Fragmenting molecules for quantum chemistry torsion scans

This manuscript (permalink) was automatically generated from ChayaSt/fragmenter-manuscript@7ed9a61 on November 15, 2019.

Authors

- **Chaya D Stern**
  0000-0001-6200-3993 · ChayaSt · SternChaya
  Tri-Insitutional PhD Program in Chemical Biology, Memorial Sloan Kettering Cancer Center, New York, New York 10065, USA; Computational and Systems Biology Program, Memorial Sloan Kettering Cancer Center, New York, New York 10065 USA · Funded by Grant XXXXXXXX

- **Christopher I Bayly**
  XXXX-XXXX-XXXX-XXXX · johndoe · johndoe
  Department of Something, University of Whatever · Funded by Grant XXXXXXXX

- **Daniel G A Smith**
  XXXX-XXXX-XXXX-XXXX · dgasmith · johndoe
  Department of Something, University of Whatever · Funded by Grant XXXXXXXX

- **Yudong Qui**
  XXXX-XXXX-XXXX-XXXX · yudongqiu · johndoe
  Department of Something, University of Whatever · Funded by Grant XXXXXXXX

- **Josh Fass**
  XXXX-XXXX-XXXX-XXXX · maxentile · johndoe
  Department of Something, University of Whatever · Funded by Grant XXXXXXXX

- **Lee-Ping Wang**
  XXXX-XXXX-XXXX-XXXX · leeping · johndoe
  Department of Something, University of Whatever · Funded by Grant XXXXXXXX

- **David L Mobley**
  XXXX-XXXX-XXXX-XXXX · davidlmobley · johndoe
  Department of Something, University of Whatever · Funded by Grant XXXXXXXX

- **John D Chodera**
  XXXX-XXXX-XXXX-XXXX · jchodera · johndoe
  Department of Something, University of Whatever · Funded by Grant XXXXXXXX

Abstract

Accurate small molecule molecular mechanics force fields are essential for predicting protein-ligand binding affinities in drug discovery and understanding the biophysics of biomolecular systems. The accuracy of torsion parameters is important for determining the conformational distribution of molecules, and can have a large effect on computed properties like binding affinities. Torsion parameters are usually fit to computationally costly quantum chemical (QC) torsion scans that scale poorly with molecule size. To reduce computational cost and avoid the complications of distant intramolecular interactions, molecules are generally fragmented into smaller entities to carry out QC torsion scans. Poor fragmentation schemes, however, have the potential to significantly disrupt the electronic properties of the region around the torsion, leading to poor representation of the real chemical environment. Here, we show that a rapidly computing quantity, the fractional Wiberg bond order (WBO), is sensitive to the chemical environment of bonds, and can be used as a useful surrogate to assess the robustness of fragmentation schemes and identify conjugated bond sets. We use this concept to construct a validation set consisting of combinatorial fragmentations of druglike organic molecules (and their corresponding WBO distributions) that can be used to evaluate fragmentation schemes. To illustrate the utility of the WBO in assessing fragmentation schemes that preserve the chemical environment, we propose a new fragmentation scheme that uses WBO to maximize the chemical equivalency of the fragment and the substructure in the larger molecule.

Introduction

Small molecules molecular mechancis (MM) force fields are essential to the design of small molecules for chemical biology and drug discovery, as well as the use of molecular simulation to understand the behavior of biomolecular systems. However, small molecule force fields have lagged behind protein force fields given the larger chemical space small molecule force fields need to cover [1,2]. Torsion parameters are particularly problematic because they do not generalize very well [3]. It is possible to significantly improve the force field accuracy by refitting torsion parameters for individual molecules in a bespoke fashion [4,5,6]. In many molecular mechanics force fields (e.g., Amber [7], CHARMM [8], OPLS [9]) a low-order Fourier series, such as a cosine series, is often used to represent the contribution of torsion terms to the potential energy. The torsion potential energy parameters such as amplitudes and phase angles for each Fourier term, are generally fit to the residual difference between gas phase quantum chemistry (QC) torsion energy profile and the non-torsion MM parameters [10]. The QC torsion energy profile is generated by fixing the torsion atoms and geometry minimizing all other atomic positions. Neighboring torsions can have correlated conformational preferences the low-order Fourier series does not capture[11]. 2D spline fits, such as CMAP [12,13], have become a popular way to model non-local correlations by fitting residuals between the 2D QC torsion energy profile and the 2D MM torsion energy profile.

Molecules are generally reduces to smaller model entities containing the torsion of interest for QC torsion scans [2] for two main reasons as illustrated in figure 1.

1. Generating one dimensional QC torsion profiles are computationally expensive and become increasingly inefficient for larger molecules and/or higher dimensional QC torsion profiles. QC calculations scale badly with the number of heavy atoms N, like $O(N^M)$ where $M \leq 3$ for ab initio QC methods [CITE]. To adequately fit the torsions, constrained geometry optimizations need to be calculated at $\leq 15^0$ intervals for a minimum of 24 constrained geometry optimizations. To avoid hysteresis in the energy profile due to orthogonal degrees of freedom [14], methods like wavefront propagation [CITE] are used. This adds a factor of 2D, where D is the dimension of the QC scan, to the computational cost. Figure 1A illustrates the average CPU time of a torsion scan for an average drug-like molecules. The shaded histogram is the distribution of the number of heavy atoms in FDA approved small molecules taken from DrugBank [15]. The average molecules size is

[N] heavy atom which corresponds to an average of [t] CPU seconds per energy and gradient evaluation at B3LYP-D3(BJ)/DZVP [16,17]. An average constrained geometry optimization takes 20 energy and gradient evaluations to converge. The average cost for a 1D QC torsion scan is t*24*20*2 = s.

