

Capturing non-local effects when fragmenting molecules for quantum chemical torsion scans

This manuscript ([permalink](#)) was automatically generated from [ChayaSt/fragmenter-manuscript@2ddbcab](#) on January 18, 2020.

Authors

- **Chaya D Stern**

 [0000-0001-6200-3993](#) ·  [ChayaSt](#) ·  [SternChaya](#)

Tri-Institutional 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

- **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 mechanics (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 reduced to smaller model entities containing the torsion of interest for QC torsion scans [2] for two main reasons as illustrated in [hold for 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 $\leq 15^\circ$ 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 [hold for CITE] are used. This adds a factor of 2D, where D is the dimension of the QC scan, to the computational cost. [hold for 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 $t2420*2 = s$.

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 [hold for 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 [30]. 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 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 organized 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 [31,7,8,9]. 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 1, A illustrates such a case with a series of biphenyls in different protonation states. While the MM torsion profiles are all the same (1 D), the QC torsion profiles are different for each protonation state (1 E). 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 slightly 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 (1 B) 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 which is reflected in the same torsion parameters for all four molecules (1, D) 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 [1,32,33] 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.

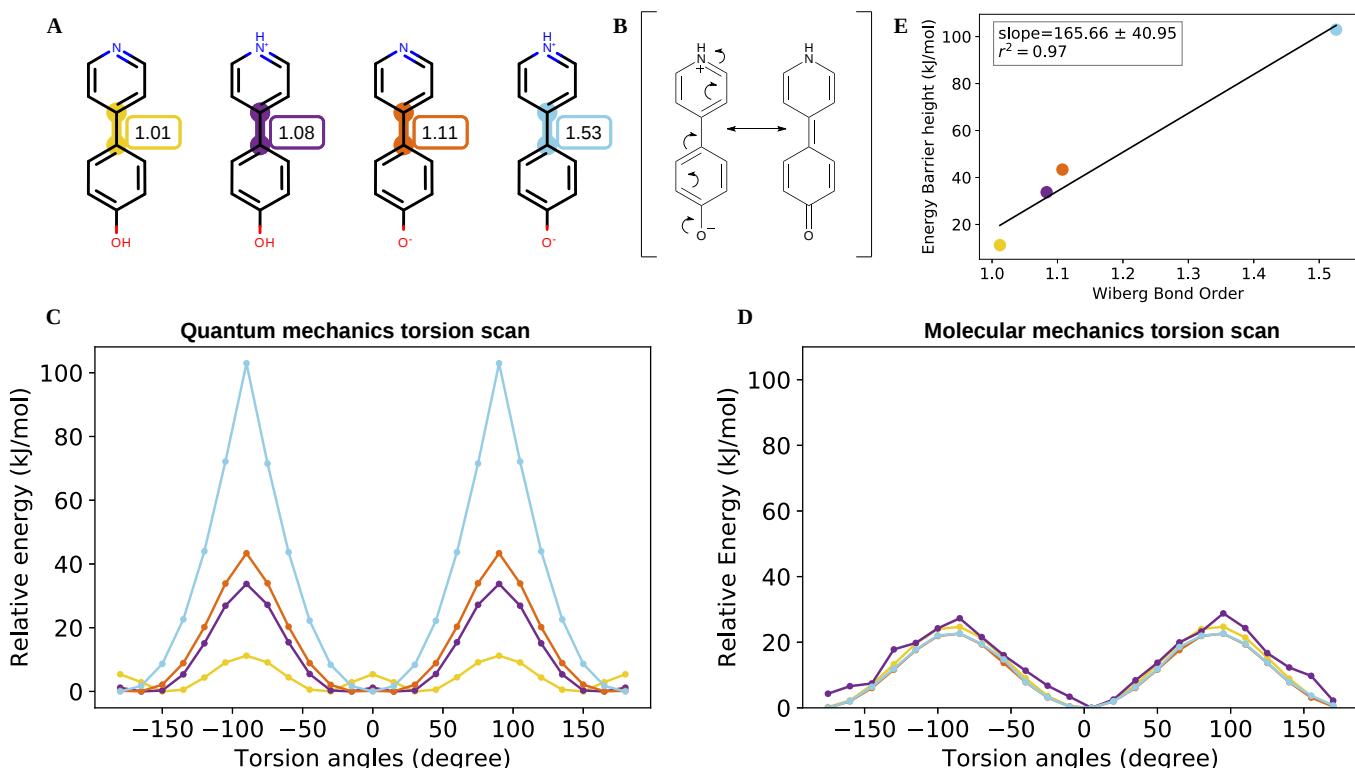


Figure 1: Torsion profiles can be sensitive to remote chemical changes in a molecule [A] Biphenyl protonation states and tautomers with increasing Wiberg bond order for the central bond. [B] The resonance structure of the biphenyl zwitterion shows that the central bond is conjugated. The Wiberg bond 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] Relative MM energy as a function of torsion angle of the central bond. The colors in the MM scan correspond to the highlighted bonds in A. [E] Torsion barrier heights vs WBOs. The color of the data points correspond to the highlighted bonds in A. The QC torsion barrier height is linear WBO.

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[34]. Wiberg originally formulated it for the CNDO basis set [27] but it can be easily extended to other semi-empirical QC methods such as AM1 [35] and PM3 [36]. 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

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

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 [37,38] which is how it is calculated in Psi4.

We calculated the WBO from AM1 calculations for the biphenyl series as shown in figure 1 A. 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 1 E), the relationship is linear with an R^2 of 0.97.

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

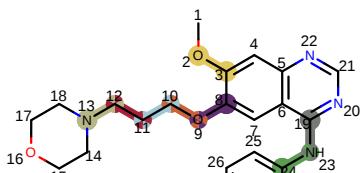
Since the WBO can be calculated from a cheap AM1 calculation, is indicative of a bond's conjugation, and is correlated with torsion energy barrier height, it is an attractive measure to use as a surrogate when automating fragmentation or interpolating torsion force constants. However, WBOs are conformational dependent [39,40] 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 have higher variance with respect to conformation and that bonds involved in conjugated systems have WBOs that are correlated with each other.

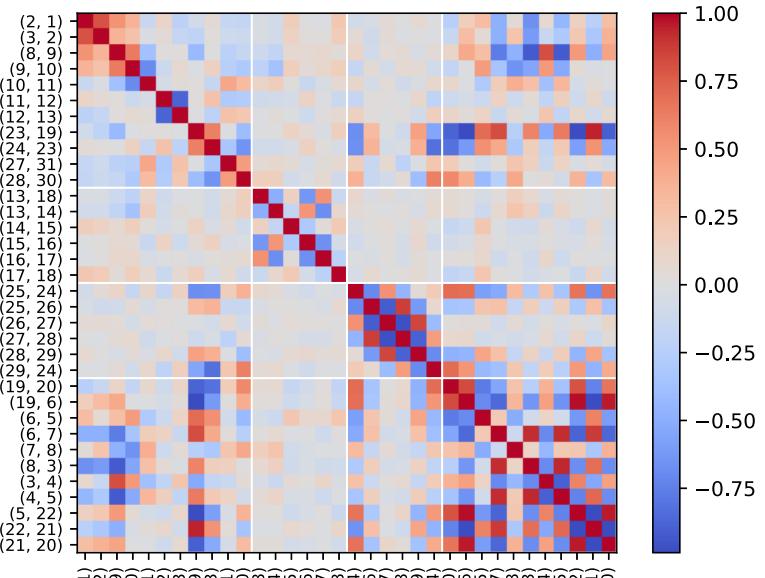
A

Gefitinib with rotatable bonds highlighted

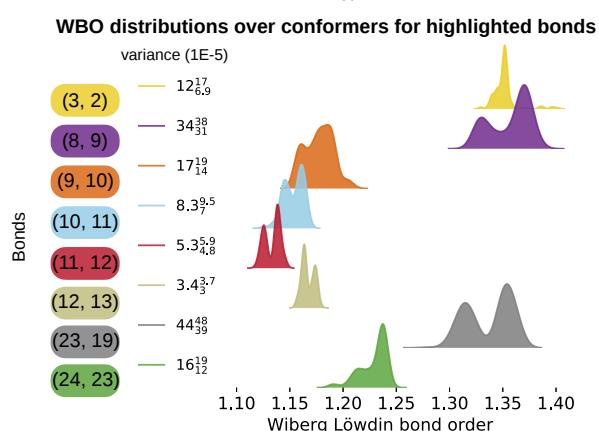


C

Pearson correlation coefficients of WBOs over conformations



B



D

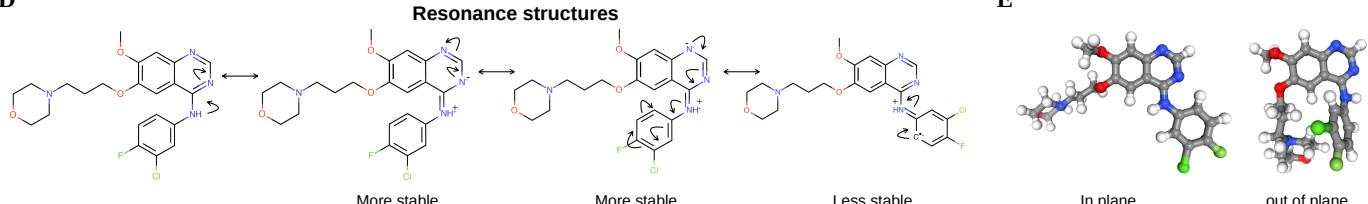


Figure 2: Variance and correlations of Wiberg bond order distributions with respect to conformations are higher for conjugated bonds [A] Gefitinib, with its rotatable bonds highlighted and numbered to correspond with figures B and C. [B] WBO distributions over 232 conformations of the highlighted, rotatable bonds. The colors of the distributions correspond to the colors in the highlighted bonds in A. The variance and their 95% confidence interval are shown on the left (with exponent base of $1E-5$). The single non conjugated bonds (blue, (10, 11), red (11, 12), and olive (12, 13)) have lower variance than conjugated bonds (yellow (3, 2), purple (8, 9), orange (9, 10), grey (23, 19), and green(24, 23)). [C] Correlation plot of WBOs every bond in Gefitinib against WBOs of all other bonds over 232 conformations. The white lines indicate ring systems. Bonds in conjugated systems have higher correlations between their WBOs (see the aromatic ring systems in the two lower right diagonal squares). Both bonds (23, 19) (grey) and (24, 23) (green) have WBOs that are correlated with their neighboring ring systems, but bond (23, 19) are more correlated with the ring systems than the green bond (24, 23). [D] Resonance structures of Gefitinib illustrates why the grey bond (23, 19) has higher variance than the green bond (24, 23) even if both bonds are conjugated. When the double bond is on bond (23, 19), the negative charge is on a nitrogen which is the more stable form, vs the resonance structure where the double bond is on (24, 23) with the negative charge on a carbon. [E] The conformations of the molecule for the highest WBO and lowest in the distribution. The mode with higher WBOs has bond (23, 19) in plane with quinazoline which allows for stronger conjugation while the mode with lower WBOs has the bond out of plane so there is less electron population overlap in out of plane conformation.

