Published in final edited form as:

Anal Chem. 2019 April 02; 91(7): 4346–4356. doi:10.1021/acs.analchem.8b04567.

ISiCLE: A Quantum Chemistry Pipeline for Establishing in Silico Collision Cross Section Libraries

Sean M. Colby[†], **Dennis G. Thomas**[†], **Jamie R. Nuñez**[†], **Douglas J. Baxter**[†], **Kurt R. Glaesemann**[‡], **Joseph M. Brown**[†], **Meg A. Pirrung**[§], **Niranjan Govind**[†], **Justin G. Teeguarden**[†,‖], **Thomas O. Metz**[*,†], and **Ryan S. Renslow**[*,†]

[†]Earth and Biological Sciences Directorate, Pacific Northwest National Laboratory, Richland, Washington 99352, United States

[‡]Communications and Information Technology Directorate, Pacific Northwest National Laboratory, Richland, Washington 99352, United States

[§]National Security Directorate, Pacific Northwest National Laboratory, Richland, Washington 99352, United States

[‖]Department of Environmental and Molecular Toxicology, Oregon State University, Corvallis, Oregon 97331, United States

Abstract

High-throughput, comprehensive, and confident identifications of metabolites and other chemicals in biological and environmental samples will revolutionize our understanding of the role these chemically diverse molecules play in biological systems. Despite recent technological advances, metabolomics studies still result in the detection of a disproportionate number of features that cannot be confidently assigned to a chemical structure. This inadequacy is driven by the single most significant limitation in metabolomics, the reliance on reference libraries constructed by analysis of authentic reference materials with limited commercial availability. To this end, we have developed the in silico chemical library engine (ISiCLE), a high-performance computing-friendly cheminformatics workflow for generating libraries of chemical properties. In the instantiation described here, we predict probable three-dimensional molecular conformers (i.e., conformational isomers) using chemical identifiers as input, from which collision cross sections (CCS) are derived. The approach employs first-principles simulation, distinguished by the use of molecular

[*]**Corresponding Author**: thomas.metz@pnnl.gov. [*]**Corresponding Author**: ryan.renslow@pnnl.gov.

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.8b04567.

Materials and methods, results, and discussion about ISiCLE methods; comparison of putative methods; computational burden of calculation steps; mass vs CCS outliers; MOBCAL-SHM benchmark; ISiCLE computational cost; environmental samples; and a comparison of theoretical CCS values (PDF)

Validation set molecules, measured CCS, associated CCS relative standard deviation, observed mass, IUPAC International Chemical Identifiers (InChI), SMILES string, formula, chemical name, source citation DOI, and chemical class information (XLSX)

ISiCLE source code and MOBCAL-SHM source code and binaries (ZIP, ZIP)

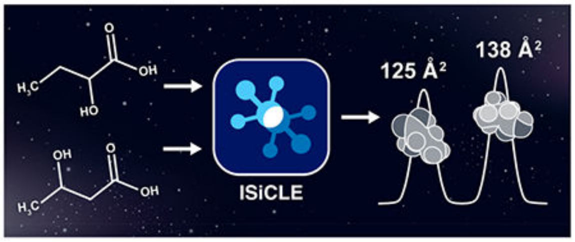
The authors declare no competing financial interest.

NOTE ADDED AFTER ASAP PUBLICATION

This paper was originally published ASAP on March 6, 2019. Due to a production error, the Supporting Information ZIP files were not properly uploaded. Minor text corrections were also made to the Article, and the corrected version was reposted on April 2, 2019.

dynamics, quantum chemistry, and ion mobility calculations, to generate structures and chemical property libraries, all without training data. Importantly, optimization of ISiCLE included a refactoring of the popular MOBCAL code for trajectory-based mobility calculations, improving its computational efficiency by over 2 orders of magnitude. Calculated CCS values were validated against 1983 experimentally measured CCS values and compared to previously reported CCS calculation approaches. Average calculated CCS error for the validation set is 3.2% using standard parameters, outperforming other density functional theory (DFT)-based methods and machine learning methods (e.g., MetCCS). An online database is introduced for sharing both calculated and experimental CCS values (metabolomics.pnnl.gov), initially including a CCS library with over 1 million entries. Finally, three successful applications of molecule characterization using calculated CCS are described, including providing evidence for the presence of an environmental degradation product, the separation of molecular isomers, and an initial characterization of complex blinded mixtures of exposure chemicals. This work represents a method to address the limitations of small molecule identification and offers an alternative to generating chemical identification libraries experimentally by analyzing authentic reference materials. All code is available at github.com/pnnl.

Graphical Abstract



The capability to routinely measure and identify even a modest fraction of biologically and environmentally important small molecules within all of chemical space, greater than 10^{60} potential compounds[4], remains one of the grand challenges in biology and environmental monitoring. This long-term challenge is best met by analytical approaches capable of measuring broad classes of molecular species, referred to here as untargeted metabolomics. The technologies and driving concepts behind metabolomics have existed for nearly 40 years and have their origins in early metabolic profiling[5–11] and metabolic flux studies,[12,13] as well as detection of metabolic defects and diagnosis of associated inborn errors of metabolism.[14–16] However, despite the solid foundation and the great strides made in metabolomic approaches over the past 20 years, present capabilities still fall short of comprehensive and unambiguous chemical identification of detected metabolites.

