HHS Public Access

Author manuscript

Anal Chem. Author manuscript; available in PMC 2021 May 21.

Published in final edited form as:

Anal Chem. 2021 March 02; 93(8): 3830–3838. doi:10.1021/acs.analchem.0c04341.

Exploring the impacts of conformer selection methods on ion mobility collision cross section predictions

Felicity F. Nielson, **Sean M. Colby**, **Dennis G. Thomas**, **Ryan S. Renslow***, **Thomas O. Metz***

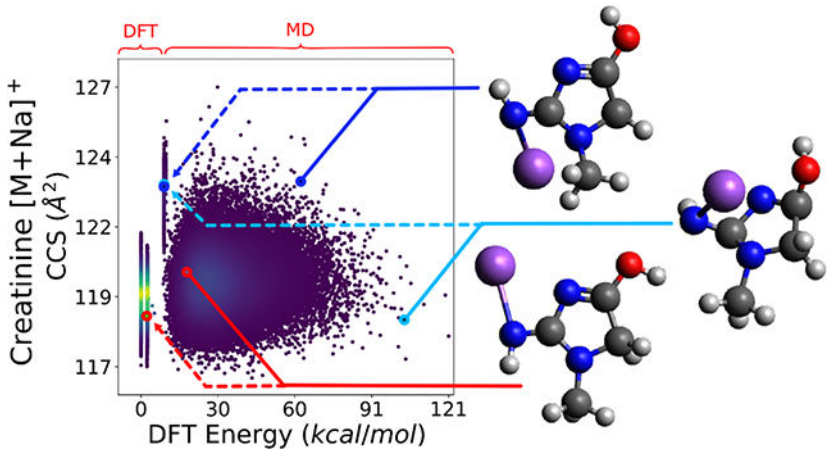
Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA, USA

Abstract

The prediction of structure dependent molecular properties, such as collision cross sections as measured using ion mobility spectrometry, are crucially dependent on the selection of the correct population of molecular conformers. Here, we report an in-depth evaluation of multiple conformation selection techniques, including simple averaging, Boltzmann weighting, lowest energy selection, low energy threshold reductions, and similarity reduction. Generating 50,000 conformers each for 18 molecules, we used the *In Silico* Chemical Library Engine (ISiCLE) to calculate the collision cross sections for the entire dataset. First, we employed Monte Carlo simulations to understand the variability between conformer structures as generated using simulated annealing. Then we employed Monte Carlo simulations to the aforementioned conformer selection techniques applied on the simulated molecular property—the ion mobility collision cross section. Based on our analyses, we found Boltzmann weighting to be a good tradeoff between precision and theoretical accuracy. Combining multiple techniques revealed that energy thresholds and root-mean-squared deviation-based similarity reductions can save considerable computational expense while maintaining property prediction accuracy. Molecular dynamic conformer generation tools like AMBER can continue to generate new lowest energy conformers even after tens of thousands of generations, decreasing precision between runs. This reduced precision can be ameliorated and theoretical accuracy increased by running density functional theory geometry optimization on carefully selected conformers.

Graphical Abstract

*Corresponding authors: ryan.renslow@pnnl.gov; thomas.metz@pnnl.gov.



INTRODUCTION

The identification and quantification of small molecules – metabolomics – has a broad range of applications, from forensics[1–3] to human health and disease[3–5], soil microbiology[6–9] and materials science.[10,11] The current gold standard methods for identifying small molecules in complex samples rely on comparing experimental data (i.e., observed “features”) to libraries derived from pure chemicals analyzed using the same experimental platform. Such reference materials are limited in availability and can be costly to acquire en masse, especially at high purity, and can require significant time to process and analyze. The vast majority of molecules in the universe are yet undiscovered, and even of those that are known, most are not readily available for purchase.[12–15] It has therefore become crucial to develop computational methods for building reliable libraries of predicted molecular properties that are validated against empirical experiments in order to reduce reliance on authentic reference materials. Many groups have developed methods for predicting chemical properties measured in several identification platforms[16–32] as elaborated in the Supporting Information.

Chemical properties and molecular behavior are a consequence of inter- and intra-molecular forces, as governed primarily by the electron distribution surrounding the constituent nuclei. Therefore, the conformer, or specific 3D structure of a molecule that otherwise has the same atoms and bonds (see SI for further definition), has significant impact on the outcome of chemical interactions. Many chemical properties (e.g. CCS and NMR chemical shifts) are highly sensitive to the underlying conformational populations, and consequently, nearly all of the computational approaches listed above require the initial step of generating conformers. Suitable conformer(s) must be chosen for accurate *in silico* molecular simulations or the results of chemical property predictions may be open to significant error.

To date, there has not been a clear study that has evaluated tradeoffs between various conformer sampling techniques—including analysis of the appropriate number to be used or the best method of selection. In this study, we explore several methods for conformer selection to assess the impact these approaches have on the prediction of the molecular property collision cross section (CCS), as measured using ion mobility spectrometry (IMS).

In IMS, a sample of molecule(s) is ionized (e.g., via electrospray ionization) and then propelled by an electric field through a drift region populated by a neutral buffer gas (commonly nitrogen or helium). The momentum transfer and chemical interactions between the molecular ions (adducts) and the buffer gas result in changing the net drift velocity of ion packets, leading to the separation of molecular adducts, including adducts for isomers[33] and isotopologues.[34] The measured arrival time of the ions at the end of the drift region can be used to calculate the CCS. As CCS is a property of both the ion and the buffer gas, different buffer gases will yield different CCS values for the same ion.

During an IMS separation, a single molecular adduct is not associated with a single spike in arrival time, but rather a distribution of arrival times as seen in Fig. 1a. This distribution is due to ion packet diffusion and multiple, interconverting conformers existing simultaneously within each packet. Commonly, the arrival time associated with the peak apex of the arrival time distribution is used to calculate a single experimental CCS for each molecular adduct. *In silico* predictions of CCS that are based on molecular structure (as opposed to 3D structure-naïve approaches[20,31]) therefore ideally choose a conformer or group of conformers that will result in CCS that are as close as possible to the experimental values represented by the arrival time peak maxima. The more accurate the predicted CCS, the more useful they will be for identification libraries and for reducing molecular identification false positive rates.

In this study, we used Monte Carlo, energy cutoff, and similarity downselection methods, as well as a variety of averaging methods, to explore how varying the number and type of conformers considered in a modified ISiCLE pipeline relate to final CCS predictions. We used a benchmark set of molecules with experimentally determined CCS, spanning various size and molecular flexibility, to compare the different methods and to correlate structure variability with chemical properties.

Similar to current literature methods, we found Boltzmann weighting yielded the best result among evaluated averaging techniques. We also found the conformers generated from MD simulations using AMBER insufficiently covered the low energy region of conformational space, leading to lower precision between simulations. DFT geometry optimization helps resolve this sparsity issue, which may also be present with other conformer generation tools. While this study focuses on the effect of conformer selection on CCS, we believe the results and the methods of evaluation can be generalized to other molecular modeling applications.