To investigate how WBOs change with conformation, we used Omega [41] to generate conformers for a set of kinase inhibitors [hold for SI] and calculated the Lowdin-Wiberg bond order for each conformation from a B3LYP-D3(BJ) / DZVP [16,17,42,43] geometry optimized calculation using Psi4 [44]. Omega is a knowledge-based conformer generator that uses a modified version of MMFF94s [45] to score conformations. It has been shown to accurately reproduce experimentally observed crystallography conformers in the Platinum benchmark dataset [46]. Figure 2 illustrates the results for Gefitinib (2, A), a representative molecule. Figure 2, B shows the distribution of WBOs for all rotatable bonds color coded with the colors used to highlight the bonds in Gefitinib (2, A). Single carbon-carbon bonds, and carbon-nitrogen bonds formed by atoms numbered 10 - 13 are freely rotating. This is reflected by the tighter distribution (lower variance) of WBOs around closer to one 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. It is interesting to note the difference in the WBOs for the conjugated bonds formed by the nitrogen between the quinazoline and chloro fluoro 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 23-19 (the grey distribution) has two modes clear modes of almost equal weights, the WBO distribution for bond 24-23 has lower variance. This is in agreement with the resonance structures shown in figure 2. 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 [hold for SI]. In addition, when we inspected the conformations associated with the highest and lowest WBO in the grey distribution (fig 2, E) we found that conformations with lowest WBO on bond 19-23 had that bond out of plane while the conformation with the highest WBO value had the bond in plane which allows conjugation. We found similar results from WBOs calculated form QC torsion scans. Fig 6 shows Lowdin-Wiberg bond order for each point in the QC corresponding torsion scans. The WBOs are 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 [29]. 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. We found that the trends discusses above are similar when using semi-empirical methods such as AM1 and HF3C (hold for SI)

Bonds in conjugated systems have highly correlated conformation-dependent WBOs

We found that certain bond orders are strongly correlated or anti-correlated with each other, indicating strong electronic coupling. Figure 2 C 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 (25-29) are strongly anti-correlated with each other. This trend is different in the ring formed by atoms 13-18 which is not aromatic. In this ring, bonds 13-18, 13-14, 16-15 and 16-17 (which involve electron rich atoms O and N) have Pearson correlation coefficients with absolute values higher than for the other bonds in the ring, but lower than the bonds in the aromatic ring. The bonds involved in the methoxy groups (atoms 1-3 and 8-10) are 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. The trends are similar for other kinase inhibitors tested as shown in (hold for SI).

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

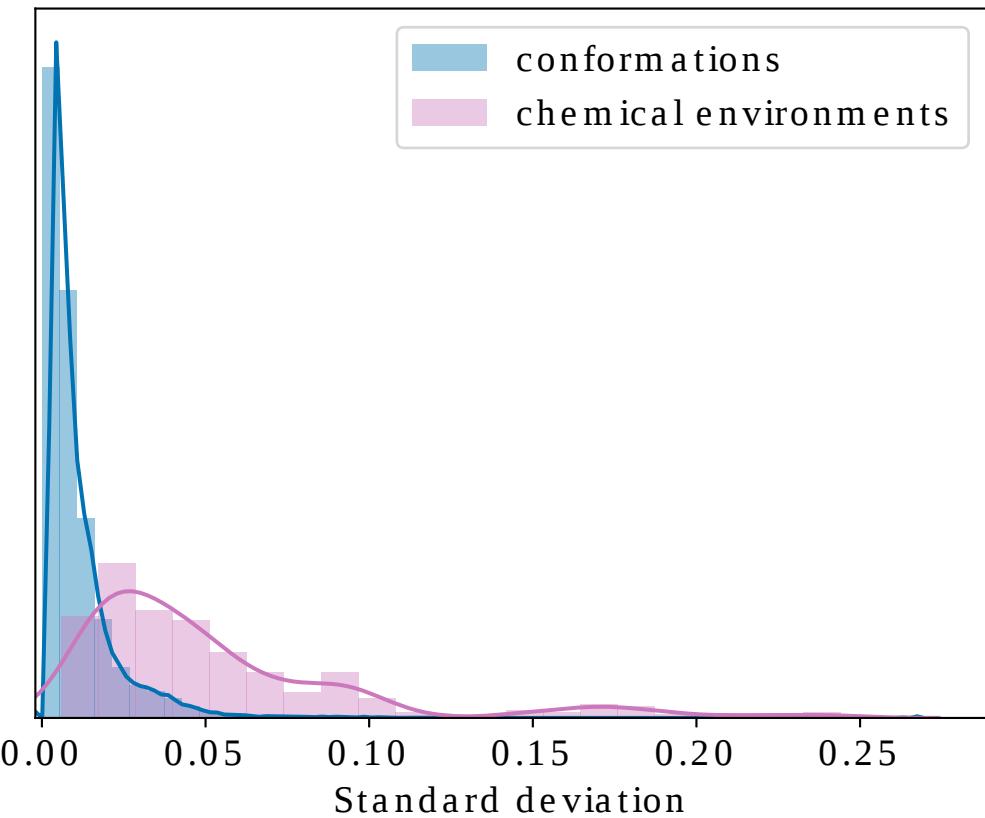


Figure 3: Standard deviations of conformation dependent WBO distributions are smaller than standard deviations of WBO distribution of the same bond in different chemical environments The distribution of standard deviations of WBO distributions over conformations is shown in blue. The distribution of standard deviations of ELF10 WBO distributions for the same bond in different chemical environments is shown in pink. The changes in WBO due to conformations are smaller than the changes in WBO due to chemical changes around the bond. The validation set (SI 9), was used to generate these distributions. The blue std distribution was calculated over 140602 WBO dependent distributions (this is the number of individual fragments in the dataset). The pink std distribution was calculated over 366 distributions of 366 bonds in different chemical environments (of the 140602 fragments)

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 3 shows the distribution of standard deviations of WBO distributions with respect to conformation in blue. Most of the standard deviations fall below 0.03, 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 [47] resolves the issue of sensitivity of molecular mechanics electrostatic energies from QM derived charges.

Leave to Christopher Bayly to describe

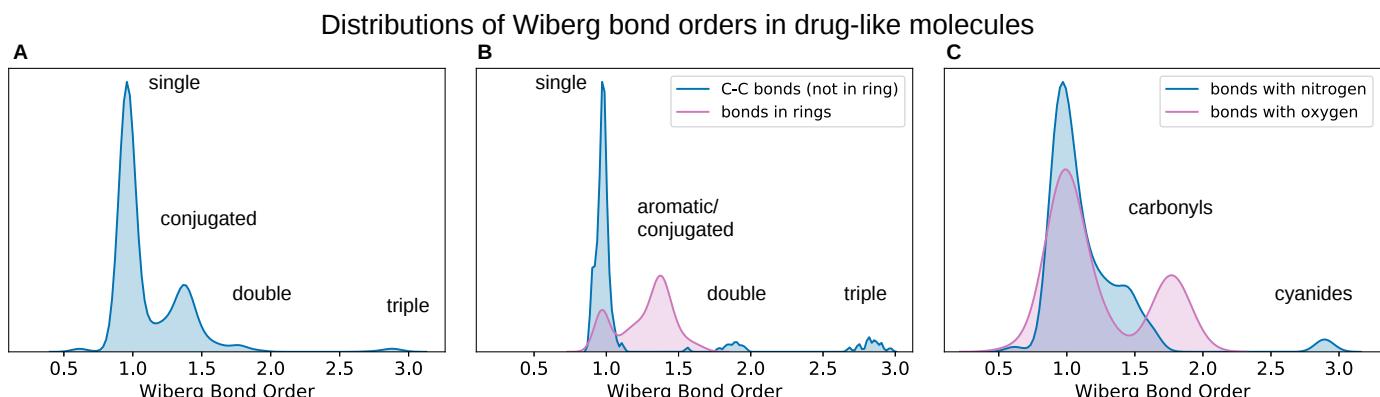


Figure 4: Distribution of WBO in drug-like molecules have peaks at expected values [A] The distribution of all WBOs for all bonds in molecules in set. The peak at one, close to two and close to three correspond to single, double and triple bonds. the density between one and two correspond to aromatic and conjugated bonds. The peak at ~0.7 correspond to bonds that include sulfur and phosphorous which are longer, weaker bonds. [B] The blue distribution includes carbon - carbon bonds that are not in rings. The peaks at one, two and three correspond to single, double and triple bonds. The pink distribution include bonds that are in rings. The peak at one corresponds to single bonds and the density between one and 1.5 are aromatics. [C] The blue distribution includes bonds that have either one or two nitrogen. Many of these bonds are conjugated as demonstrated by the density around 1.5. The density at three corresponds to nitriles. The pink distribution include bonds that have oxygens. The peak at two corresponds to carbonyls.

This method can also be applied to derive WBOs that are insensitive to conformers. To check how well ELF10 estimated W_{AB} recapitulates the multiplicity of bonds, we calculated W_{AB} from ELF10 AM1 calculations for all bonds in a set of molecules shown in [hold for SI]. The distribution in figure 4 corresponds closely with bond multiplicity. The density at ~ 0.7 correspond bonds involving sulfur and phosphorous since these are weaker and longer bonds. The peak at 1.0 corresponds to C-H and C-C bonds, the peak close to 1.5 corresponds to bonds in aromatic rings, the peak close to 2.0 corresponds to double bonds, and finally the triple bonds form the last peak. Figures 4 B and D separate out different kinds of bonds to more clearly illustrate what the WBO captures. Figure 4 B shows carbon - carbon bonds not in rings (blue) and bonds in rings (pink). The carbon - carbon distribution has distinct peaks at one, two and three corresponding to single, double and triple bonds. There is also a smaller peak at 1.5 that corresponds to conjugated bonds. The pink distribution includes bonds in rings and has a peak at one and 1.5 which corresponds to aliphatic and aromatic rings, respectively. Figure 4 D shows distributions with bonds that have nitrogens (blue) and oxygens (pink). The peaks occur at the expected values; 1, 1.5 and 3 for nitrogen which corresponds to single, conjugated and triple bonds and 1 and 2 for oxygens which correspond to single and carbonyl 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 [48] 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 substituted phenyls and pyridine 5 A with 25 functional groups that cover a wide range of electron donating and withdrawing groups. We then calculated the ELF10

WBO for the bond attaching the functional group to the aromatic ring for all functional groups which resulted in 128 ($25 \times 5 + 3$) data points for each functional group. 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. The resulting distributions are in figure 5, B. It is interesting to note that the trend of decreasing WBOs for more electron donating groups are anti correlates with increasing Hammett substituent constants. In { [???] } C and D, the AM1 ELF10 WBO of the bond between the functional group and benzoic acid is plotted against their Hammett para and meta substituent constants. 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.