2. In larger molecules, there is a greater potential for the torsion atoms to interact with other degrees of freedom and convolute the energy profile. While this can also happen in smaller molecules such as ethylene glycol this problem is reduced when a minimal model molecule is used as illustrated in figure 1B.

Many fragmentation algorithms exist, but they are not appropriate for torsion scans in particular and are insufficiently automated. These algorithms fall into two categories: 1. fragmentation for synthetic accessibility [18,19,20] and 2. fragmenting molecules to achieve linear scaling for QC calculations [21,22,23,24]. Fragmentation schemes for synthetic accessibility find building blocks for combinatorial and fragment based drug design. Cleavage happens at points where it makes sense for chemical reactions and do not consider how those cuts affect the electronic properties of the fragments. For retrosynthetic applications, many cleavage points are at functional groups because those are the reactive sites of molecules. However, for our application, we especially do not want to fragment at these reactive points given how electron rich they are and how much the electronic density changes when they are altered. Fragmentation algorithms for linear scaling such as Divide-and-Conquer methods [25], effective fragment potential method [26] and systematic molecular fragmentation methods [24] require the users to manually specify where the cuts should be or which bonds not to fragment. Furthermore, none of these methods address the needs specific to fragmenting molecules for QC torsion scans. Fragments need to include all atoms involved in 1-4 interactions, since they are incorporated in the fitting procedure. We also need a systematic way to determine if remote substituents change the barrier to rotation significantly for the central bond of interest.

In this work, we use the Wiberg Bond Order (WBO) [27], which is both simple to calculate from semi-empirical QC methods and is sensitive to the chemical environment around a bond. WBOs are correlated with bond vibrational frequencies [28] and is used to predict trigger bonds in high energy-density material because it is correlated with the strength of the bond [29]. Here, we develop an approach that uses the WBO to validate whether a fragmentation scheme corrupts the local chemical environment of interest, with a particular focus on fragmentation schemes suitable for QC torsion drives. Our approach uses simple heuristics to arrive at a minimal fragment for QM torsion scan that is representative of of the torsion scan of the substructure in the parent molecules. For a central bond, include all atoms that are involved in the sterics of a torsion scan. Then use the WBO as a surrogate signal to determine if the fragment needs to be grown out more to restore the correct electronics around the central bonds.

The paper is organizes as follows: Section 2 provides a mathematical and physical definition of the problem. Section 3 provides the motivation for using the WBO as a surrogate, evaluates how robust it is, proposes a minimal fragmentation scheme and describes a rich validation set that can be used to benchmark fragmentation schemes Section 4 provides a discussion of the implications of this study and section 5 provides detailed methods.

Theory

TBD

- Define problem mathematically

- Define the problem physically

Results

Torsion energy barriers are sensitive to the chemical environment, which can be influenced by remote substituents

In most forcefields, torsions are defined by the quartet of atom types involved in the dihedral [CITE]. However, the quartet of atom types do not always capture the relevant chemistry, especially when the effects are non local i.e., atoms contributing to hyperconjugation, delocalization or other non classical effects, are not part of the quartet involved in the torsion [3]. Figure 1A illustrates such a case with a series of biphenyls in different protonation states. While the MM torsion profiles are all the same (Fig 2B), the QC torsion profiles are different for each protonation state (Fig 2C). The torsion energy barrier increases relative to the neutral state for the cation, anion and zwitterion, in that order. The profile changes qualitatively as well. For the neutral molecule, the lowest energy conformer is slighlty out of plane, at 150° and 120°. For the zwitterion, the lowest energy conformer is at 180°. In the neutral molecule, the slightly out of plane conformer is preferred to accommodate the hydrogens. However, the central bond in the zwitterion is part of the larger conjugated system between the two aromatic rings (Fig 2D) so the planar conformer is preferred to enable conjugation. This trend poses several problems to automatic forcefield parametrization. Most forcefields consider the central bond in the zwitterion rotatable while the QC scan clearly shows that is it not. This illustrates one of the fundamental limits of atom types in classical forcefields. At what point in this series should a new atom type be introduced? The Open Force Field Initiative's effort on automating data driven direct chemical perception [1,30,31] addresses this problem by using SMIRKS patterns to assign parameters, and providing a framework to sample over SMIRKS space in a data driven way. In addition, this example illustrates why fragmenting molecules appropriately for QC torsion scans requires human expertise and is difficult to automate. In this case, a small change three bonds away from the torsion central bond changed the bond from a rotatable bond to a non-rotatable conjugated bond. When fragmenting molecules we need to avoid destroying a bond's chemical environment by naively removing an important remote substituent.