METHODS:

Conformer Generation and Processing

To test sampling methods on large sets of conformers, a modified ISiCLE pipeline was used to generate ~50,000 conformers for each adduct in a benchmark molecule set (see Fig. 2 and Table S1 for set details). Specifically, the generalized amber force field (GAFF) and the AmberTools17[35] simulated annealing MD tool, Sander, was used with simulated temperatures of 300 K to 600 K for 1000 annealing cycles, from which 50 conformers were randomly selected out of each cycle at the 300 K level. After conformer generation, CCS values for each conformer were calculated using MOBCAL-SHM,[21] a shared-memory

version of MOBCAL written in C and optimized for HPC resources, yielding 135× speed up over the original MOBCAL. Finally, DFT energies were calculated for each conformer using NWChem[19] (v 6.8), with B3LYP exchange-correlation and 6–31G* basis set, via ISiCLE. B3LYP and 6–31G* were chosen for their ~~purport and~~ prevalence in recent literature. For hardware requirements and for parameters for all relevant tools, please see the SI.

For three molecular adducts (mandelonitrile $[M+H]^+$, creatinine $[M+Na]^+$, and sucrose $[M-H]^-$), 25k–50k of their conformers were additionally geometry optimized using DFT in NWChem, with CCS subsequently calculated by MOBCAL-SHM to enable comparison of conformer selection techniques on quantum chemistry optimized structures.

To briefly compare simulated annealing against other conformer generation methods, 50k conformers were generated for one molecular adduct, mandelonitrile $[M+H]^+$ using RDKit (v 2019.03.1, rdkit.org), and the lowest energy conformer was generated for each adduct in the set using the Conformer-Rotamer Ensemble Sampling Tool (CREST, v 2.7.1)[36] with the GFN2-xTB method.

Conformer Geometry Variability

Custom Python scripts (available for download on GitHub at https://github.com/pnnl/conformer_selection and provided in the **SI**) were created for performing Monte Carlo (MC) simulations in order to understand the root-mean-square deviation (RMSD) variability between conformer geometries as produced by simulated annealing. RMSD were calculated using the OpenBabel (v 2.4.1) OBAlign function[37,38] to align the conformers and calculate RMSD between corresponding heavy atoms (i.e. non-hydrogen atoms). The goal using MC was to simulate random draws from the true population of conformers for a range of sample sizes, where at each sample size, or step, conformers were randomly sampled and their RMSD averaged. Each MC simulation step was run for 10,000 iterations (see Fig. S1 for justification) to produce a simulation average -- analogous to the most probable result when choosing a sample of that size -- and a standard deviation. Because simulated annealing works in cycles, we investigated two approaches when sampling: (1) The full 50k conformer set treated as a single pool (with a bias to sample across cycles) and (2) each cycle sampled as a group. This allowed for assessing possible correlations between adjacent cycles (Fig. S2). Details describing how sampling was applied to allow direct comparison between the two approaches is given in the SI.

Using Pearson product-moment correlation coefficient, three characteristics of the complete MC simulations, namely RMSD convergence point, final converged average RMSD value, and maximum standard deviation from the average RMSD, were correlated against 71 molecular properties calculated using ChemAxon's tool cxcalc[39] (v 17.17.0), as well as against experimental CCS. The converged value is the final MC result when sampling the full population, and the convergence point is the sample size when the maximum standard deviation is within 0.01% of the converged value, as shown in Fig. 4a. We note the molecular property calculations were done on the parent (non-adduct) molecules and the MC convergence was measured on the ionized (adduct) molecules.

Conformer Selection and CCS Averaging Methods

Our goal was to sample from the full 50,000 conformer population of each adduct in order to simulate a situation in which a researcher had only generated the sampled conformers. The foundation for this decision was a hypothesis that the full 50,000 conformer population would represent the vast majority of the possible conformational space for the adducts in our molecule set. MC methods were used to simulate the result of CCS calculations after conformers were chosen from increasingly large conformer populations using a variety of selection techniques (described in more detail below): (1) simple average, (2) Boltzmann weighted average, (3) lowest energy, and (4) averaging below an energy threshold. These were chosen based on their prevalence in the literature.[40–45]

In addition to these methods, a fifth technique, which preemptively down-selects from the full sample to the m most similar and n most dissimilar set of conformers, builds off of an approach introduced by Colby *et al.*,[21] which provides a more computationally efficient method of sampling while maintaining high precision. This method and the previous four methods were used in tandem to analyze every possible combination of the methods for a set of parameters, as described in Results and Discussion, Section 5.

A schematic of the following selection techniques is shown in Fig. 3.

1. **Simple average (SA)—**The simple average CCS is the arithmetic mean of all CCS values for a sample of conformers. Because the samples are randomly drawn from the full population, this simulates random conformer selection.

2. **Boltzmann Weighted (BW) average—**BW weights each conformer according to its Boltzmann probability distribution given by the equation,

$$p_i = \frac{e^{-\frac{E_i}{kT}}}{\sum_{i=1} e^{-\frac{E_i}{kT}}}$$

where p_i is the probability or weight of the i^{th} conformer, E_i is the energy, k is the Boltzmann constant, and T is temperature. Conformers that are lower in energy will have a higher weight when the CCS is averaged. This aims to reflect their time-averaged existence as a thermodynamic property and is heavily biased toward low energies. BW is currently considered the gold standard of conformer averaging used for CCS, NMR chemical shift calculations[41], as well as for other chemical properties.

3. **Lowest Energy (LE)—**Only the conformer with the single lowest energy is selected.

4. **Energy Threshold (ET)—**Only conformers with energy under the threshold are selected, and their CCS are simple averaged. Here, we apply 5, 2, 1, and 0.5 *kcal/mol* thresholds.

5. **Similarity downselection (SDS)—**The goal of SDS is to sample conformational space with fewer conformers while still being representative of the larger population, thus

saving on computational expense. SDS uses RMSD-based similarity metrics and a heuristic selection algorithm freely available at https://github.com/pnnl/sds. SDS is described in more detail in the SI.

MC simulation was run on BW, LE, SA, and ET (in combination with SA) at 1,000 iterations for each MC step. As with the RMSD analysis, the across- versus within-cycle approaches are also assessed here. MC analysis was done separately using AMBER potential energies and DFT energies.

RESULTS AND DISCUSSION

Our goal for this work was to evaluate many of the methods found in recent literature for sampling molecular conformations, especially with consideration for those methods that have been used for CCS calculations. Toward this end, we performed Monte Carlo analysis and various sampling techniques to assess conformational coverage (using RMSD) as well as the impact on CCS as a function of conformer sampling methods. This was done with a validation set of protonated, deprotonated, and sodiated adducts ($[M+H]^+$, $[M-H]^+$, $[M+Na]^+$) of various chemical classes spanning about 100–700 Da.

Convergence of RMSD as a function of Monte Carlo sampled conformers

MC simulations were run for 1,000 or 10,000 iterations (for CCS or RMSD analysis respectively) per data point (e.g. per number of conformers sampled) for each molecule in the validation set to ensure convergence. Fig. 4 demonstrates an example MC convergence plot of the variability between conformer geometries as defined by average RMSD. The convergence plots of all molecules are almost indistinguishable when viewed separately (see Fig. S3). We note three characteristics of these plots distinct to each molecule: the final converged RMSD average at full population (which we refer to as the “converged value”), the maximum standard deviation, and the “convergence point”, which we have defined as the sample size when the standard deviation reaches 0.01% of the final converged RMSD average.