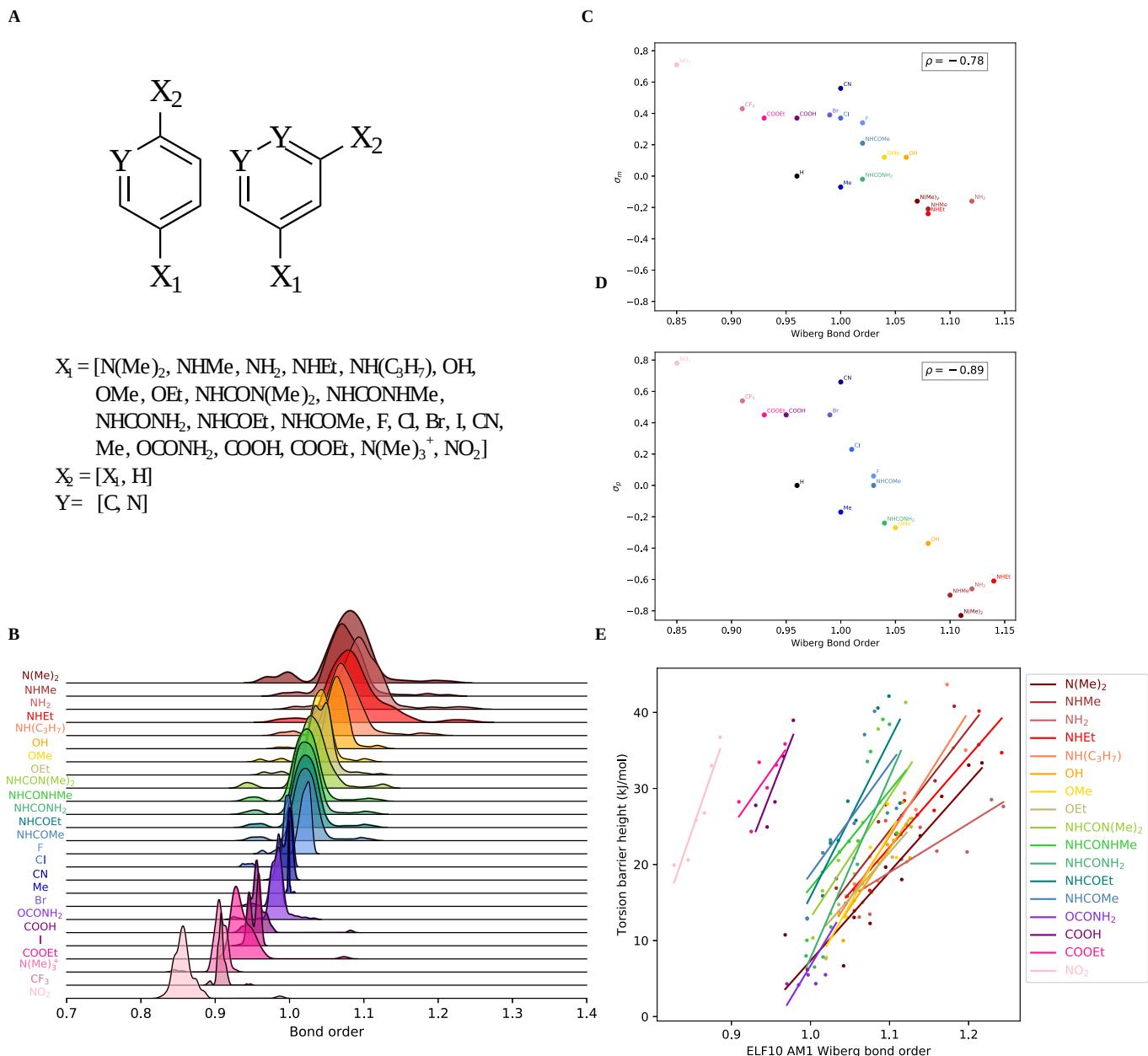


Figure 5: Change in AM1 ELF10 WBOs correlates with barrier heights in torsion profiles [A] Systems and functional groups used in the substituted phenyl set. The functional groups were chosen to span a large range of electron donating and withdrawing groups. [B] AM1 ELF10 WBO distributions for the bond between the phenyl ring and X_1 in different chemical environments [C] Hammett sigma para parameters vs AM1 ELF10 WBOs of X_1 para to carboxylic acid [D] Hammett sigma meta parameters vs AM1 ELF10 WBOs of X_1 meta to carboxylic acid [E] Selected QC torsion scan barrier heights vs ELF10 AM1 WBOs

Table 1: slope and associated statistics for torsion barrier height vs WBO for selected functional groups

functional group	slope	r^2	P value	standard error
N(Me) ₂	116.916250	0.880571	0.000019	14.352479
NHMe	134.517241	0.896103	0.000033	16.194060
NH ₂	64.266814	0.577819	0.017438	20.763035
NHEt	119.513117	0.836392	0.000552	19.978559
NH(C ₃ H ₇)	163.763962	0.871119	0.000236	23.808080
OH	154.824936	0.729133	0.003393	35.666975
OMe	185.309951	0.800852	0.006494	41.326214
OEt	119.662499	0.479470	0.038706	47.124984
NHCON(Me) ₂	159.312781	0.579507	0.010533	47.979491
NHCONHMe	127.645256	0.434585	0.053430	55.030363
NHCONH ₂	238.119990	0.734429	0.003158	54.120536
NHCOEt	205.802258	0.692740	0.005374	51.804680
NHCOMe	144.320372	0.457787	0.065376	64.121783
OCONH ₂	172.716947	0.508785	0.111518	84.854215
COOH	267.218639	0.739937	0.061417	91.463658
COOEt	149.012778	0.581298	0.077956	63.233457
NO ₂	302.072242	0.909174	0.003192	47.737899

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 (hold for SI). 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 ammonium and functional 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 [hold for SI]. We show a representative series of torsion scan for the nitro functional group in figure 6A. The torsion energy barrier height increase with increasing ELF10 WBO of the bond. In addition, 6B 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 biphenyl set discussed in the previous section. We also found that the linear relationship between WBOs and torsion energy barrier height shown in 1D generalizes to all functional groups tested in this set 5E. Table 1 lists the slopes and associated statistics for the fitted lines.

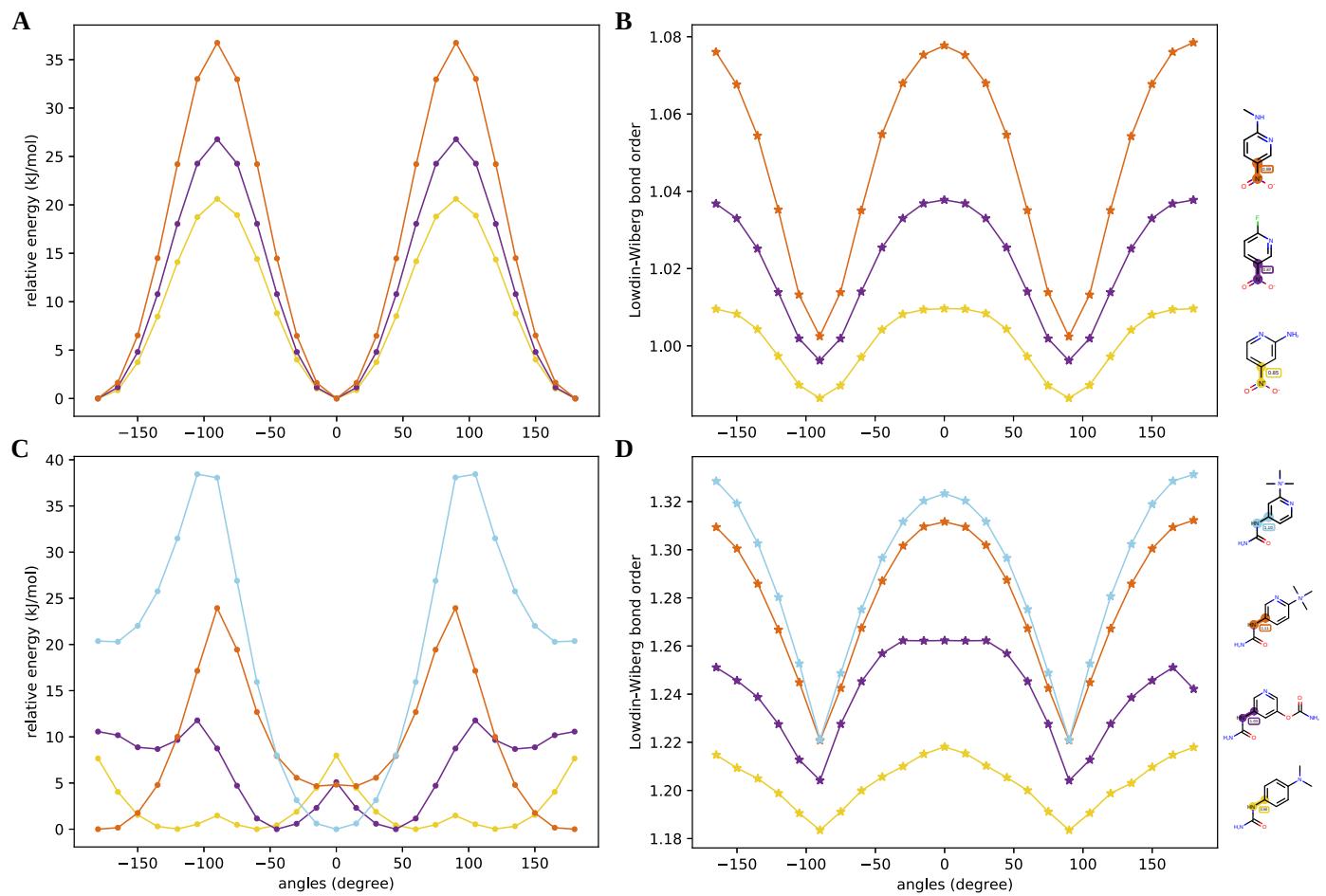


Figure 6: Löwdin-Wiberg bond orders are anti correlated with QC torsion scans [A] QC torsions for methylamino in series of different chemical environment. Barrier heights increase with increasing ELF10 AM1 WBOs [B] Löwdin-Wiberg bond orders calculated at each point in the QC torsion scan using the same level of theory. The bond orders are anti correlated with QC torsion scans [C] QC scans for urea in a series of different chemical environment. Both profiles and energy barriers change with ELF10 AM1 WBOs [D] Löwdin-Wiberg bond orders are not perfectly anti correlated to QC scans

For most functional groups, the change in WBOs correspond to changes in torsion barrier heights. [hold for SI]. However, for some functional groups, the change in WBO does not fully capture the differences in torsion scans because not only do the torsion energy barrier heights increase, but the profile changes considerably as shown in 6C for urea. Interestingly, the Löwdin-Wiberg bond order scans do have the same profiles 6 D.

When we compare the standard deviations of WBO distributions with respect to conformation versus with respect to changes in chemical space (figure 3, pink distribution), we find that the changes in ELF10 WBO for remote chemical environment changes are usually 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

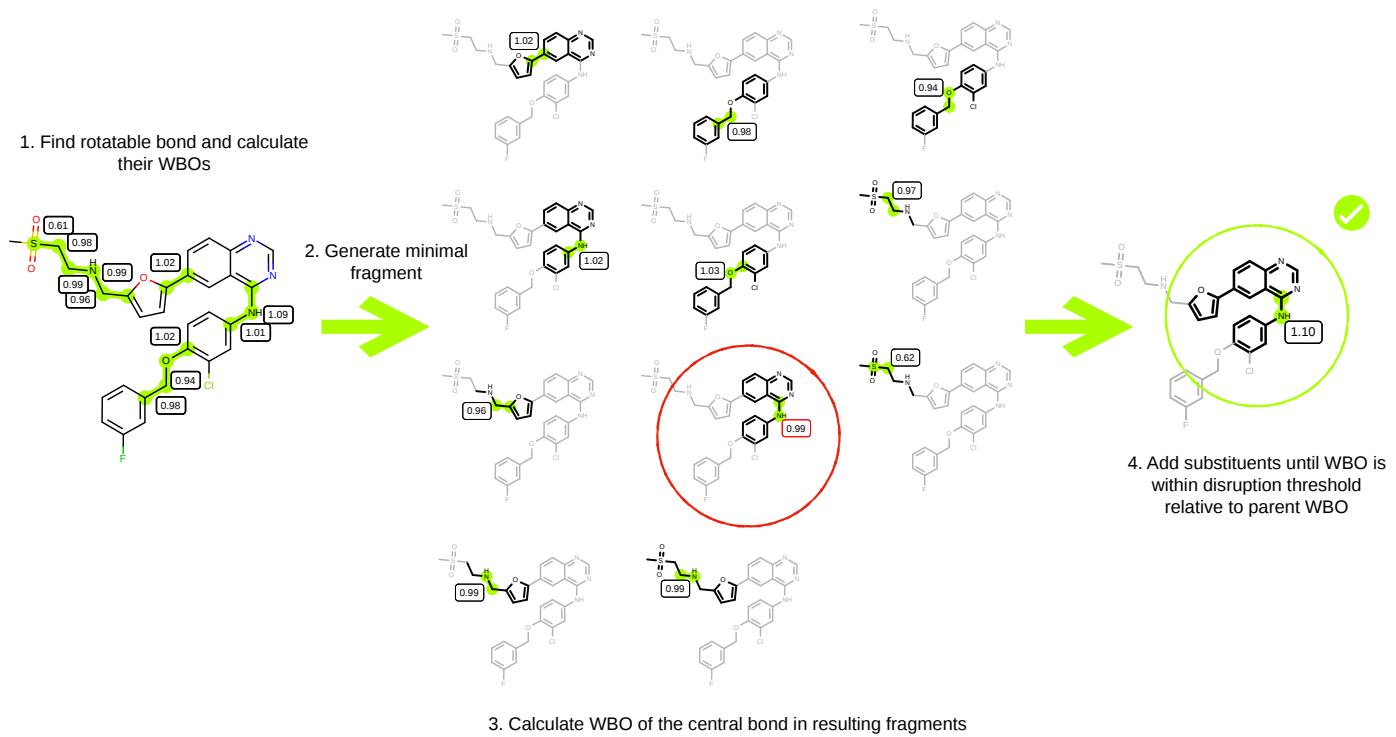


Figure 7: Illustration of fragmentation scheme using WBOs First, we find the rotatable bonds and calculate their ELF10 WBOs. Then, for each rotatable bond, we find the minimal fragment as described in the text. Then we calculate WBOs for the bonds around which the minimal fragments were generated and check if the new WBO is within the disruption threshold relative to the parent's WBO. If the absolute difference is greater than the disruption threshold, substituents are added, one at a time, and the WBO is recalculated. The fragment continues to grow until the central bond's WBO is within the disruption threshold of the WBO of the bond in the parent molecule

