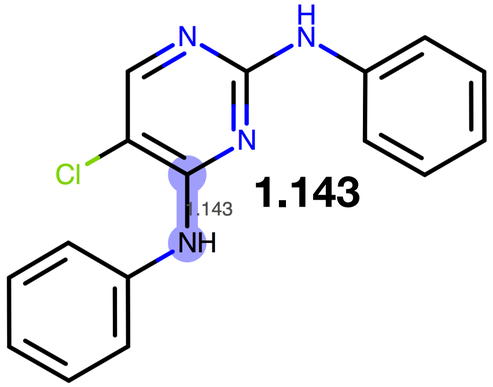
1.107



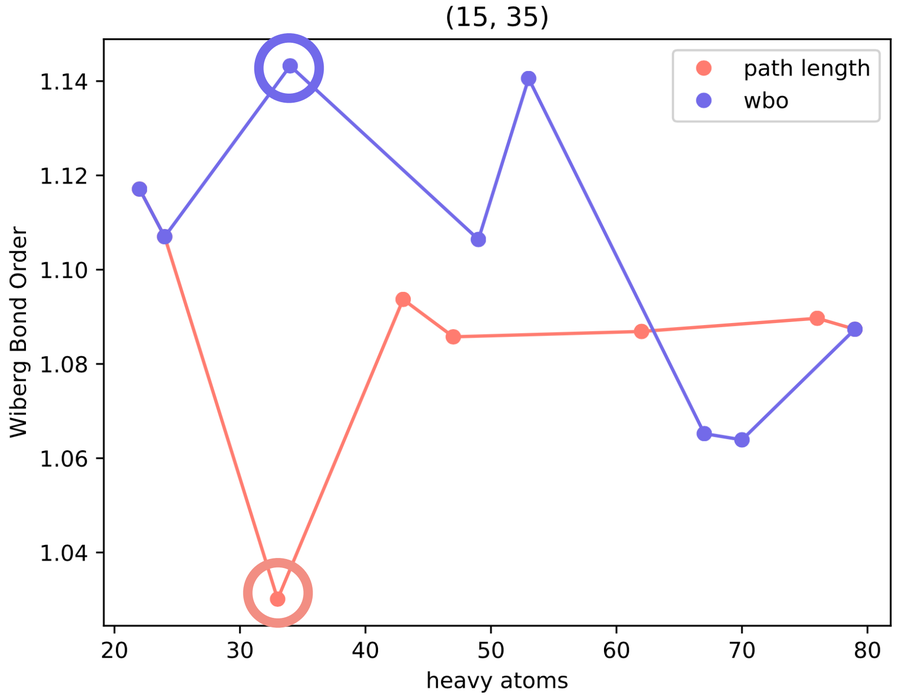
1.087



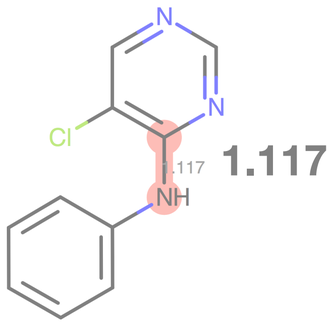
1.03

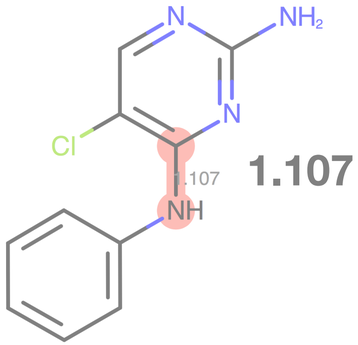


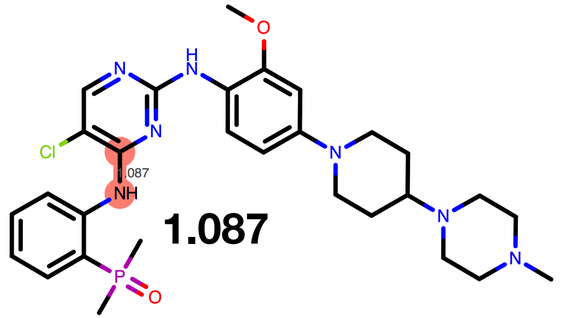
1.143

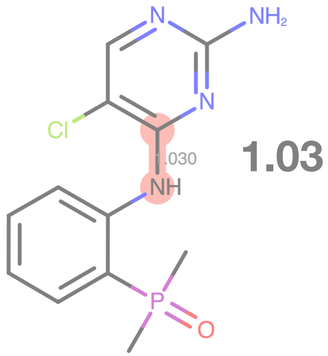


