Supporting Information for: High-Throughput Measurement and Machine Learning-Based Prediction of Collision Cross Sections for Drugs and Drug Metabolites

Dylan H. Ross[#,1], Ryan P. Seguin[#], Allison M. Krinsky[#], and Libin Xu[#, *]

[#], Department of Medicinal Chemistry, University of Washington, Seattle, WA, USA, 98195.

[1], current address: Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA, USA, 99352.

Corresponding Author: *Libin Xu, Ph.D., Email: libinxu@uw.edu. Tel: (206) 543-1080.

Contents

1. Supplementary Experimental Section and Results

1.1 – Partial Least-Squares Regression Analysis on dmCCS

Figure S1 shows the results from partial least-squares regression analyses (PLS-RA) computed on dmCCS using 2D, 3D and combined feature sets with CCS as the target variable. Figures S1A-C show the PLS-RA projections of the dmCCS database computed using 2D, 3D, and combined molecular descriptors, respectively. When compared against the corresponding PCA projections (Figures 2E-G in the main text), the overall distributions are similar. Further, the x-loadings from the PLS-RAs (Figures S1D-E) display nearly identical rank-ordering relative to the PC1 feature loadings (Figures 2H-J in the main text). Taken together, these results show that a targeted analysis of dmCCS reproduces the same basic conclusions as those garnered from PCA for all feature sets tested, which is to be expected given the high degree of alignment between CCS and PC1 observed in the PCAs.

1.2 – Feature Selection for CCS Prediction

Starting from a complete combined feature set (2D + 3D molecular descriptors, 50 features total), a set of tests were per-formed (using only the training set data) to determine the minimal feature set necessary to make robust and accurate CCS predictions. First, the relative importance of all individual features was determined by three methods: PLS-RA, gradient boosting regression (GBR, sklearn.ensemble.GradientBoostingRegressor), and a permutation feature importance function built into Scikit-Learn (PER, sklearn.inspection.permutation_importance). PLS-RA gives an indication of feature importance based on the magnitude of the loadings in the x-dimension (i.e., the multidimensional axis that explains the maximal variance in the target variable). GBR is an ensemble method in which successive decision tree models are fitted to the residuals of previous models, and relative feature importance can be inferred from the frequency with which individual features are used for decision tree splits. In the PER method, feature importance is related to the decrease in prediction performance when a feature is randomly shuffled relative to a baseline (unshuffled) performance. Once feature importance had been calculated, sequential feature removal tests were performed using the importance from each method. In the feature removal tests, the least important features were successively removed, and new predictive models were trained and evaluated on the smaller feature sets. This process was repeated until only a single feature (with the highest importance) remained (Figure S2A-C). For each method, a reduced feature set was selected as the set of features for which the prediction error (RMSE) increased above 5 Å^2 upon their removal. Finally, a minimal feature set was selected as those common among 2 sets of features remaining after feature removal tests using the PLS-RA, GBR, and PER feature importance (Figure S2D).

1.3 – Comparison of CCS Prediction Model to Theory-Based Conventional Methods

Computational modeling to produce 3D structures at a low theory level is the primary bottleneck in training and application of ML models for CCS prediction based on 3D molecular descriptors. Given that production of such structures is also a bottleneck for some of the faster theory-based CCS prediction methods (*e.g.* projection approximation, PA, and exact hard-sphere scattering, EHS),[1] we sought to compare the accuracy of CCS values predicted using both approaches for compounds in dmCCS. Figure S4A shows measured and calculated CCS for compounds from dmCCS, colored according to calculation method. The ML values were predicted using the model trained on the MIN feature set described in the previous section. Both of the theory-driven methods (PA and EHS) display significant systematic errors; however, these systematic errors are likely attributable to the parameterization of these methods, which were originally optimized for He as the drift gas. When systematic errors were corrected using linear regression, the residuals of the fit for PA or EHS-generated values were significantly larger than the ML-predicted values (Figure S4). Taken together, it is clear that ML-based CCS prediction produces higher quality CCS values with this dataset than comparable theory-based methods, likely attributable to the nuanced structural trends that such ML model can capture when provided with appropriate training data.

1.4 – Generation of 2-Dimensional Molecular Descriptors

Molecular quantum numbers[2] (MQNs) were used as 2D molecular descriptors for analysis of the drug and metabolite CCS database. MQNs are graph properties of a 2D molecular structure (e.g. a SMILES structure), which include counts of atoms, bonds, and topological features. MQNs were computed from the neutral SMILES structures for all entries in the drug and metabolite CCS database using the RDKit library (https://www.rdkit.org). The computed MQNs were added as a separate table to the drug and metabolite CCS database.

1.5 – Generation of 3-Dimensional Molecular Descriptors

Principal moments of inertia (PMI) and binned radial mass distributions (RMD) were used as molecular descriptors for 3D molecular structures. PMI are derived from the eigendecomposition of the inertia tensor of a rigid body computed relative to its center of mass. Physically, this computation produces a set of orthogonal axes within a body, such that the radial distribution of mass about each successive axis is minimized; the magnitude of the PMIs reflects the extent of radial mass distribution about their corresponding axes (Figure S6B). Given a 3D molecular structure defined by N atoms having masses (m) and positions (x, y, z) with center of mass located at the origin, the body frame inertia tensor (I) was computed as follows:

$$I = \begin{bmatrix} I_{xx} & I_{xy} & I_{xz} \\ I_{yx} & I_{yy} & I_{yz} \\ I_{zx} & I_{zy} & I_{zz} \end{bmatrix}$$

where the diagonal elements (I_{xx}, I_{yy}, I_{zz}) were computed as:

$$I_{xx} = \sum_{i=1}^{N} m_i(y_i^2 + z_i^2)$$

$$I_{yy} = \sum_{i=1}^{N} m_i(x_i^2 + z_i^2)$$

$$I_{zz} = \sum_{i=1}^{N} m_i(x_i^2 + y_i^2)$$

and the off-diagonal elements (I_{xy}, I_{yx}, I_{xz}, I_{zx}, I_{yz}, I_{zy}) were computed as:

$$I_{xy} = I_{yx} = \sum_{i=1}^{N} m_i x_i y_i$$

$$I_{xz} = I_{zx} = \sum_{i=1}^{N} m_i x_i z_i$$

$$I_{yz} = I_{zy} = \sum_{i=1}^{N} m_i y_i z_i$$

An eigendecomposition (as implemented in the SciPy Python library: *scipy.linalg.eigh*) was then performed on the inertia tensor, yielding the principal moments of inertia (PMI_1, PMI_2, PMI_3):

$$I = Q\Lambda Q^T$$

$$\Lambda = \begin{bmatrix} PMI_1 & 0 & 0 \\ 0 & PMI_2 & 0 \\ 0 & 0 & PMI_3 \end{bmatrix}$$

RMDs reflect the proportions of a structure's mass that lie within specific distances radially from its center of mass. Specifically, RMDs are normalized, mass-weighted histograms of atomic distances relative to the center of mass. The histograms were binned at specific distance intervals (0-2 Å, 2-4 Å, 4-6 Å, 6-8 Å, and >8 Å) in order to reduce the total number of features. The binning intervals were chosen based on the combined distribution of mass-weighted radial distances from all 3D structures in the drug and metabolite CCS database (Figure S6C). The computed 3D molecular descriptors were added as a separate table to the drug and metabolite CCS database.