As expected, a large converged value (a measure of the degree of variability between conformers for a molecule) is positively correlated with molecular mass (r^2: 0.67; p-value: < 1e-4), but also with properties such as chain atom/bond count (r^2: 0.91 / 0.93; p-value: < 1e-9 / < 1e-9), and rotatable bond count (r^2: 0.90; p-value: < 1e-08), and negatively correlated with the second acidic pKa site (r^2: 0.66; p-value: 1.3e-4). The converged value was also negatively correlated with having high ring counts; however, our data was insufficient for assessing statistical significance for these properties, and we would need to perform a study with a larger molecule set to verify this (Fig. S5). Note the positively correlated properties also correlate with molecular mass, reflecting how larger molecules typically have higher degrees of freedom than smaller molecules. A volcano plot showing the statistical significance and the magnitude of correlation for several other properties (Fig. S6) also revealed other properties such as the 3D Van der Waals surface area (r^2: 0.81; p-value: < 1e-6), and water accessible surface area (r^2: 0.80; p-value: < 1e-6) were highly correlated and significant. Also interesting are the convergence point results, because while a molecule may have a high converged value (high average pairwise RMSD of the entire

50,000 population), it could have a relatively small convergence point. In some cases, this may be because the variability of conformer space is sampled in relatively few conformers despite the large RMSD between those conformers. More correlations between MC convergence characteristics and molecular properties as a heatmap of Pearson r correlations for 71 properties are found in Fig. S5.

The Monte Carlo Sampling Methods section in the SI discusses additional details that consider convergence for within versus across simulated annealing cycles for molecular dynamics, revealing, as expected, that sampling across cycles resulted in better conformational space coverage. Interestingly, sampling within versus across cycles had little effect on the MC convergence of calculated CCS except to lower the precision of simple averaging methods when sampling within cycles, as seen in Fig. S4.

Effect of conformer sampling on calculated CCS—Ultimately, the desire is to assess the appropriate conformer sampling method producing stable and accurate chemical property predictions. In this manuscript, our application focused on CCS values. Table S2 provides a summary of CCS calculated by each conformer sampling technique when all 50k conformers are sampled. The table also includes values from the best combination of these methods, ISiCLE,[21] CREST,[36] and experimental values. Table S3 shows the mean absolute percent error relative to ISiCLE.

1. SA, BW, and LE selection techniques on AMBER generated conformers—Fig. S7 demonstrates convergence plots of the BW, LE, and SA sampling techniques. Consistently for all molecules, LE had the widest standard deviation. Both LE and BW had averages that sometimes skewed dramatically and had much higher standard deviations than the simple average. This happens whenever the conformer generation randomly produces a small population of one or more conformers with energies significantly lower than the rest. Because LE and BW are heavily biased toward lower energies, their selection was affected by the increasing probability of the MC simulation “generating” lower energies as the sample size increased. A non-linear dependency between CCS and energy means the average CCS will keep changing as sample size increases to include more conformers with lower energies, and two samples of the same size may have widely different outcomes leading to high standard deviation. Therefore, the results of selection techniques dependent on energy are essentially functions of the CCS versus energy landscape. More specifically, they are functions of the low energy region, which is sparsely populated by AMBER. This sparsity leads to lower precision for low-energy dependent selection techniques. Thus, in order to understand how the selection techniques will behave, it is crucial to first understand how the conformer generation and optimization techniques shape the CCS (or other calculated value) versus energy landscape. Conformer generators like AMBER can produce significantly lower energy conformers even after thousands of generations (see Table S5), reducing precision for BW and LE between simulations.

We note although the standard deviations of BW and LE seem wide relative to SA, it is misleading to think SA is the better option. SA may give the best precision in a computational model, but its choice does not reflect the underlying physics. For this reason, BW is recommended as it has better precision than LE and is still reflective of known

physics. BW is supported in the literature as the current gold standard approach, even for other properties such as NMR chemical shift calculations.[41,42,46] In IMS in particular, a molecule does not exist as a single conformer, but rather interconverts between low-energy-barrier conformations rapidly during flight at room temperature. The final arrival time captured is then a weighted average of these conformers according to their duration of existence. It is reasonable to assume BW is more accurate because this is what is intended to be captured with Boltzmann's energy and temperature-dependent probability equation. At this time, however, we cannot directly say anything about the accuracy of the techniques with high confidence. Other underlying conditions, such as the ionization site location or tautomer form, can significantly influence conformer formation, changing the CCS versus energy landscape, and thus altering the predicted CCS. Unique tautomers, for instance, can have CCS differences significantly larger than the CCS differences of the conformers for a single tautomer, with CCS of individual conformation populations of two unique tautomers possibly not even overlapping. To achieve optimal accuracy, a wholistic approach needs to be taken, optimizing all aspects that could significantly change conformation simultaneously. Regardless, Table S3 and Table S4 have been provided as comparisons of the various selection techniques to the method implemented in ISiCLE (DFT geometry optimized conformers) and experimental values, respectively. Both tables suggest BW and LE have lower mean absolute percent errors, and therefore better accuracy, than SA.

1.a SA, BW, and LE selection techniques on RDKit generated and DFT geometry optimized conformers: A molecule's CCS vs energy space is shaped differently by different conformer generation and optimization methods, which can lead to significantly varied final calculated properties, even when using the same conformer selection techniques. Fig. 6 compares the same selection techniques (SA, BW, and LE) as discussed above, but applied to RDKit generated conformers and DFT geometry optimized conformers (starting from AMBER generated structures). This is shown for 50k mandelonitrile $[M+H]^+$ conformers. Like AMBER, RDKit sparsely captured the low energy region of conformer space, leading to lower precision for LE. In this example, BW had a precision more comparable to SA's high precision, but this is likely an anomaly due to a split conformer population, since the low energy region is still sparsely populated. We note the RDKit conformers were generated using distance geometry and were not optimized using RDKit's universal force field (UFF) optimization tool. See Fig. S8 for a discussion on how this may affect CCS.

DFT geometry optimization, on the other hand, significantly lowers the energy of all conformers and clusters them into "bars" where the energies are very close but the range of CCS remains wide (e.g. ~0.05 *kcal/mol* versus ~4 $Å^2$ for the example shown in Fig. 6). It appears that small changes, even the rotation of a methyl group on a rotamer, can lead to strikingly different CCS (e.g. one rotation by ~39 deg on creatinine's methyl group yielded a difference of ~1.1 $Å^2$, or ~0.95%). DFT geometry optimized conformers for creatinine $[M+Na]^+$ and sucrose $[M-H]^-$ are plotted in Fig. S9, showing similar results. DFT optimization densely populates the low energy region of CCS versus energy space, allowing BW and LE to have better precision. For example, the max standard deviation of BW for mandelonitrile dropped from σ 0.99 $Å^2$ to σ 0.08 $Å^2$, suggesting a few DFT geometry-optimized

conformers are more effective than a large series of MD-based structures for small rigid molecules. How far this translates to larger, more flexible molecules is yet unknown. For sucrose, BW and LE had significant increases in precision, but the standard deviation of BW did not drop to 1% of the converged value until 650 conformers were randomly sampled, whereas mandelonitrile and creatinine were below 1% at the first sample size (50 conformers).