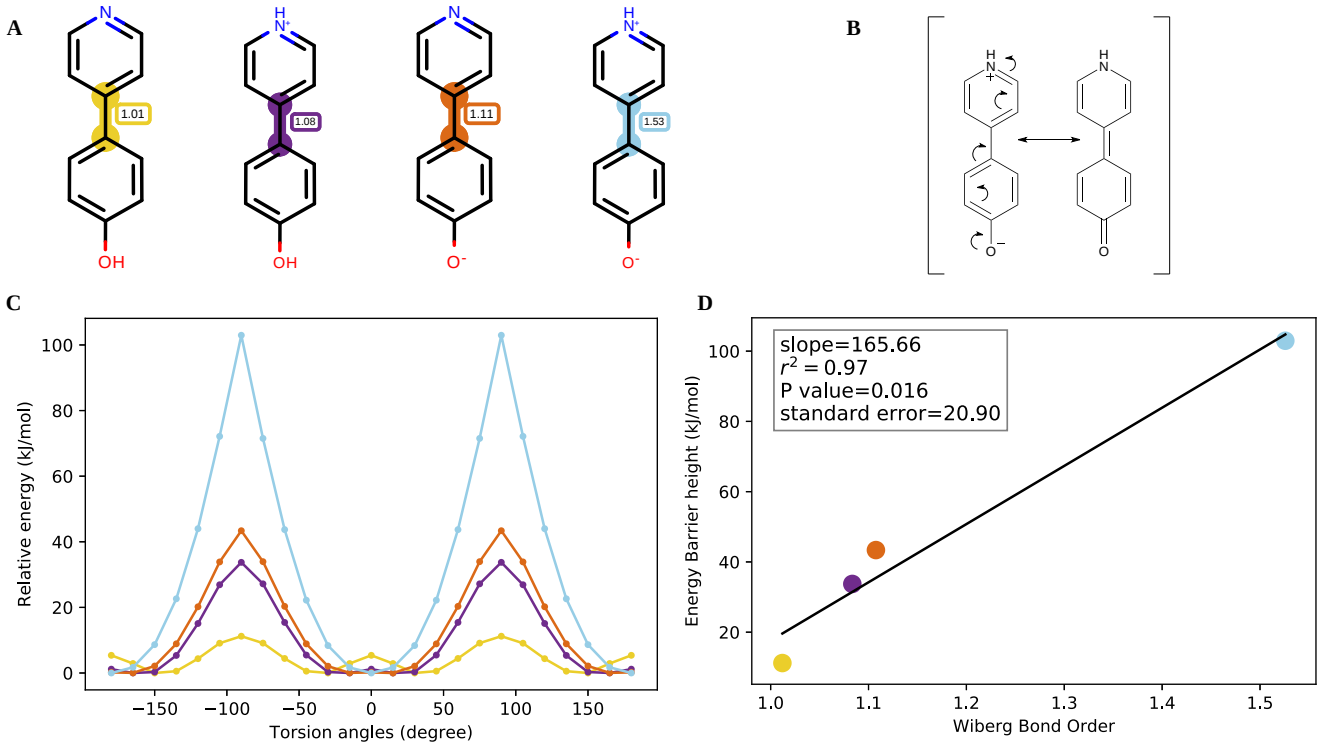

Figure 1: Illustration of the sensitivity of torsion profiles to remote chemical changes in a molecule [A] Biphenyl protonation states and tautomers with increasing Wiberg bond order for the central bond. **[B]** The resonance structure of the biphenyl zwitterion shows that the central bond is conjugated. The Wiberg bond order and torsion scan for this bond (see **A** and **C**) are reflective of a conjugated bond. **[C]** Relative QC energy as a function of torsion angle of the central bond. The colors of the QC scan corresponds to the highlighted bonds in **A**. **[D]** Torsion barrier heights vs WBOs. The color of the data points correspond to the highlighted bonds in **A**. The WBO and QC torsion barrier heights are correlated.

The Wiberg Bond Order quantifies the electronic population overlap between two atoms and captures bond conjugation

The Wiberg bond order (WBO) is a bond property that is calculated using orthonormalized atomic orbitals that are used as basis sets in semi-empirical QC methods[CITE]. Wiberg originally formulated it for the CNDO basis set [CITE] but it can be easily extended to other semi-empirical QC methods such as AM1 [32] and PM3 [CITE]. The WBO is a measure of electron density between two atoms in a bond and is given by the quadratic sum of the density matrix elements over occupied atomic orbitals on atoms A and B

$$W_{AB} = \sum_{\mu \in A} \sum_{\nu \in B} P_{\mu\nu}^2$$

To check how well W_{AB} recapitulates the multiplicity of bonds, we calculated W_{AB} from AM1 calculations for all bonds in FDA approved molecules in DrugBank [15]. The distribution [fig 3] corresponds closely with bond multiplicity. The plateau at 0.8 correspond to bonds involving S and P since these are weaker and longer bonds. The plateau at 1.0 corresponds to C-H and C-C bonds, the plataea at 1.5 corresponds to bonds in aromatic rings, the plateau at 2.0 corresponds to double bonds, and finally the triple bonds form the last plateau. The density between 1.0-1.4 correspond to the conjugated bonds not inside rings.

The Wiberg bond order assumes atomic orbitals (AOs) are orthonormal. In ab initio calculations, however, AOs are not orthogonal but the WBO can be calculated via Löwdin normalization [33,34] which is how it is calculated in Psi4.

We calculated the WBO from AM1 calculations for the biphenyl series as shown in figure 2A. The increase in the WBO corresponds to increasing conjugation and torsion energy barrier height of the bond. When the torsion energy barrier heights are plotted against the WBO [fig 2E], the relationship is linear with an R^2 of [hold].

The WBO is an inexpensive surrogate for the chemical environment around a bond

Since the WBO can be calculated from a cheap AM1 calculation, is indicative of a bond's conjugation, and is correlated with torsion energy barrier height, it is an attractive measure to use as a surrogate when automating fragmentation or interpolating torsion force constants. However, WBOs are conformationally dependent [35] so we investigated this dependence to understand if WBOs will be a robust descriptor. In addition, we also investigated the generality of the torsion energy barrier and WBO linear relationship. In this section we will first discuss our findings and solution to the conformational dependency and then discuss how general the linear relationship is.

Conformation dependent variance of WBOs are higher for conjugated bonds

Since WBOs are a function of the electronic density, which is conformational dependent, WBOs change with conformation. However, not all bonds' WBOs change the same way with conformation.