1.6 – Analysis of Mass and CCS Shifts for Metabolites

Mass and CCS shifts relative to parent compounds were computed for all metabolites in the database. Figure S9A shows the distribution of mass shifts for all metabolites, with annotations reflecting the corresponding metabolic modifications. Figures S9B and S9C are arrow plots for specific metabolic modifications with increased or decreased mass relative to the parent compound, respectively, which demonstrate the absolute and relative *m/z* and CCS shifts for these modifications. For Phase-II metabolites +GSH and +Glc, the effect of metabolism on the structure of the parent is mostly consistent across all compounds with a large increase in *m/z* and a corresponding increase in CCS. Dealkylation reactions (-Me, -2Me/-Et) similarly display consistent decreases in CCS, with the exception of a few -Me metabolites. For the oxygenated metabolites (+O, +2O), the structural effect of metabolism is more complex with some modifications resulting in an expected increase in CCS while others leading to a decrease. The desaturated metabolites (-2H) similarly display somewhat complex structural characteristics, leading to increase or decrease in CCS depending on the parent compound, although these effects are much smaller in magnitude than most of the other metabolic modifications.

2. Supplementary Figures and Tables

2.1 – Figure S1: Partial Least-Squares Regression Analysis on dmCCS

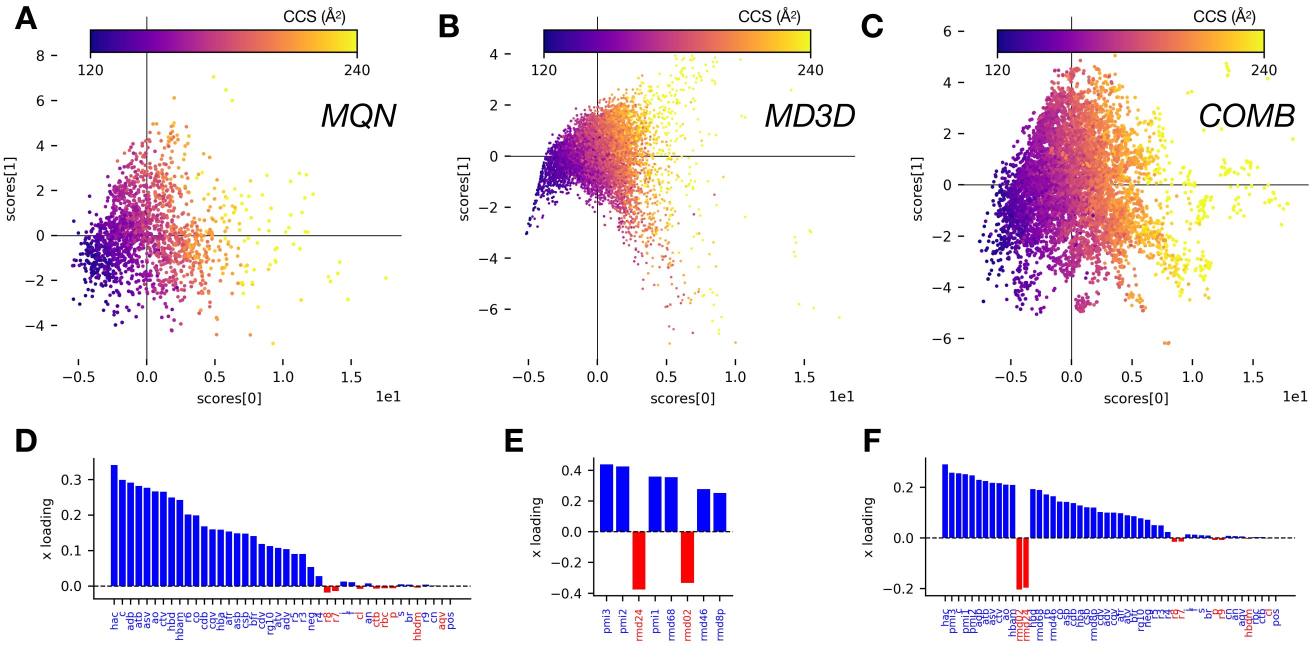
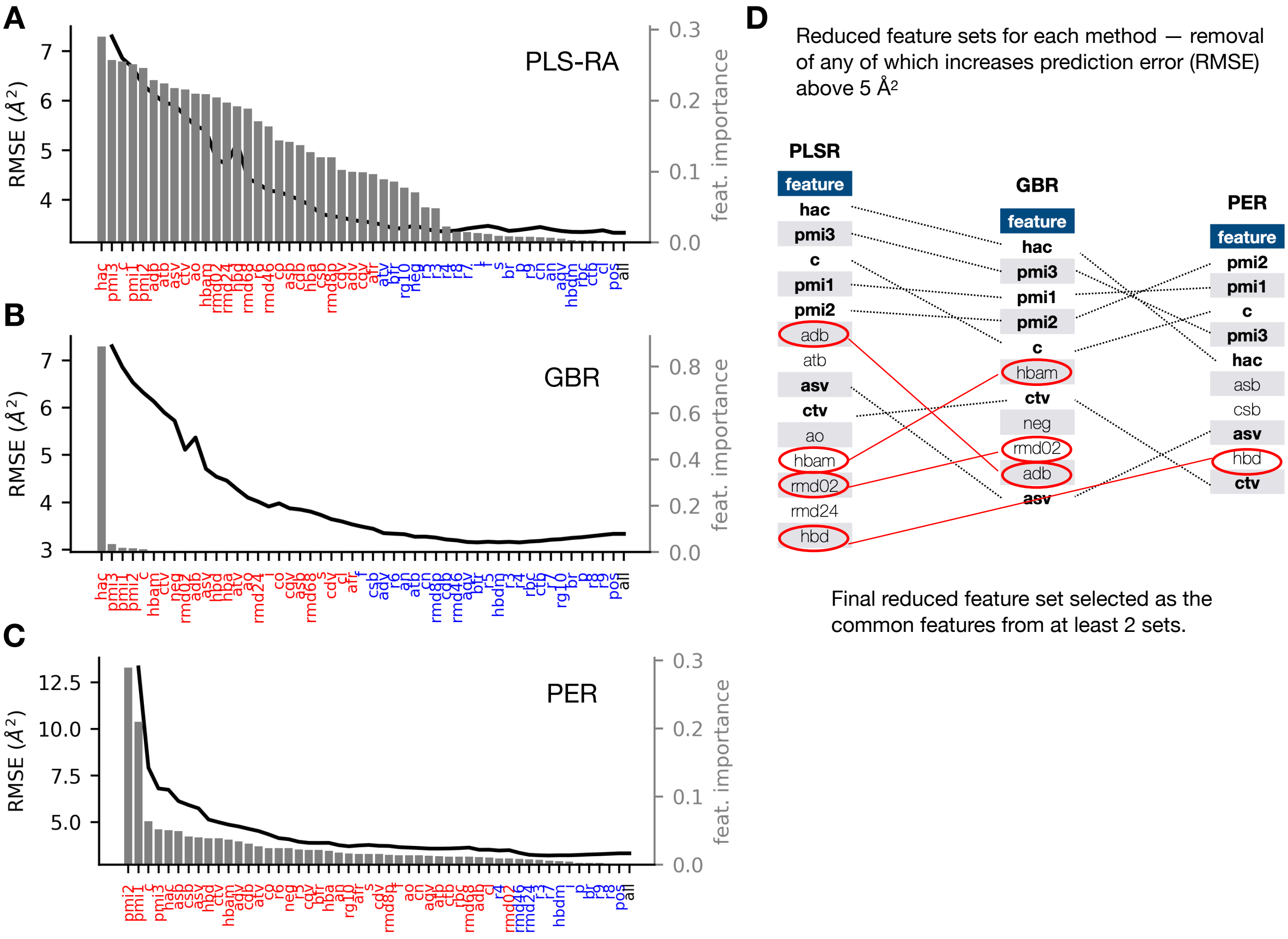