2. Averaging the CCS of conformers under energy thresholds—Performing simple averages of all conformers under energy thresholds has historically been used as an alternative to choosing only the lowest energy conformer or performing Boltzmann weighting. Fig. S10 compares convergence plots for 5, 2, 1, and 0.5 *kcal/mol* energy thresholds for all molecules in our set. At certain thresholds, ET mimics the other selection techniques. This is no surprise because SA is the same as a threshold so large it encompasses every conformer, and LE is the same as a threshold so small it captures only one conformer. Thus, ET is bound by SA and LE methods. ET suffers from the same undersampling of the low energy region of CCS vs energy space that BW and LE do; as many as 1,051 or as few as 1 conformer were found for 5 *kcal/mol* threshold depending on the molecule, as shown in Fig. S11. The sparsity of the low energy region further complicates how to recommend which energy threshold is best, but there is a general trend that higher thresholds give higher precision between simulations at the expense of theoretical accuracy, and lower thresholds sacrifice precision when using MD conformers.

After selecting the conformers under an energy threshold, more than just SA can be applied, such as BW or SDS, as described in Section 5 below.

3. RMSD based Similarity Downselection—In our previous work, we found choosing the two most dissimilar conformers and the single most similar conformer from simulated annealing cycles (based on RMSD), yielded final CCS results within 99% of the result obtained from using all conformers from each cycle.[21] Building on this idea, we wanted to test how incrementally adding increasingly dissimilar conformers would impact the final property prediction. The idea behind SDS is to cover conformational space with fewer structures, thus maintaining accuracy while saving on computational expense. Ermanis et al. recently employed an RMSD-based similarity downselection method to exclude structures that were already very similar to each other and found using the 25 most dissimilar conformers was sufficient to minimize computational costs for NMR structure elucidation on small, rigid molecules.[47] We use SDS in an evaluation of different technique combinations as described below in Section 5.

4. Using MD vs DFT energy on MD structures—Conformers not at energy minima or in strict transition states are not well defined by quantum mechanical methods, and so calculating DFT energies on MD structures that have not been optimized by DFT are thought to be untrustworthy. However, we found DFT energy on MD-generated structures, before DFT geometry optimization, has better correlation to the CCS and energy of conformers after DFT geometry optimization (Fig. 6). Whereas MD energies have almost no correlation, DFT energies cluster the MD structure space in a way that can be mapped to the DFT geometry optimized space. If one can predict which cluster will map to the lowest

energy DFT geometry optimized “bar,” one can then select those structures to perform full DFT geometry optimization (assuming the goal is to get the lowest energy conformers at a given temperature). Since both MD and DFT energy calculations (on MD structures) run orders of magnitude faster than full DFT geometry optimization (Fig. S12 and Table S6), predicting beforehand which conformers to geometry optimize would result in considerable speed up. For molecules like creatinine, a low DFT energy threshold would suffice to secure conformers from this cluster. For sucrose, the cluster mapping to the lowest energy DFT geometry optimized “bar” was located closer to the middle of the MD “cloud,” making it unclear how to successfully select those conformers without hindsight. Even so, there remains a general trend that higher energy clusters mapped to higher energy geometry-optimized “bars” and lower energies to lower geometry-optimized “bars” when using DFT energies.

Similarly, Kanal et al.[48] found poor correlation between energies of commonly used classical force fields and DFT and semiempirical methods. They likewise found DFT energies calculated on MMFF94 geometry optimized conformers had better correlation to DFT geometry optimized energies than MMFF94 energies. Different methods for calculating geometry and energies optimize different potential energy surfaces. Dependent on the application, selecting different levels of theory will change cost versus accuracy tradeoffs. While using DFT energies on non-DFT-optimized structures is not best practice in general, we feel our findings not only validate the use of DFT energies on MD structures for energy-based conformer space reduction methods such as energy thresholds, but also are better than using MD energies, especially with the goal of performing DFT geometry optimization on the structures afterwards. Analyses using MD energies, which showed similar trends between conformer selection techniques, were also done and can be found in the SI. Additionally, Fig S13 plots CCS vs energy space for all molecules, using AMBER energies, DFT energies, and DFT geometry optimized structures.

6. Exploring the best combinations of space-reduction and selection techniques—We ran a method-combinations search to find an optimal combination of the selection techniques for accuracy, precision, and computational expense when using non-DFT-optimized AMBER generated conformers. We explored over 1,700 combinations using many different energy thresholds (both DFT and MD energies, alternatively or in combination), RMSD-based downselection (SDS), and conformer selection method combinations, including how many conformers were initially generated. The BW value of DFT geometry optimized conformers in the same manner as described in Colby et al. (2019)[21] was used to define the baseline in which to assess each combination. For sucrose, creatinine, and mandelonitrile, where we had 25k–50k DFT geometry optimized structures, all of their optimized conformers were used for the baseline. Of the conformer downselection techniques (ET, SDS, random), SDS and ET appeared to give lower MAPE than random methods. The best method combination, AMBER for 10 cycles (50 conformers generated per cycle), using a 10 *kcal/mol* AMBER energy threshold followed by SDS to choose the 1 most similar and 10 most dissimilar, and then selecting the lowest energy conformer by DFT energy, resulted in 1.2% (σ: 0.9%) MAPE and is estimated to take 200 (σ 157) minutes.

Please refer to the SI for an extended limitation of this study.

CONCLUSION

Using Monte Carlo analysis, we have shown the relative precision and behavior of various conformer selection techniques on AMBER generated conformers. Of the averaging or consolidation techniques--Boltzmann weighting (BW), lowest energy (LE), simple average (SA)--BW had better precision than LE, and is physics-based, unlike SA, and is therefore expected to be more accurate. Example analysis on RDKit and DFT geometry optimized conformers confirms this trend, and also demonstrates the need for more efficient conformer generation tools that more thoroughly target the low energy region of conformer space. MD-based conformer generation tools like AMBER sparsely populate the low energy region of conformer space, leading to lower precision between simulations for energy-based selection methods, such as BW, LE, and ET. Applying robust structure optimization methods like DFT geometry optimization can help ameliorate this problem, greatly increasing the precision and expected accuracy of e.g. BW, but this comes at the cost of greater computational expense. For this reason, we have hopes for tools like CREST[36] and BOKEI[49] which use mixtures of methods (e.g. MTD, genetic z-matrix crossing, and semi-empirical methods) to specifically target low energy conformer space. A preliminary test of CREST showed it consistently generated conformers with lower (DFT) energies than AMBER for all molecules, half of the examples having energies as low as the DFT geometry optimized conformers. Subsequent research rigorously testing CREST against other methods is needed.

For single field experimental CCS methods, an interlaboratory study demonstrated an average uncertainty of 0.54%,[50] yet a recent study reported measured CCS uncertainty estimates of 4.7–9.1%.[51] For building *in silico* chemical libraries, we hope to achieve <1% MAE to meet future improvements of experimental platforms. The best method combination found had a MAPE greater than 1%. This further suggests the precursory conformer generation step was insufficient for our purposes. A more thorough analysis with a larger molecule set and tighter parameters would be needed to confirm this.