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 (with slight changes) proposed by [49] by considering resonance via WBOs. The scheme, illustrated in figure 2 is as follows:

1. Find acyclic bond. For this step we use the SMARTS pattern `[!$(*##)&!D1]-,=;@[$(*##)&!D1]`.
2. Keep the four atoms in the torsion quartet and all atoms bonded to those atoms (this ensures that all 1-5 atoms are included in the minimal fragment).
3. If any of the atoms are part of a ring or functional group shown in table 2, include ring and functional groups atoms to avoid ring breaking and fragmenting functional groups that contain more than one heteroatom.
4. Keep ortho substituents to the rotatable bond and the bonds directly bonded to the rotatable bond it is involved in the sterics of the torsion. This rule is slightly different than described in [49].
5. N, O and S are capped with methyl. All other open valence atoms are capped with hydrogen.
6. Calculate WBO for fragment.
7. 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.

Table 2: Functional groups that are not fragmented. This list is not comprehensive and users can add their own functional groups they do not want to fragment

Chemical group	SMARTS pattern
azo	'[NX3][N]' = '
nitric oxide	'[N]-[O]' '

Chemical group	SMARTS pattern
amide	'[#7][#6](=[#8])', [#7][#6](-[O-])'
urea	'[NX3][CX3](=[OX1])[NX3]'
aldehyde	'[CX3H1](=O)[#6]'
sulfoxide	[#16X3]=[OX1], '[#16X3+][OX1-]'
sulfonyl	'[#16X4](=[OX1])=([OX1])'
sulfinic acid	'[#16X3](=[OX1])[OX2H,OX1H0-]'
sulfinamide	'[#16X4](=[OX1])=([OX1])([NX3R0])'
sulfonic acid	'[#16X4](=[OX1])(=[OX1])[OX2H,OX1H0-]'
phosphine oxide	'[PX4](=[OX1])([#6])([#6])([#6])'
phosphonate	'P(=[OX1])([OX2H,OX1-])([OX2H,OX1-])'
phosphate	'[PX4](=[OX1])([#8])([#8])([#8])'
carboxylic acid	'[CX3](=O)[OX1H0-,OX2H1]'
nitro	'([NX3+](=O)[O-])', '([NX3](=O)=O)'
ester	'[CX3](=O)[OX2H0]'

(Add discussion on the changes I made to the Pfizer scheme for our minimal fragment before we start building out and why. Add SI figure justifying it)

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

This fragmentation scheme improves upon [49], however, it leaves some parameters up to the user. In order to assess 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 has less than 10 rotatable bonds
5. Molecule must have at least one aromatic ring
6. Molecule has only one connected component

This left us with 730 small molecules [hold for SI]. 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 [hold for SI]. 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. The resulting dataset is very rich where exquisitely nuanced, long distance chemical changes are detected.

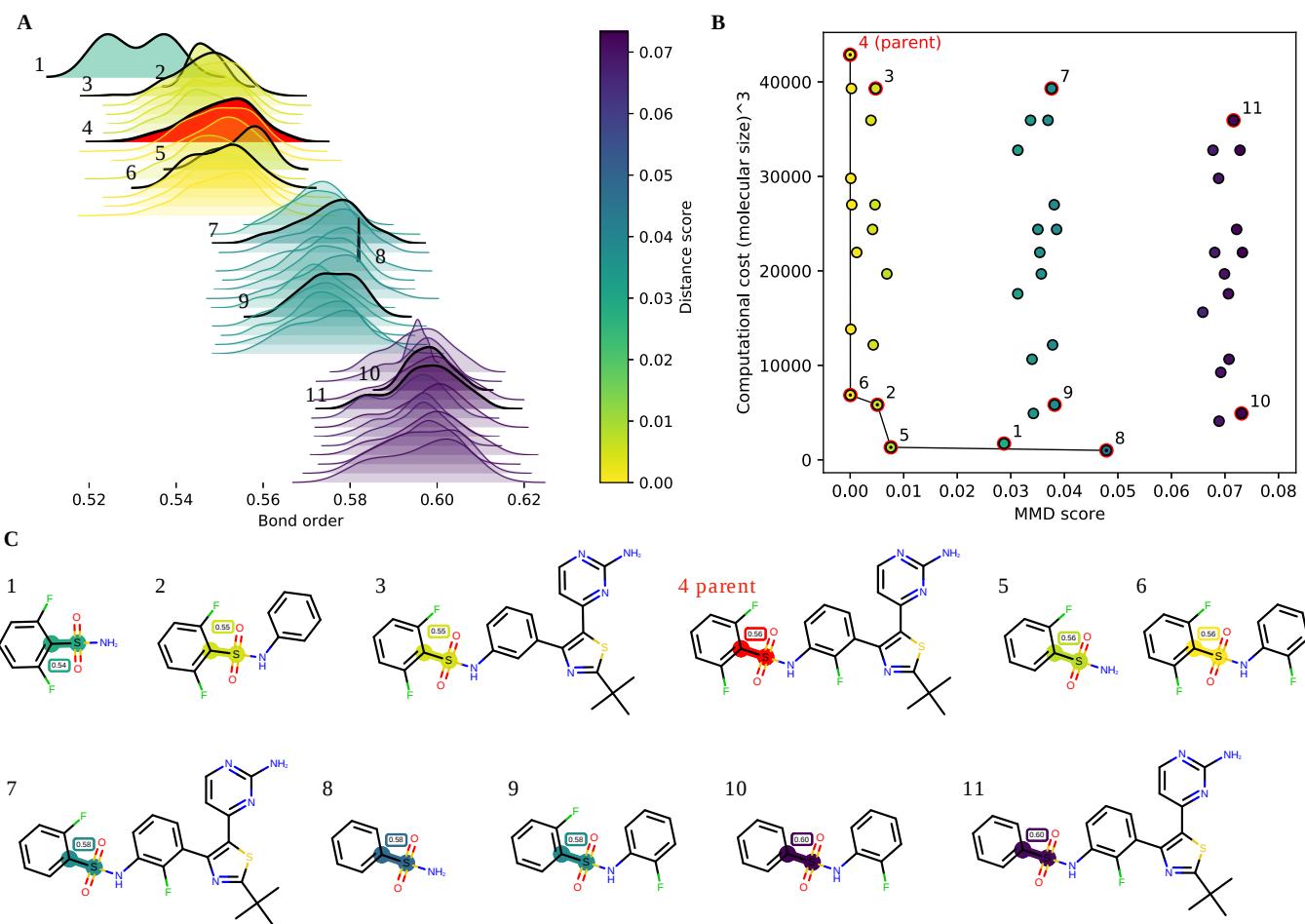


Figure 8: Changes in WBO distributions are a good indicator of remote substituent electronic effects [A] An illustrative example of the shift in the conformationally dependent WBO distributions due to crucial chemical changes such as the loss of Fluorine. The distributions are shaded with their corresponding distance score shown in the colorbar on the right. The parent molecule WBO distribution (numbered 4) is shaded red. Selected distributions are outlined and the corresponding fragments are shown in C. [B] Computational cost of fragment (heavyatoms^3) vs distance score (MMD) of the fragment indicates that it is possible to reduce the cost of torsion scans without destroying the torsion profile. The black line is the Pareto frontier, or the cheapest fragment with the best score at that size. The selected fragment should be on the Pareto front at the lower left corner. [C] Selected fragments. Bonds are highlighted with their distance score. The ELF10 WBO is shown in the boxes above the highlighted bonds.

Figure 8 shows an example of the results of combinatorial fragmentation and how this data can be used to benchmark fragmentation schemes. All rotatable bonds in the parent molecule, Dabrafenib (8C 4) were fragmented into 11 fragments (in this example, the trimethyl was not fragmented) resulting in 108 connected fragments when all connected combinations were generated. Of those 108 fragments, 44 fragments contained the bond between the sulfur in the sulfonamide and phenyl ring highlighted in fragments in 8C. When the WBOs were calculated for all Omega generated conformers for each of the 44 fragments, the resulting WBO distributions clustered into 4 distinct bins (8A). Upon closer inspection we found that the shifts of the distributions corresponded to specific remote substituent changes, in this case the loss of fluorine and the phenyl ring bonded to the nitrogen in the sulfonamide. Here, these two changes cause the distributions to shift in opposite directions. While the loss of a fluorine on the phenyl bonded to the sulfure shifts the distribution to the right, the loss of the ring bonded to the nitrogen shifts the distributions to the left illustrating that the changes are multi dimensional. Fragments 2, 3, 4, and 6 (8C) all contain two fluorine and fall in the same cluster as

the parent molecule, regardless if the rest of the molecule is included in the fragment. Fragments 7 and 9 only have one fluorine on the phenyl ring and both of their distributions are shifted to the right relative to the parent WBO distribution. Fragments 10 and 11 have no fluorine on the ring and are shifted to the right even more. Since removing the ring bonded to the nitrogen shifts the WBO distribution in the opposite direction, fragment 1, while having two fluorine, is shifted to the left of the parent distribution, fragment 5 WBO distribution overlaps with the parent WBO distribution even if it only has one fluorine, and fragment 8 is only shifted slightly to the right of the parent WBO distribution with no fluorine.

Scoring how well fragments preserve chemical environments using WBO distributions

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

$$MMD(P, Q) = \|\mathbb{E}_{X \sim P}[\varphi(X)] - \mathbb{E}_{Y \sim Q}[\varphi(Y)]\|_{\mathcal{H}}$$

where the feature map $\varphi : \mathcal{X} \rightarrow \mathcal{H}$ we use is squared $\varphi(x) = (x, x^2)$ and the MMD becomes:

$$MMD = \sqrt{(\mathbb{E}[X] - \mathbb{E}[Y])^2 + (\mathbb{E}[X^2] - \mathbb{E}[Y^2])^2} \quad (1)$$

where X is the parent WBO distribution and Y is the fragment WBO distribution. Including the squared mean incorporates the variance of the distribution and helps distinguish distributions both with different means and variances. It is important incorporate changes in variance given how the variance of the WBO distributions change for different chemical environments (see figure 6B and D. Change in variance corresponds to change in relative barrier heights).

In figure 8, the MMD score, which we call the distance score, is shown with the color map. The distributions in 8A are shaded with the distance score. The scores clearly differentiates the shifted distributions.