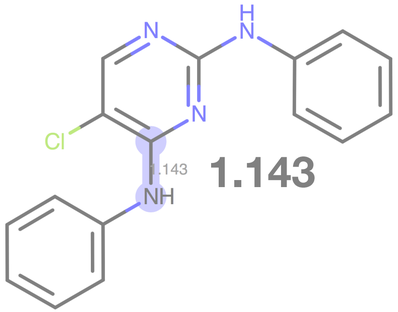
Many paths exist to grow out a fragment

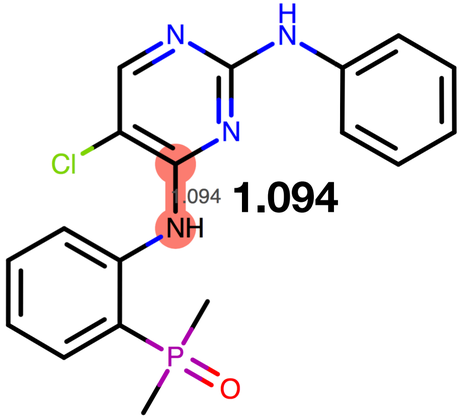


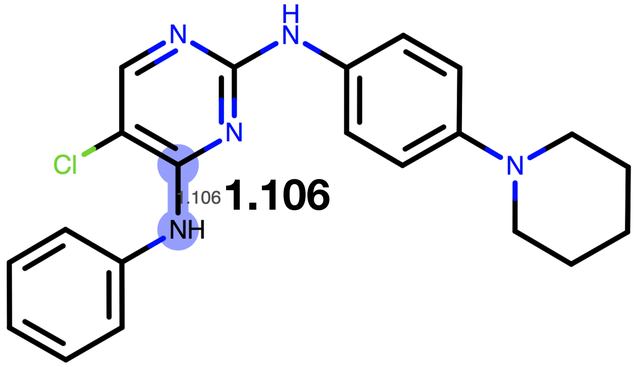


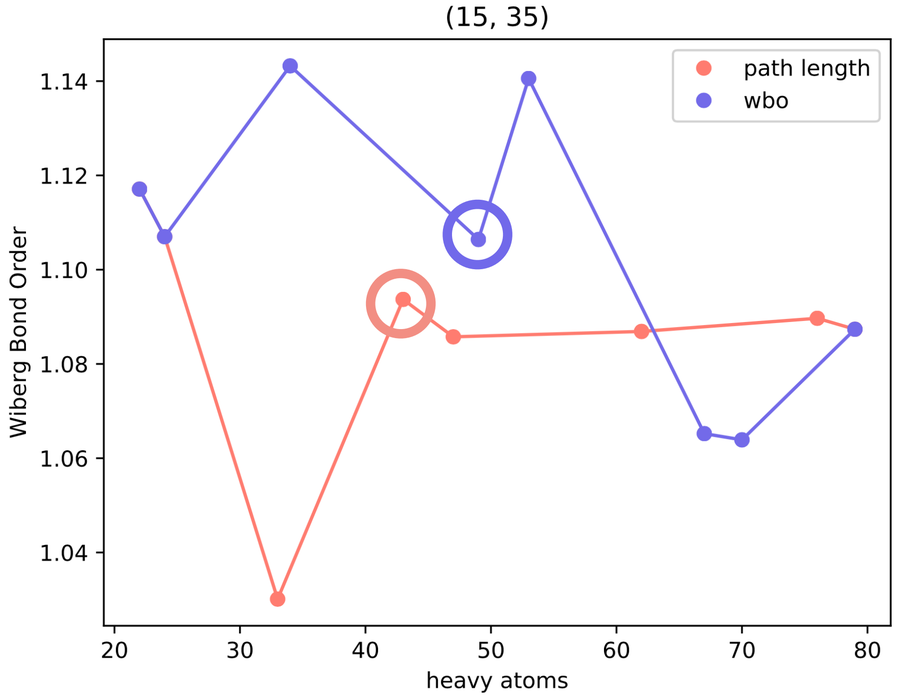




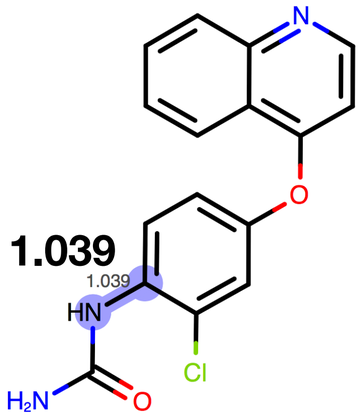