Many of our conclusions have been assumed in the literature, but here we have provided evidence for them. Already, researchers have been gravitating toward improved conformational sampling methods (e.g. those emerging from Prof. Stefan Grimme's group), showing a shift away from older, less relevant methods. In summary, we recommend Boltzmann weighting conformers as generated and optimized by tools that sufficiently populate the low energy region of conformer space (e.g. CREST and DFT geometry optimization). Doing so is expected to increase accuracy and precision while minimizing computational expense.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health, National Institute of Environmental Health Sciences grant U2CES030170. Pacific Northwest National Laboratory (PNNL) is operated for the U.S. Department of Energy by Battelle Memorial Institute under contract DE-AC05–76RL01830.

CITATIONS

(1). Ifa DR; Gumaelius LM; Eberlin LS; Manicke NE; Cooks RG Forensic analysis of inks by imaging desorption electrospray ionization (DESI) mass spectrometry; Analyst 2007, 132, 461–467. [PubMed: 17471393]

(2). Lee JR; Choi J; Shultz TO; Wang SX Small Molecule Detection in Saliva Facilitates Portable Tests of Marijuana Abuse; Anal Chem 2016, 88, 7457–7461. [PubMed: 27434697]

(3). Musshoff F; Hess C; Madea B Disorders of glucose metabolism: post mortem analyses in forensic cases--part II; Int J Legal Med 2011, 125, 171–180. [PubMed: 20927632]

(4). Davis VW; Bathe OF; Schiller DE; Slupsky CM; Sawyer MB Metabolomics and surgical oncology: Potential role for small molecule biomarkers; J Surg Oncol 2011, 103, 451–459. [PubMed: 21400531]

(5). Velasquez LS; Sutherland LB; Liu Z; Grinnell F; Kamm KE; Schneider JW; Olson EN; Small EM Activation of MRTF-A-dependent gene expression with a small molecule promotes myofibroblast differentiation and wound healing; Proc Natl Acad Sci U S A 2013, 110, 16850–16855. [PubMed: 24082095]

(6). Jones OA; Sdepanian S; Lofts S; Svendsen C; Spurgeon DJ; Maguire ML; Griffin JL Metabolomic analysis of soil communities can be used for pollution assessment; Environ Toxicol Chem 2014, 33, 61–64. [PubMed: 24122881]

(7). Desai C; Pathak H; Madamwar D Advances in molecular and “-omics” technologies to gauge microbial communities and bioremediation at xenobiotic/anthropogen contaminated sites; Bioresour Technol 2010, 101, 1558–1569. [PubMed: 19962886]

(8). Miura M; Hill PW; Jones DL Impact of a single freeze-thaw and dry-wet event on soil solutes and microbial metabolites; Applied Soil Ecology 2020, 153.

(9). Ghosal D; Ghosh S; Dutta TK; Ahn Y Current State of Knowledge in Microbial Degradation of Polycyclic Aromatic Hydrocarbons (PAHs): A Review; Front Microbiol 2016, 7, 1369. [PubMed: 27630626]

(10). Yu T; Liu L; Xie Z; Ma Y Progress in small-molecule luminescent materials for organic light-emitting diodes; Science China Chemistry 2015, 58, 907–915.

(11). Zhang Q; Kan B; Liu F; Long G; Wan X; Chen X; Zuo Y; Ni W; Zhang H; Li M; Hu Z; Huang F; Cao Y; Liang Z; Zhang M; Russell TP; Chen Y Small-molecule solar cells with efficiency over 9%; Nature Photonics 2014, 9, 35–41.

(12). Dobson CM Chemical space and biology; Nature 2004, 432, 824–828. [PubMed: 15602547]

(13). Reymond JL The chemical space project; Acc Chem Res 2015, 48, 722–730. [PubMed: 25687211]

(14). Reymond JL; Awale M Exploring chemical space for drug discovery using the chemical universe database; ACS Chem Neurosci 2012, 3, 649–657. [PubMed: 23019491]

(15). Ruddigkeit L; van Deursen R; Blum LC; Reymond JL Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17; J Chem Inf Model 2012, 52, 2864–2875. [PubMed: 23088335]

(16). Bouwmeester R; Martens L; Degroeve S Comprehensive and Empirical Evaluation of Machine Learning Algorithms for Small Molecule LC Retention Time Prediction; Anal Chem 2019, 91, 3694–3703. [PubMed: 30702864]

(17). Ruttkies C; Schymanski EL; Wolf S; Hollender J; Neumann S MetFrag relaunched: incorporating strategies beyond in silico fragmentation; J Cheminform 2016, 8, 3. [PubMed: 26834843]

(18). Djoumbou-Feunang Y; Pon A; Karu N; Zheng J; Li C; Arndt D; Gautam M; Allen F; Wishart DS CFM-ID 3.0: Significantly Improved ESI-MS/MS Prediction and Compound Identification; Metabolites 2019, 9.

(19). Apra E; Bylaska EJ; de Jong WA; Govind N; Kowalski K; Straatsma TP; Valiev M; van Dam HJJ; Alexeev Y; Anchell J; Anisimov V; Aquino FW; Atta-Fynn R; Autschbach J; Bauman NP; Becca JC; Bernholdt DE; Bhaskaran-Nair K; Bogatko S; Borowski P, et al. NWChem: Past, present, and future; J Chem Phys 2020, 152, 184102. [PubMed: 32414274]

(20). Colby SM; Nunez JR; Hodas NO; Corley CD; Renslow RR Deep Learning to Generate in Silico Chemical Property Libraries and Candidate Molecules for Small Molecule Identification in Complex Samples; Anal Chem 2020, 92, 1720–1729. [PubMed: 31661259]

(21). Colby SM; Thomas DG; Nunez JR; Baxter DJ; Glaesemann KR; Brown JM; Pirrung MA; Govind N; Teeguarden JG; Metz TO; Renslow RS ISiCLE: A Quantum Chemistry Pipeline for Establishing in Silico Collision Cross Section Libraries; Anal Chem 2019, 91, 4346–4356. [PubMed: 30741529]

(22). Ewing SA; Donor MT; Wilson JW; Prell JS Collidoscope: An Improved Tool for Computing Collisional Cross-Sections with the Trajectory Method; J Am Soc Mass Spectrom 2017, 28, 587–596. [PubMed: 28194738]

(23). Frisch MJ; Trucks GW; Schlegel HB; Scuseria GE; Robb MA; Cheeseman JR; Scalmani G; Barone V; Petersson GA; Nakatsuji H; Li X; Caricato M; Marenich AV; Bloino J; Janesko BG; Gomperts R; Mennucci B; Hratchian HP; Ortiz JV; Izmaylov AF, et al.: Wallingford, CT, 2016.

(24). Heinonen M; Rantanen A; Mielikainen T; Kokkonen J; Kiuru J; Ketola RA; Rousu J FiD: a software for ab initio structural identification of product ions from tandem mass spectrometric data; Rapid Commun Mass Spectrom 2008, 22, 3043–3052. [PubMed: 18763276]

(25). Larriba-Andaluz C; Hogan CJ Jr. Collision cross section calculations for polyatomic ions considering rotating diatomic/linear gas molecules; J Chem Phys 2014, 141, 194107. [PubMed: 25416874]

(26). Mesleh MF; Hunter JM; Shvartsburg AA; Schatz GC; Jarrold MF Structural Information from Ion Mobility Measurements: Effects of the Long-Range Potential; The Journal of Physical Chemistry 1996, 100, 16082–16086.

