Supporting Information

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*
Pacific Northwest National Laboratory, Richland, WA, USA
*Corresponding authors: ryan.renslow@pnnl.gov; thomas.metz@pnnl.gov

Table of Contents

1. Example Chemical Property Prediction Methods

Many groups have developed methods for predicting chemical properties measured in several identification platforms including quantitative structure-retention relationship and machine learning models to predict liquid chromatography retention times,[1,2] combinatorial approaches to predict MS/MS fragmentation patterns,[3-5] quantum chemical calculations and artificial neural networks for NMR chemical shift predictions,[6-9] and classical scattering and machine/deep learning to predict CCS for IMS.[10-15] Similarly, our group recently developed the *In Silico* Chemical Library Engine (ISiCLE), which is an automated workflow for molecular property calculation. It has shown preliminary success for calculating collision cross sections (CCS) and NMR chemical shifts.[16,17]

2. Conformer Definition

We define a conformer as returned by conformer generation tools: Each structure is a conformer, regardless of energy or energy minima. This is important in applications like IMS, where any valid structure can contribute to the CCS. This is in contrast to the IUPAC definition, where a conformer is only a structure that sits at the minimum of a potential energy well. [18] This latter definition makes no reference to transition state structures that, although fleeting, are present during experiments and impact measured properties such as CCS.

Table S1. Summary of the small molecule test set. We used a benchmark set of 18 small molecules reported in Colby et al. (2019) [17] with masses ranging from 113 to 687 Da. Experimental CCS values for benchmark set adducts ($[M+H]^+$, $[M-H]^+$, or $[M+Na]^+$) were obtained using an Agilent 6560 Ion Mobility Q-TOF MS (Agilent Technologies, Santa Clara) with nitrogen buffer gas, as described in Zheng et al. [19]. This adduct set was also processed through ISiCLE ("Standard" calculation methods) to create an initial predicted CCS baseline. **Fig. 2** in the main plots the m/z vs CCS for the benchmark set molecules.

Molecule	Formula	Adduct	Mass	Experimental CCS	Superclass	Class
Harmine	C13H12N2O	+H	212.095	146.033	Alkaloids and derivatives	Harmala alkaloids
1-Methylguanosine	C11H15N5O5	+H	297.107	168.803	Nucleosides, nucleotides, and analogues	Purine nucleosides
Sphingosine	C18H37NO2	+H	299.282	185.998	Organic nitrogen compounds	Organonitrogen compounds
riboflavin	C17H20N4O6	+H	376.138	188.27	Organoheterocyclic compounds	Pteridines and derivatives
Mandelonitrile	C8H7NO	+H	133.053	128.871	Benzenoids	Benzene and substituted derivatives
Creatinine	C4H7N3O	+Na	113.059	133.413	Organic acids and derivatives	Carboxylic acids and derivatives
Methyleugenol	C11H14O2	+Na	178.099	160.357	Benzenoids	Benzene and substituted derivatives
N6-methyladenosine	C11H15N5O4	+Na	281.112	170.398	Nucleosides, nucleotides, and analogues	Purine nucleosides
Cholic Acid	C24H40O5	+Na	408.288	197.349	Lipids and lipid-like molecules	Steroids and steroid derivatives
Astilbin	C21H22O11	+Na	450.116	212.637	Phenylpropanoids and polyketides	Flavonoids
SDGRG	C17H30N8O9	+Na	490.214	203.5	Organic acids and derivatives	Carboxylic acids and derivatives
Biliverdin	C33H34N4O6	+Na	582.248	246.731	Organoheterocyclic compounds	Tetrapyrroles and derivatives
Anthranilic acid	C7H7NO2	-H	137.048	123.994	Benzenoids	Benzene and substituted derivatives
Aminohippuric acid	C9H10N2O3	-H	194.069	147.552	Benzenoids	Benzene and substituted derivatives
3'-O-methylguanosine	C11H15N5O5	-H	297.107	163.776	Nucleosides, nucleotides, and analogues	Purine nucleosides
Sucrose	C12H22O11	-H	342.116	168.467	Organic oxygen compounds	Organooxygen compounds
Naringin	C27H32O14	-H	580.179	217.329	Phenylpropanoids and polyketides	Flavonoids
PE 16:1/16:1	C37H70NO8P	-H	687.484	256.3	Lipids and lipid-like molecules	Glycerophospholipids

3. Hardware Architecture and Software Parameters

RDKit, CREST, and software used in the ISiCLE pipeline (AMBER, NWChem, MOBCAL-SHM) were run on PNNL supercomputing platforms, Cascade and Constance. Cascade has 1,440 compute nodes with 16 Intel Xeon cores (E5-2670, 2.6 GHz), 128 GB memory per node, and 14Gb/s data rate per lane (FDR InfiniBand). Constance has 464 dual socket compute nodes with 12-core Intel Haswell processors (E5-2670v3, 2.3 GHz), and 64 GB of DDR3-1600 memory per node.

CREST (v2.7.1):
Under the iMTD-GC workflow, CREST uses a mixture of meta-dynamics (MTD), MD, z-matrix crossing, and other methods to iteratively search for low energy conformers and fill out their conformations by finding their rotamers (conformers in this case being understood under the IUPAC definition, i.e. a conformer is only the lowest energy structure of a potential energy well). We used the following parameters and other default options.

GFN2-xTB (very tight or "vtight" optimization level)
z-matrix sorting
6 kcal/mol energy threshold
40 ps MD simulations with 5 fs time step
and other default options

RDKit (v 2019.03.1):
RDKit randomly generates conformers using distance geometry, where constraints bound the minimum and maximum pairwise distances between any two atoms [20]. We used the default parameters with and without UFF optimization.

MOBCAL-shm:
SEED_I2 5013489
BUFFER_GAS NITROGEN
BUFFER_GAS_MASS 28.0134
TEMPERATURE 300
IPR 1000
ITN 10
INP 48
IMP 1024
NUM_THREADS 24

NWChem:
XC: B3LYP
Basis: * library 6-31G*
task dft energy

AmberTools17, Sanders:
As described by Colby et al. (2019) [17]

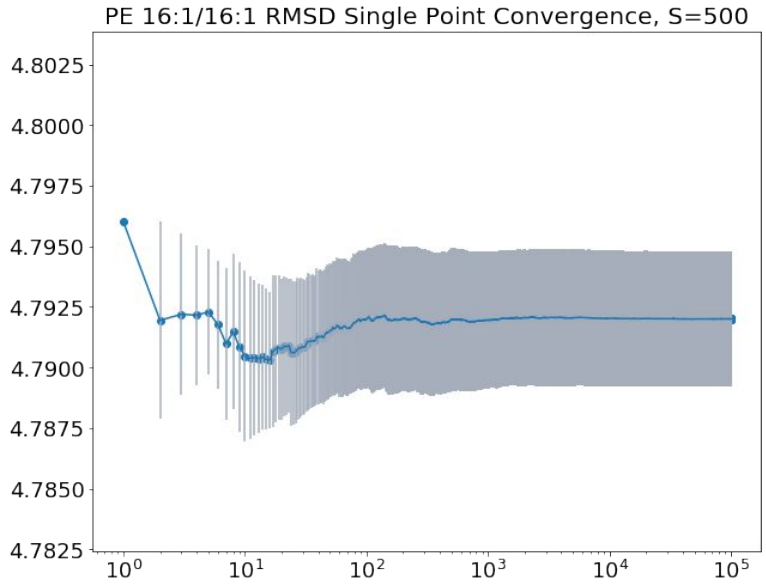


Fig. S1 Single point convergence plot on conformer variability (RMSD) for PE 16:1/16:1, the most flexible molecule in our set. At sample size S=500, the convergence plot shows the thoroughness of a MC simulation at increasing MC iterations. We used 10,000 which we found to be sufficient.

4. Monte Carlo Sampling Across Versus Within Cycles

Because simulated annealing works in cycles, conformers were sampled from cycles using two different methods to distinguish possible cycle correlation. As shown in **Fig. S3**, in one method the cycles were effectively pooled together by sampling across cycles. One conformer was randomly selected from each of the 1000 cycles for a total of 50 conformers. Their RMSD was then calculated for every conformer pair and then averaged. The second method was to keep the annealing cycles isolated, or to select all of the 50 conformers within a cycle, calculate their pairwise RMSD, and average the result. MC was then performed on these averages to simulate generating random cycles. Indeed, the lower average RMSD in **Fig. S3** for the within cycle method, as well as the clustering shown in **Fig. S2**, shows the conformers of a single cycle are, in some cases, more correlated with each other than they are with conformers of another cycle, as expected.

For BW and LE on CCS, there was no noticeable difference between sampling across cycles vs within cycles. SA would occasionally have a wider standard deviation (as shown in **Fig. S4**) when sampling within cycles, suggesting conformational space is more thoroughly covered when sampling across cycles. We confirm it is best to sample across many cycles to achieve the higher variability between conformer geometries.

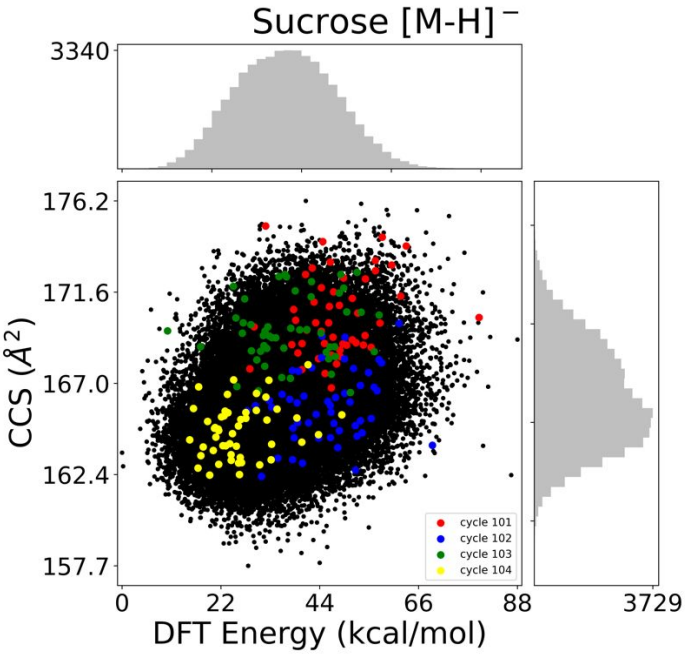


Fig. S2 A particularly distinct example of four sequential AMBER simulated annealing cycles clustering separately.

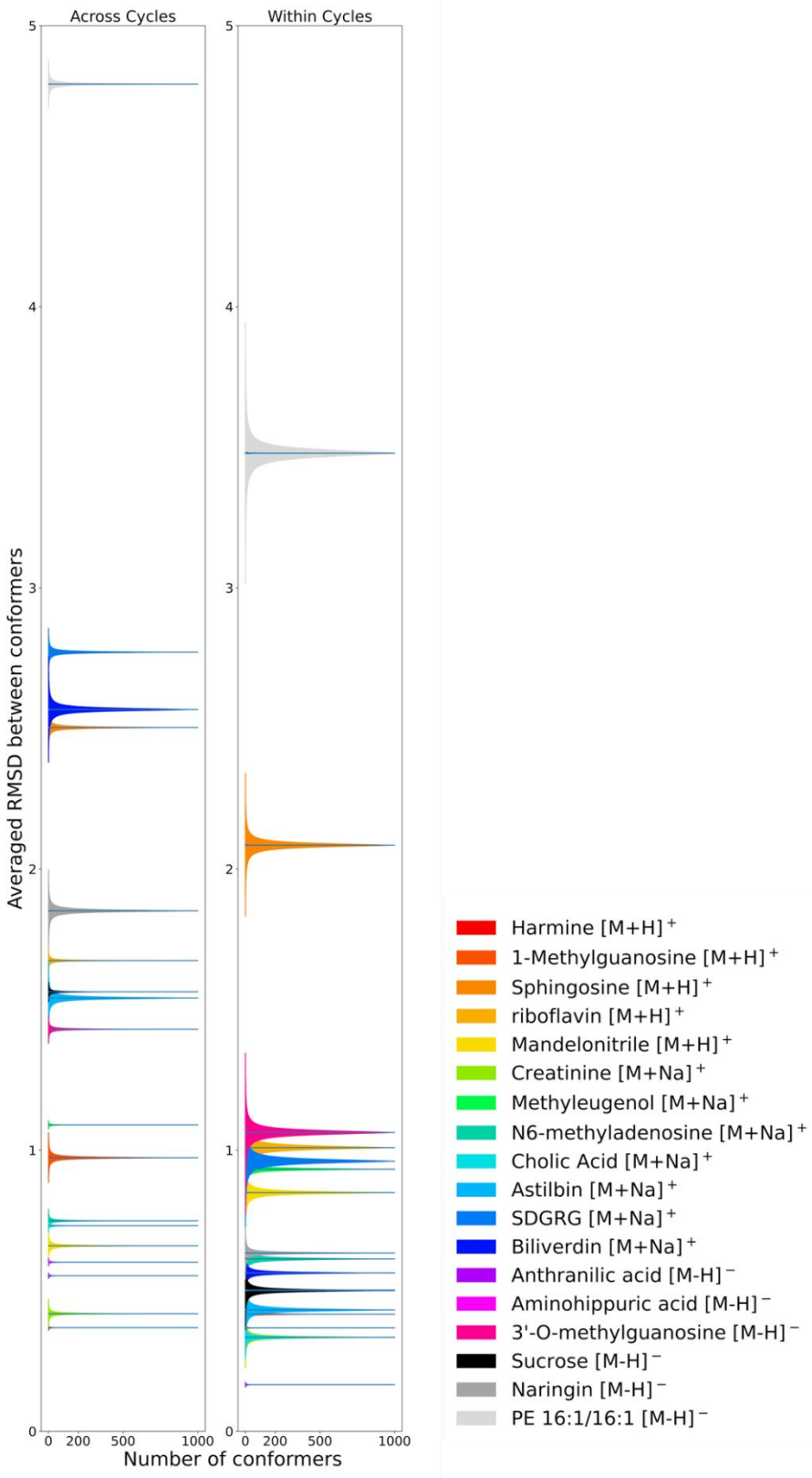


Fig. S3 Monte Carlo convergence plots of the RMSD between conformer geometries for 18 small molecules. The left shows random sampling across AMBER simulated annealing cycles, or treating the whole conformation as a single pool. The right shows sampling within cycle, or sampling a number of whole cycles together.