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 computational cost of the fragment. We estimate the computational cost of a fragment by cubing its number of heavy atoms because DFT calculations grow by $O(n^3)$. The distance score calculated with MMD indicates how far the fragment's WBO distribution is or how much the chemical environment changed from its parent. 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 8B shows an illustrative example of this. The fragments data points on the Pareto front have a black dot in the center. The numbers on the annotated data points correspond the the numbered fragments in 8C. Fragment 6 has is the smallest fragment with the smallest distance to the parent molecule. It has the important chemical moieties, such as all three fluorine and the ring bonded to the nitrogen. While fragments 2 and 5 are also on the Pareto front, the missing ring and fluorine increase the distance score, however, it is not clear if this difference is significant. It is interesting to note that fragment 3, which is also missing the fluorine on the ring bonded to the nitrogen, is shifted in the distance score relative to the parent by the same amount as fragment 2 from 6, even if it has all other parts of the

molecule adding credence to the fact that the small difference in the distance score does pick up on this chemical change. The trend is in the opposite direction for molecules missing a fluorine on the ring bonded to the sulfur. Fragment 9 and 10 both contain fluorine on the ring bonded to the nitrogen, but have greater distance scores than the fragments without that fluorine (data points to the bottom left of 9 and 10. Fragments not shown). Data points 7 and 11 illustrate that having larger fragments will not improve the distance score if the important remote substituents are not in the fragment. Fragment 9, while a lot smaller than fragment 7, has the same distance score because they both are missing the important fluorine. Data points 10 and 11 show the same trend for the fragments missing both fluorine. While fragments 1, 5 and 8 are all small, the loss of the ring results in larger distance scores.

In molecule 12, both the amide and ester bond are sensitive to the same circled negatively charged oxygen.

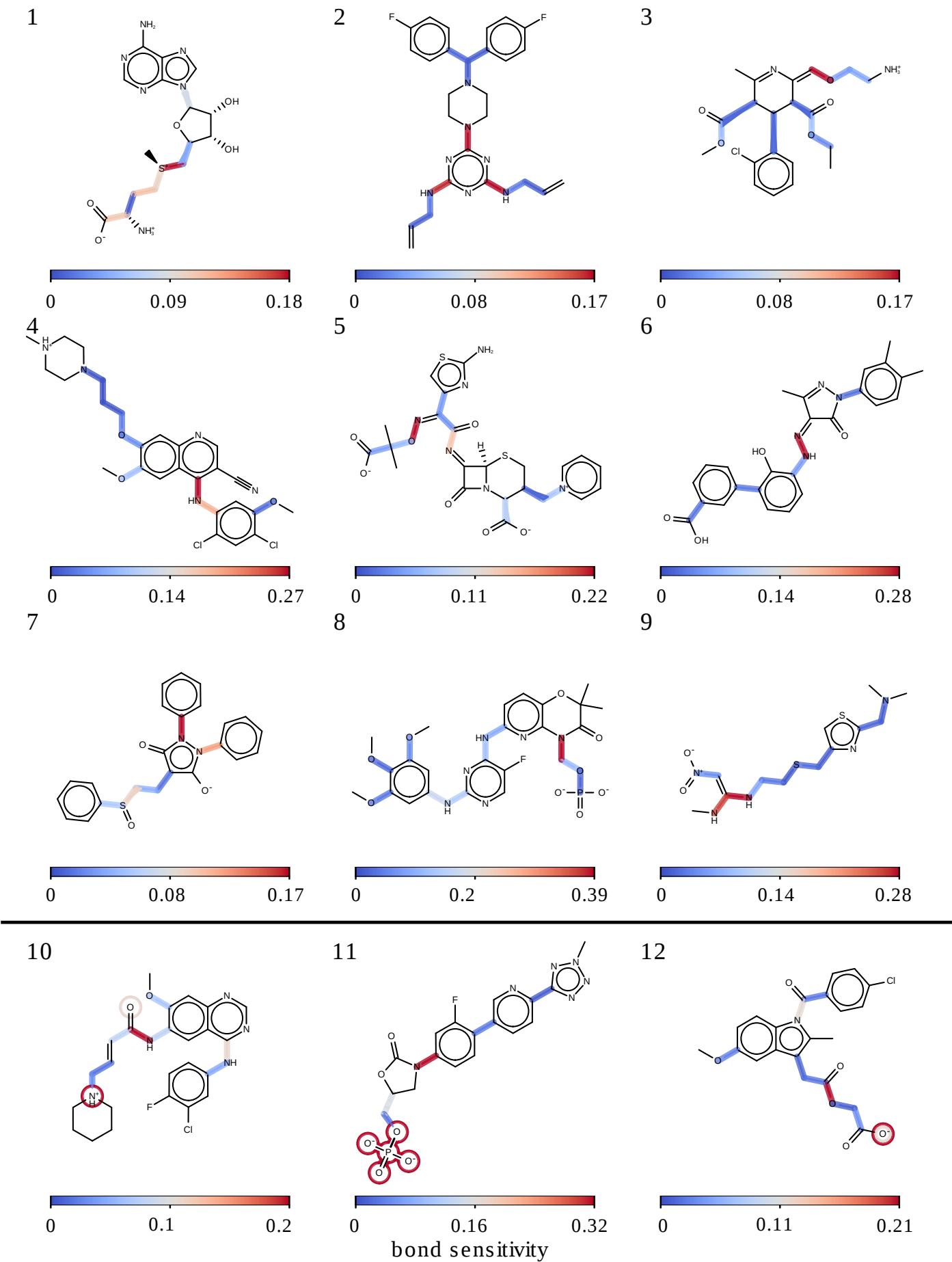


Figure 9: Some bonds are more sensitive to peripheral fragmentation and are therefore challenging to fragment Selected molecules of the validation set. Bonds are highlighted by how sensitive they are to remote substituent changes. The bond sensitivity score is given by finding the maximum distance score of WBO distributions of the bond in fragments generated by exhaustive fragmentation, to the WBO distribution of the same bond in the parent

molecule. Fragments used to get the maximum distances include all 1-5 atoms around the highlighted bonds. Molecules 10-12 also show the atoms the bonds are sensitive to. The atoms are circled with the same color as the bond that is sensitive to it. The rest of the molecules used in the validation set are shown in SI figure [14](#)

The goal of any fragmentation scheme is to find fragments on the Pareto front that minimize both the changes in the chemical environment of the bond and fragment size. In other words, they should be on the lower left corner of the plot. To test our fragmentation scheme, we wanted to find the molecules that are challenging to fragment. To do that, we scored the WBO distributions of all resulting fragments from our exhaustive fragmentation experiment using equation [1](#) and chose 100 molecules that had bonds where fragments that included all 1-5 atoms around the central bond had the highest distance scores. Selected molecules with the bonds highlighted according to their sensitivity are shown in figure [9](#). The rest of the molecules are shown in SI figure [14](#). This set included many molecules in charged states. The sensitivity score of the bonds are given by taking the MMD of the fragment where the WBO distribution of that bond has the greatest distance relative to the WBO distribution of the bond in the parent molecule. This is a good indication of a bond's sensitivity to peripheral fragmentation because the more its WBO distribution shifts relative to the parent when fragmented, the more the electronic population overlap around that bond changes with remote chemical changes.

Not all bonds are equally sensitive to such changes. This is shown by how different the sensitivity score is for different bonds in the same molecule in figure [9](#). The general trend observed is that conjugated bonds, and bonds including open valence atoms such as N, O, and S, are more sensitive to peripheral cuts. Molecules 10-12 (fig[9](#)) also show which chemical moiety the bond is sensitive to, indicated by circles around the atoms which are colored with the corresponding bond's sensitivity score. In molecule 10, the WBO distribution of the amide bond shifts significantly if the positively charged nitrogen is removed regardless if the rest of the molecule is intact (data not shown). In molecule 11, the removal of the phosphate group shifts the distribution of the red bond. In molecule 12, both the amide and ester bond are sensitive to the same negatively charged oxygen indicated by two circles around the oxygen.

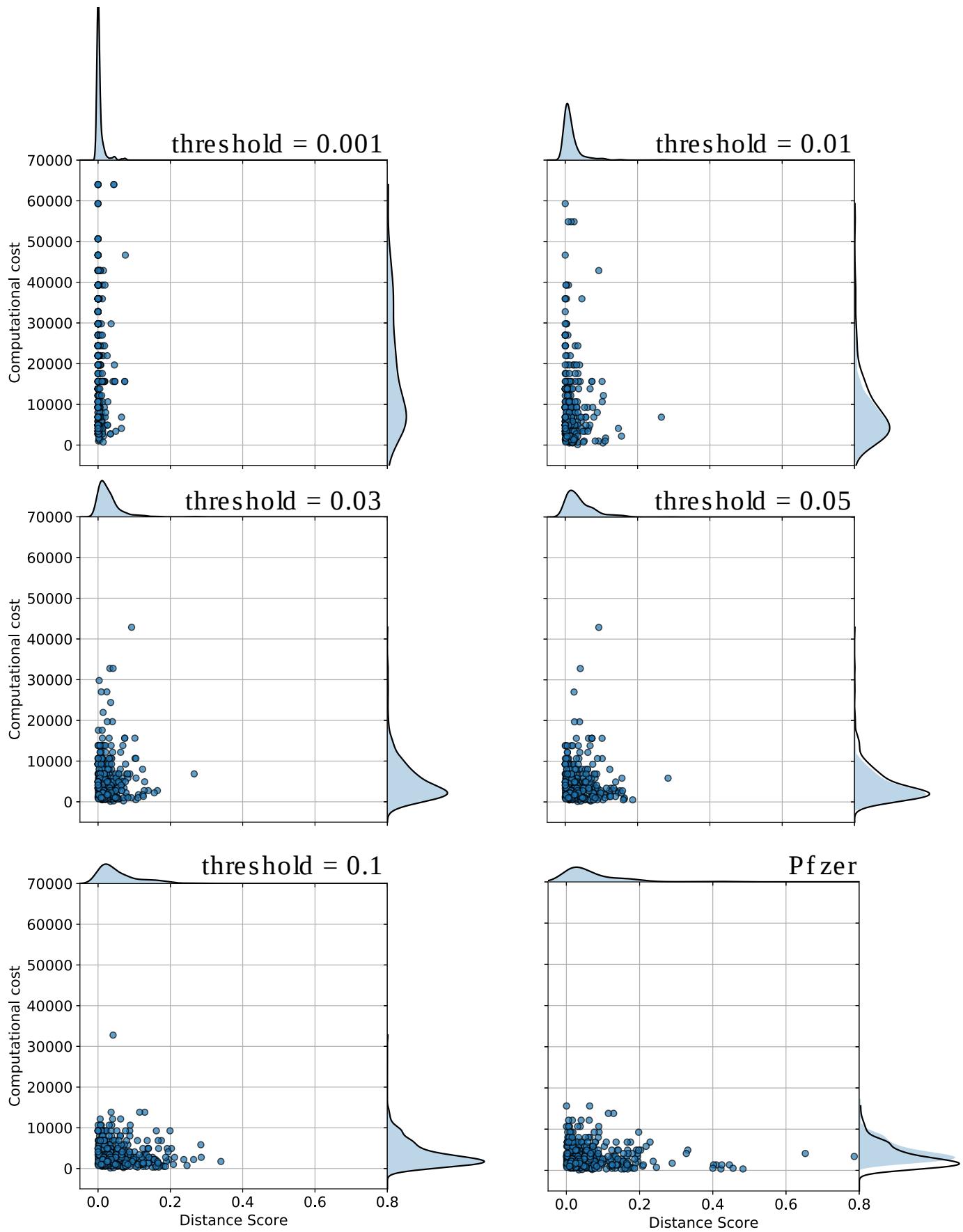


Figure 10: Assessment of fragmentation scheme illustrate an accuracy vs cost trade-off Computational cost vs distance score of the fragments for the bonds in the benchmark set shown in figure 14. Computational cost is the cubed of heavy atoms in the fragment. The threshold is the maximum amount of change allowed in the ELF10 WBO relative to the parent's ELF10 WBO during fragmentation.