Figure S1. (**A**) PLS-RA projections of dmCCS database computed using MQNs as molecular descriptors and CCS as the target variable, colored by CCS. (**B**) PLS-RA projections of dmCCS database computed using MD3Ds as molecular descriptors and CCS as the target variable, colored by CCS. (**C**) PLS-RA projections of dmCCS database computed using the combination of MQNs and MD3Ds as molecular descriptors and CCS as the target variable, colored by CCS. (**D**) Individual feature loadings for component 1 from PLS-RA computed on dmCCS using MQNs as molecular descriptors and CCS as the target variable. (**E**) Individual feature loadings for component 1 from PLS-RA computed on dmCCS using MD3Ds as molecular descriptors and CCS as the target variable. (**F**) Individual feature loadings for component 1 from PLS-RA computed on dmCCS using the combination of MQNs and MD3Ds as molecular descriptors and CCS as the target variable.

2.2 – *Figure S2: Feature Selection Trials*



Reduced feature sets for each method — removal of any of which increases prediction error (RMSE) above 5 Å^2

Final reduced feature set selected as the common features from at least 2 sets.

Figure S2. (**A**-**C**) Results from feature selection trials. Features were removed in descending order of feature importance (from right to left, grey bars), and resulting predictive model performance was recorded (RMSE, black line). Blue labels indicate the features that could be removed without model performance increasing RMSE more than 5% relative to the baseline (*all*). (**D**) Selected features from individual trials, selected as those for which removal increased error above 5 Å^2. The features selected in at least two of the individual tests were retained as the final minimal feature set.

2.3 – Figure S3: Minimal Feature Set Correlation Matrix

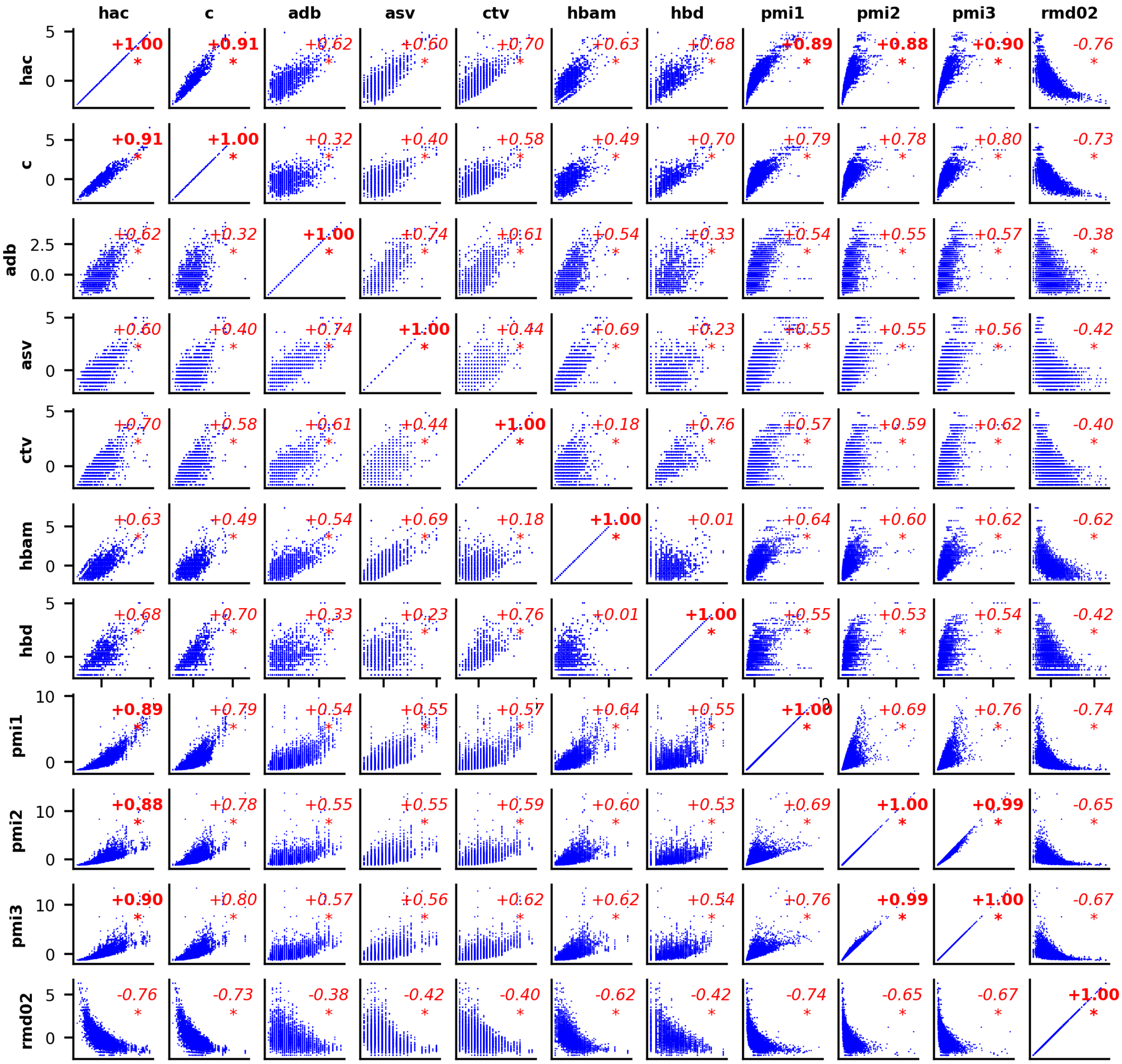


Figure S3. Correlation matrix of minimal feature set from feature selection trials. Red numbers correspond to spearman rank test correlation coefficients (coefficients with magnitude > 0.85 are in bold). Asterisks denote a p-value < 0.01 for the correlation.