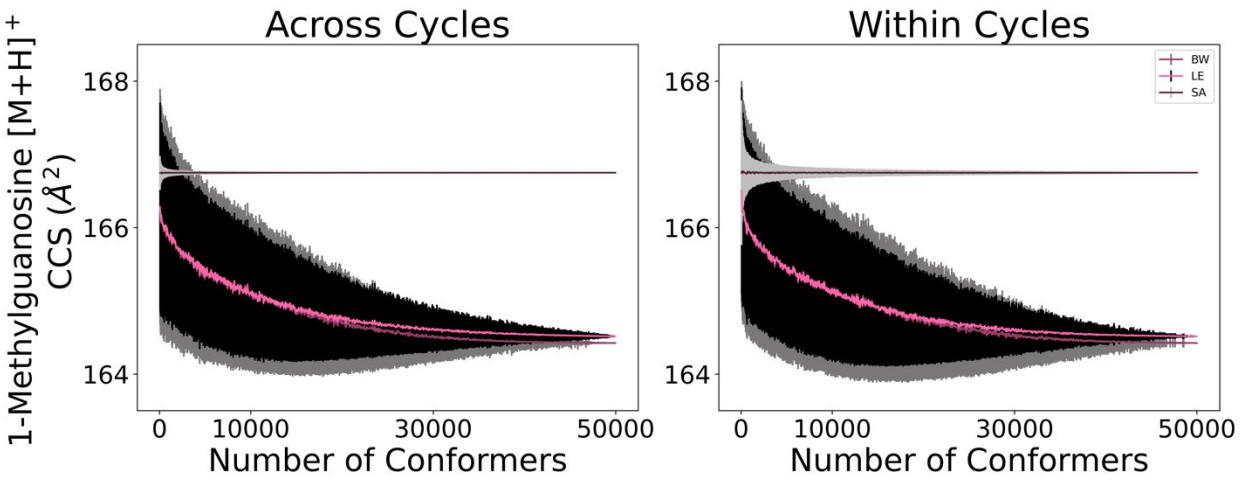


Fig. S4 Monte Carlo convergence plots of BW, LE, and SA demonstrating a lower standard deviation (higher precision) for SA when sampling across cycles (left) than when sampling within cycles (right).

5. Similarity Downselection Description

The goal of similarity downselection (SDS) is to sample conformational space with fewer conformers while still being representative of the larger population, thus saving on computational expense. Pairwise RMSD between conformations are used as a reciprocative similarity metric – the smaller the RMSD, the greater the similarity. SDS downselects based on this structural similarity to choose a subset of representative similar and most dissimilar conformers.

We developed a heuristic algorithm for performing SDS and created an open source Python package that can be found at https://github.com/pnnl/sds. The package includes relevant functions for performing SDS on conformers, but the SDS algorithm can also be generalized to any set of items where the items can be described as arrays whose elements are composed of the pairwise relations between the item in question and all other items of the set. Here, we employed the SDS algorithm to find the set of the *n* conformers most dissimilar from each other. To choose the most similar conformer, the pairwise RMSD between all conformers was summed, and the conformer with the smallest total RMSD was considered the most similar conformer.

6. Molecular Property Correlations

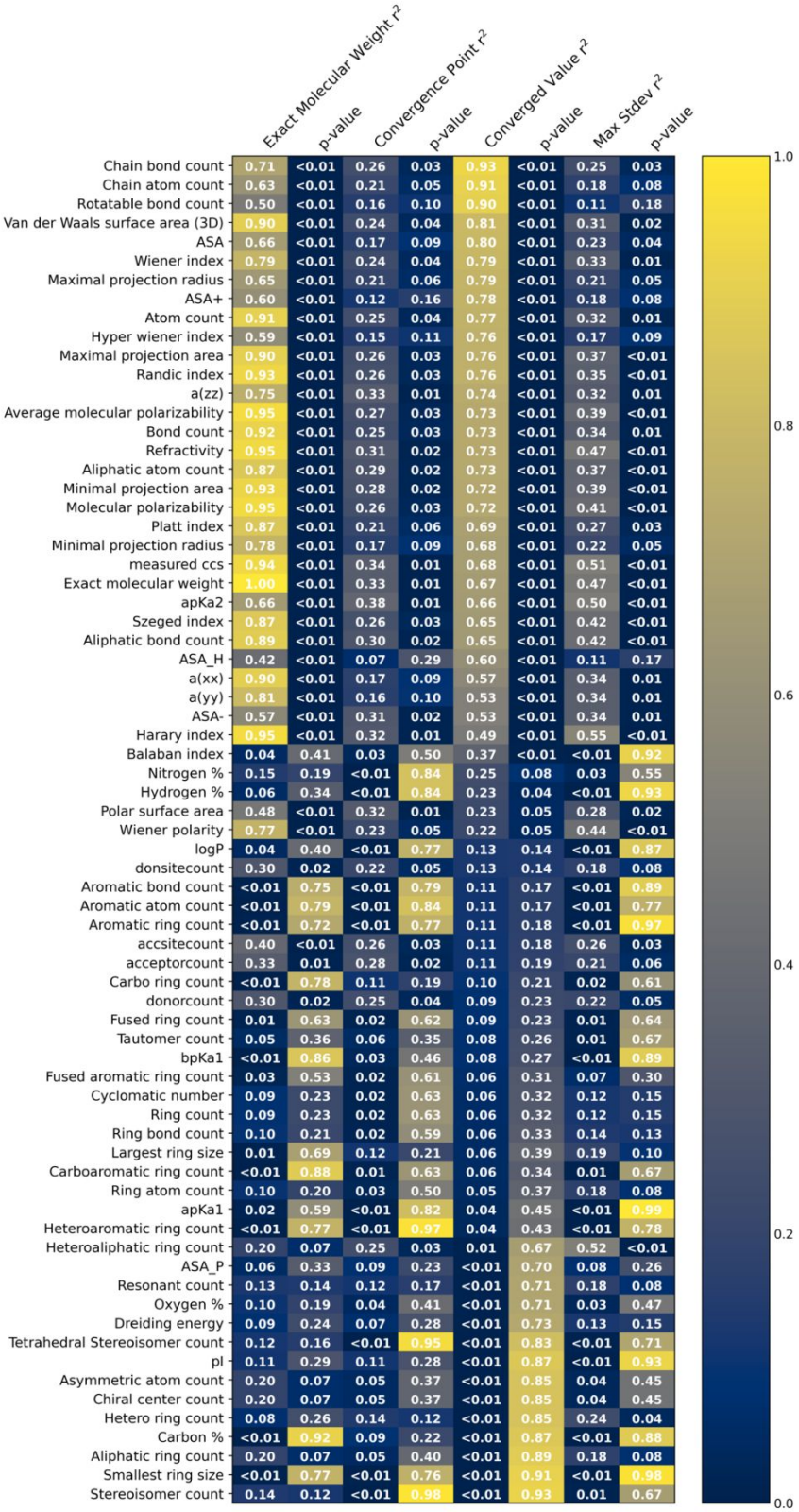


Fig. S5 Heat map of molecular properties (y-axis) correlated with RMSD MC convergence properties and exact molecular weight (x-axis), showing the r^2 correlation with its associated p-value.

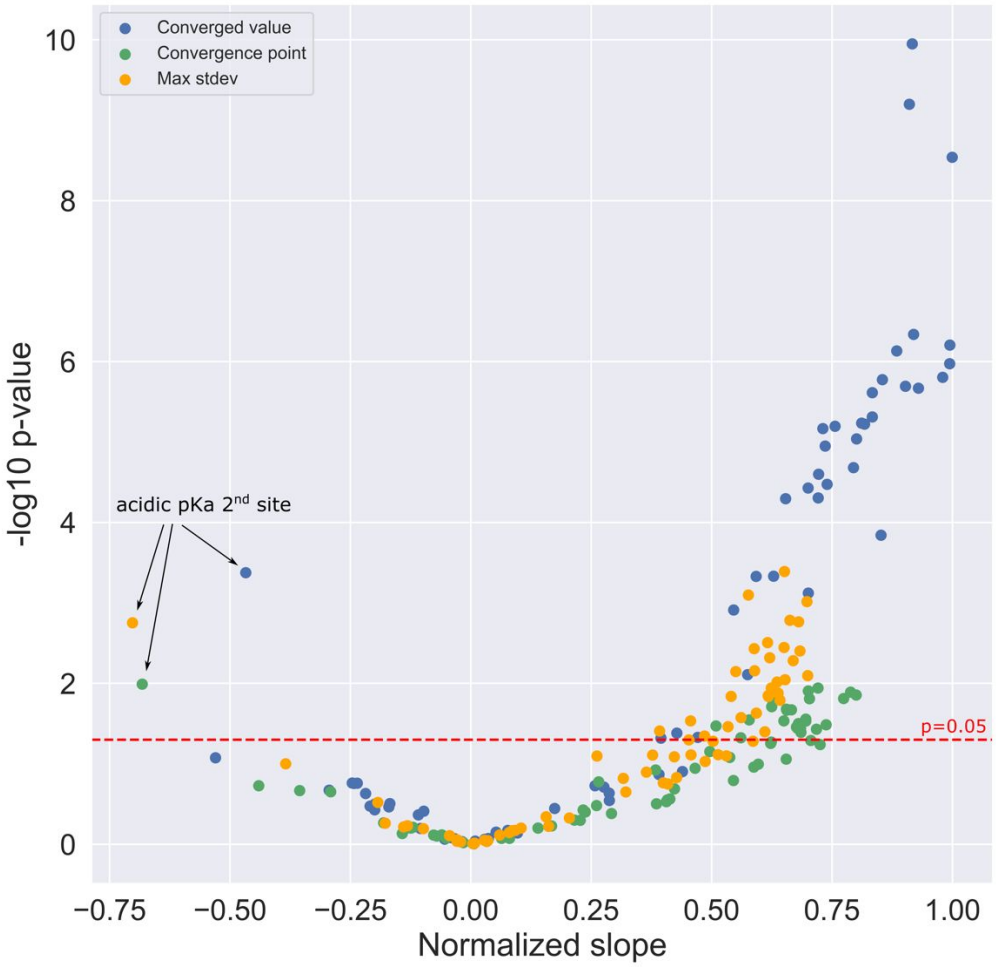


Fig. S6 Volcano plot of MC convergence properties (converged value, convergence point, maximum standard deviation) correlated against calculated molecular properties. The plot shows possible statistical significance of these correlations, although the p-value of 0.05 was arbitrarily chosen, and having a lower p-value does not necessarily mean higher statistical significance once the target p-value threshold his crossed. Interestingly, we note the one molecular property most negatively correlated with all three MC convergence properties (metrics of the variability between a molecule's conformers) was the second acidic pKa site being more weakly acidic.

Table S2 CCS values for various conformer selection methods. See **Table S3** for description of columns.

Molecule	BW	LE	SA	ET 5	ET 2	ET 1	ET 0.5	CREST	ISiCLE	Best Combo	Experimental
Harmine +H	149.2	148	149.7	149.4	149.1	149.3	149.3	149.2	148.7	149.7	146
1-Methylguanosine +H	164.5	164.4	166.7	165.4	165	164.5	164.5	164	165.7	165.8	168.8
Sphingosine +H	201.8	207.4	207.4	204.9	202.8	202.8	196.8	170.9	180.6	184.5	186
riboflavin +H	192.8	193.1	197.3	193.7	193.9	191.6	191.6	182.4	192.9	193.4	188.3
Mandelonitrile +H	129	128.5	130.4	130.3	129.7	128.5	128.5	131.3	127.1	129.2	128.9
Creatinine +Na	119.4	119.5	119.7	119.3	119.3	119.4	119.6	118.7	118.9	119.7	133.4
Methyleugenol +Na	143.3	143	143	143	143	143.4	144.2	139.4	141.8	143.1	160.4
N6-methyladenosine +Na	166.8	168	165.1	165.7	166.2	166.7	167.3	164.9	166	165.5	170.4
Cholic Acid +Na	186.5	186.6	186.9	186.2	186.4	186.7	187.4	183.6	185.8	187.1	197.3
Astilbin +Na	206.5	208.3	201.2	202.7	204	204	208.3	202.8	200.9	203.3	212.6
SDGRG +Na	227.5	227.5	218.4	227.5	227.5	227.5	227.5	209.6	226.6	224.3	203.5
Biliverdin +Na	256.8	256.4	251.8	250	263.4	256.4	256.4	258.3	257.6	265.2	246.7
Anthranilic acid -H	127.8	127.7	128.1	127.9	127.7	127.6	127.7	126.7	127.9	126.8	124
Aminohippuric acid -H	154.7	154.7	155.4	154.7	154.6	154.7	154.7	152.3	152.1	157.1	147.6
3'-O-methylguanosine -H	169	167.8	176.1	169.4	169.2	169.4	169.4	168.7	168.7	171.9	163.8
Sucrose -H	163.2	163.5	166.3	165.3	163.1	163.1	163.1	157.6	163.4	162.9	168.5
Naringin -H	238.8	238.1	235.9	238.8	240.5	238.1	238.1	219.3	240.9	234.7	217.3
PE 16:1/16:1 -H	292.4	292.4	309.7	305.2	291.9	291.9	291.9	NaN	311.3	314.5	256.3

Table S3 Mean absolute percent error of various conformer selection method results relative to ISiCLE CCS. The ISiCLE method selects the most similar and two most dissimilar conformers out of 10 AMBER simulated annealing cycles (for a total of 30 conformers), applies DFT geometry optimization, and averages the CCS with Boltzmann weighting. Boltzmann weighting (BW), lowest energy (LE), simple average (SA), and simple averaging under energy thresholds 5, 2, 1, and 0.5 *kcal/mol* (ET 5, 2, 1, 0.5) were applied to 50k AMBER conformers and using DFT energies. CREST is the single lowest energy CREST conformer. The best combo is the statistical best combination we found on the AMBER conformers for the 18 molecules—10 AMBER cycles, selecting the most similar and 10 most dissimilar set under an AMBER energy threshold of 10 *kcal/mol,* and choosing the lowest energy according to the conformer's DFT energy.