We found that WBOs for conjugated bonds are multimodal with respect to conformation and that bonds involved in conjugated systems have WBOs that are correlated with each other.

To investigate how WBOs change with conformation, we used Omega [36] to generate conformers for a set of kinase inhibitors [Supplementary figure 1] and calculated the Lowdin-Wiberg bond order for each conformation from an hf3c [37] geometry optimized calculation using Psi4 [38].Omega is a knowledge-based conformer generator that uses a modified version of MMFF94s [39] to score conformations. It has been shown to accurately reproduce experimentally observed crystallography conformers in the Platinum benchmark dataset [40]. Figure 4 illustrates the results for Gefitinib [Figure 4A], a representative molecule. Figure 4B shows the distribution of WBOs for all rotatable bonds color coded with the colors used to highlight the bonds in Gefitinib [Fig 4A]. Single carbon-carbon bonds, and carbon-nitrogen bonds formed by atoms numbered 10 - 13 are freely rotating. This is reflected by the tight distribution of WBOs around 1.00 for those bonds. The bonds involving the ether oxygens and aromatic rings (formed by atoms numbered 1-3, 8-10, 19, 23-24) exhibit higher variance and multimodality. It is interesting to note the difference in the WBOs for the conjugated bonds formed by the nitrogen between the quinazoline and chloro fluro phenyl (bonds formed by atoms numbered 19, 23 and 23, 24). Both of these bonds are conjugated with their neighboring ring systems, however, While the distribution of WBOs for bond 19-23 (the purple distribution) is clearly bimodal, the WBO distribution for bond 23-24 has lower variance. This is in agreement with the resonance structures shown in figure 4D. The resonance strictures that have the double bond on the bond closer to the quinazoline (bond 19-23) are more stable because the negative charge is on a nitrogen. When the double bond is on the neighboring 23-24 bond, the negative charge is on an aromatic carbon which is less stable. The results are similar for other kinase inhibitors tested shown in supplementary figure 1. In addition, when we inspected the conformations associated with each mode in the purple distribution [figure 4B] we found that conformations with lower WBOs on bond 19-23 had that bond out of plane while the conformations in the higher mode of the distribution had the bond in plane which allows conjugation. We found similar results from WBOs calculated form QC torsion scans. Fig 2D shows the Lowdin-Wiberg bond order for each point in the QC torsion scans of the biphenyl zwitterion. The WBOs are perfectly anti-correlated with the torsion potential energy which is in line with chemical intuition. Conjugation stabilizes conformations and leads to more electronic population overlap in bonds [CITE]. At higher energy conformers, the aromatic rings are out of plan and cannot conjugate. Therefore the WBO is lower for those conformers. At lower energy conformations, the rings are in plane and can conjugate so the WBO is higher.

Bonds in conjugated systems have highly correlated conformation-dependent WBOs

We found that certain bond orders are strongly correlated or anticorrelated with each other, indicating strong electronic coupling. Figure 4C shows the Pearson correlation coefficient for each bond WBO distribution against all other bond WBO distributions. There is a clear structure in this correlation plot. The square formed by bonds from atoms 24-29 shows that the alternating bonds in the aromatic ring are strongly anticorrelated with each other. While the ring formed by atoms 13-18 also exhibit this trend, the absolute Pearson correlation coefficients are not as high given that it is not an aromatic ring. The bonds involved in the ethers (atoms 1-3 adn 8-10) are strongly correlated with each other and also correlated to the quinazoline, albeit not as strongly. And the bonds between the chloro fluoro phenyl and quinazoline follow the same trend as their WBO distribution and resonance structures. The bond closer to the quinazoline (bond 23-19) has WBO distribution correlated with the quinazoline while the bond closer to the chloro fluoro phenyl (bond 23-24) is not as strongly coupled with the quinazoline or phenyl ring. The results are similar for other kinase inhibitors tested as shown in supplementary figure 1

ELF10 provides a useful way to capture informative conformation-independent WBOs

As we have shown, the WBO is conformation dependent and this dependency can also be highly informative of the electronic coupling in a system. Figure 5 shows the distribution of standard deviations of WBO distribution with respect to conformation in blue. Most of the standard deviations fall below 0.02, which is encouragingly small. However, it can become computationally expensive to calculate the WBO for all conformations. If we want to use WBOs as a surrogate to determine if our fragment is representative of the parent molecule in a reproducible way, we need a way to capture informative conformation-independent WBOs. Electronically Least-interacting Functional groups (ELF) conformation selection implemented in quacpac [CITE] resolves the issue of sensitivity of molecular mechanics electrostatic energies from QM derived charges.

Leave to Christopher Bayly to describe

This method can also be applied to deriving WBOs that are insensitive to conformers. To test how well ELF10 WBOs correspond to expected WBOs, we calculated ELF10 WBOs for the kinase inhibitor set shown in supplementary figure 1. Figure 3A shows the distribution of all ELF WBO. To gain insight how the ELF10 WBOs corresponds to bond multiplicity, we separated the distributions by element. Figure 3B shows the distributions of carbon-carbon bonds. The blue distribution shows the carbon-carbon bonds not in rings. There are peaks at one, two and three which correspond to single, double and triple bonds. Single bonds are the most abundant followed by double and then triple. The blue distribution shows carbon-carbon bonds in rings. There is a peak as 1.0 and 1.5 which corresponds to single and aromatic rings. Longer, weaker bonds involving Sulfur and Phosphorous [Fig 3C] both have peaks at ~0.6. Oxygen has a peak at 1.0 and 1.8 which corresponds to single and double bonds. For the rest of this section we will be focusing on the robustness and generalizability of ELF10 WBOs.