2.4 – Figure S4: Comparison of measured CCS and CCS predicted using PA/EHS methods

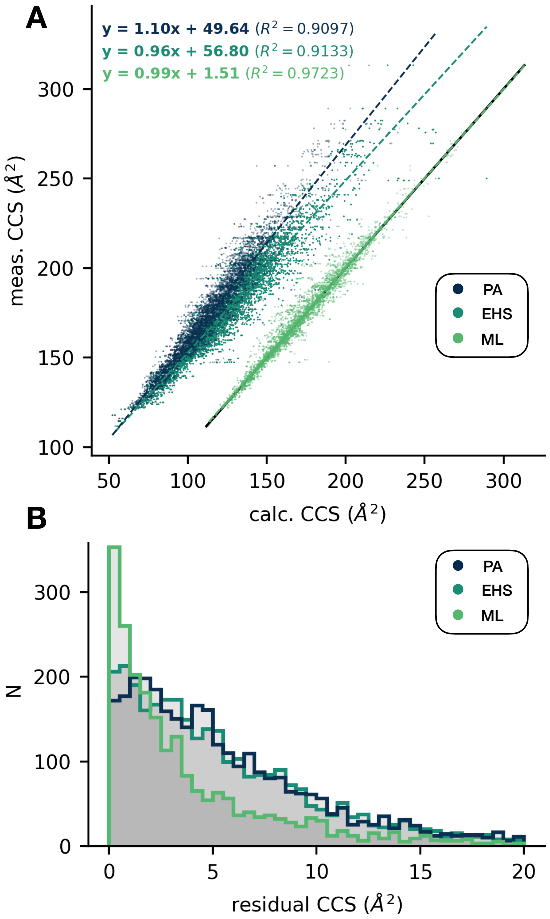


Figure S4. (**A**) Comparison of measured CCS and CCS predicted using PA/EHS methods or by a ML model trained on the dmCCS database. Dotted lines correspond to linear fits on each set of values. (**B**) Distributions of residuals from each linear fit.

2.5 – Figure S5: MetFrag Fragmenter Score Cutoff

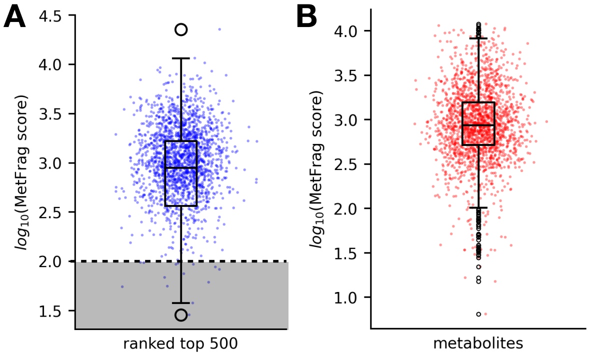
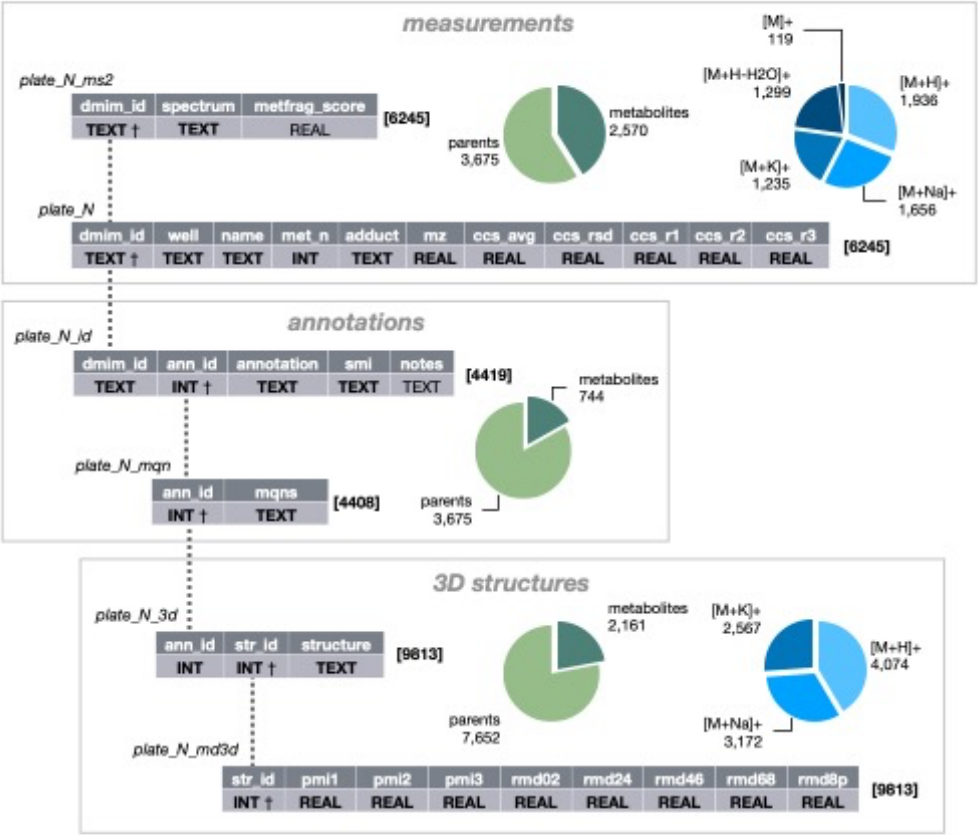


Figure S5. (**A**) Distribution of log-transformed MetFrag fragmenter scores for all parent compounds with true annotations ranked in the top 500 from the parent rank test. The dashed line indicates the empirically determined cutoff used to filter out metabolite annotations during construction of the dmCCS database. (**B**) Distribution of log-transformed MetFrag fragmenter scores for metabolites in dmCCS prior to filtering. The center line is the median, the box edges are the upper/lower quartiles (*i.e.*, Q1 and Q3), the whiskers are 1.5x the interquartile range, and the points are outliers beyond the whiskers.