Molecule	BW	LE	SA	ET 5	ET 2	ET 1	ET 0.5	CREST	ISiCLE	Best Combo
Harmine +H	0.33	0.48	0.70	0.49	0.27	0.45	0.44	0.35	0.66	0.48
1-Methylguanosine +H	0.69	0.75	0.66	0.16	0.39	0.72	0.72	1.01	0.10	0.75
Sphingosine +H	11.75	14.83	14.85	13.47	12.32	12.30	8.95	5.36	2.13	14.83
riboflavin +H	0.10	0.10	2.26	0.40	0.51	0.69	0.69	5.46	0.23	0.10
Mandelonitrile +H	1.54	1.09	2.59	2.54	2.08	1.09	1.09	3.32	1.63	1.09
Creatinine +Na	0.41	0.56	0.66	0.35	0.34	0.45	0.65	0.17	0.67	0.56
Methyleugenol +Na	1.10	0.89	0.86	0.88	0.86	1.17	1.69	1.66	0.96	0.89
N6-methyladenosine +Na	0.51	1.22	0.54	0.17	0.14	0.40	0.77	0.66	0.32	1.22
Cholic Acid +Na	0.41	0.42	0.61	0.21	0.33	0.47	0.89	1.15	0.72	0.42
Astilbin +Na	2.76	3.68	0.14	0.88	1.55	1.55	3.68	0.94	1.17	3.68
SDGRG +Na	0.42	0.42	3.60	0.42	0.42	0.42	0.42	7.51	1.02	0.42
Biliverdin +Na	0.33	0.47	2.28	2.97	2.22	0.47	0.47	0.24	2.95	0.47
Anthranilic acid -H	0.07	0.09	0.16	0.03	0.16	0.23	0.16	0.89	0.80	0.09
Aminohippuric acid -H	1.66	1.70	2.17	1.66	1.61	1.70	1.70	0.11	3.25	1.70
3'-O-methylguanosine -H	0.20	0.50	4.40	0.45	0.32	0.43	0.43	0.02	1.92	0.50
Sucrose -H	0.11	0.06	1.78	1.17	0.16	0.16	0.16	3.52	0.29	0.06
Naringin -H	0.86	1.14	2.07	0.86	0.15	1.14	1.14	8.95	2.56	1.14
PE 16:1/16:1 -H	6.06	6.07	0.51	1.97	6.24	6.24	6.24	NaN	1.04	6.07
MAPE	1.63	1.91	2.27	1.62	1.67	1.67	1.68	2.43	1.25	1.91

Table S4 Same **Table S3** except with mean absolute percent error calculated relative to experimental CCS.

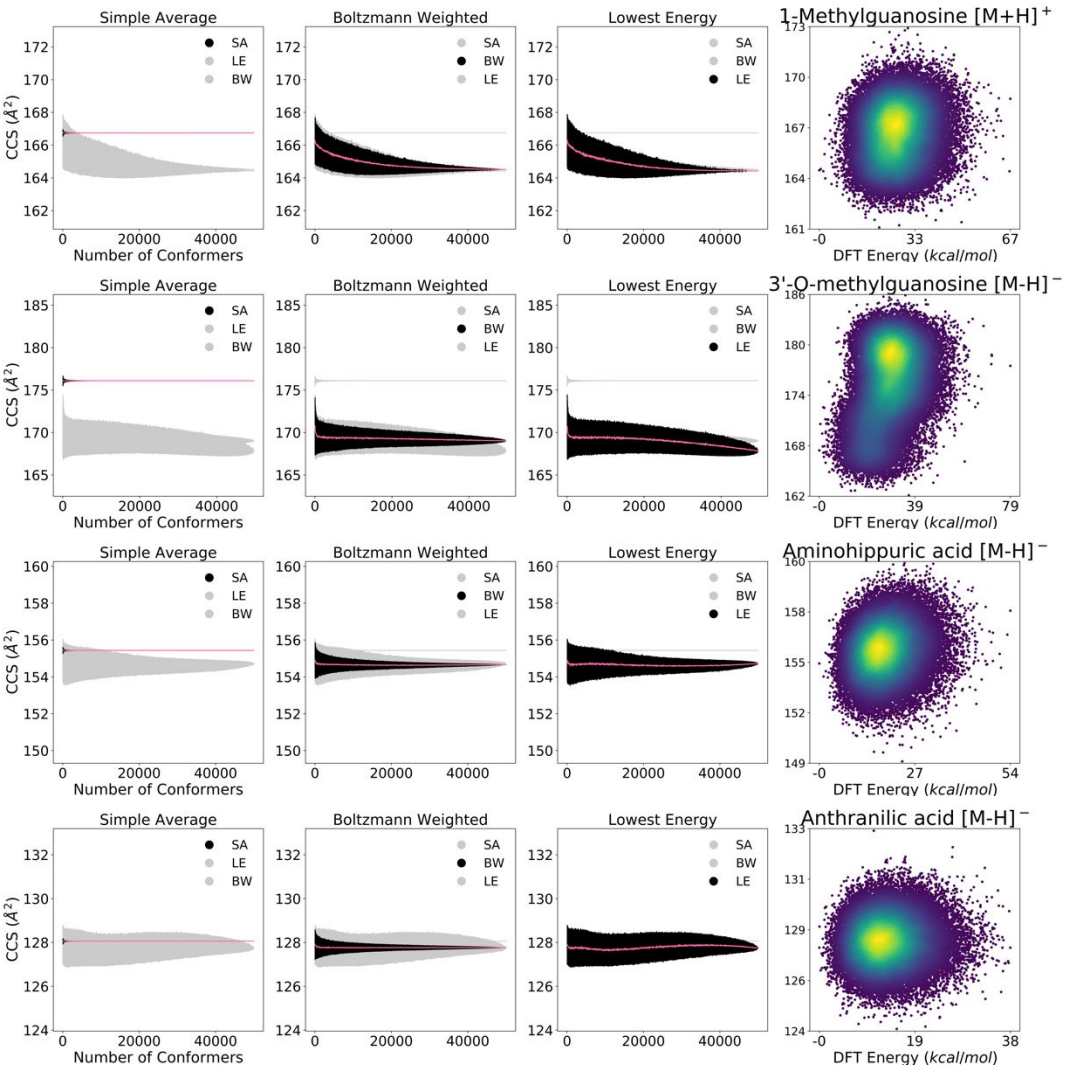
Molecule	BW	LE	SA	ET 5	ET 2	ET 1	ET 0.5	CREST	ISiCLE	Best Combo
Harmine +H	2.14	1.32	2.52	2.31	2.08	2.27	2.26	2.17	1.81	2.48
1-Methylguanosine +H	2.54	2.59	1.22	2.02	2.24	2.57	2.57	2.86	1.86	1.76
Sphingosine +H	8.50	11.49	11.51	10.17	9.06	9.04	5.78	8.11	2.90	0.83
riboflavin +H	2.38	2.59	4.80	2.89	3.01	1.77	1.77	3.12	2.48	2.72
Mandelonitrile +H	0.13	0.32	1.16	1.11	0.66	0.32	0.32	1.88	1.39	0.22
Creatinine +Na	10.53	10.40	10.31	10.59	10.60	10.51	10.32	11.06	10.90	10.31
Methyleugenol +Na	10.62	10.80	10.82	10.81	10.83	10.55	10.09	13.06	11.59	10.74
N6-methyladenosine +Na	2.10	1.40	3.12	2.76	2.45	2.20	1.84	3.24	2.59	2.90
Cholic Acid +Na	5.48	5.46	5.29	5.66	5.55	5.42	5.02	6.95	5.86	5.19
Astilbin +Na	2.90	2.03	5.37	4.67	4.04	4.04	2.03	4.62	5.51	4.40
SDGRG +Na	11.81	11.81	7.34	11.81	11.81	11.81	11.81	2.99	11.34	10.20
Biliverdin +Na	4.08	3.93	2.05	1.32	6.74	3.93	3.93	4.67	4.42	7.50
Anthranilic acid -H	3.04	3.02	3.28	3.14	2.95	2.87	2.95	2.19	3.11	2.29
Aminohippuric acid -H	4.82	4.86	5.34	4.82	4.77	4.86	4.86	3.22	3.11	6.46
3'-O-methylguanosine -H	3.20	2.48	7.52	3.46	3.32	3.43	3.43	3.01	2.99	4.96
Sucrose -H	3.13	2.97	1.29	1.88	3.17	3.17	3.17	6.43	3.02	3.31
Naringin -H	9.88	9.57	8.54	9.88	10.67	9.57	9.57	0.91	10.83	7.99
PE 16:1/16:1 -H	14.10	14.09	20.84	19.07	13.88	13.88	13.88	NaN	21.46	22.72
MAPE	5.63	5.62	6.24	6.02	5.99	5.68	5.31	4.73	5.96	5.94

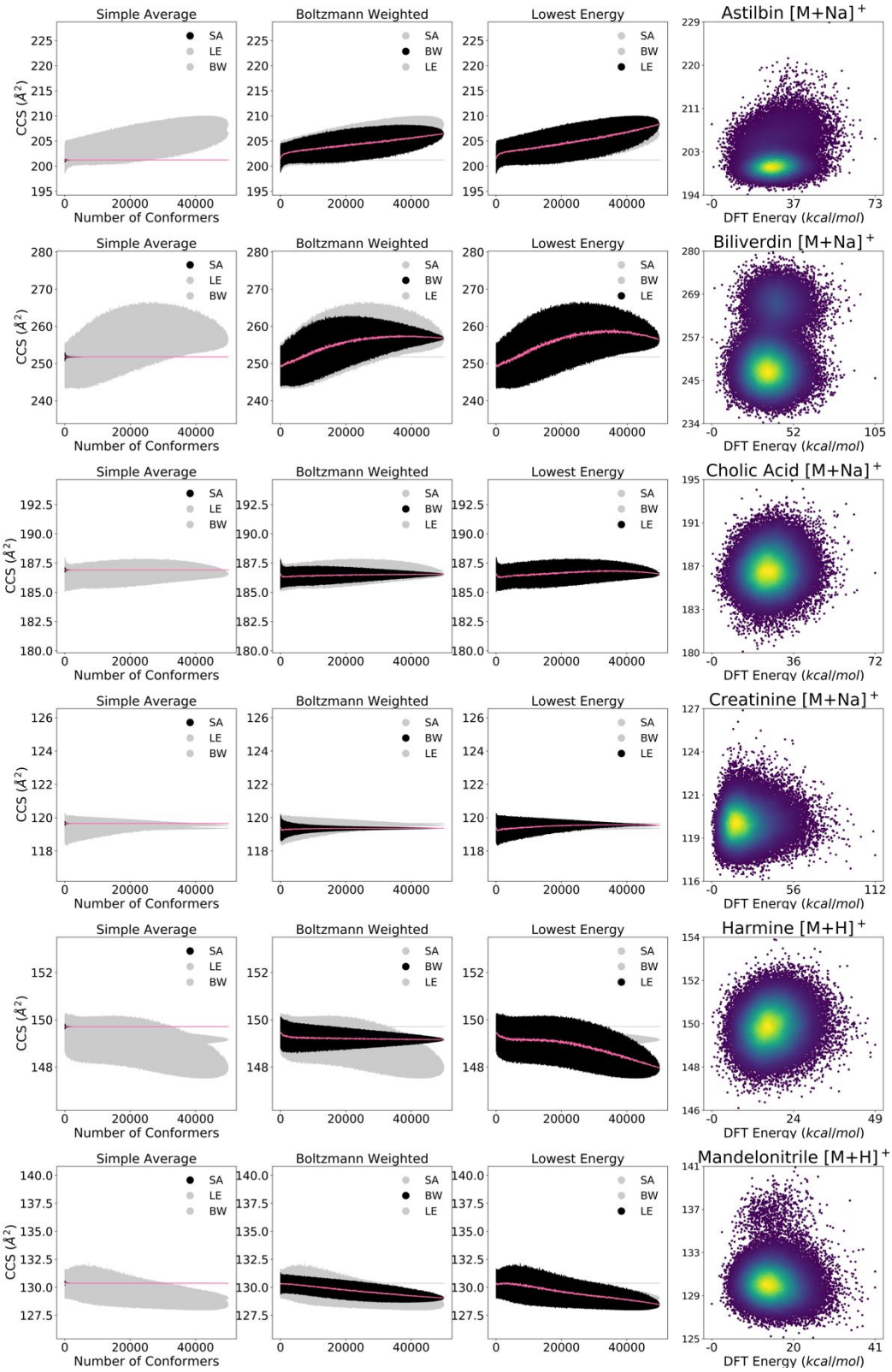
Table S5. Shows at which Sanders (AmberTools17) simulated annealing cycle the n^{th} lowest energy conformer was generated for n=1-10. The maximum number of cycles generated for this project was 1000. Based on this data it is reasonable to assume new low energy conformers would be generated past 1000 cycles.