We want to find the parameters for our fragmentation scheme that maximizes the number of fragments that end up in that lower left corner (illustrated in figure 8, B). To do that, we generated

fragments for the red bonds in the 100 molecules shown in figure 14 set using different disruption thresholds. For every fragment, we found the distance score of their fragments' WBO distribution and their computational cost. We then plotted all fragments from the validation set for different thresholds {figure {fig;joint_plots}}. When the threshold is low, the fragmentation scheme will generate fragments which have very good distance 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 scores are too big. For the molecules we tested, a threshold of 0.03 leads to the most fragments in the lower left quadrant (defined as cost < 10000 and score < 0.05) as shown in table 3. This threshold is similar to what we found when we looked at the distribution of standard deviations for WBO distributions with respect to conformations (3, blue). Most of them fall under 0.03. Both of these data points lead us to recommend a disruption threshold of 0.03 for our fragmentation scheme. While the current scheme does not provide a perfect solution, plots in figure 10 shows less fragments outside of the lower left region for thresholds 0.01, 0.03 and 0.05. This scheme performs better than other schemes such as the scheme Pfizer used in [49] (figure 10, lower right and table 3).

Table 3: Number of fragments in the lower left quadrant in figure 10 defined as a distance score less than 0.1 and computational cost less than 10000.

disruption threshold	fragments in lower left quadrant
0.001	153
0.005	197
0.01	229
0.03	259
0.05	231
0.07	214
0.1	200
scheme from 49	189

Benchmark results reveal chemical groups that induce long range effects

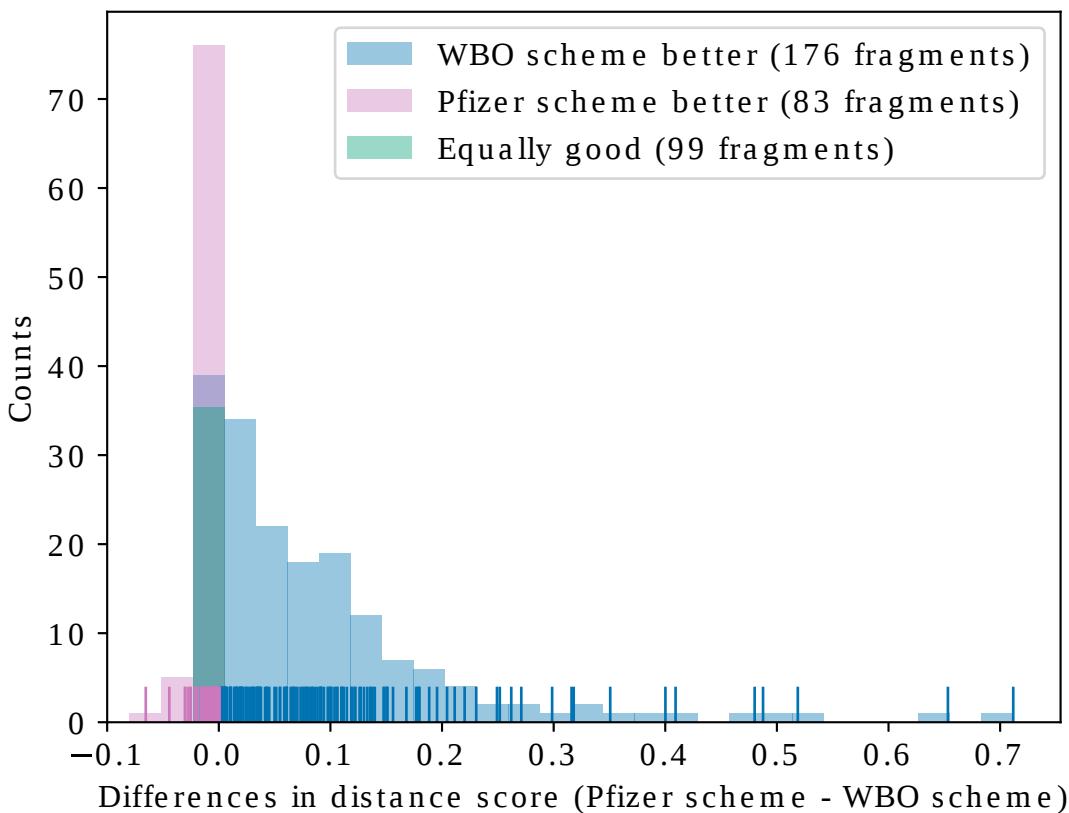
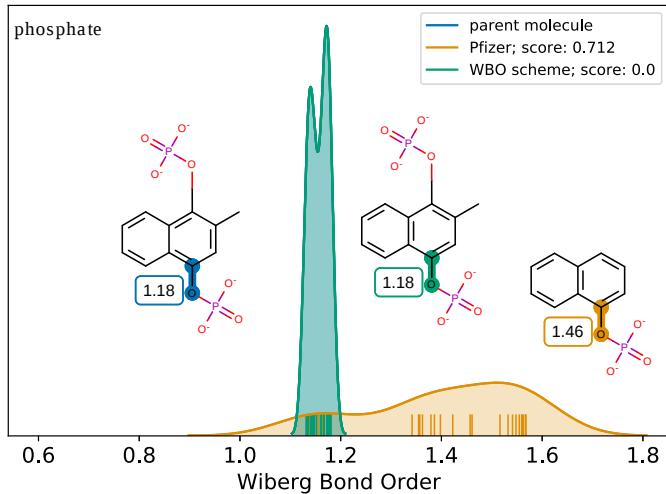
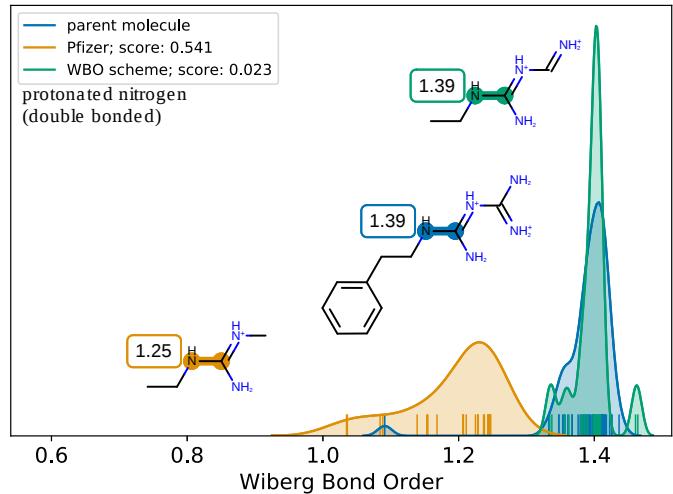
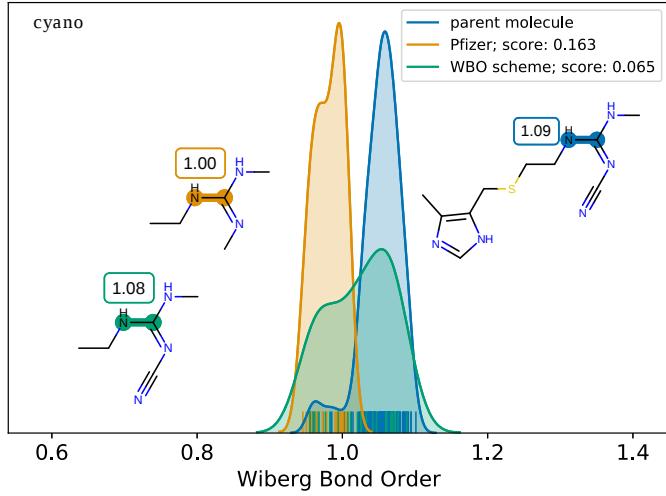
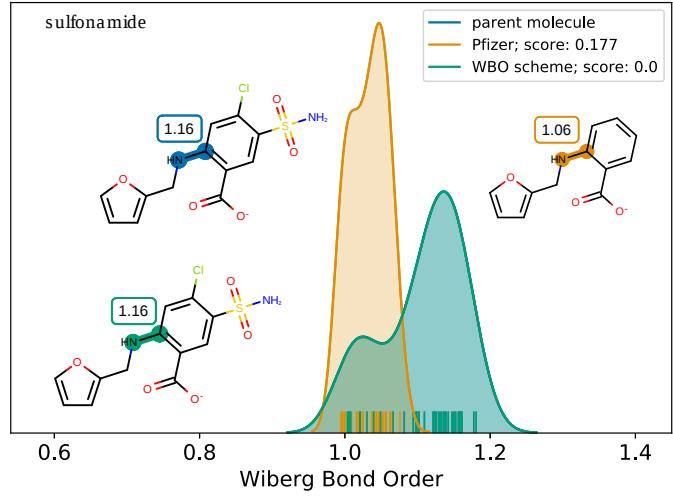
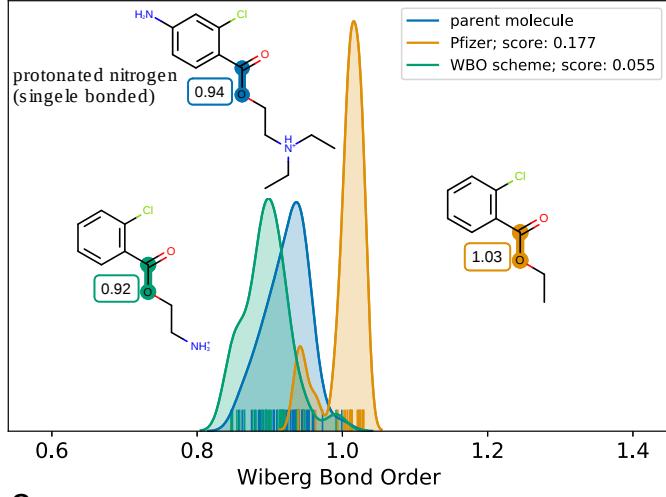
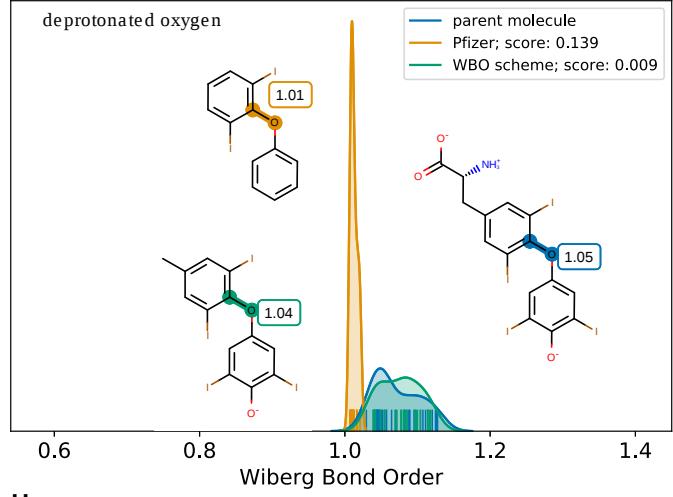
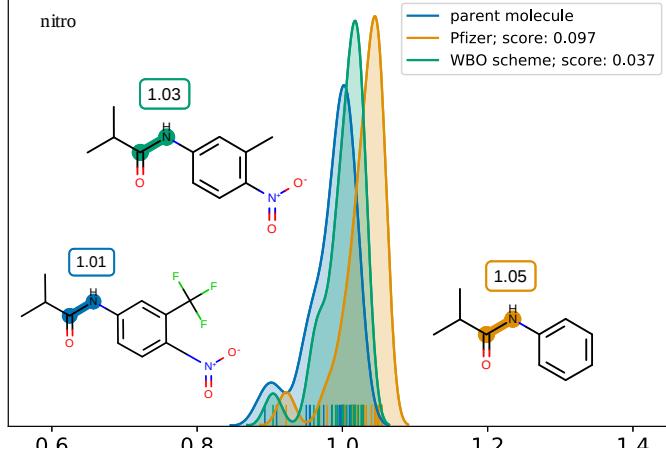
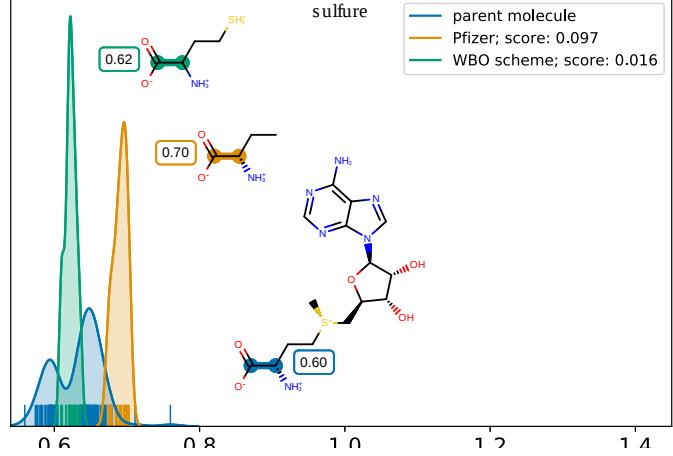