WBOs are a robust signal to how remote substituent changes alter a bond's torsion barrier height

To investigate how resonance and electronic effects from remote substituents change the torsion energy of a bond, we took inspiration from the Hammett equation [41] of reactions involving benzoic acid derivatives. The Hammett equation relates meta and para benzoic acid substituents to the acid's ionization equilibrium constants

$$log\frac{K}{K_0} = \sigma\rho$$

Where σ is a substituent constant and ρ is a reaction constant. It aims to isolate the resonance and inductive effects of substituents from the sterics effects of a reaction. Here, we generated a combinatorial set of meta and para subsituted phenyls and pyridine [Fig 6A] with functional groups that cover a wide range of electron donating and withdrawing groups as shown in figure 6A. We then calculated the ELF10 WBO for the bond attaching the functional group to the aromatic ring for all functional groups [highlighed bonds in figure 6A]. This allowed us to isolate the effect on a bond's WBO from remote chemical environment changes, defined as a change more than two bonds away, from other effects such as sterics and conformations. For each functional group, we get a distribution of WBOs, where each point in the distribution is the WBO at the bond connecting that functional group to the meta or para substituted phenyl or pyridine ring for all other functional groups. The resulting distributions are shown in supplementary figure 2. It is interesting to note that the trend of decreasing WBOs for more electron donating groups anti correlates with with increasing substituent constant for the para effect [Supplementary figure 2B]. Functional groups that are more electron donating will have more electron density on the bond attaching the functional group to the benzoic acid. The resonance and/or inductive effect destabilize the benzoate, increases its pKa, which corresponds to lower substituent constants.

To investigate how these long range effects observed in the WBOs capture changes in the bonds' torsion potential energy, we ran representative QC torsion scans for 17 of the functional groups [Figure 6A]. We did not run QC torsion scans for functional groups that either did not have a torsion such as halogens, were congested such as trimethyl amonium and factional groups where the WBOs did not change by more than 0.01 for different functional groups at the meta or para position such as methyl. We chose the representative molecules for the 17 functional groups by sorting them by their WBO and selecting molecules with minimum WBO difference of 0.02. All of the resulting QC torsion scans are shown in supplementary figure 3. We show a representative series of torsion scan for the nitro functional group in figure 6D. The torsion energy barrier height increase with increasing ELF10 WBO of the bond. In addition, Figure 6E shows that the Wiberg-Lowdin bond orders are anti-correlated with the QC torsion scan which is the same result we saw for the initial bipheynl set discussed in the previous section. We also found that the trend shown in figure 2D that the torsion energy barrier height is correlated with WBO generalizes to all functional groups tested in this set [Fig 6C].

For most functional groups, the change in WBOs correspond to changes in torsion barrier heights, but not in the torsion energy profile [supplementary figure 3]. However, for some functional groups the change in WBO only captures one aspect of the electronic changes because not only do the torsion energy barrier heights increase, but the profile changes considerably as shown in figure 7 for [hold for functional group and some way to explain this].

When we compare the standard deviations of WBO distributions with respect to conformation versus with respect to changes in chemical space [Figure 5 red distribution], we find that the changes in ELF10 WBO for remote chemical environment changes are bigger than the changes in WBO that arise from change in conformation. This allows us to use the difference in ELF10 WBO of parent and fragment as a good surrogate to the level of disruption of the chemical environment.

A simple fragmentation scheme can use the WBO to preserve the chemical environment around a torsion

The WBO is a robust indicator of changes in torsion energy barrier heights for related torsions. Therefore, if a fragment's WBO changes too much from its parent WBO at the same bond, the fragmentation is probably inadequate. Using this concept, we extended the fragment-and-cap scheme proposed by [CITE Pfizer paper] by considering resonance via WBOs. The scheme, illustrated in figure 8 is as follows:

1. Find rotatable bond. For this step we use the SMARTS pattern [hold for SMARTS pattern]
2. Build out one bond in each direction
3. If the next atom is part of a ring or part of a functional group listed in figure 8B, keep the ring and functional group
4. Keep meta substitutents to the rotatable bond of interest because it is involved in the sterics of the torsion
5. Cap with hydrogen and recalculate WBO
6. If the fragment's WBO differs by more than a user defined threshold, continue grow out one bond at a time until the fragment's WBO is within the threshold of the parent WBO.

Fragmentation schemes can be assessed by their ability to preserve the chemical environment while minimizing fragment size

This fragmentation scheme improves upon Pfizers scheme, however, it leaves some parameters up to the user. In order to asses various thresholds and different fragmentation schemes in general, we generated a diverse set of FDA-approved drug molecules that can be a useful validation set. The goal of this set was to find molecules that are challenging to fragment. In other words, molecules that have

bonds that are sensitive to remote substituent changes. To find these molecules, we first filtered DrugBank (version 5.1.3 downloaded on 2019-06-06) with the following criteria:

1. FDA approved small molecules
2. Largest ring size has less than 14 heavy atoms
3. Smallest ring size has at least 3 heavy atoms
4. Molecule must have at least one aromatic ring
5. Molecule has only one connnected component

This left us with 730 small molecules [Supplementary figure 3]. Charged molecules exacerbates remote substituent sensitivity and many molecules are in charged states at physiological pH. To ensure that our dataset is representative of drugs at physiological pH, we used the OpenEye `EnumerateReasonableTautomers` to generate tautomers that are highly populated at pH ~7.4. This tautomer enumeration extended the set to 1234 small molecules [Supplementary figure 3B] We then generated all possible fragments of these molecules by using a combinatorial fragmentation scheme. In this scheme, every rotatable bond is fragmented and then all possible connected fragments are generated where the smallest fragment has 4 heavy atoms and the largest fragment is the parent molecule. This scheme generated ~300,000 fragments. For each fragment, Omega was used to generate conformers and the AM1 WBO was calculated for every bond in every conformer. This resulted in a distribution of WBOs for all bonds in all fragments as shown in figure 9A.