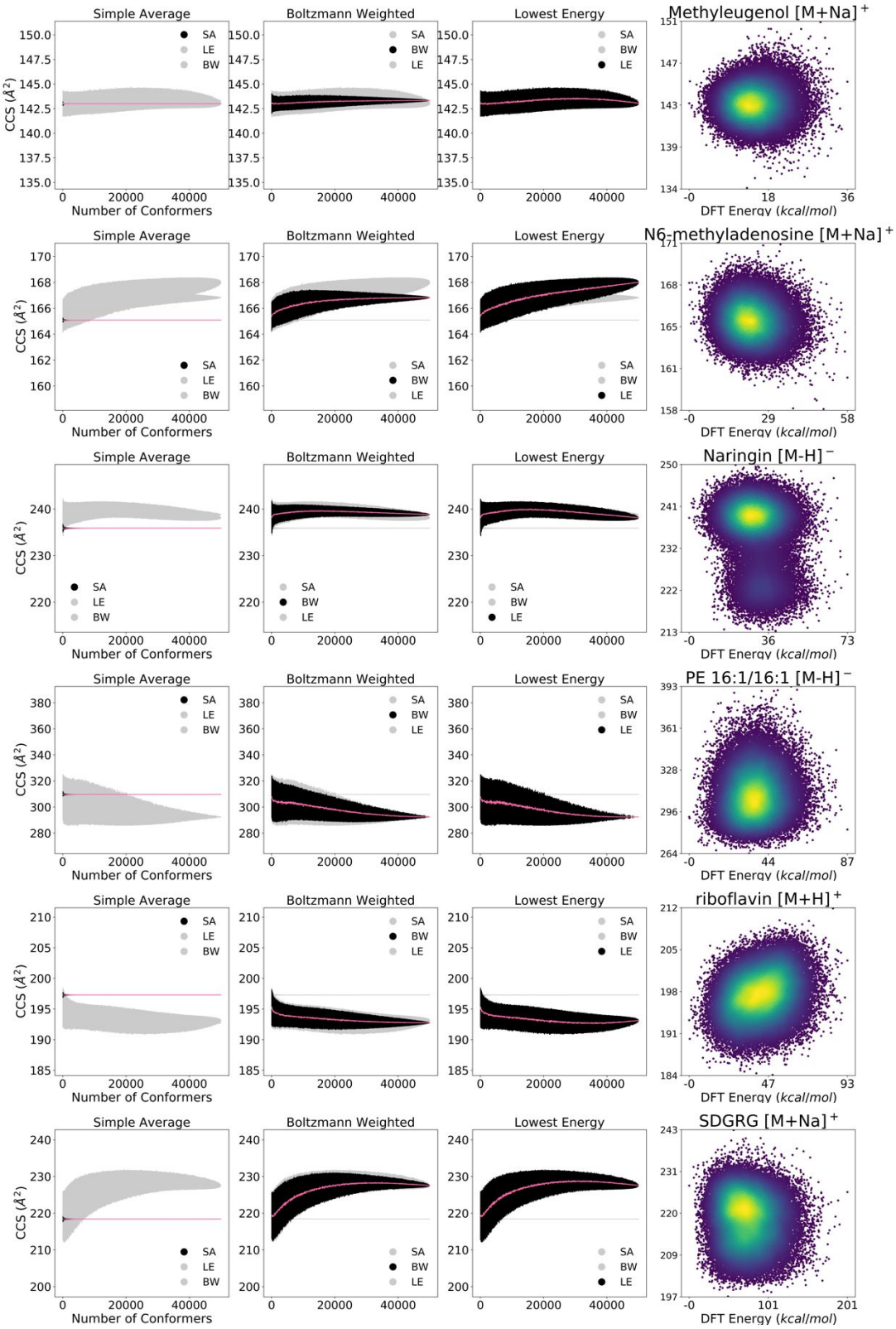
	AMBER cycles which generated the nth lowest energy conformer									
Molecular Adduct	1st LE	2nd LE	3rd LE	4th LE	5th LE	6th LE	7th LE	8th LE	9th LE	10th LE
Harmine	765	832	903	47	767	798	814	45	703	776
1-Methylguanosine	145	285	922	145	312	490	147	486	685	81
Sphingosine	362	449	656	947	479	273	92	180	991	60
riboflavin	343	66	442	934	501	781	303	792	87	134
Mandelonitrile	152	975	649	972	462	720	348	160	223	785
Creatinine	198	944	3	565	761	162	594	951	374	926
Methyleugenol	234	292	580	865	902	612	62	655	997	544
N6-methyladenosine	409	583	409	380	202	975	181	560	563	176

Cholic Acid	117	955	789	246	766	246	190	29	418	658
Astilbin	704	594	574	866	361	854	83	494	11	133
SDGRG	422	815	283	50	521	202	34	152	822	206
Biliverdin	590	95	505	351	986	406	590	641	491	224
Anthranilic acid	547	644	652	655	922	227	374	259	120	930
Aminohippuric acid	366	10	607	733	643	930	372	718	882	583
3'-O-methylguanosine	114	31	708	434	635	929	541	249	754	633
Sucrose	316	540	117	241	630	505	960	37	77	540
Naringin	153	664	621	801	681	663	39	490	328	338
PE 16:1/16:1	398	870	400	728	407	311	219	227	705	883
Average	**352**	**536**	**546**	**553**	**608**	**560**	**330**	**396**	**513**	**478**
stdev	199	336	243	322	220	282	273	284	324	311
Max	**765**	**975**	**922**	**972**	**986**	**975**	**960**	**951**	**997**	**930**
Min	**114**	**10**	**3**	**47**	**202**	**162**	**34**	**29**	**11**	**60**

7. Monte Carlo Simulations and CCS vs Energy Space




Simple Average
Boltzmann Weighted
Lowest Energy
Astilbin [M+Na]+
Biliverdin [M+Na]+
Cholic Acid [M+Na]+
Creatinine [M+Na]+
Harmine [M+H]+
Mandelonitrile [M+H]+
SA
BW
LE
CCS (Å2)
Number of Conformers
DFT Energy (kcal/mol)


Simple Average
Boltzmann Weighted
Lowest Energy
Methyleugenol [M+Na]+
N6-methyladenosine [M+Na]+
Naringin [M-H]−
PE 16:1/16:1 [M-H]−
riboflavin [M+Na]+
SDGRG [M+Na]+
SA
BW
LE
CCS (Å2)
Number of Conformers
DFT Energy (kcal/mol)

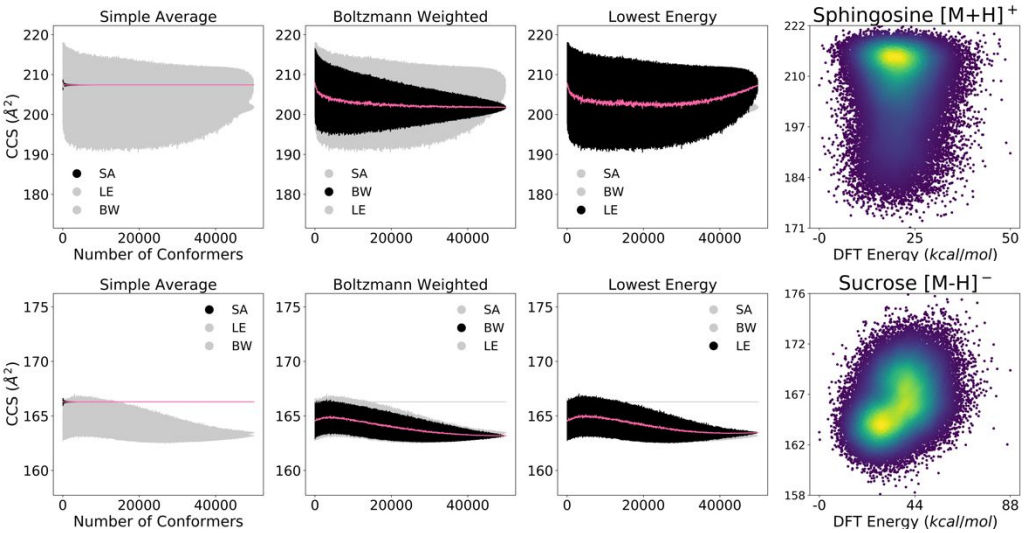


Fig. S7 MC simulation convergence plots on CCS (left) for SA, BW, and LE, and how they relate to CCS vs energy space (right). Black and gray represent standard deviation from the average (pink). SA converges to the CCS where the conformers are most dense in the CCS versus energy space, BW convergences to the average of low energy conformers or clusters of conformers, and LE converges to the single lowest energy CCS after 50k conformers are selected.

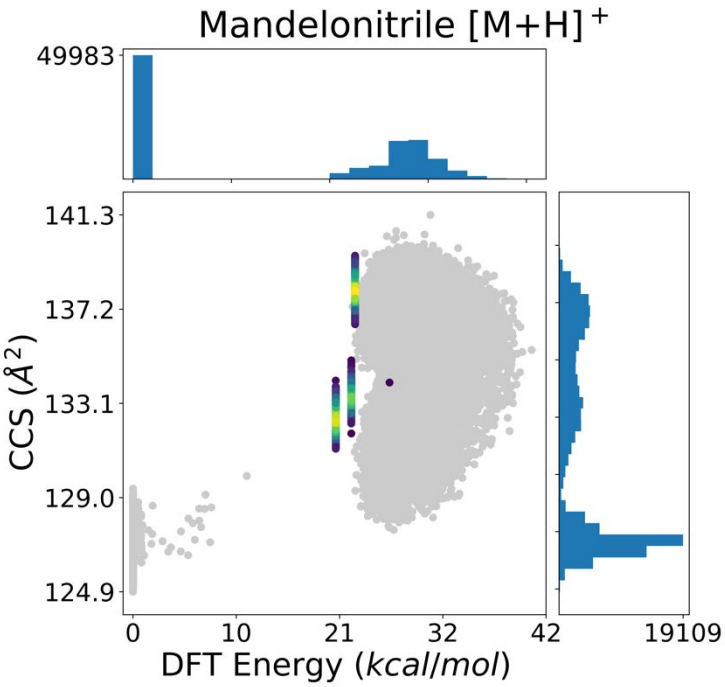


Fig. S8 RDKit with UFF optimization on ~6k mandelonitrile [M+H]⁺ conformers. Left gray cluster shows ~50k DFT geometry optimized AMBER structures, right gray cluster shows ~50k RDKit without UFF, and middle clusters with density coloring indicate the ~6k RDKit structures with UFF optimization. In this example, UFF optimization clustered the RDKit conformers into tight energy intervals, which would likely greatly increase precision for BW, LE, and other low energy dependent conformer selection methods. However, the UFF conformers have energies much higher than the DFT geometry optimized conformers, and different CCS as well, making it unclear how this affects accuracy.