2.6 – Figure S6: dmCCS Database Architecture and Composition



plate_N_ms2

dmim_id	spectrum	metfrag_score
TEXT †	TEXT	REAL

[6245]

plate_N

dmim_id	well	name	met_n	adduct	mz	ccs_avg	ccs_rsd	ccs_r1	ccs_r2	ccs_r3
TEXT †	TEXT	TEXT	INT	TEXT	REAL	REAL	REAL	REAL	REAL	REAL

[6245]

plate_N_id

dmim_id	ann_id	annotation	smi	notes
TEXT	INT †	TEXT	TEXT	TEXT

[4419]

plate_N_mqn

ann_id	mqns
INT †	TEXT

[4408]

plate_N_3d

ann_id	str_id	structure
INT	INT †	TEXT

[9813]

plate_N_md3d

str_id	pmi1	pmi2	pmi3	rmd02	rmd24	rmd46	rmd68	rmd8p
INT †	REAL	REAL	REAL	REAL	REAL	REAL	REAL	REAL

[9813]

Figure S6. Overview of the structure of the dmCCS SQLite3 database. Each grey box represents the general type of information contained within each table, and pie charts reflect characteristics of these grouped tables.

The names and data types are shown for each table, with bold datatypes indicating a required column and † indicating the primary key of the table. The dashed lines indicate the related columns between each table.

2.7 – Figure S7: Description of 3D Molecular Descriptors

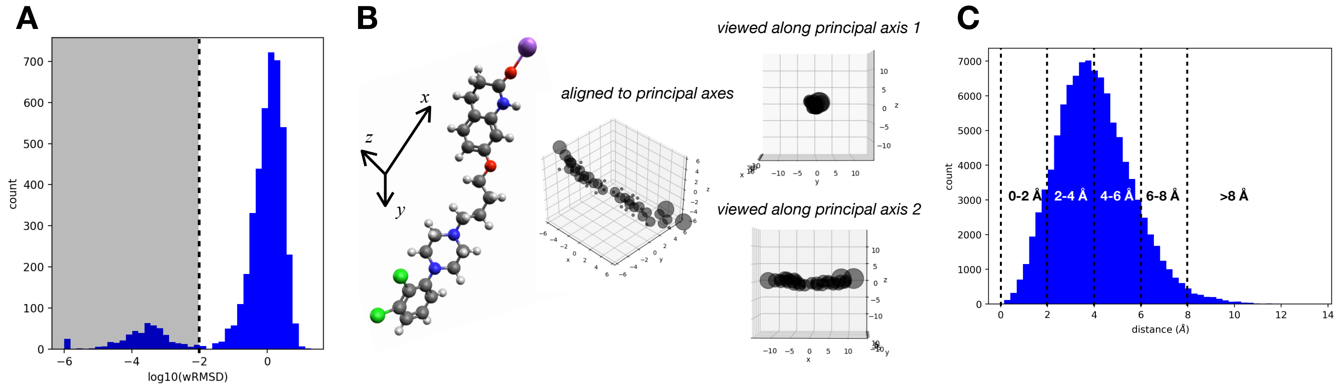


Figure S7. (**A**) Distribution of log-transformed mass-weighted RMSD for all pairwise combinations of multiple 3D structures for all compounds in the dmCCS database. The dashed line indicates an empirically determined cutoff used for determination of whether individual 3D structures are distinct enough to be kept when assembling the final database. (**B**) Demonstration of the physical interpretation of principal axes in a 3D molecular structure. The principal axes x, y, and z are defined such that they each minimize the radial distribution of mass about successive orthogonal axes. The center image is a representation of the atomic positions from the structure on the left, with radii proportional to atomic masses. In this example, when viewed along the first principal axis (x, top right), there is very little radial distribution of masses about the central axis. In contrast, when viewed along the second principal axis (y, bottom right) the radial distribution of masses is in greater. The PMI are related to the magnitude of radial mass distribution about the respective principal axes, where increased radial mass distribution results in a higher moment. (**C**) Mass-weighted radial atomic distance distribution for all 3D structures in the dmCCS database. Dashed lines indicate binning intervals used to compute binned radial mass distributions for individual structures as part of the MD3D features.

2.8 – Figure S8: AllCCS and CCSbase Predicted CCS for Compounds With Multimodal ATDs