Scoring how well fragments preserve chemical environments using WBO distributions

Each fragment needs to be assigned a score of how well is preserves its parent chemical environment. To score each fragment, we compare the conformer dependent WBO for a bond in a fragment context against the conformer-dependent distribution of WBO for the same bond in the parent molecule context. To compare these distributions, we compute the maximum mean discrepancy [CITE] with a squared kernel for the fragment distribution to the parent as follows:

$$MMD = \sqrt{(E[X] - E[Y])^2 + (E[X^2] - E[Y^2])^2}$$

where X is the parent WBO distribution and Y is the fragment WBO distribution. Including the squared mean incorporates the variance of the distribution. It is important incorporate changes in variance given how the variance of the WBO distributions change for different chemical environments. The final validation set was chosen by finding the fragments that have bonds with the top 50 scoring MMD scores. The 50 molecules are shown in figure 10. The sensitive bonds that are high scoring are highlighted.

Good fragmentation schemes minimize both chemical environment disruption and fragment size

The goal of our fragmentation scheme is to find fragments that have a WBO distribution of the bond of interest closest the the parent while minimizing the size of the fragment. We can use the distance score calculated with MMD as an indicator of how far the fragment's WBO distribution is. When we plot the fragment size against this score, the points that fall on the Pareto front [CITE] are the ones where the distance score is the best for for a given fragment size or vice versa. Figure 9B shows an illustrative example of this. The aim of any fragmentation scheme is to find fragments on the Pareto front that minimize both the WBO distance and fragment size. In other words, they should be on the lower left corner of the plot. We want to find the parameters for our fragmentation scheme that maximizes the number of fragments that end up in that lower left corner. To do that, we fragmented the 50 molecules in the validation set using different disruption thresholds. For every molecule, we found the distance score of their fragments' WBO distribution and their size. We then plotted all

fragments from the validation set for different thresholds [figure 11] When the threshold is low, the fragmenation scheme will generate fragments which have very good scores, but many of them will be too big for computational efficient QC torsion scan. On the other hand, when the disruption threshold is too low, the scheme generates fragments that are small but the distance score is too big. For the molecules we tested, a threshold of 0.01 leads to the most fragments on the Pareto front as shown in table 1. This threshold is similar to what we found when we looked at the distribution of standard deviations for WBO distributions with respect to conformations [figure 5 blue distribution]. Most of them fall under 0.02. Both of these data points leads us to recommend a distruption threshold of 0.01 for our fragmentation scheme. While the current scheme does not provide a perfect solution, the plot in figure 11B shows a few fragments outside of the lower left region, it preforms considerably better than other schemes such as the Pfizer scheme [Figure 11D] and Bemis Murko [Figure 11E]. The Bemis Murko scheme was not designed for this application so it is not surprising how poorly it preforms.

References

1. **Toward Learned Chemical Perception of Force Field Typing Rules**
Camila Zanette, Caitlin C. Bannan, Christopher I. Bayly, Josh Fass, Michael K. Gilson, Michael R. Shirts, John D. Chodera, David L. Mobley
Journal of Chemical Theory and Computation (2018-12-04) https://doi.org/gft4hf
DOI: 10.1021/acs.jctc.8b00821 · PMID: 30512951 · PMCID: PMC6467725

2. **OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins**
Edward Harder, Wolfgang Damm, Jon Maple, Chuanjie Wu, Mark Reboul, Jin Yu Xiang, Lingle Wang, Dmitry Lupyan, Markus K. Dahlgren, Jennifer L. Knight, … Richard A. Friesner
Journal of Chemical Theory and Computation (2015-12) https://doi.org/f76wpm
DOI: 10.1021/acs.jctc.5b00864 · PMID: 26584231

3. **Accuracy evaluation and addition of improved dihedral parameters for the MMFF94s**
Joel Wahl, Joel Freyss, Modest von Korff, Thomas Sander
Journal of Cheminformatics (2019-08-07) https://doi.org/gf6rz2
DOI: 10.1186/s13321-019-0371-6 · PMID: 31392432 · PMCID: PMC6686419

4. **Paramfit: Automated optimization of force field parameters for molecular dynamics simulations**
Robin M. Betz, Ross C. Walker
Journal of Computational Chemistry (2014-11-21) https://doi.org/f6svdh
DOI: 10.1002/jcc.23775 · PMID: 25413259

5. **Rapid parameterization of small molecules using the force field toolkit**
Christopher G. Mayne, Jan Saam, Klaus Schulten, Emad Tajkhorshid, James C. Gumbart
Journal of Computational Chemistry (2013-09-02) https://doi.org/f5ggrj
DOI: 10.1002/jcc.23422 · PMID: 24000174 · PMCID: PMC3874408

6. **Fitting of Dihedral Terms in Classical Force Fields as an Analytic Linear Least-Squares Problem**
Chad W. Hopkins, Adrian E. Roitberg
Journal of Chemical Information and Modeling (2014-07-09) https://doi.org/f6cffs
DOI: 10.1021/ci500112w · PMID: 24960267

7. **A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules**
Wendy D. Cornell, Piotr Cieplak, Christopher I. Bayly, Ian R. Gould, Kenneth M. Merz, David M. Ferguson, David C. Spellmeyer, Thomas Fox, James W. Caldwell, Peter A. Kollman
Journal of the American Chemical Society (1995-05) https://doi.org/dbzh27
DOI: 10.1021/ja00124a002