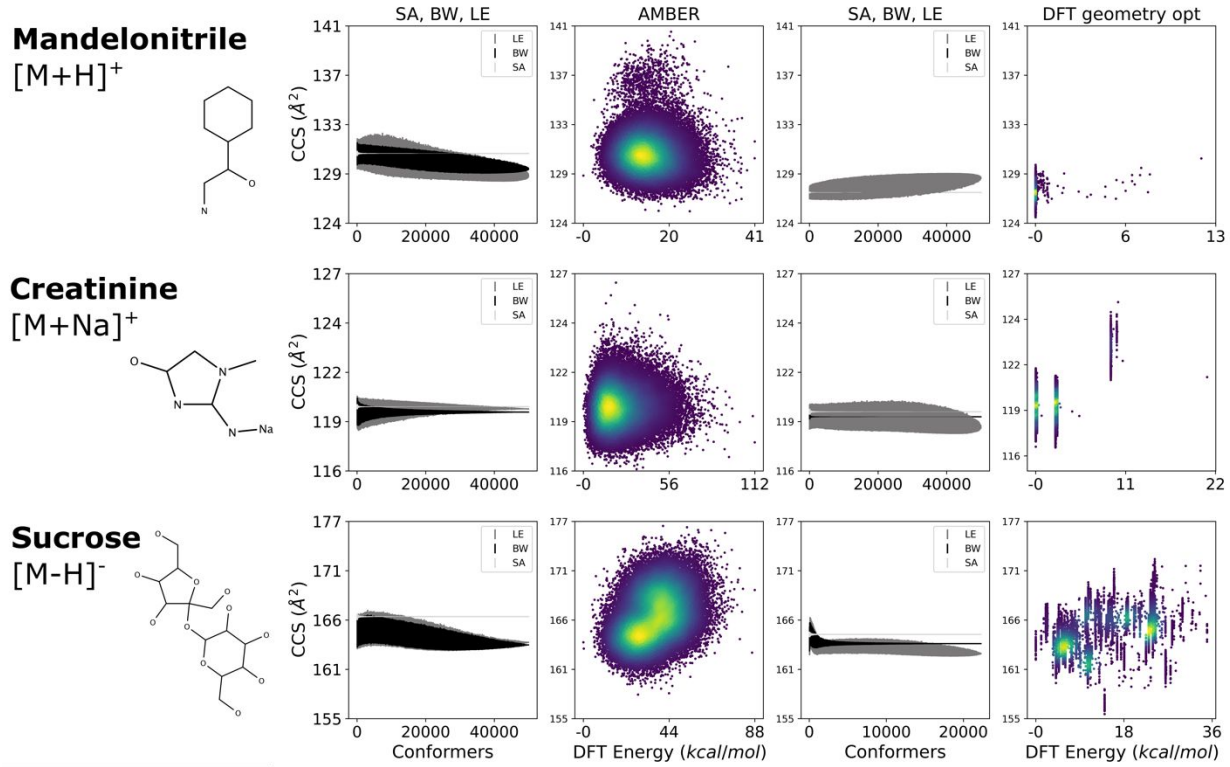
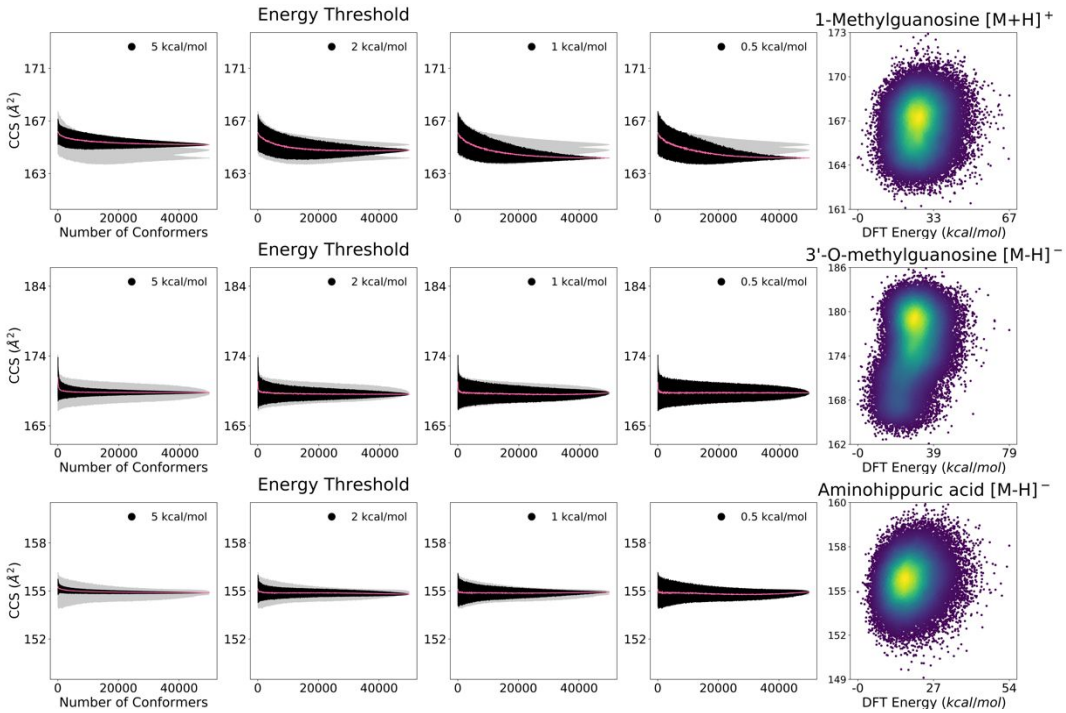
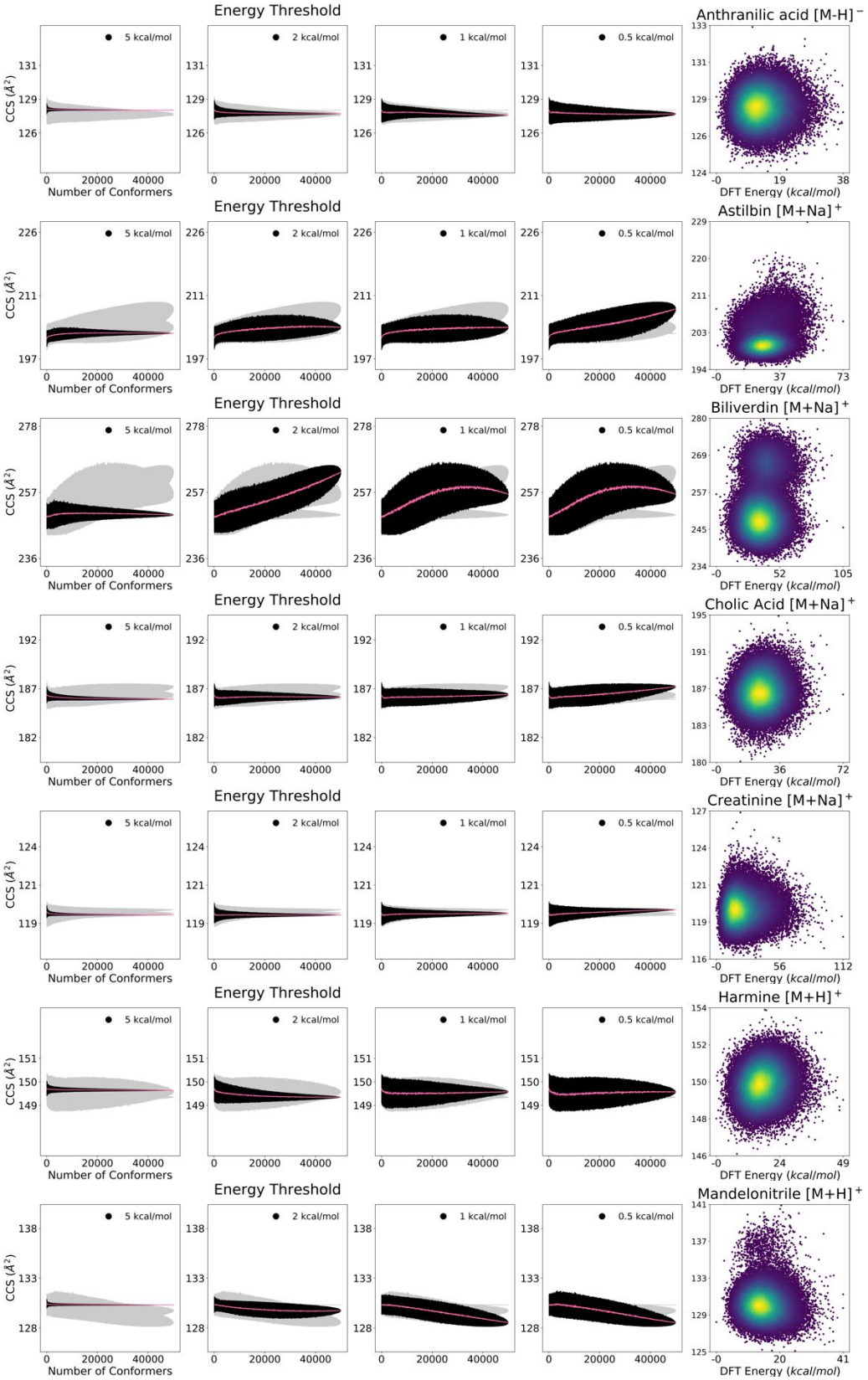
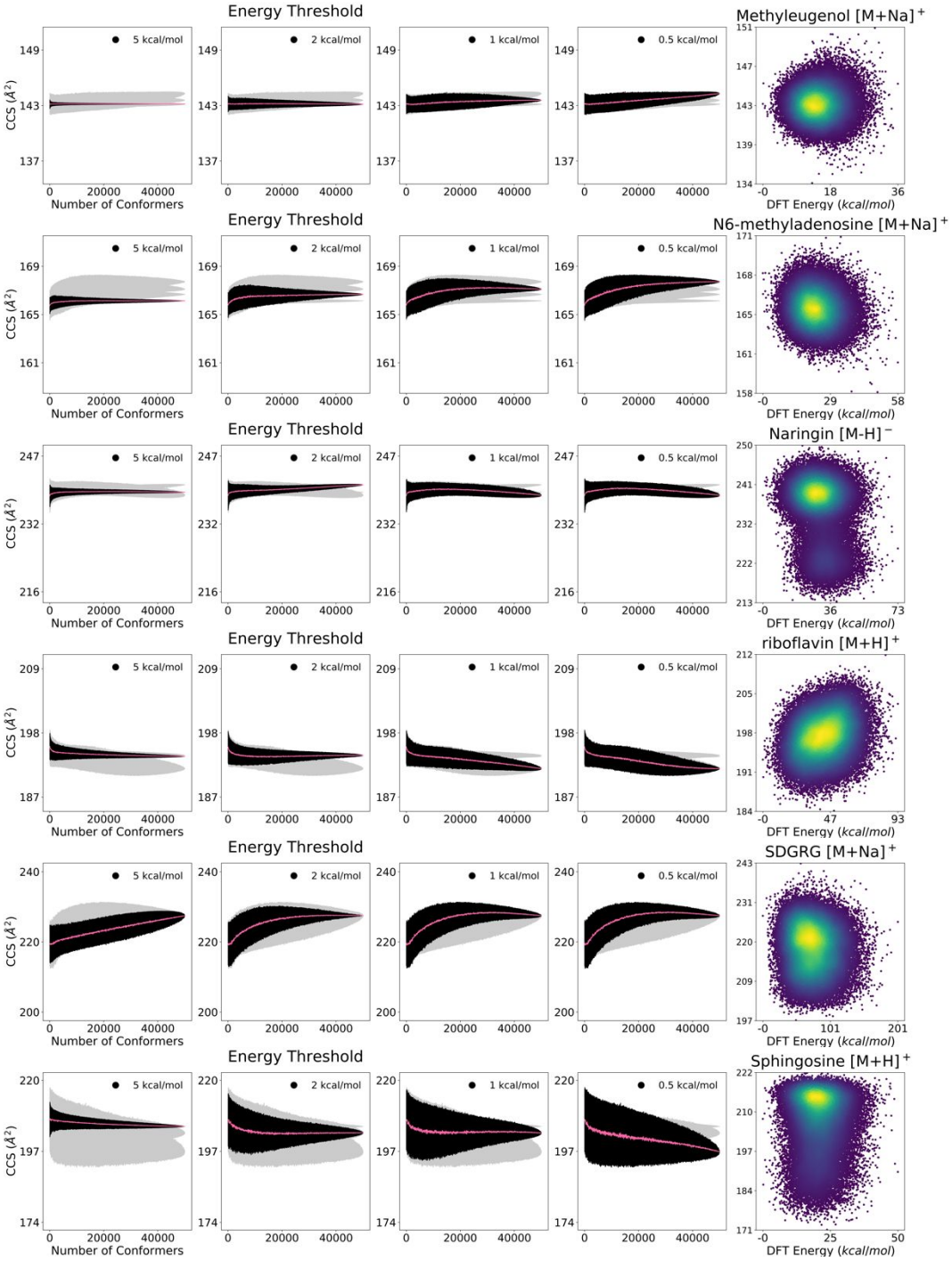


Fig. S9 MC convergence plots on CCS using three sampling techniques (SA, BW, LE) for conformers generated in AMBER and the same AMBER conformers after a DFT geometry optimization for mandelonitrile $[M+H]^+$, creatinine $[M+Na]^+$, and sucrose $[M-H]^-$. Note that for sucrose, only about 25k of the 50k AMBER conformers were DFT geometry optimized here.




Energy Threshold
5 kcal/mol
2 kcal/mol
1 kcal/mol
0.5 kcal/mol
CCS (Å²)
Number of Conformers
Anthranilic acid [M-H]⁻
Astilbin [M+Na]⁺
Biliverdin [M+Na]⁺
Cholic Acid [M+Na]⁺
Creatinine [M+Na]⁺
Harmine [M+H]⁺
Mandelonitrile [M+H]⁺
DFT Energy (kcal/mol)


Energy Threshold
5 kcal/mol
2 kcal/mol
1 kcal/mol
0.5 kcal/mol
CCS (Å²)
Number of Conformers
Methyleugenol [M+Na]⁺
N6-methyladenosine [M+Na]⁺
Naringin [M-H]⁻
riboflavin [M+H]⁺
SDGRG [M+Na]⁺
Sphingosine [M+H]⁺
DFT Energy (kcal/mol)