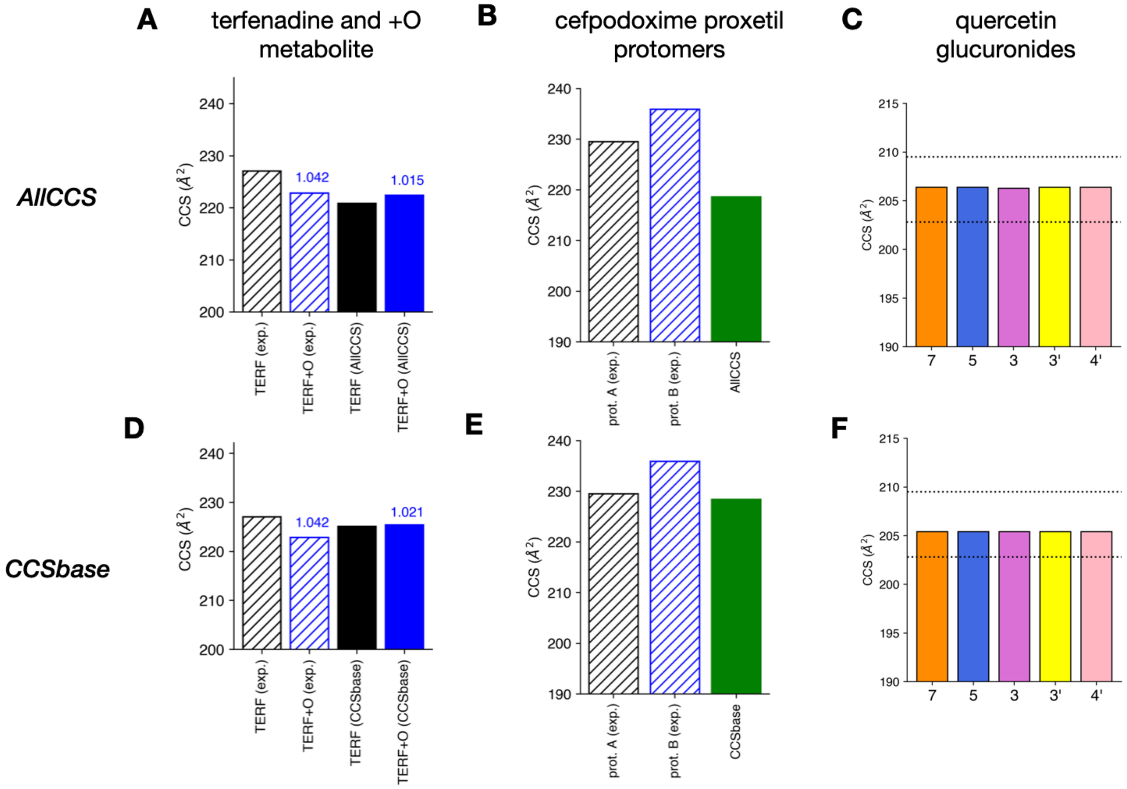


Figure S8. (**A**) Comparison of measured (hatched) and AllCCS predicted (solid) CCS for terfenadine and its +O metabolite. (**B**) Comparison of measured (hatched) and AllCCS predicted (solid) CCS for two protomers of cefpodoxime proxetil. (**C**) Comparison of measured (dashed lines) and AllCCS predicted (solid bars) CCS for the positional isomers of quercetin glucuronide. (**D**) Comparison of measured (hatched) and CCSbase predicted (solid) CCS for terfenadine and its +O metabolite. (**E**) Comparison of measured (hatched) and CCSbase predicted (solid) CCS for two protomers of cefpodoxime proxetil. (**F**) Comparison of measured (dashed lines) and CCSbase predicted (solid bars) CCS for the positional isomers of quercetin glucuronide.

2.9 – Table S1: Molecular Quantum Numbers (MQNs)

MQN	description	MQN	description
c	carbon atom count	hbdm	H-bond donor sites
f	fluorine atom count	hdb	H-bond donor atoms
cl	chlorine atom count	negc	negative charges
br	bromine atom count	posc	positive charges
i	iodine atom count	asv	acyclic monovalent nodes
s	sulfur atom count	adv	acyclic divalent nodes
p	phosphorus atom count	atv	acyclic trivalent nodes
an	acyclic nitrogen atom count	aqv	acyclic tetravalent nodes
cn	cyclic nitrogen atom count	cdv	cyclic divalent nodes
ao	acyclic oxygen atom count	ctv	cyclic trivalent nodes
co	cyclic oxygen atom count	cqv	cyclic tetravalent nodes
hac	heavy (non-hydrogen) atom count	r3	3-membered ring count
asb	acyclic single bonds	r4	4-membered ring count
adb	acyclic double bonds	r5	5-membered ring count
atb	acyclic triple bonds	r6	6-membered ring count
csb	cyclic single bonds	r7	7-membered ring count
cdb	cyclic double bonds	r8	8-membered ring count
ctb	cyclic triple bonds	r9	9-membered ring count
rbc	rotatable bond count	rg10	≥10-membered ring count
hbam	H-bond acceptor sites	afrc	nodes shared by ≥2 rings
hba	H-bond acceptor atoms	bfrc	edges shared by ≥2 rings

2.10 – Table S2: 3D Molecular Descriptors (MD3D)

MD3D	**description**
pmi1	first principal moment of inertia
pmi2	second principal moment of inertia
pmi3	third principal moment of inertia
rmd02	proportion of mass between 0 and 2 Å of center of mass
rmd24	proportion of mass between 2 and 4 Å of center of mass
rmd46	proportion of mass between 4 and 6 Å of center of mass
rmd68	proportion of mass between 6 and 8 Å of center of mass
rmd8p	proportion of mass more than 8 Å from center of mass