8. **CHARMM: A program for macromolecular energy, minimization, and dynamics calculations**
Bernard R. Brooks, Robert E. Bruccoleri, Barry D. Olafson, David J. States, S. Swaminathan, Martin Karplus
Journal of Computational Chemistry (1983) https://doi.org/bqh7f2
DOI: 10.1002/jcc.540040211

9. **Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids**
William L. Jorgensen, David S. Maxwell, Julian Tirado-Rives

Journal of the American Chemical Society (1996-01) https://doi.org/fvftxj
DOI: 10.1021/ja9621760

10. **Automated conformational energy fitting for force-field development**
Olgun Guvench, Alexander D. MacKerell Jr.
Journal of Molecular Modeling (2008-05-06) https://doi.org/bzphqw
DOI: 10.1007/s00894-008-0305-0 · PMID: 18458967 · PMCID: PMC2864003

11. **Machine learning of correlated dihedral potentials for atomistic molecular force fields**
Pascal Friederich, Manuel Konrad, Timo Strunk, Wolfgang Wenzel
Scientific Reports (2018-02-07) https://doi.org/gczmpn
DOI: 10.1038/s41598-018-21070-0 · PMID: 29416116 · PMCID: PMC5803249

12. **Empirical force fields for biological macromolecules: Overview and issues**
Alexander D. Mackerell
Journal of Computational Chemistry (2004) https://doi.org/dbhsbb
DOI: 10.1002/jcc.20082 · PMID: 15264253

13. **ff19SB: Amino-Acid Specific Protein Backbone Parameters Trained Against Quantum Mechanics Energy Surfaces in Solution**
Chuan Tian, Koushik Kasavajhala, Kellon Belfon, Lauren Raguette, He Huang, Angela Migues, John Bickel, Yuzhang Wang, Jorge Pincay, Qin Wu, Carlos Simmerling
American Chemical Society (ACS) (2019-06-17) https://doi.org/gf6rz8
DOI: 10.26434/chemrxiv.8279681

14. **Building a More Predictive Protein Force Field: A Systematic and Reproducible Route to AMBER-FB15**
Lee-Ping Wang, Keri A. McKiernan, Joseph Gomes, Kyle A. Beauchamp, Teresa Head-Gordon, Julia E. Rice, William C. Swope, Todd J. Martínez, Vijay S. Pande
The Journal of Physical Chemistry B (2017-04-06) https://doi.org/f92nv5
DOI: 10.1021/acs.jpcb.7b02320 · PMID: 28306259

15. **DrugBank 5.0: a major update to the DrugBank database for 2018**
David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, … Michael Wilson
Nucleic Acids Research (2017-11-08) https://doi.org/gcwtzk
DOI: 10.1093/nar/gkx1037 · PMID: 29126136 · PMCID: PMC5753335

16. **A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu**
Stefan Grimme, Jens Antony, Stephan Ehrlich, Helge Krieg
The Journal of Chemical Physics (2010-04-21) https://doi.org/bnt82x
DOI: 10.1063/1.3382344 · PMID: 20423165

17. **Optimization of Gaussian-type basis sets for local spin density functional calculations. Part I. Boron through neon, optimization technique and validation**
Nathalie Godbout, Dennis R. Salahub, Jan Andzelm, Erich Wimmer
Canadian Journal of Chemistry (1992-02) https://doi.org/c78qjn
DOI: 10.1139/v92-079

18. **RECAPRetrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry**

Xiao Qing Lewell, Duncan B. Judd, Stephen P. Watson, Michael M. Hann
Journal of Chemical Information and Computer Sciences (1998-04-11) https://doi.org/d4z4pf
DOI: 10.1021/ci970429i · PMID: 9611787

19. **Break Down in Order To Build Up: Decomposing Small Molecules for Fragment-Based Drug Design with eMolFrag**
Tairan Liu, Misagh Naderi, Chris Alvin, Supratik Mukhopadhyay, Michal Brylinski
Journal of Chemical Information and Modeling (2017-04-04) https://doi.org/f9x9bg
DOI: 10.1021/acs.jcim.6b00596 · PMID: 28346786 · PMCID: PMC5433162

20. **The Properties of Known Drugs. 1. Molecular Frameworks**
Guy W. Bemis, Mark A. Murcko
Journal of Medicinal Chemistry (1996-01) https://doi.org/fshj3p
DOI: 10.1021/jm9602928 · PMID: 8709122

21. **pyEFP: Automatic decomposition of the complex molecular systems into rigid polarizable fragments**
Alexey V. Odinokov, Nikita O. Dubinets, Alexander A. Bagaturyants
Journal of Computational Chemistry (2017-12-26) https://doi.org/gcq4qs
DOI: 10.1002/jcc.25149 · PMID: 29280158

22. **Approximateab initioenergies by systematic molecular fragmentation**
Vitali Deev, Michael A. Collins
The Journal of Chemical Physics (2005-04-15) https://doi.org/ch4zhg
DOI: 10.1063/1.1879792 · PMID: 15945620

23. **Fragmentation Methods: A Route to Accurate Calculations on Large Systems**
Mark S. Gordon, Dmitri G. Fedorov, Spencer R. Pruitt, Lyudmila V. Slipchenko
Chemical Reviews (2011-08-26) https://doi.org/b8tc8n
DOI: 10.1021/cr200093j · PMID: 21866983

24. **Systematic fragmentation of large molecules by annihilation**
Michael A. Collins
Physical Chemistry Chemical Physics (2012) https://doi.org/gf6v2d
DOI: 10.1039/c2cp23832b · PMID: 22373545