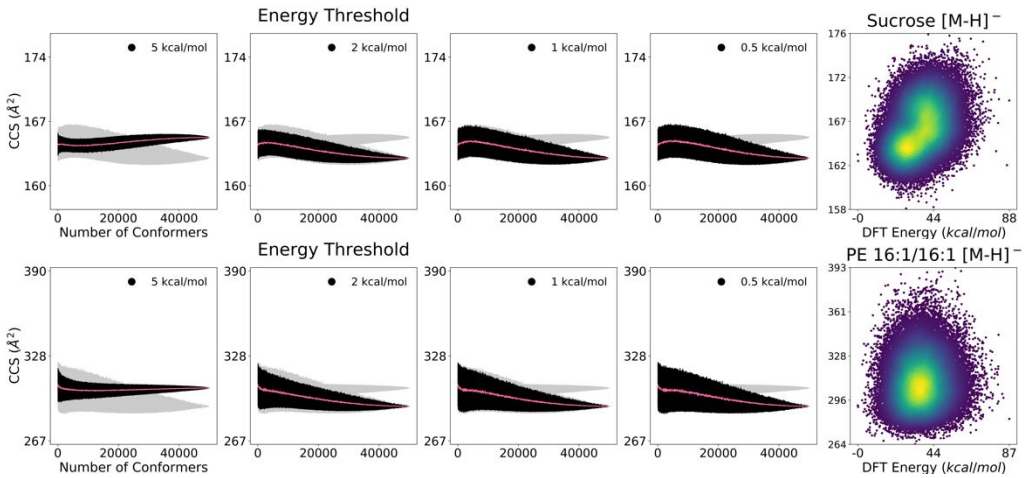
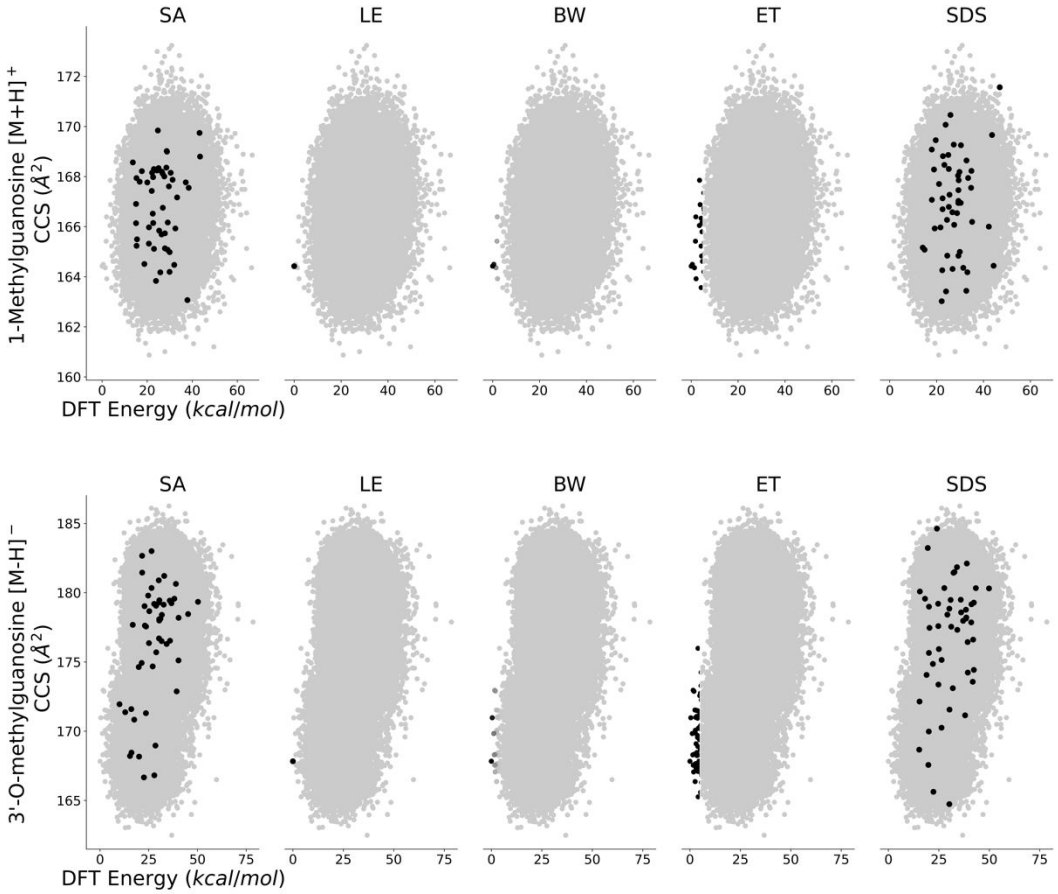
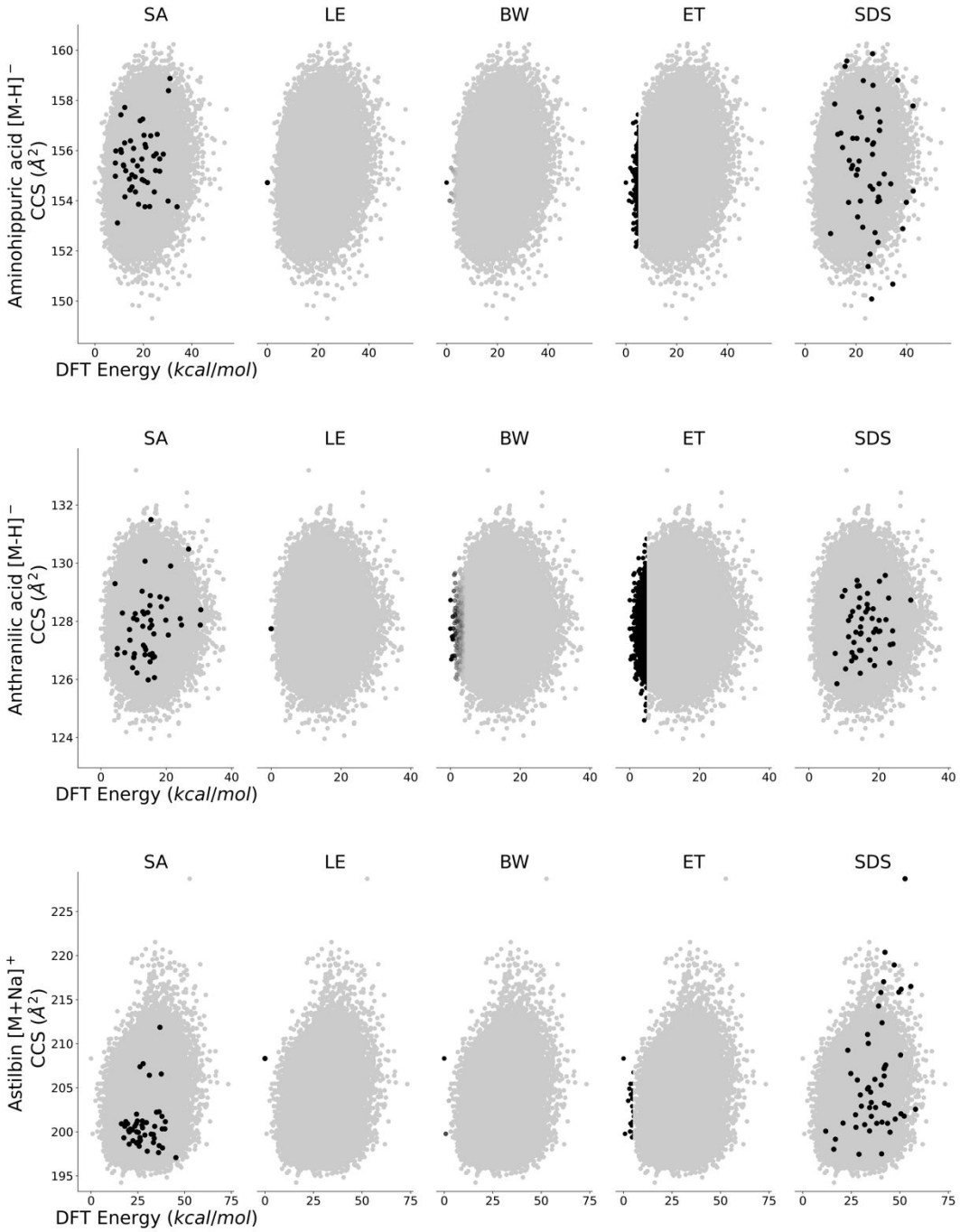
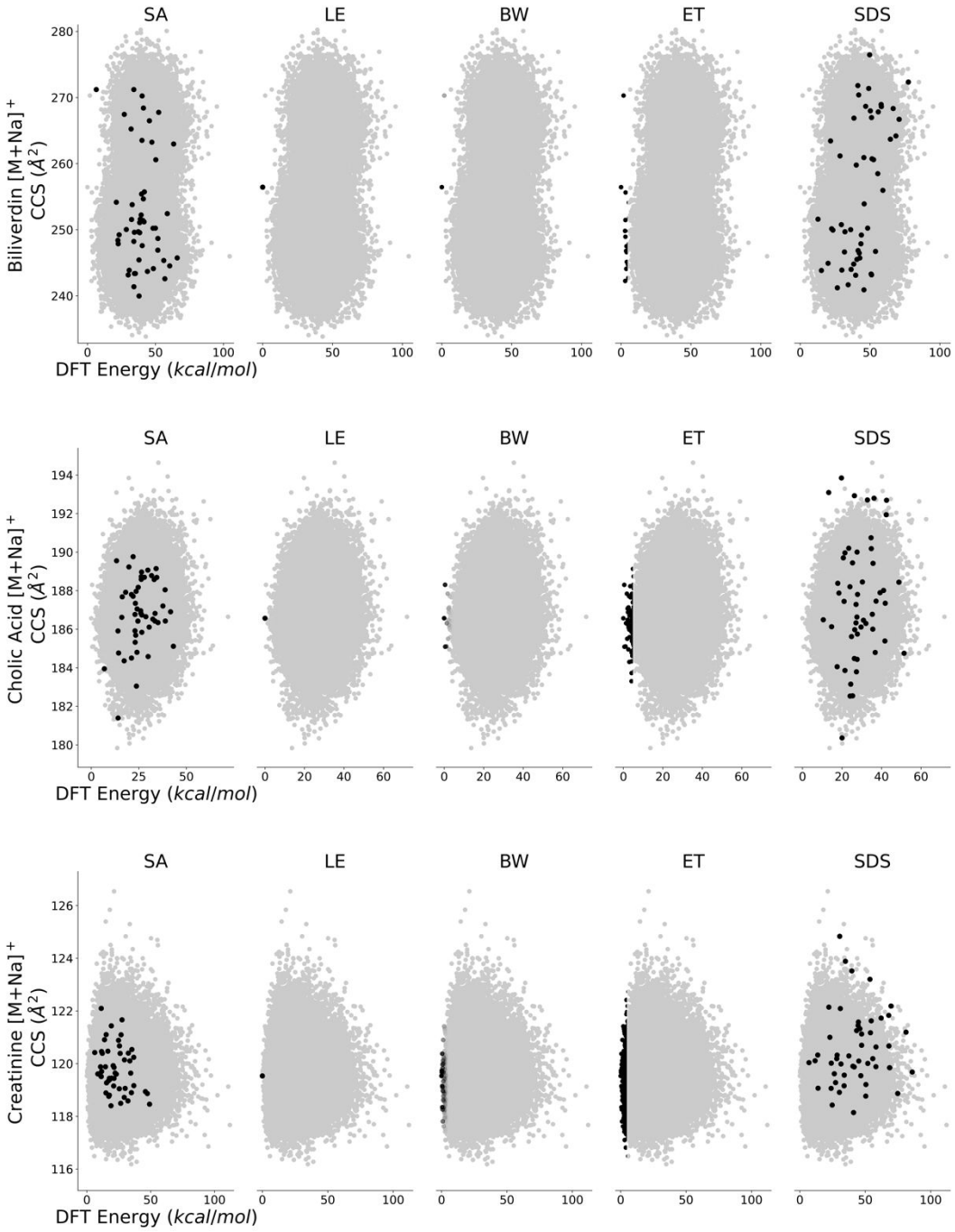
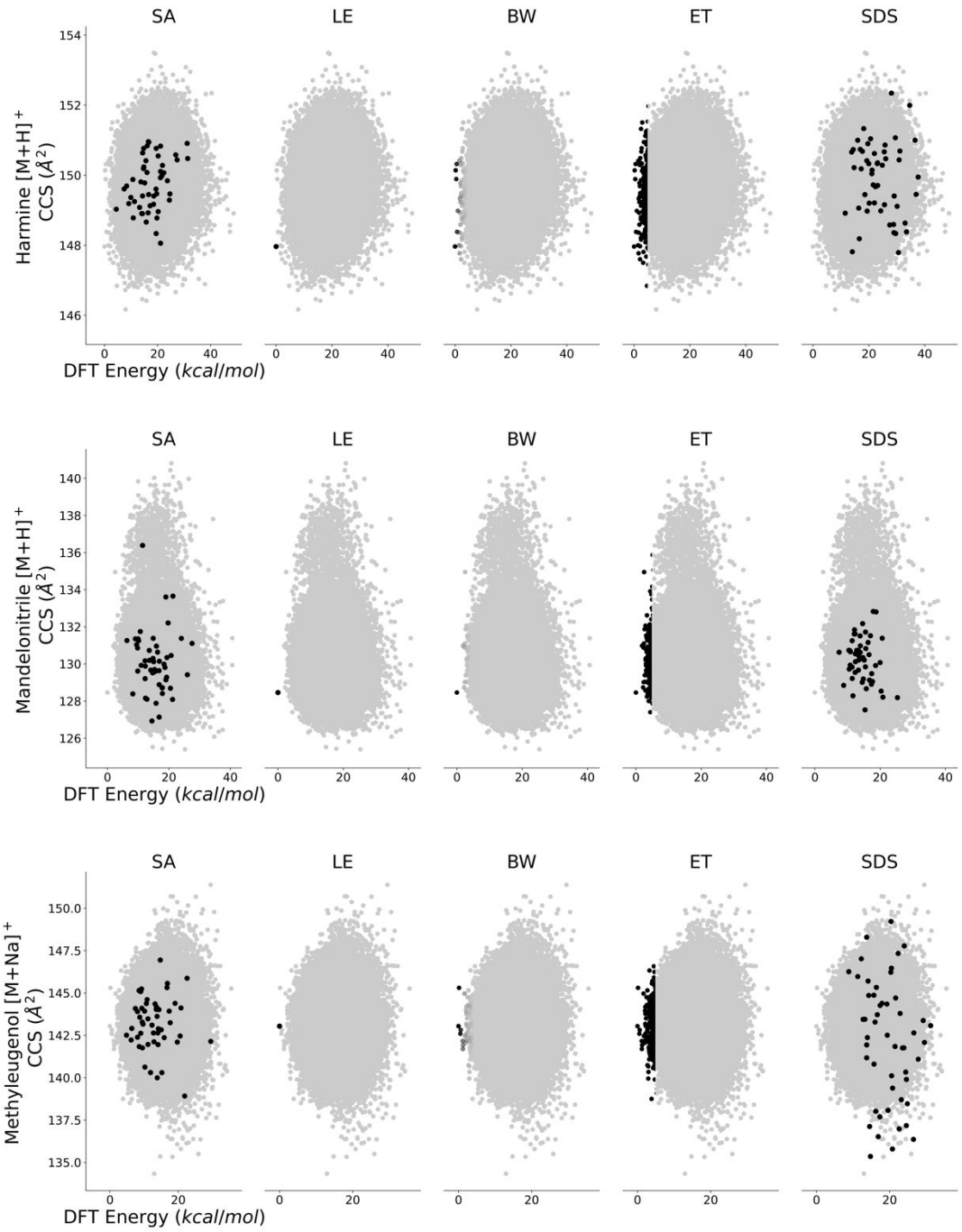