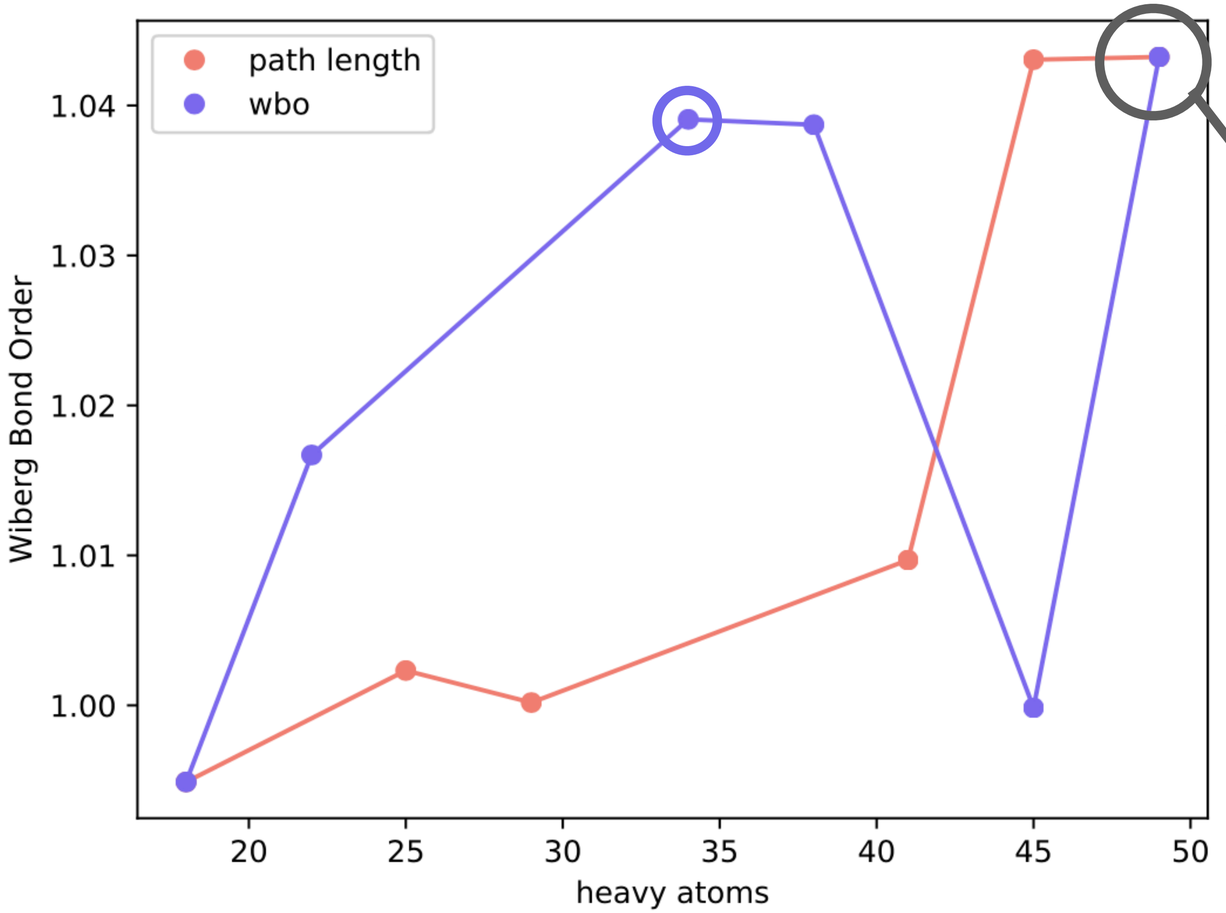


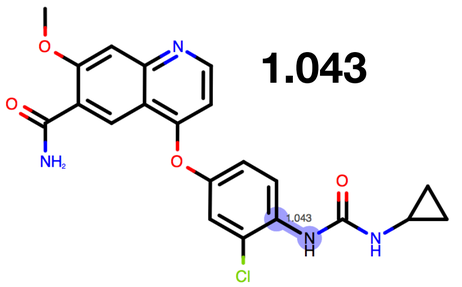




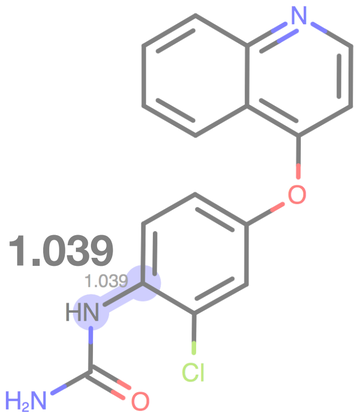
Larger fragments are not always better

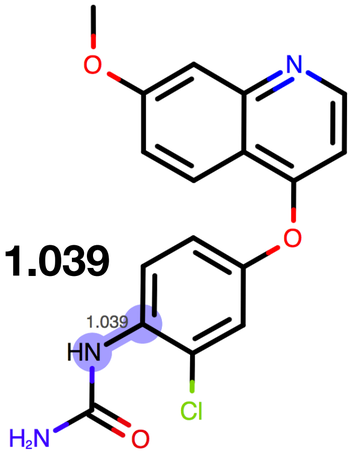


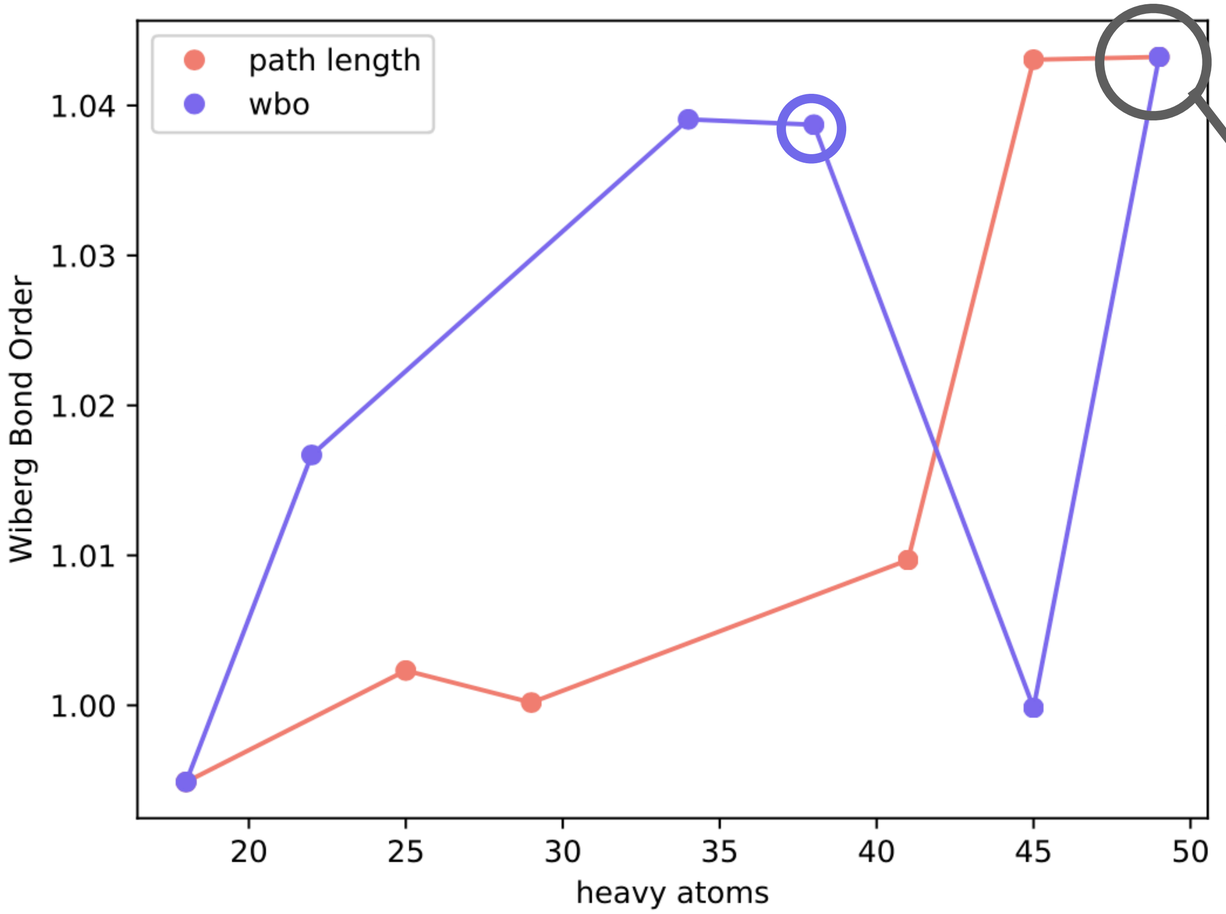