25. **Linear-scaling semiempirical quantum calculations for macromolecules**
Tai-Sung Lee, Darrin M. York, Weitao Yang
The Journal of Chemical Physics (1996-08-15) https://doi.org/bdtpqw
DOI: 10.1063/1.472136

26. **Flexible effective fragment QM/MM method: Validation through the challenging tests**
A. V. Nemukhin, B. L. Grigorenko, I. A. Topol, S. K. Burt
Journal of Computational Chemistry (2003-07-11) https://doi.org/dpwk5b
DOI: 10.1002/jcc.10309 · PMID: 12868106

27. **Application of the pople-santry-segal CNDO method to the cyclopropylcarbinyl and cyclobutyl cation and to bicyclobutane**
K. B. Wiberg
Tetrahedron (1968-01) https://doi.org/fvwkkh
DOI: 10.1016/0040-4020(68)88057-3

28. **Bond Order Analysis Based on the Laplacian of Electron Density in Fuzzy Overlap Space**
Tian Lu, Feiwu Chen
The Journal of Physical Chemistry A (2013-04-02) https://doi.org/f4t9v3
DOI: 10.1021/jp4010345 · PMID: 23514314

29. **Predicting Trigger Bonds in Explosive Materials through Wiberg Bond Index Analysis**
Lenora K. Harper, Ashley L. Shoaf, Craig A. Bayse
ChemPhysChem (2015-11-06) https://doi.org/f3jt5h
DOI: 10.1002/cphc.201500773 · PMID: 26458868

30. **Escaping Atom Types in Force Fields Using Direct Chemical Perception**
David L. Mobley, Caitlin C. Bannan, Andrea Rizzi, Christopher I. Bayly, John D. Chodera, Victoria T. Lim, Nathan M. Lim, Kyle A. Beauchamp, David R. Slochower, Michael R. Shirts, … Peter K. Eastman
Journal of Chemical Theory and Computation (2018-10-11) https://doi.org/gffnf3
DOI: 10.1021/acs.jctc.8b00640 · PMID: 30351006 · PMCID: PMC6245550

31. **ChemPer: An Open Source Tool for Automatically Generating SMIRKS Patterns**
Caitlin C. Bannan, David Mobley
American Chemical Society (ACS) (2019-06-21) https://doi.org/gf66hw
DOI: 10.26434/chemrxiv.8304578.v1

32. **Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model**
Michael J. S. Dewar, Eve G. Zoebisch, Eamonn F. Healy, James J. P. Stewart
Journal of the American Chemical Society (1985-06) https://doi.org/fd8bwp
DOI: 10.1021/ja00299a024

33. **On the Non-Orthogonality Problem Connected with the Use of Atomic Wave Functions in the Theory of Molecules and Crystals**
Per-Olov Löwdin
The Journal of Chemical Physics (1950-03) https://doi.org/dj2c35
DOI: 10.1063/1.1747632

34. **On the quantum theory of valence and bonding from the ab intio standpoint**
Mario A. Natiello, Jorge A. Medrano
Chemical Physics Letters (1984-03) https://doi.org/bdfk5f
DOI: 10.1016/0009-2614(84)85645-6

35. **Can we treat ab initio atomic charges and bond orders as conformation-independent electronic structure descriptors?**
T. Yu. Nikolaienko, L. A. Bulavin, D. M. Hovorun
RSC Advances (2016) https://doi.org/gf66tp
DOI: 10.1039/c6ra17055b

36. **Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database**
Paul C. D. Hawkins, A. Geoffrey Skillman, Gregory L. Warren, Benjamin A. Ellingson, Matthew T. Stahl
Journal of Chemical Information and Modeling (2010-03-17) https://doi.org/d4rb6g
DOI: 10.1021/ci100031x · PMID: 20235588 · PMCID: PMC2859685

37. **Corrected small basis set Hartree-Fock method for large systems**
Rebecca Sure, Stefan Grimme

Journal of Computational Chemistry (2013-05-14) https://doi.org/f42979
DOI: 10.1002/jcc.23317 · PMID: 23670872

38. **Psi4 1.1: An Open-Source Electronic Structure Program Emphasizing Automation, Advanced Libraries, and Interoperability**
Robert M. Parrish, Lori A. Burns, Daniel G. A. Smith, Andrew C. Simmonett, A. Eugene DePrince III, Edward G. Hohenstein, Uğur Bozkaya, Alexander Yu. Sokolov, Roberto Di Remigio, Ryan M. Richard, … C. David Sherrill
Journal of Chemical Theory and Computation (2017-06-06) https://doi.org/gcz64j
DOI: 10.1021/acs.jctc.7b00174 · PMID: 28489372

39. **MMFF VI. MMFF94s option for energy minimization studies**
Thomas A. Halgren
Journal of Computational Chemistry (1999-05) https://doi.org/brxdg7
DOI: 10.1002/(sici)1096-987x(199905)20:7<720::aid-jcc7>3.0.co;2-x

40. **Benchmarking Commercial Conformer Ensemble Generators**
Nils-Ole Friedrich, Christina de Bruyn Kops, Florian Flachsenberg, Kai Sommer, Matthias Rarey, Johannes Kirchmair
Journal of Chemical Information and Modeling (2017-10-18) https://doi.org/gb4v2v
DOI: 10.1021/acs.jcim.7b00505 · PMID: 28967749

41. **The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives**
Louis P. Hammett
Journal of the American Chemical Society (1937-01) https://doi.org/dz8d4r
DOI: 10.1021/ja01280a022