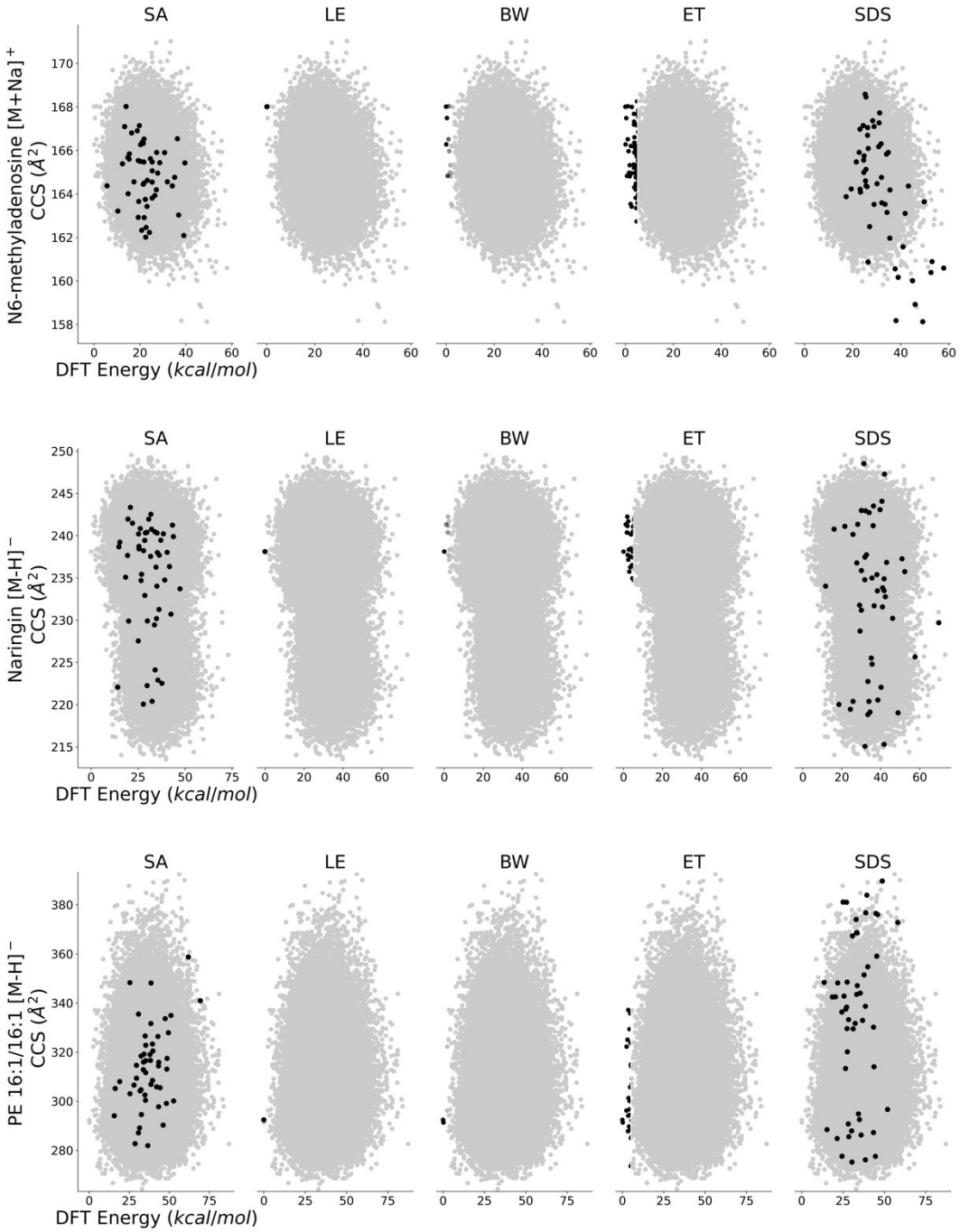
Fig. S10 MC convergence plots on CCS applying SA under 5, 2, 1, and 0.5 *kcal/mol* energy thresholds. Black represents standard deviation from the average (pink).

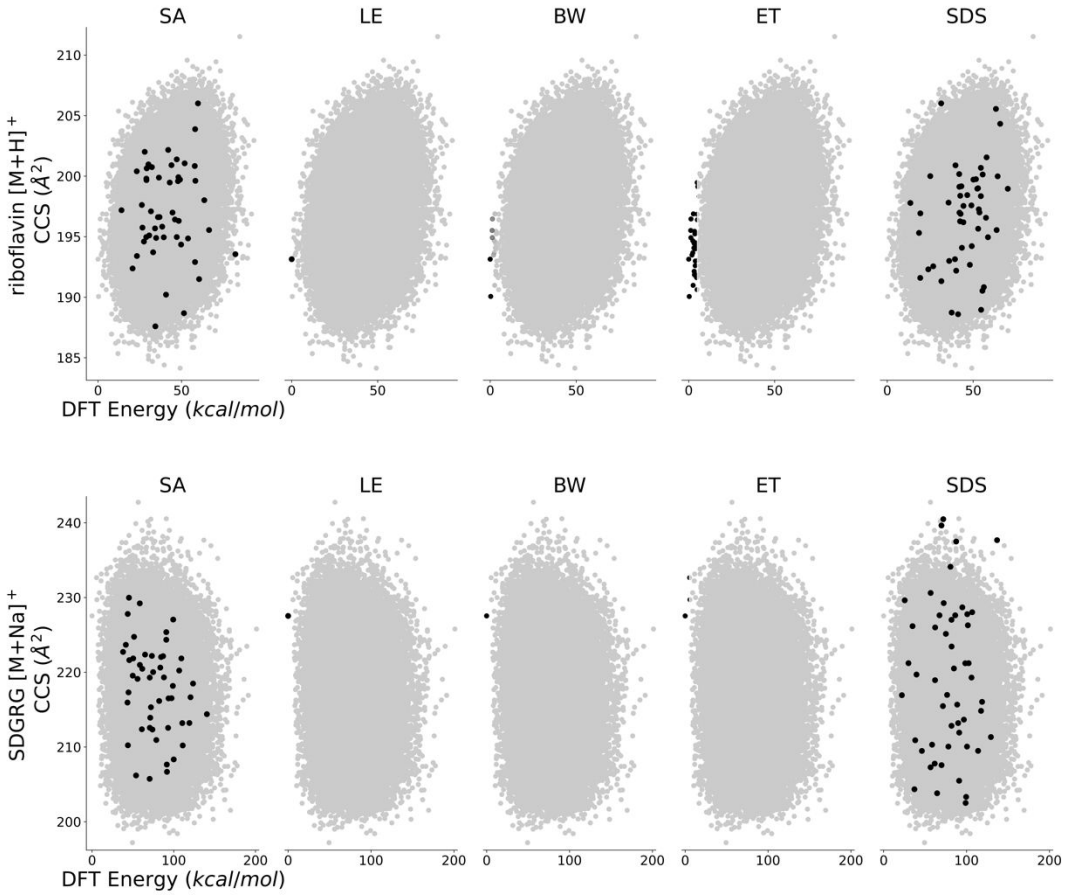



SA
LE
BW
ET
SDS
Aminohippuric acid [M-H]⁻
CCS (Å²)
Anthranilic acid [M-H]⁻
CCS (Å²)
Astilbin [M+Na]⁺
CCS (Å²)
DFT Energy (kcal/mol)


SA
LE
BW
ET
SDS
Biliverdin [M+Na]+ CCS (Å2)
Cholic Acid [M+Na]+ CCS (Å2)
Creatinine [M+Na]+ CCS (Å2)
DFT Energy (kcal/mol)


SA
LE
BW
ET
SDS
Harmine [M+H]+ CCS (Å2)
Mandelonitrile [M+H]+ CCS (Å2)
Methyleugenol [M+Na]+ CCS (Å2)
DFT Energy (kcal/mol)


SA
LE
BW
ET
SDS
N6-methyladenosine [M+Na]+
CCS (Å2)
DFT Energy (kcal/mol)
Naringin [M-H]−
CCS (Å2)
DFT Energy (kcal/mol)
PE 16:1/16:1 [M-H]−
CCS (Å2)
DFT Energy (kcal/mol)


SA
LE
BW
ET
SDS
riboflavin [M+H]+
CCS (Å2)
DFT Energy (kcal/mol)
SDGRG [M+Na]+
CCS (Å2)
DFT Energy (kcal/mol)

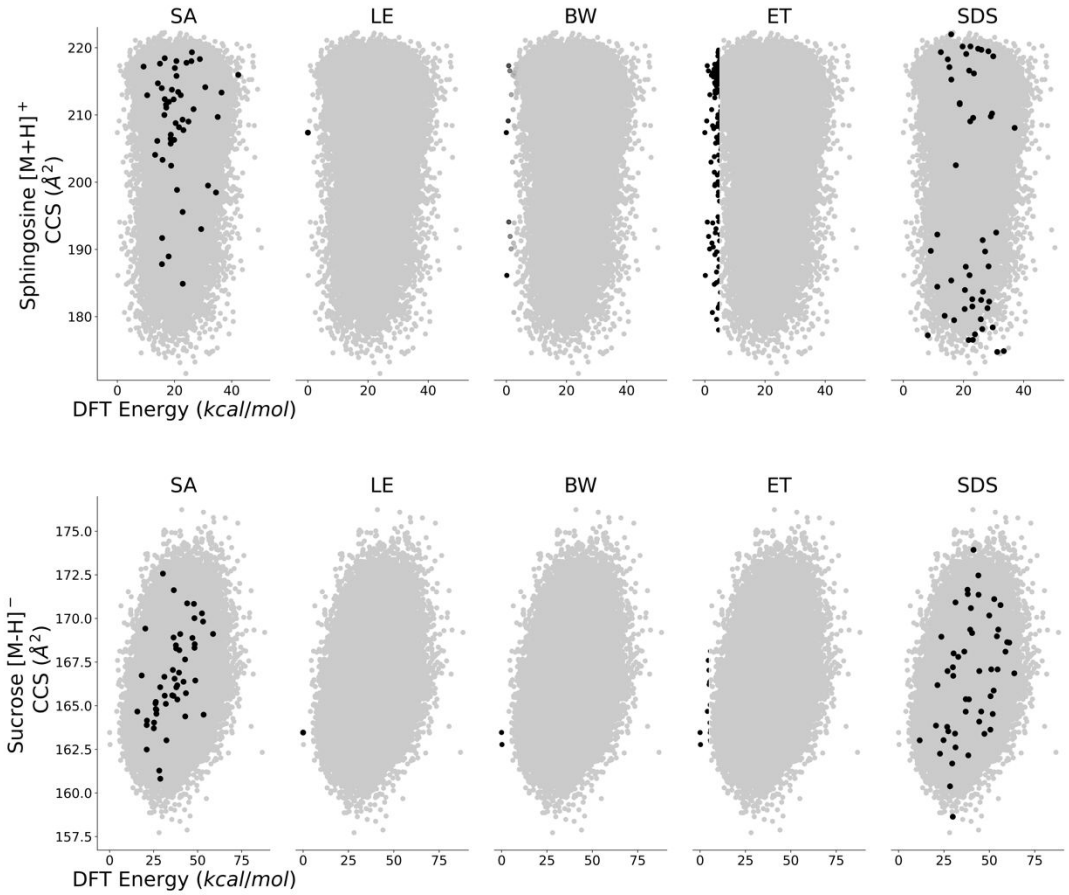


Fig. S11 Diagram of conformer selection and down selection methods for all molecules in the test set. Simple average (SA), lowest energy (LE), Boltzmann weighting (BW), energy threshold (ET), and similarity down selection (SDS). SA shows 50 randomly selected conformers, BW is shaded based off real weighted values, ET is a 5 kcal/mol threshold, and SDS shows the one most similar and 49 most dissimilar conformers.

8. Using MD vs DFT energy on MD structures

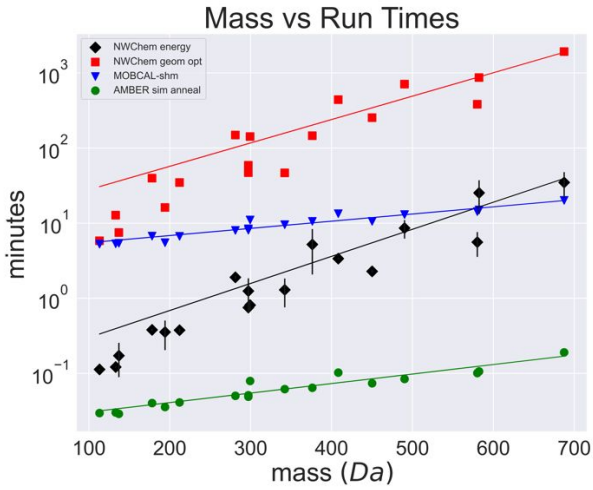


Fig. S12 Comparison of run times for NWChem energy optimization on MD structures (node minutes), NWChem geometry optimization (node minutes), MOBCAL-shm (node minutes), and AMBER simulated annealing (averaged per conformer, wall minutes).