Figure 11: Using the WBO as an indicator of chemical environment disruption improves fragmentation

Distribution of differences in distance scores for fragments in the validation set (SI fig 14) generated using Pfizer's rules and our scheme using 0.03 as the disruption threshold. For many bonds, both approaches yield equally preforming fragments (shown in green). In some cases, Prizer's rules preforms better than our scheme (shown in red), however, the differences are usually very small. In most cases, using the WBO as an indicator improves the distance score (shown in blue)

In the benchmark experiment (figure 10), the distance scores measured the distance between WBO distributions generated from Omega generated conformers of the parents and fragments. Omega aims to generate low energy conformers [41] and in some cases, fragments only have one or two low energy conformers so it is not clear how accurate the distances measured are. In addition, only comparing low energy conformers do not fully capture torsion energy barriers which we also want to ensure remain accurate relative to their parent's torsion energy scan. To mitigate the above mentioned issues when validating our scheme, we also added WBOs calculated from conformers generated on a grid of torsion angles about the bonds which included higher energy conformers that are closer to conformers generated in canonical torsion scans. Furthermore, since we know that WBOs from structures in a QC torsion scan are anti correlated with the QC torsion energy scan (6), adding these WBOs to the distributions provided a better validation of our method than only looking at the distance between omega generated WBO distributions. The differences in distances of these distributions from fragments generated with our scheme and [49] is shown in figure 11.

For many molecules, using a common sense rule based approach, such as the one used in [49] to fragmenting molecules, will yield fragments that are the same fragments generated with our scheme shown in green in figure 11, and sometimes can even preform slightly better than using the WBO as an indicator (fig 11, red). However, in many cases, especially if certain chemical groups are involved, using the WBO as an indicator significantly improves the electron population overlap about the bonds and brings them closer to their parent's chemical environment (fig 11, blue). It is important to note that when the fragment generated from both scheme are the same (green in figure 11), they are not necessarily the optimal fragment and both schemes can preform equally poorly (see SI).

A**B****C****D****E****F****G****H**

v.v

Wiberg Bond Order

Wiberg Bond Order

Figure 12: Some chemical groups induce non local effects that are captured in fragments when using the WBO as an indicator of chemical environments Wiberg bond order distributions for parent molecules (shown in blue) and fragments generated with Pfizer rules (shown in orange) and our scheme (shown in green). This figure shows eight chemical groups where the WBO distributions of the highlighted bonds change when those groups are removed. These changes are consistent across the validation set.

Upon closer inspection of the validation set, we found eight chemical groups that induce long range effects to sensitive bonds shown in figure 12. These chemical groups with representative examples are shown in figure 12. The groups are ordered by how strongly they induce long range effect, in decreasing order. The most dramatic change happens when a phosphate group is removed (figure 12, A). The variance of the WBO distribution increases which conveys an increase in relative energies of conformers in the QC torsion scans. In other molecules where phosphates are removed, the variance can decrease even if the phosphate group is ten bonds away (figure 13, F and SI). In figure 12, B, removing a protonated nitrogen that is double bonded causes the WBO distribution to shift and the variance to increase. Long range effects are seen in other molecules with similar chemical patterns up to eight bonds away (SI). Removing a nitrile group (fig 12, C) and sulfonamide group (12, D) have similar effects on the WBO distributions which is also consistent with other molecules that contain these groups up to three bonds away (SI). A protonated nitrogen and deprotonated oxygen (12 E and F) can effects bonds between 3-6 bonds away (SI). While the changes in distributions for removing a nitro group and sulfur (12, G and H) are not as big as other chemical groups, they are mostly consistent across other molecules in the validation set (SI).

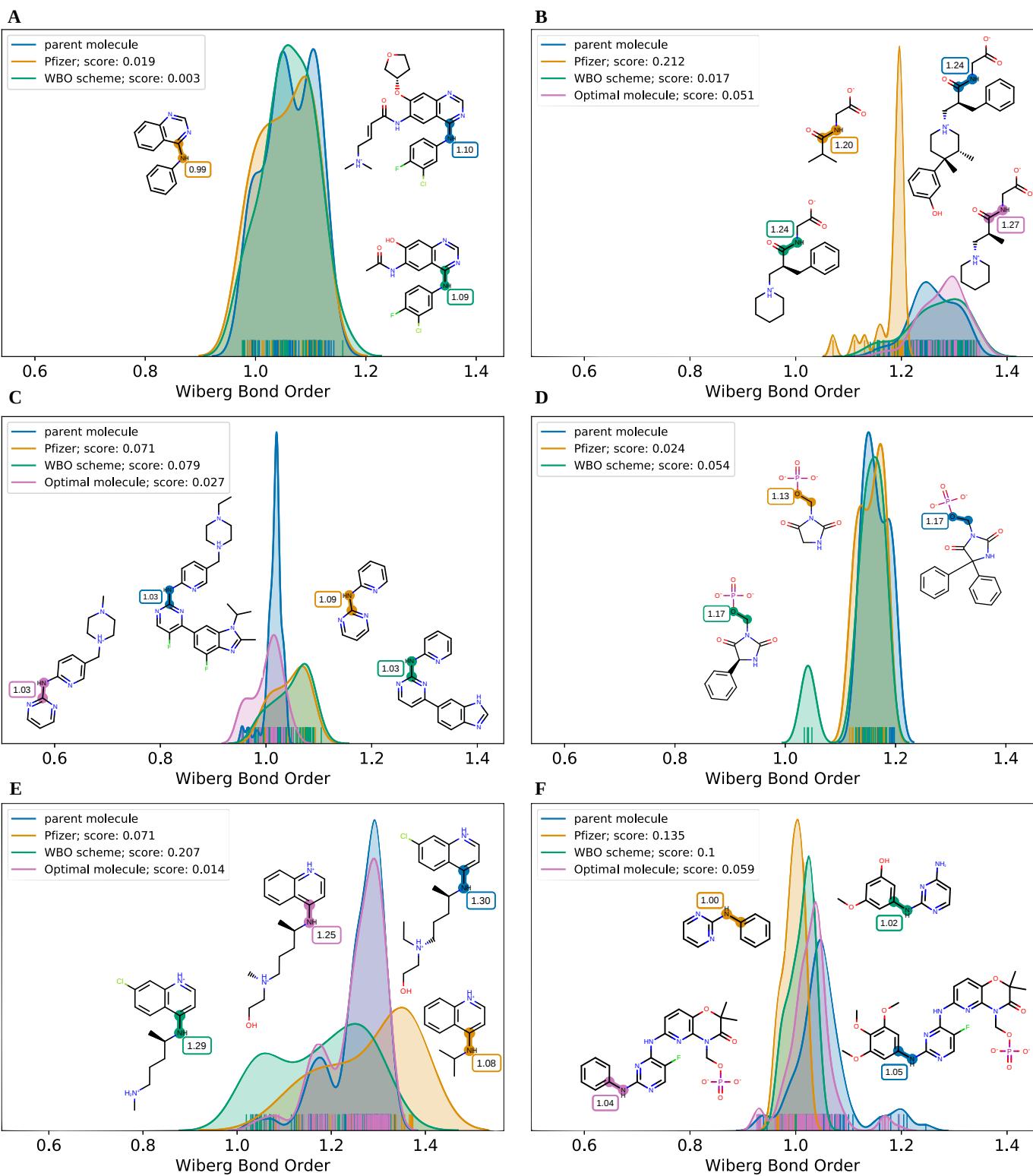


Figure 13: Using the WBO as an indicator when fragmenting can still fail to find the optimal fragment Our scheme can fail in several ways. A. A smaller fragment (shown in orange) is just as good as a larger fragment (shown in green) even if the ELF10 WBO estimate of the bond in the smaller fragment relative to its parent (shown in blue) is more than the disruption threshold. B. While our scheme finds a fragment with good overlap of the WBO distributions (shown in green), it is not the smallest fragment possible with good distributions overlap (smallest fragment with good overlap is shown in purple). C. The fragment we find is bigger than what the simple scheme finds (shown in orange) but without improving the WBO overlap (green). The optimal fragment that neither scheme generates is shown in purple. D. Our scheme finds a larger fragment that has worse WBO distribution overlap. E and F. Sometimes, almost the entire molecule is needed to achieve good WBO distribution overlap between the fragment and the parent. This is not a failure mode but inherent to the challenge of fragmenting molecules for QC calculations.

While our scheme captures long range effects that a simple rule based approach does not, it is not an optimal solution and will sometimes fail to find the most optimal fragment. By optimal we mean the

smallest fragment that retains the torsion potential of the bond in the parent molecule. Our scheme can fail in multiple ways as illustrated in figure [13](#) and listed below.

1. We find a fragment with good WBO distributions overlap but do not find the smallest fragment. This is shown in both [13](#) A and B. In A, the fragment that Pfizer scheme find is smaller and has a WBO distribution that is close to the parent's WBO distribution (MMD 0.019). In this case, the ELF10 WBO estimate of the bond in the fragment is 0.11 lower than the ELF10 WBO estimate in the parent molecule. In B, our fragment has better WBO distribution overlap with the parent WBO distribution vs using Pfizer's scheme (0.017 vs 0.212), but it is not the smallest fragment. According to the fragment highlighted in purple, the benzene ring is not required to achieve good overlap of the WBO distributions (0.051)
2. We find a fragment that is bigger than Pfizer's scheme fragment and the remote substituents do not improve the WBO distribution overlap (MMD 0.079 vs 0.071) ([13](#) C). The better fragment is shown in purple. It is both smaller and has better overlap (MMD 0.027) than the orange and green fragment.
3. We find a fragment that is both larger and has worse overlap (0.054 vs 0.024) than what the Pfizer's scheme generates (fig [13](#))

While it is usually possible to find a fragment that is significantly smaller than the parent and retains remote substituent effects, the effects are sometimes more than 3-6 bonds away and a large fragment is needed to accurately represent the chemical environment of the parent molecule. Two such examples are shown in figure [13](#) E and F. In E, not only is the protonated nitrogen needed (shown in green), but the alcohol group is also needed to achieve good WBO distribution overlap (shown in purple). In F, the phosphate group nine bonds away from the bond of interest is needed to get the density of of the mode at 1.2 (shown in blue and purple).