(27). Shen Y; Bax A SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network; J Biomol NMR 2010, 48, 13–22. [PubMed: 20628786]

(28). Shen Y; Bax A Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks; J Biomol NMR 2013, 56, 227–241. [PubMed: 23728592]

(29). Shvartsburg AA; Jarrold MF An exact hard-spheres scattering model for the mobilities of polyatomic ions; Chemical Physics Letters 1996, 261, 86–91.

(30). Yesiltepe Y; Nunez JR; Colby SM; Thomas DG; Borkum MI; Reardon PN; Washton NM; Metz TO; Teeguarden JG; Govind N; Renslow RS An automated framework for NMR chemical shift calculations of small organic molecules; J Cheminform 2018, 10, 52. [PubMed: 30367288]

(31). Zhou Z; Shen X; Tu J; Zhu ZJ Large-Scale Prediction of Collision Cross-Section Values for Metabolites in Ion Mobility-Mass Spectrometry; Anal Chem 2016, 88, 11084–11091. [PubMed: 27768289]

(32). Amos RIJ; Haddad PR; Szucs R; Dolan JW; Pohl CA Molecular modeling and prediction accuracy in Quantitative Structure-Retention Relationship calculations for chromatography; TrAC Trends in Analytical Chemistry 2018, 105, 352–359.

(33). Zheng X; Renslow RS; Makola MM; Webb IK; Deng L; Thomas DG; Govind N; Ibrahim YM; Kabanda MM; Dubery IA; Heyman HM; Smith RD; Madala NE; Baker ES Structural Elucidation of cis/trans Dicaffeoylquinic Acid Photoisomerization Using Ion Mobility Spectrometry-Mass Spectrometry; J Phys Chem Lett 2017, 8, 1381–1388. [PubMed: 28267339]

(34). Wojcik R; Nagy G; Attah IK; Webb IK; Garimella SVB; Weitz KK; Hollerbach A; Monroe ME; Ligare MR; Nielson FF; Norheim RV; Renslow RS; Metz TO; Ibrahim YM; Smith RD SLIM Ultrahigh Resolution Ion Mobility Spectrometry Separations of Isotopologues and Isotopomers Reveal Mobility Shifts due to Mass Distribution Changes; Anal Chem 2019, 91, 11952–11962. [PubMed: 31450886]

(35). Pearlman D; Case D; Caldwell J; Seibel G; Singh UC; Weiner P; Kollman P; University of California: San Francisco, CA, 2017.

(36). Pracht P; Bohle F; Grimme S Automated exploration of the low-energy chemical space with fast quantum chemical methods; Phys Chem Chem Phys 2020, 22, 7169–7192. [PubMed: 32073075]

(37). O'Boyle NM; Morley C; Hutchison GR Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit; Chem Cent J 2008, 2, 5. [PubMed: 18328109]

(38). O'Boyle NM; Banck M; James CA; Morley C; Vandermeersch T; Hutchison GR Open Babel: An open chemical toolbox; J Cheminform 2011, 3, 33. [PubMed: 21982300]

(39). ChemAxon. 2017.

(40). Campuzano I; Bush MF; Robinson CV; Beaumont C; Richardson K; Kim H; Kim HI Structural characterization of drug-like compounds by ion mobility mass spectrometry: comparison of theoretical and experimentally derived nitrogen collision cross sections; Anal Chem 2012, 84, 1026–1033. [PubMed: 22141445]

(41). Willoughby PH; Jansma MJ; Hoye TR A guide to small-molecule structure assignment through computation of ((1)H and (1)(3)C) NMR chemical shifts; Nat Protoc 2014, 9, 643–660. [PubMed: 24556787]

(42). Reading E; Munoz-Muriedas J; Roberts AD; Dear GJ; Robinson CV; Beaumont C Elucidation of Drug Metabolite Structural Isomers Using Molecular Modeling Coupled with Ion Mobility Mass Spectrometry; Anal Chem 2016, 88, 2273–2280. [PubMed: 26752623]

(43). Williams JP; Grabenauer M; Holland RJ; Carpenter CJ; Wormald MR; Giles K; Harvey DJ; Bateman RH; Scrivens JH; Bowers MT Characterization of simple isomeric oligosaccharides and the rapid separation of glycan mixtures by ion mobility mass spectrometry; International Journal of Mass Spectrometry 2010, 298, 119–127.

(44). Voronina L; Masson A; Kamrath M; Schubert F; Clemmer D; Baldauf C; Rizzo T Conformations of Prolyl-Peptide Bonds in the Bradykinin 1–5 Fragment in Solution and in the Gas Phase; J Am Chem Soc 2016, 138, 9224–9233. [PubMed: 27366919]

(45). Voronina L; Rizzo TR Spectroscopic studies of kinetically trapped conformations in the gas phase: the case of triply protonated bradykinin; Phys Chem Chem Phys 2015, 17, 25828–25836. [PubMed: 25940085]

(46). Graton J; Hernandez-Mesa M; Normand S; Dervilly G; Le Questel JY; Le Bizec B Characterization of Steroids through Collision Cross Sections: Contribution of Quantum Chemistry Calculations; Anal Chem 2020, 92, 6034–6042. [PubMed: 32212634]

(47). Ermanis K; Parkes KEB; Agback T; Goodman JM The optimal DFT approach in DP4 NMR structure analysis - pushing the limits of relative configuration elucidation; Org Biomol Chem 2019, 17, 5886–5890. [PubMed: 31147659]

(48). Kanal IY; Keith JA; Hutchison GR A sobering assessment of small-molecule force field methods for low energy conformer predictions; International Journal of Quantum Chemistry 2018, 118, e25512.

(49). Chan L; Hutchison GR; Morris GM BOKEI: Bayesian optimization using knowledge of correlated torsions and expected improvement for conformer generation; Phys Chem Chem Phys 2020, 22, 5211–5219. [PubMed: 32091055]

(50). Stow SM; Causon TJ; Zheng X; Kurulugama RT; Mairinger T; May JC; Rennie EE; Baker ES; Smith RD; McLean JA; Hann S; Fjeldsted JC An Interlaboratory Evaluation of Drift Tube Ion Mobility-Mass Spectrometry Collision Cross Section Measurements; Anal Chem 2017, 89, 9048–9055. [PubMed: 28763190]

(51). Causon TJ; Hann S Uncertainty Estimations for Collision Cross Section Determination via Uniform Field Drift Tube-Ion Mobility-Mass Spectrometry; Journal of the American Society for Mass Spectrometry 2020, 31, 2102–2110. [PubMed: 32812758]