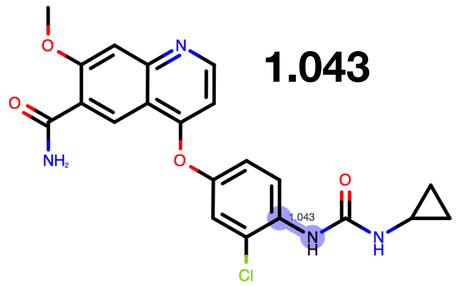


Larger fragments are not always better

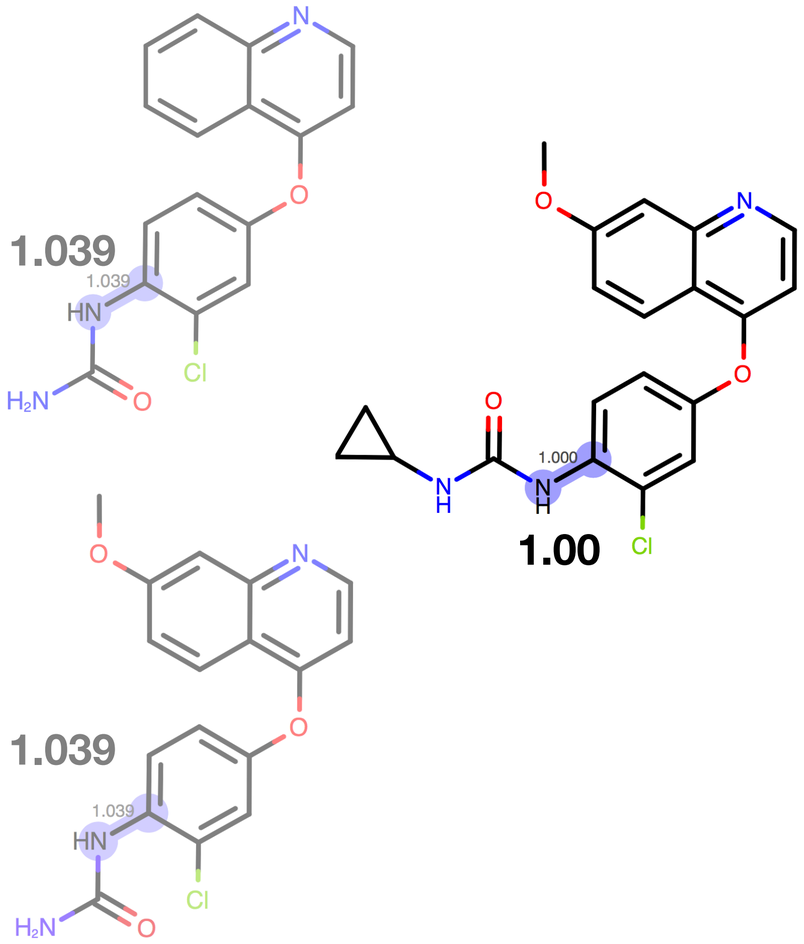


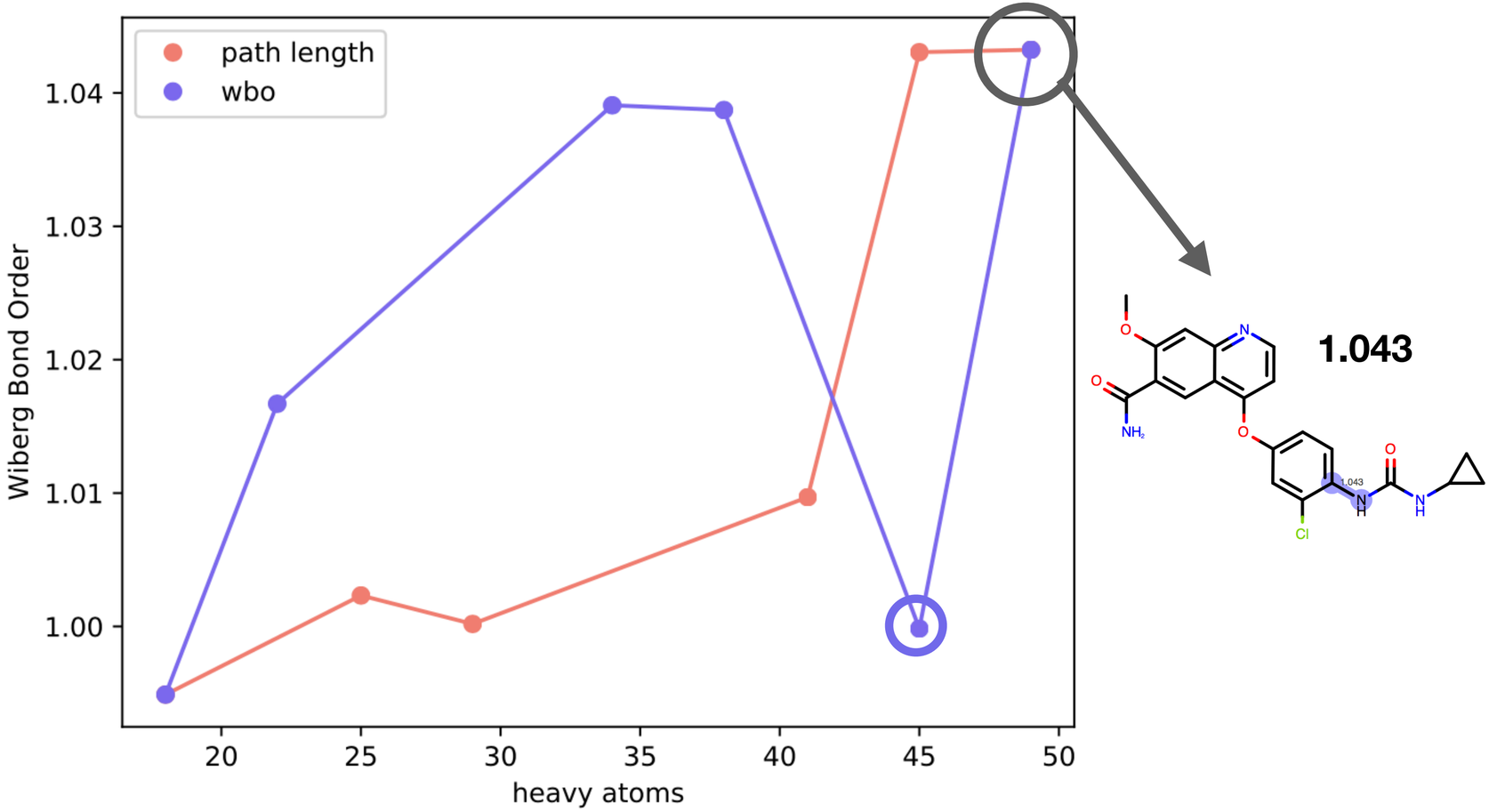






Larger fragments are not always better

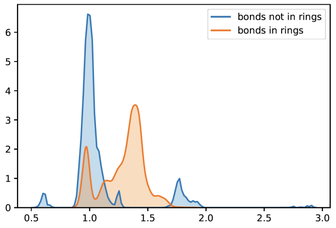




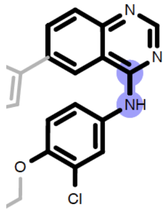