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](https://doi.org/10.1021/acs.jctc.8b00821) · PMID: [30512951](https://pubmed.ncbi.nlm.nih.gov/30512951/) · PMCID: [PMC6467725](https://pubmed.ncbi.nlm.nih.gov/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](https://doi.org/10.1021/acs.jctc.5b00864) · PMID: [26584231](https://pubmed.ncbi.nlm.nih.gov/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](https://doi.org/10.1186/s13321-019-0371-6) · PMID: [31392432](https://pubmed.ncbi.nlm.nih.gov/31392432/) · PMCID: [PMC6686419](https://pubmed.ncbi.nlm.nih.gov/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](https://doi.org/10.1002/jcc.23775) · PMID: [25413259](https://pubmed.ncbi.nlm.nih.gov/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](https://doi.org/10.1002/jcc.23422) · PMID: [24000174](https://pubmed.ncbi.nlm.nih.gov/24000174/) · PMCID: [PMC3874408](https://pubmed.ncbi.nlm.nih.gov/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](https://doi.org/10.1021/ci500112w) · PMID: [24960267](https://pubmed.ncbi.nlm.nih.gov/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](https://doi.org/10.1021/ja00124a002)

8. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations

Bernard R. Brooks, Robert E. Brucolieri, Barry D. Olafson, David J. States, S. Swaminathan, Martin Karplus

Journal of Computational Chemistry (1983) <https://doi.org/bqh7f2>

DOI: [10.1002/jcc.540040211](https://doi.org/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/fvftx>
DOI: [10.1021/ja9621760](https://doi.org/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](https://doi.org/10.1007/s00894-008-0305-0) · PMID: [18458967](https://pubmed.ncbi.nlm.nih.gov/18458967/) · PMCID: [PMC2864003](https://pubmed.ncbi.nlm.nih.gov/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](https://doi.org/10.1038/s41598-018-21070-0) · PMID: [29416116](https://pubmed.ncbi.nlm.nih.gov/29416116/) · PMCID: [PMC5803249](https://pubmed.ncbi.nlm.nih.gov/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](https://doi.org/10.1002/jcc.20082) · PMID: [15264253](https://pubmed.ncbi.nlm.nih.gov/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](https://doi.org/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](https://doi.org/10.1021/acs.jpcb.7b02320) · PMID: [28306259](https://pubmed.ncbi.nlm.nih.gov/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](https://doi.org/10.1093/nar/gkx1037) · PMID: [29126136](https://pubmed.ncbi.nlm.nih.gov/29126136/) · PMCID: [PMC5753335](https://pubmed.ncbi.nlm.nih.gov/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](https://doi.org/10.1063/1.3382344) · PMID: [20423165](https://pubmed.ncbi.nlm.nih.gov/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](https://doi.org/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/ci970429j](https://doi.org/10.1021/ci970429j) · PMID: [9611787](https://pubmed.ncbi.nlm.nih.gov/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](https://doi.org/10.1021/acs.jcim.6b00596) · PMID: [28346786](https://pubmed.ncbi.nlm.nih.gov/28346786/) · PMCID: [PMC5433162](https://pubmed.ncbi.nlm.nih.gov/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](https://doi.org/10.1021/jm9602928) · PMID: [8709122](https://pubmed.ncbi.nlm.nih.gov/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](https://doi.org/10.1002/jcc.25149) · PMID: [29280158](https://pubmed.ncbi.nlm.nih.gov/29280158/)

22. Approximate ab initio energies 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](https://doi.org/10.1063/1.1879792) · PMID: [15945620](https://pubmed.ncbi.nlm.nih.gov/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](https://doi.org/10.1021/cr200093j) · PMID: [21866983](https://pubmed.ncbi.nlm.nih.gov/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](https://doi.org/10.1039/c2cp23832b) · PMID: [22373545](https://pubmed.ncbi.nlm.nih.gov/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](https://doi.org/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](https://doi.org/10.1002/jcc.10309) · PMID: [12868106](https://pubmed.ncbi.nlm.nih.gov/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](https://doi.org/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](https://doi.org/10.1021/jp4010345) · PMID: [23514314](#)

29. Resonance interactions in acyclic systems

K. B. Wiberg, C. M. Breneman, K. E. Laidig, R. E. Rosenberg

Pure and Applied Chemistry (1989-01-01) <https://doi.org/c6rd43>

DOI: [10.1351/pac198961040635](https://doi.org/10.1351/pac198961040635)

30. 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](https://doi.org/10.1002/cphc.201500773) · PMID: [26458868](#)

31. A new force field for molecular mechanical simulation of nucleic acids and proteins

Scott J. Weiner, Peter A. Kollman, David A. Case, U. Chandra Singh, Caterina Ghio, Giuliano Alagona, Salvatore Profeta, Paul Weiner

Journal of the American Chemical Society (1984-02) <https://doi.org/dcvwwx>

DOI: [10.1021/ja00315a051](https://doi.org/10.1021/ja00315a051)

32. 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](https://doi.org/10.1021/acs.jctc.8b00640) · PMID: [30351006](#) · PMCID: [PMC6245550](#)

33. 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](https://doi.org/10.26434/chemrxiv.8304578.v1)

34. Bond order and valence indices: A personal account

I. Mayer

Journal of Computational Chemistry (2006) <https://doi.org/ddssrj>

DOI: [10.1002/jcc.20494](https://doi.org/10.1002/jcc.20494) · PMID: [17066501](#)

35. 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](https://doi.org/10.1021/ja00299a024)

36. Optimization of parameters for semiempirical methods I. Method

James J. P. Stewart

Journal of Computational Chemistry (1989-03) <https://doi.org/cg5pck>

DOI: [10.1002/jcc.540100208](https://doi.org/10.1002/jcc.540100208)

37. 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](https://doi.org/10.1063/1.1747632)

38. On the quantum theory of valence and bonding from the ab initio standpoint

Mario A. Natiello, Jorge A. Medrano

Chemical Physics Letters (1984-03) <https://doi.org/bdfk5f>

DOI: [10.1016/0009-2614\(84\)85645-6](https://doi.org/10.1016/0009-2614(84)85645-6)

39. 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](https://doi.org/10.1039/c6ra17055b)

40. Resonance interactions in acyclic systems. 3. Formamide internal rotation revisited. Charge and energy redistribution along the C-N bond rotational pathway

Kenneth B. Wiberg, Curt M. Breneman

Journal of the American Chemical Society (1992-01) <https://doi.org/c85zcc>

DOI: [10.1021/ja00029a005](https://doi.org/10.1021/ja00029a005)

41. 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](https://doi.org/10.1021/ci100031x) · PMID: [20235588](https://pubmed.ncbi.nlm.nih.gov/20235588/) · PMCID: [PMC2859685](https://pubmed.ncbi.nlm.nih.gov/PMC2859685/)

42. Density-functional thermochemistry. III. The role of exact exchange

Axel D. Becke

The Journal of Chemical Physics (1993-04) <https://doi.org/d2fq9j>

DOI: [10.1063/1.464913](https://doi.org/10.1063/1.464913)

43. Effect of the damping function in dispersion corrected density functional theory

Stefan Grimme, Stephan Ehrlich, Lars Goerigk

Journal of Computational Chemistry (2011-03-01) <https://doi.org/b3ztst>

DOI: [10.1002/jcc.21759](https://doi.org/10.1002/jcc.21759) · PMID: [21370243](https://pubmed.ncbi.nlm.nih.gov/21370243/)

44. 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](https://doi.org/10.1021/acs.jctc.7b00174) · PMID: [28489372](https://pubmed.ncbi.nlm.nih.gov/28489372/)

45. 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](https://doi.org/10.1002/(sici)1096-987x(199905)20:7<720::aid-jcc7>3.0.co;2-x)

46. 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](https://doi.org/10.1021/acs.jcim.7b00505) · PMID: [28967749](https://pubmed.ncbi.nlm.nih.gov/28967749/)

47. QUACPAC 2.0.2.2 — Applications,
v2019.Nov.2 <https://docs.eyesopen.com/applications/quacpac/index.html>

48. 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](https://doi.org/10.1021/ja01280a022)

49. Comprehensive Assessment of Torsional Strain in Crystal Structures of Small Molecules and Protein-Ligand Complexes using ab Initio Calculations

Brajesh K. Rai, Vishnu Sresht, Qingyi Yang, Ray Unwalla, Meihua Tu, Alan M. Mathiowitz, Gregory A. Bakken
Journal of Chemical Information and Modeling (2019-10) <https://doi.org/ggfvzc>
DOI: [10.1021/acs.jcim.9b00373](https://doi.org/10.1021/acs.jcim.9b00373) · PMID: [31573196](#)

Supporting Information

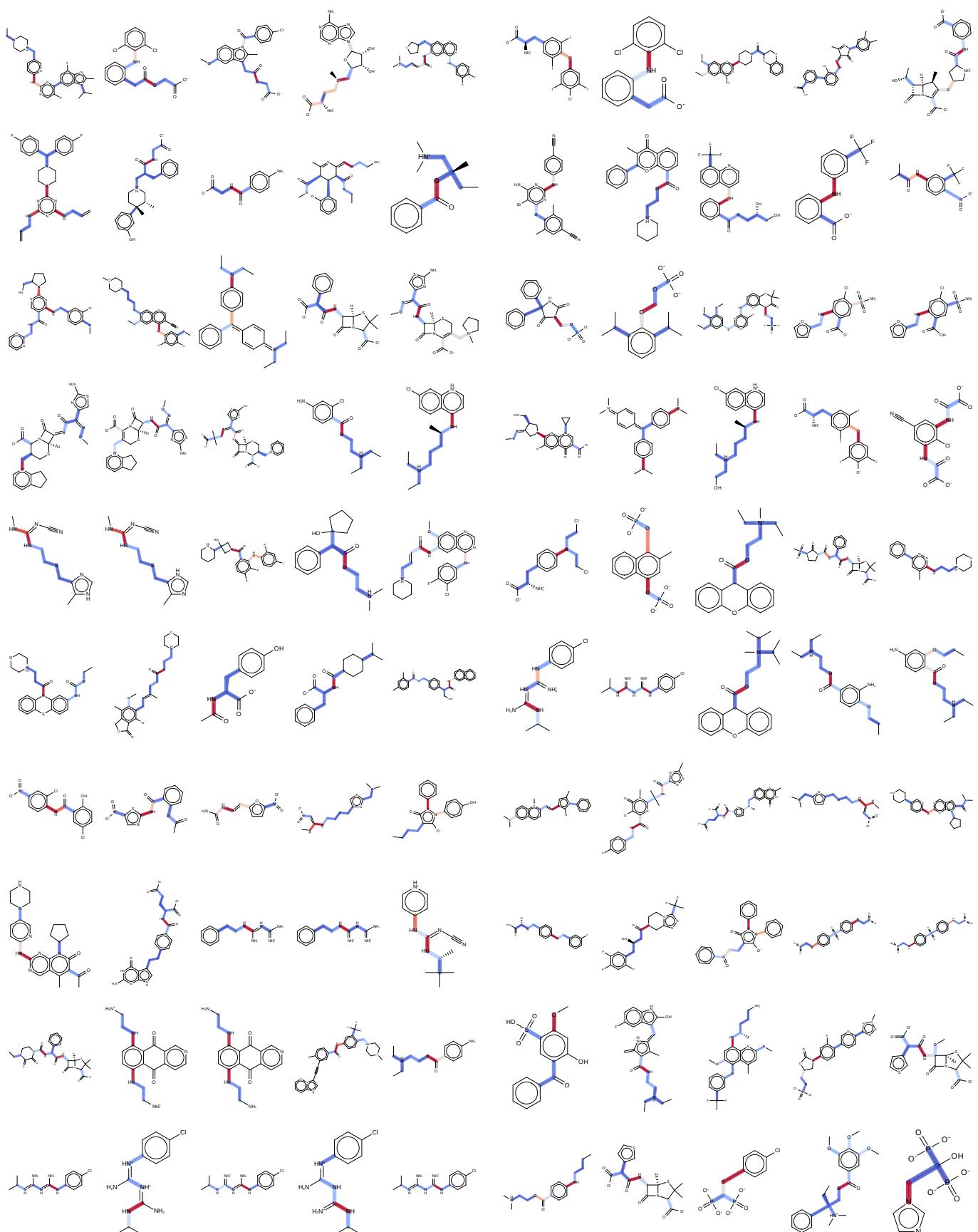


Figure 14: Validation set All molecules used in the validation set of fragmentation schemes. The bonds are highlighted by how sensitive they are to remote fragmentation. The redder bonds are more sensitive while the WBO distributions around the blue bonds do not change much with remote fragmentation.