Table S6 Average and range of runtimes for NWChem (node), MOBCAL-shm (node), and AMBER (wall).

	NWChem energy	NWChem geom opt	MOBCAL-shm	AMBER sim. anneal.
mean	3.61	209.73	9.47	0.06
stdev	6.07	254.35	3.15	0.03
max	25.29	868.93	15.03	0.11
min	0.11	5.82	5.27	0.03

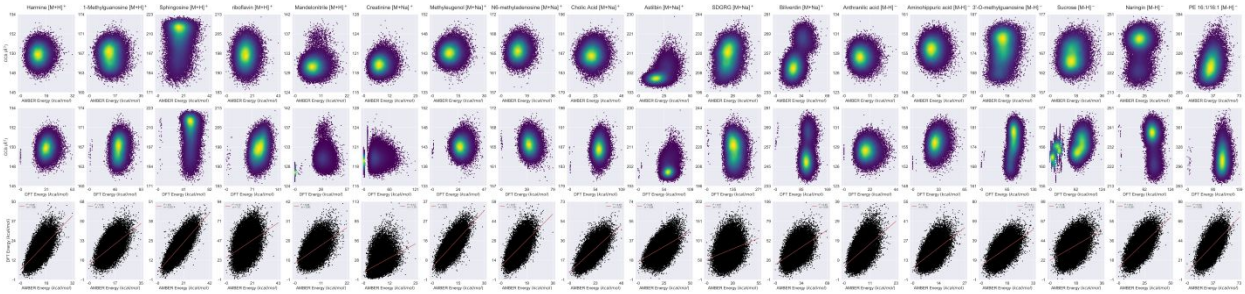


Fig. S13 Comparing CCS vs energy space for MD and DFT calculations. Top row: AMBER conformers with MD energies relative to the minimum energy. Middle row: DFT geometry optimized and non-optimized AMBER conformers with DFT energies, relative to the minimum energy of both sets. There are 30 DFT geometry optimized conformers for all molecules except for mandelonitrile and creatinine which have about 50k and sucrose which has about 25k. Bottom row: DFT vs MD energy correlations.

9. Limitations of the study

- The RMSD for SDS and geometry variability were calculated by excluding the non-backbone hydrogens. However, it appears even rotating a methyl group can lead to significantly different CCS calculations, and this appears to contribute to the CCS range of DFT geometry optimized conformers.
- Duplicate or nearly identical conformers generated by chance may give those conformers more weight than they should when Boltzmann weighting. This is a general problem for conformer

generation tools like AMBER that don't screen for duplicate conformers. Tools like CREST check for this duplicity.

- The stereochemistry of generated conformers was spot checked. Differences in stereochemistry can lead to differences in CCS and energy. We recommend using automated checking software, such as Bond Locator Utilizing Electronic Structure (BLUES, https://github.com/quantum2classical/blues) in conjunction with smiles canonicalizers.

10. Conformer selection analyses with AMBER energies

The following are example figures with the original AMBER energies instead of DFT energies, which were calculated on the AMBER structures. DFT energies were found to have better correlation with the conformers after DFT geometry optimization.

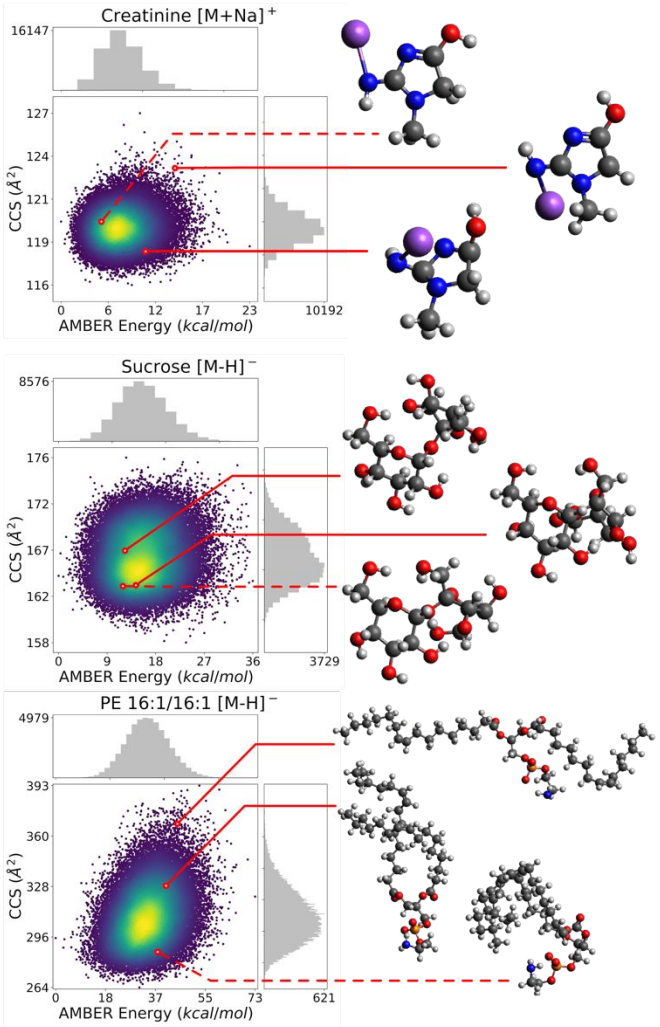


Fig. S14 CCS vs energy landscapes for 50k AMBER generated conformers for creatinine $[M+Na]^+$, sucrose $[M-H]^-$, PE 16:1/16:1 $[M-H]^-$ respectively. Highlighted are the most similar (dashed) and two most dissimilar (solid) conformers chosen heuristically with a structural RMSD metric.

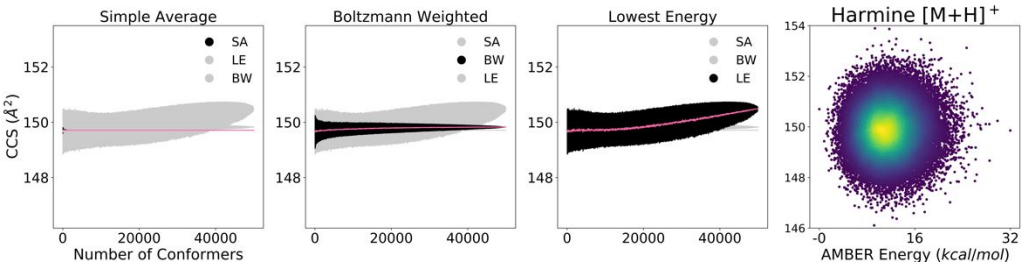


Fig. S15 MC convergence plots on CCS for harmine $[M+H]^+$.

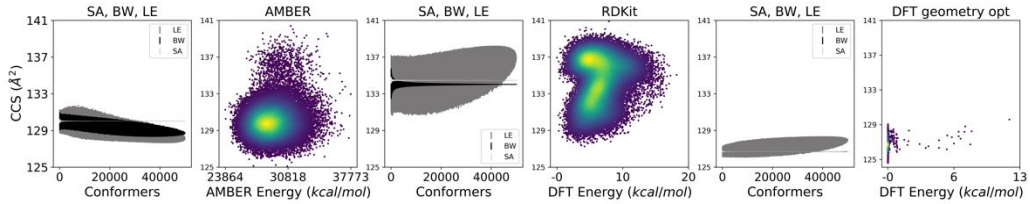


Fig. S16 MC convergence plots on CCS using three sampling techniques (SA, BW, LE) for conformers generated in AMBER, RDKit, and the AMBER conformers after a DFT geometry optimization for mandelonitrile $[M+H]^+$.

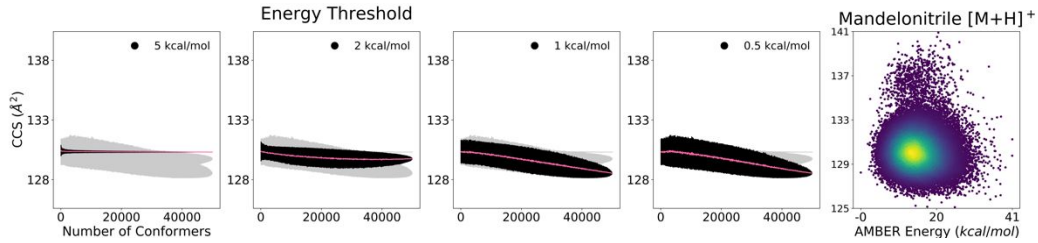


Fig. S17 MC convergence plots on CCS for mandelonitrile $[M+H]^+$ for 5, 2, 1, and 0.5 kcal/mol energy thresholds.

CITATIONS

(1) 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.
(2) Amos, R. I. J.; Haddad, P. R.; Szucs, R.; Dolan, J. W.; Pohl, C. A. Molecular modeling and prediction accuracy in Quantitative Structure-Retention Relationship calculations for chromatography; *TrAC Trends in Analytical Chemistry* **2018**, *105*, 352-359.
(3) Djoumbou-Feunang, Y.; Pon, A.; Karu, N.; Zheng, J.; Li, C.; Arndt, D.; Gautam, M.; Allen, F.; Wishart, D. S. CFM-ID 3.0: Significantly Improved ESI-MS/MS Prediction and Compound Identification; *Metabolites* **2019**, *9*.
(4) Ruttkies, C.; Schymanski, E. L.; Wolf, S.; Hollender, J.; Neumann, S. MetFrag relaunched: incorporating strategies beyond in silico fragmentation; *J Cheminform* **2016**, *8*, 3.
(5) Heinonen, M.; Rantanen, A.; Mielikainen, T.; Kokkonen, J.; Kiuru, J.; Ketola, R. A.; 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.
(6) Apra, E.; Bylaska, E. J.; de Jong, W. A.; Govind, N.; Kowalski, K.; Straatsma, T. P.; Valiev, M.; van Dam, H. J. J.; Alexeev, Y.; Anchell, J.; Anisimov, V.; Aquino, F. W.; Atta-Fynn, R.; Autschbach, J.; Bauman, N. P.; Becca, J. C.; Bernholdt, D. E.; Bhaskaran-Nair, K.; Bogatko, S.; Borowski, P., et al. NWChem: Past, present, and future; *J Chem Phys* **2020**, *152*, 184102.
(7) 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.
(8) 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.
(9) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F., et al.: Wallingford, CT, 2016.
(10) Ewing, S. A.; Donor, M. T.; Wilson, J. W.; Prell, J. S. Collidoscope: An Improved Tool for Computing Collisional Cross-Sections with the Trajectory Method; *J Am Soc Mass Spectrom* **2017**, *28*, 587-596.
(11) Mesleh, M. F.; Hunter, J. M.; Shvartsburg, A. A.; Schatz, G. C.; Jarrold, M. F. Structural Information from Ion Mobility Measurements: Effects of the Long-Range Potential; *The Journal of Physical Chemistry* **1996**, *100*, 16082-16086.
(12) Shvartsburg, A. A.; Jarrold, M. F. An exact hard-spheres scattering model for the mobilities of polyatomic ions; *Chemical Physics Letters* **1996**, *261*, 86-91.
(13) Larriba-Andaluz, C.; Hogan, C. J., Jr. Collision cross section calculations for polyatomic ions considering rotating diatomic/linear gas molecules; *J Chem Phys* **2014**, *141*, 194107.
(14) Colby, S. M.; Nunez, J. R.; Hodas, N. O.; Corley, C. D.; Renslow, R. R. 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.
(15) Zhou, Z.; Shen, X.; Tu, J.; Zhu, Z. J. Large-Scale Prediction of Collision Cross-Section Values for Metabolites in Ion Mobility-Mass Spectrometry; *Anal Chem* **2016**, *88*, 11084-11091.

(16) Yesiltepe, Y.; Nunez, J. R.; Colby, S. M.; Thomas, D. G.; Borkum, M. I.; Reardon, P. N.; Washton, N. M.; Metz, T. O.; Teeguarden, J. G.; Govind, N.; Renslow, R. S. An automated framework for NMR chemical shift calculations of small organic molecules; *J Cheminform* **2018**, *10*, 52.
(17) Colby, S. M.; Thomas, D. G.; Nunez, J. R.; Baxter, D. J.; Glaesemann, K. R.; Brown, J. M.; Pirrung, M. A.; Govind, N.; Teeguarden, J. G.; Metz, T. O.; Renslow, R. S. ISiCLE: A Quantum Chemistry Pipeline for Establishing in Silico Collision Cross Section Libraries; *Anal Chem* **2019**, *91*, 4346-4356.
(18) McNaught, A. D.; Wilkinson, A. *IUPAC. Compendium of Chemical Terminology (the "Gold Book").* 2nd ed.; Blackwell Scientific Publications: Oxford, 1997, Online version (2019-) created by S. J. Chalk.
(19) Zheng, X.; Aly, N. A.; Zhou, Y.; Dupuis, K. T.; Bilbao, A.; Paurus, V. L.; Orton, D. J.; Wilson, R.; Payne, S. H.; Smith, R. D.; Baker, E. S. A structural examination and collision cross section database for over 500 metabolites and xenobiotics using drift tube ion mobility spectrometry; *Chem Sci* **2017**, *8*, 7724-7736.
(20) Ebejer, J. P.; Morris, G. M.; Deane, C. M. Freely available conformer generation methods: how good are they?; *J Chem Inf Model* **2012**, *52*, 1146-1158.