Summary

- The Open Forcefield Initiative strives to improve all aspects of forcefields

- cmiles indexes molecules for QCArchive to ensure broad usability and long term sustainability



- WBO can inform on the chemical environment of bonds



- fragmenter uses WBO to fragment molecules without destroying important chemistry

Future work

- Integrate cmiles and fragmenter with QCArchive
- Integrate torsion parameter fitting with QCArchive and the rest of the OFF stack
- Move away from Wiberg Bond Order for faster fragmentation by learning SMARTS pattern not to fragment

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

QCArchive
A MolSSI Project

https://qcarchive.molssi.org/

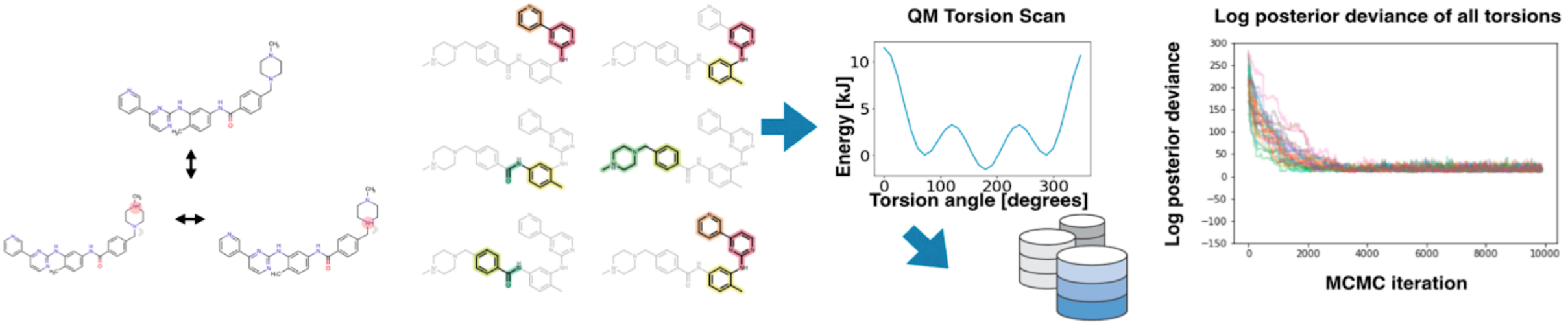
Funding

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