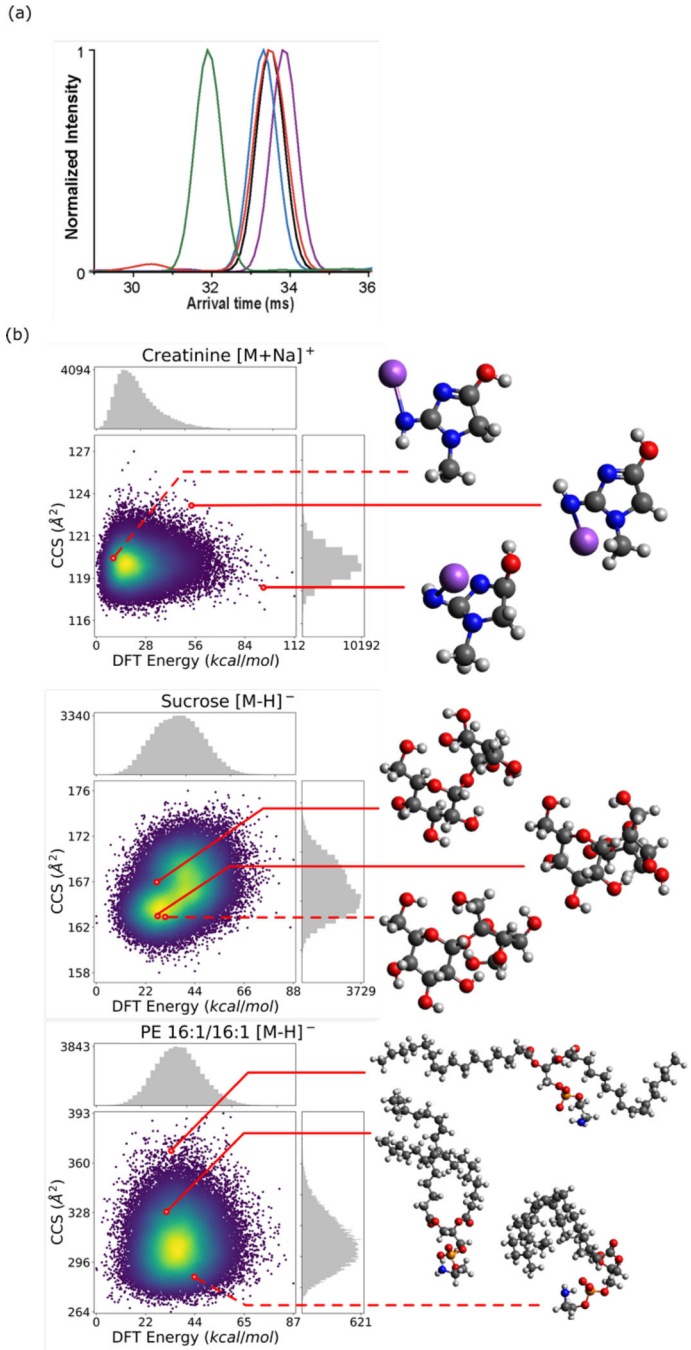


Fig. 1. Example conformer-influenced empirical arrival time and *in silico* CCS distributions.
(a) Di-CQA isomers were shown by Zheng et al. to have overlapping distributions in DTIMS.[33] Distributions in IMS are believed to be largely due to diffusion and conformers. Reprinted (adapted) with permission from Zheng et al. Copyright (2017) American Chemical Society. (b) CCS vs energy landscapes for 50k AMBER generated conformers for creatinine $[M+Na]^+$, sucrose $[M+Na]^+$, and PE 16:1/16:1 $[M-H]^-$ respectively. Highlighted are the most similar (dashed) and two most dissimilar (solid) conformers chosen heuristically with a structural RMSD metric.

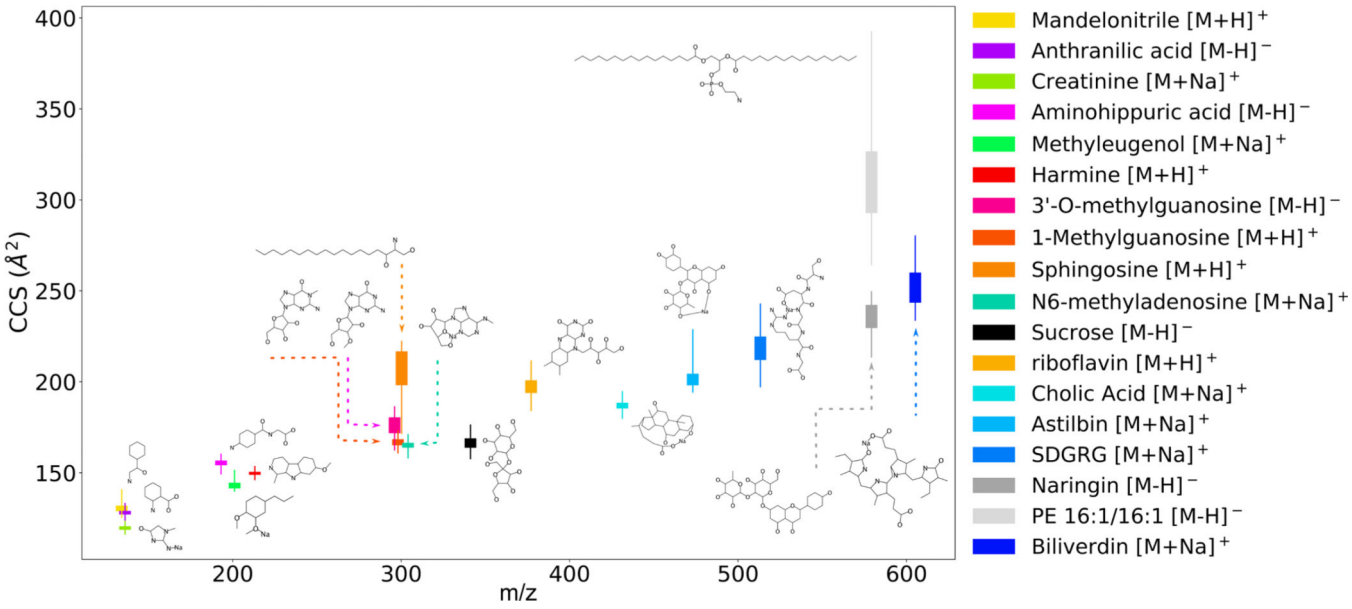


Fig. 2.

Ranges (thin line) and standard deviations (thick box) of CCS for a set of 18 small molecules.

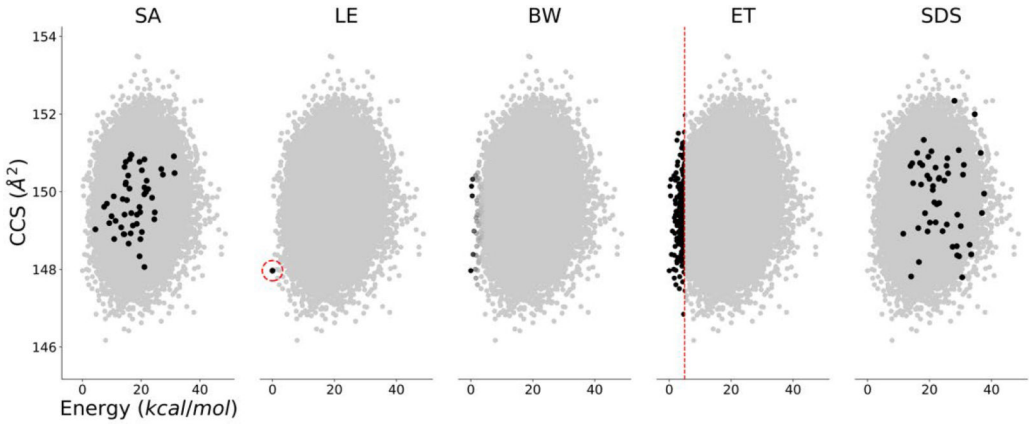


Fig. 3. Diagram of conformer selection and downselection methods.