2.11 – Figure S9: Analysis of Mass and CCS Shifts for Metabolites

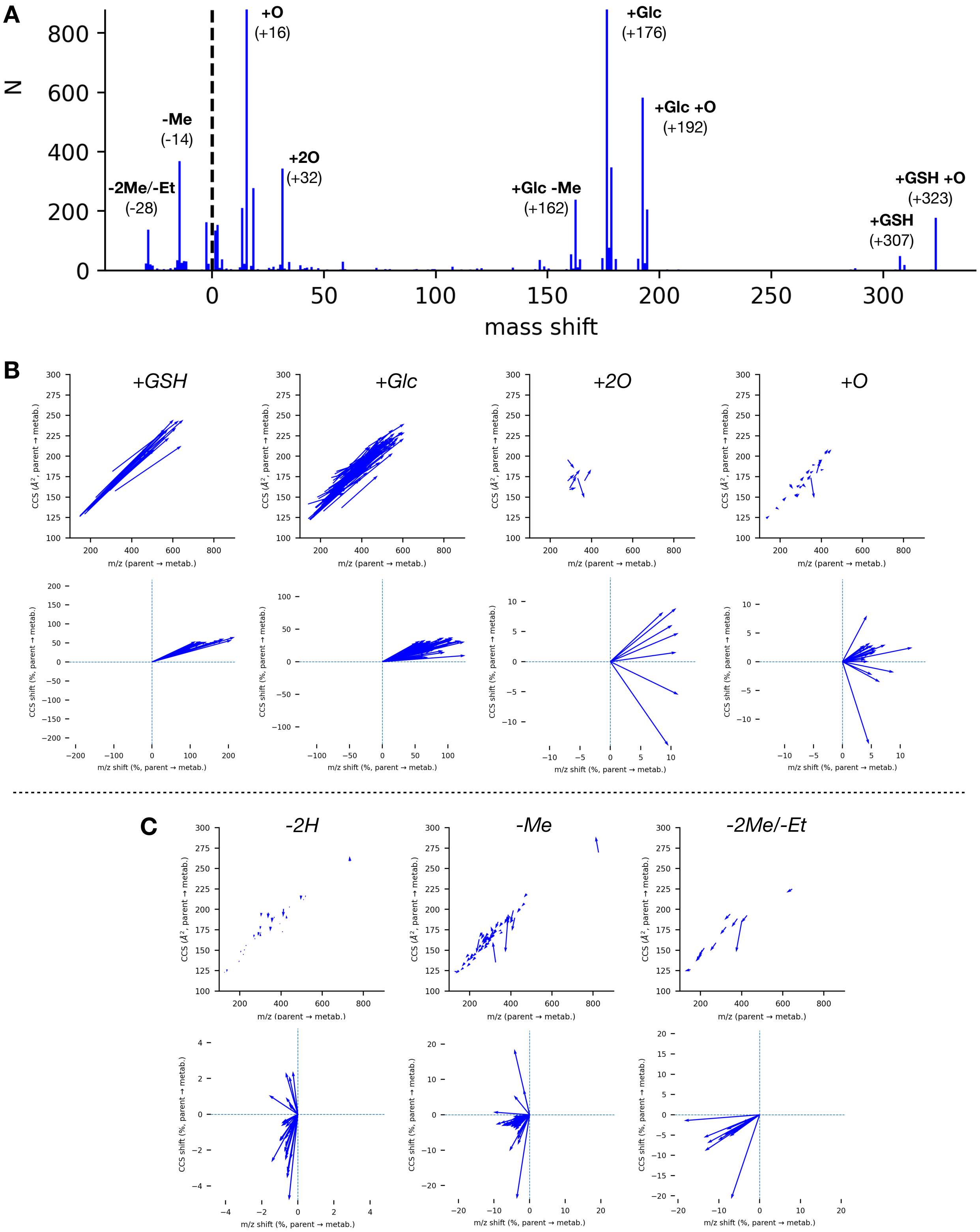
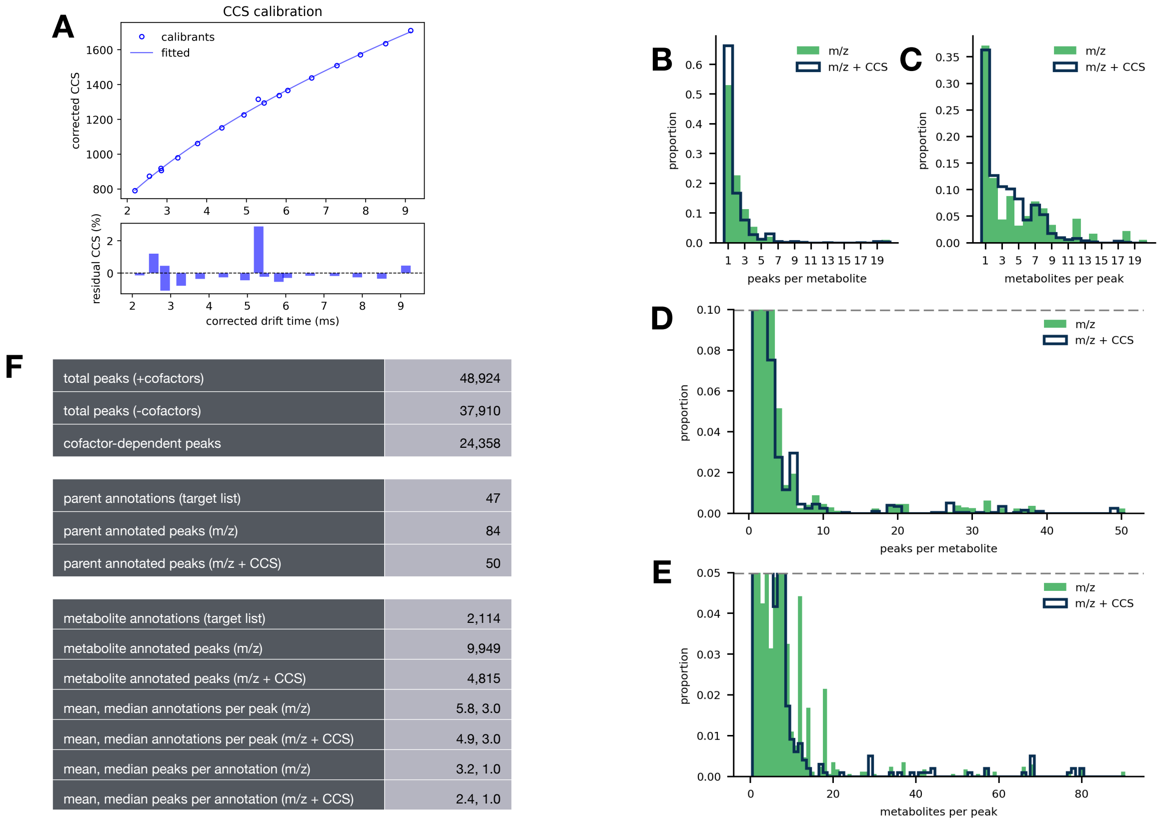


Figure S9. (**A**) Distribution of mass shifts for all metabolites, annotated with metabolic modifications. (**B**-**C**) Arrow plots showing absolute and relative *m/z* and CCS shifts for specific metabolic modifications with increased or decreased mass relative to the parent compound, respectively.

2.12 – Figure S10: LC-IM-MS Analysis of Pooled Drug Metabolism Incubations



total peaks (+cofactors)	48,924
total peaks (-cofactors)	37,910
cofactor-dependent peaks	24,358

parent annotations (target list)	47
parent annotated peaks (m/z)	84
parent annotated peaks (m/z + CCS)	50

metabolite annotations (target list)	2,114
metabolite annotated peaks (m/z)	9,949
metabolite annotated peaks (m/z + CCS)	4,815
mean, median annotations per peak (m/z)	5.8, 3.0
mean, median annotations per peak (m/z + CCS)	4.9, 3.0
mean, median peaks per annotation (m/z)	3.2, 1.0
mean, median peaks per annotation (m/z + CCS)	2.4, 1.0

Figure S10. (**A**) CCS calibration curve, calibrants: polyalanine and drug mixture. (**B**, **D**) Distribution of number of peaks per annotated metabolite, with matching based on *m/z* or *m/z* + CCS. (**C**, **E**) Distribution of number of metabolite annotations per peak, with matching based on *m/z* or *m/z* + CCS. (**F**) Summary of annotation results.