Simple average (SA), lowest energy (LE), Boltzmann weighting (BW), energy threshold (ET), and similarity downselection (SDS). SA shows 50 randomly selected conformers, LE shows the single lowest energy conformer, BW is shaded based on real weighted values, ET is a 5 *kcal/mol* threshold, and SDS shows the one most similar and 49 most dissimilar conformers.

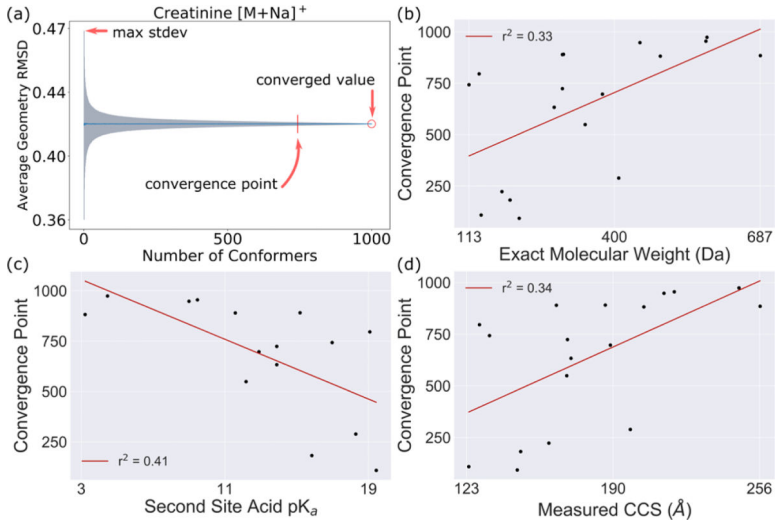


Fig. 4. Example Monte Carlo simulation results on pairwise conformer RMSD.

(a) Example Monte Carlo convergence plot on RMSD between conformers for creatinine [M +Na]$^+$. Convergence point is the sample size (number of conformers) when standard deviation reaches 0.01% of the converged value. (b–d) Convergence point correlations with exact molecular weight, acidic pKa 2nd site (apKa 2), and experimentally measured CCS.

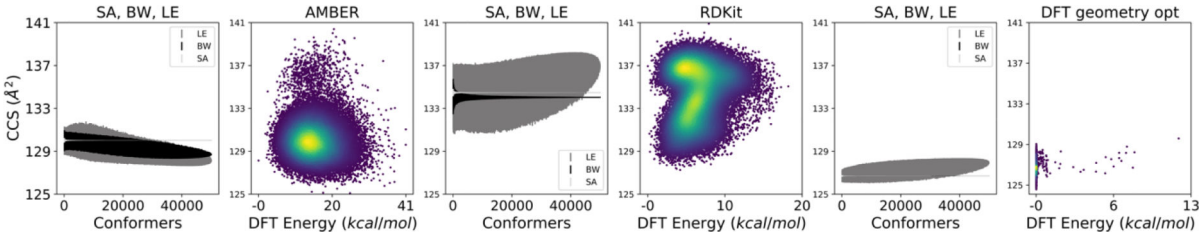


Fig. 5. MC simulation convergence plots of CCS using three sampling techniques (SA, BW, LE) for conformers generated in AMBER, RDKit, and the AMBER conformers after a DFT geometry optimization for mandelonitrile $[M+H]^+$.

Interestingly, RDKit sampled a part of the CCS vs energy landscape that AMBER under sampled, and DFT geometry optimization collapsed the AMBER landscape into a single "bar" cluster where structures had similar energy, but subtle distinctions (e.g. rotamers) led to significantly different CCS. Under all three generation/optimization techniques, LE had the least precision. For all three molecules tested under DFT geometry optimization, BW precision improved dramatically. For this example, BW and SA happened to have the same effect after DFT geometry optimization (their convergence plots exactly overlap). This was not the case for the other two molecules tested.

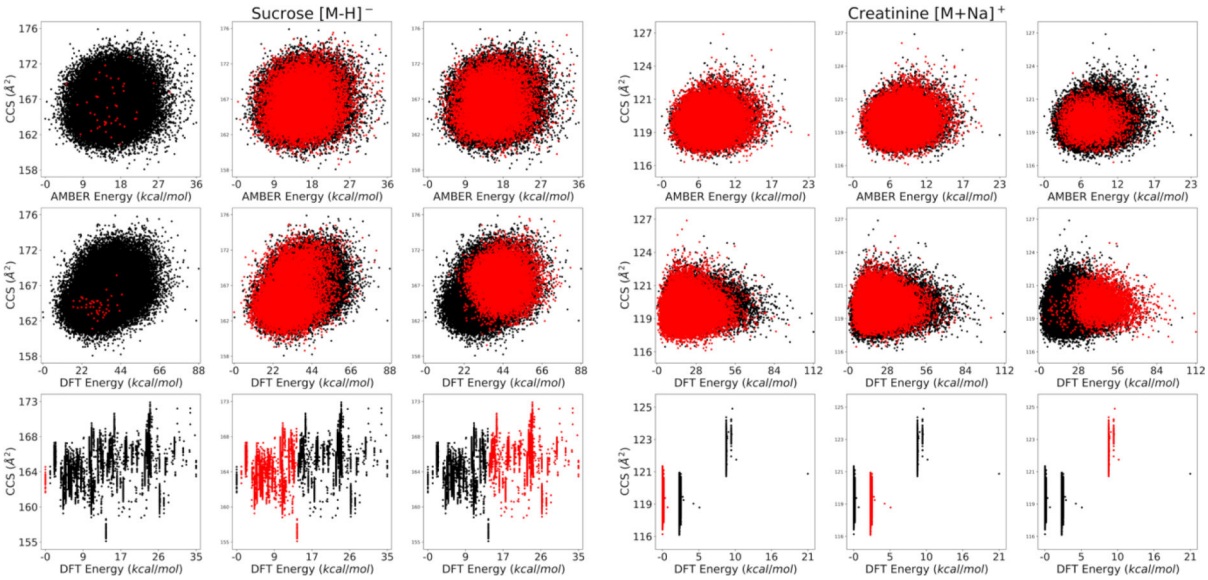


Fig. 6. Demonstrations of clustering between DFT geometry optimized and non-optimized AMBER CCS vs energy space.

Specific clusters of conformers from DFT geometry optimized space (bottom) are chosen and highlighted in red. They are compared with the corresponding source conformers before optimization, using DFT energy (middle) and MD energy (top). DFT energy on MD structures has better correlation with the DFT geometry optimized structures than MD energy. DFT energy clearly predicts the fate of the conformers after DFT geometry optimization, whereas this is not evident with MD energy.