2.13 – Figure S11: Analysis of Isobaric Metabolites Predicted by BioTransformer

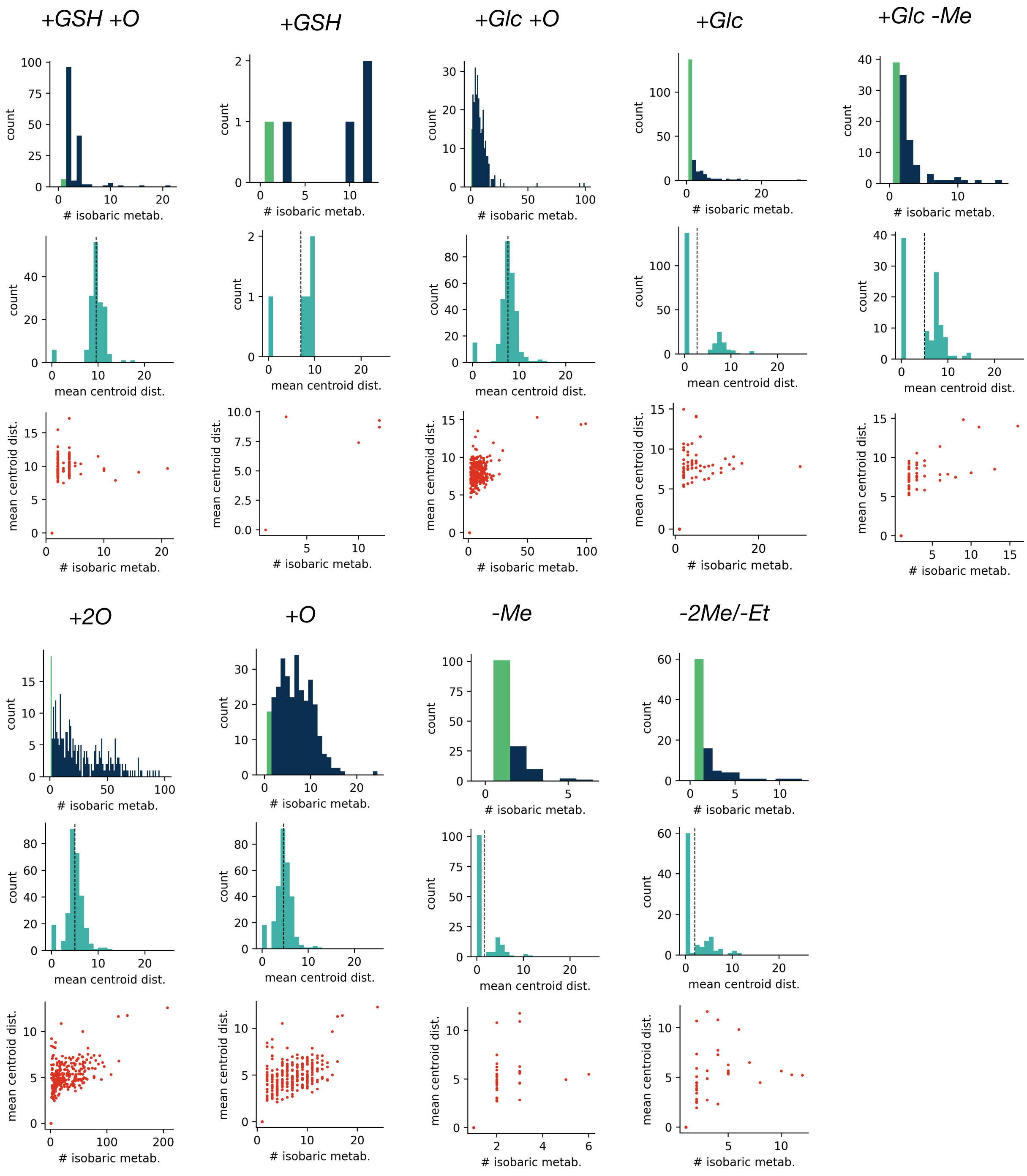


Figure S11. Distribution of isobaric metabolites predicted by BioTransoformer and corresponding dispersion of these metabolites in feature space (MQNs), grouped by metabolic modification (light green corresponds to metabolites with only one possible positional isomer). Correlation between isobaric metabolite count and dispersion in feature space are also presented. Metabolites were predicted for all parent compounds in the CCS database, without exclusion of isobaric predicted metabolites.

2.14 – Figure S12: Example Arrival Time Distribution and Extracted Ion Chromatogram from Automated Workflow

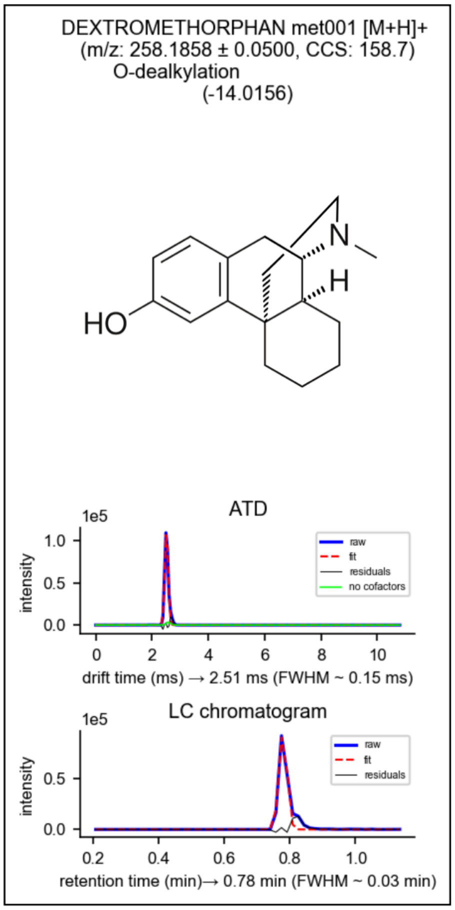


Figure S12. Example arrival time distribution and extracted ion chromatogram from automated data processing workflow.

2.15 – Figure S13: Performance of MIN Model on Normal and Reduced Metabolite Data Sets

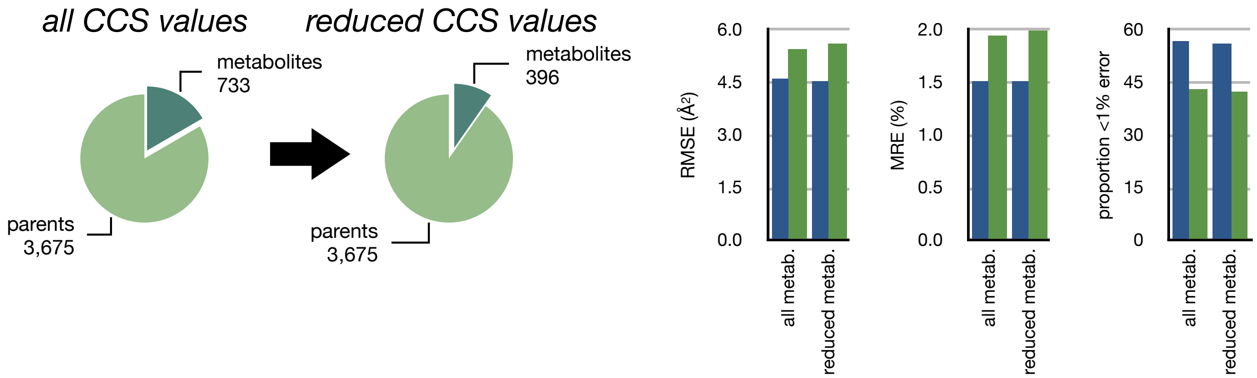


Figure S13. (Left) Composition of database before and after reducing the included metabolites to those with few potential isomers. (Right) Comparison of CCS prediction metrics of MIN model trained using normal and reduced metabolite data sets.

References

(1) Shvartsburg, A. A.; Jarrold, M. F. An exact hard-spheres scattering model for the mobilities of polyatomic ions, *Chem Phys Lett* **1996**, *261*, 86-91.
(2) Nguyen, K. T.; Blum, L. C.; van Deursen, R.; Reymond, J. L. Classification of organic molecules by molecular quantum numbers, *ChemMedChem* **2009**, *4*, 1803-1805.