For example, NMR-based structural elucidation is an established method for unambiguous chemical structure assignment of novel molecules but requires high sample concentration and purity. This limits its usefulness for high throughput and comprehensive structural elucidation. Synthesis of chemical reference standards for suspected novel molecules is another alternative but is costly, often difficult, and time-consuming.

For identification of known molecules, the analytical methodologies that have proven to be the most efficient in confident identification of large numbers of metabolites in high throughput metabolomic studies have been GC–MS, LC–MS, and NMR and comparison of experimental data to reference libraries containing data from analyses of authentic chemical standards using identical analytical methods. Such approaches adhere to the recommendations of the Metabolomics Standards Initiative of the Metabolomics Society for confident molecular identification[17,18] but depend on data from analysis of pure compounds. This represents a significant limitation, because authentic chemical standards are not available for the majority of metabolites.[19] For example, approximately 92% of the HMDB 4.0 molecules do not have authentic chemical standards (verified through custom Python scripts to search known vendors), and the HMDB only represents <5% of the estimated total metabolite space across multiple organisms.[20,21] Further, ChemSpider,[22] PubChem,[23] and American Chemical Society’s chemical abstracts service (CAS) databases[24] contain entries for tens-of-millions of chemicals, yet one of the largest repositories of authentic reference spectra, the Wiley Registry/NIST Mass Spectral Library,[25] contains data for roughly 730 000 unique compounds, <1% of known chemicals.[26]

The most practical approach for dramatically increasing the size of libraries is through in silico calculation of molecular attributes. The metabolomics community has made great strides in predictions of chromatographic retention times and tandem mass spectra.[27–31] While the associated tools and methods have demonstrated important proofs-of-concept, challenges remain with relying on these predicted attributes for metabolite identification. For example, GC and LC separations involve interactions of molecules with surfaces, and degradation of chromatographic stationary phases will result in a mismatch of experimental to predicted retention times. Tandem mass (MS/MS) spectra are gas-phase molecular properties, less susceptible to the chemical interactions and artifacts that can affect retention time stabilities in GC and LC, and have good reproducibility between laboratories. MS/MS spectra for small molecules can be predicted with reasonable accuracy given appropriate training data,[31–35] enabling the generation of short lists of molecules whose MS/MS spectra might match to experimental spectra. However, most MS/MS prediction methods rely on machine learning or deep learning approaches, and therefore, they can be limited by the size of the training data sets.[35,36] MS/MS spectra for molecules that are not chemically similar to compounds used in the training set may not be accurately predicted. New methods to accurately predict molecular properties that are also measured with high experimental reproducibility and without loss of data quality through time are required to transition metabolomics from the current paradigm to one applicable to global comprehensive chemical identification.

Quantum chemistry, i.e., the application of quantum mechanics to the understanding of molecules, holds great promise for the calculation of molecular properties in support of global chemical identification. For example, infrared spectra,[37] nuclear magnetic resonance chemical shifts,[38] and molecular collisional cross sections (CCS) can be calculated from first-principles, showing success where machine learning approaches[39] have underperformed. CCS is a measurable, calculable property of three-dimensional (3D) chemical structures that can contribute to the unambiguous identification of even positional and *cis/trans* isomers.[40,41] CCS is a measure of the apparent surface area of a chemical ion and is related to the molecular gas-phase 3D conformation of that ion. It is reported as an area in angstroms ($Å^2$). Ion mobility (IM) spectrometry separates ions based on the extent of their interactions with an inert gas (usually N_2 or He) as they travel under the influence of an electric field.[42] Ions of smaller CCS have shorter drift times, while ions of larger CCS have longer drift times. CCS is highly sensitive to molecular shape, showing measurable differences between even positional and *cis/trans* isomers.[40] Ultrahigh resolution IM separations coupled with mass spectrometry (IM–MS), such as structures for lossless ion manipulations (SLIM),[43–48] are capable of resolving compounds with only slight differences in stereochemistry and measuring mass and CCS with high accuracy. The gas-phase separations made by IM instruments have several advantages over conventional GC and LC platforms. The keys among these are extremely high reproducibility between instruments and laboratories (0.2% relative standard deviation);[49] no column degradation over time; separation principles sufficient to resolve constitutional, positional, and *cis/trans* isomers;[40] and platforms currently advancing to provide separation resolution 5-fold higher than conventional platforms.[43–48]

Currently, in silico methods for property prediction, including CCS, are limited by throughput, accuracy, and/or a reliance on large training sets. These obstacles hinder the rapid expansion of in silico libraries, particularly for molecules outside of any known training set (i.e., “out of sample”). To help advance methods of building in silico libraries, which may be used to provide evidence of the presence of molecules in samples, we introduce the in silico chemical library engine (ISiCLE), a quantum chemistry-based computational infrastructure for predicting molecular properties, including NMR chemical shifts[50] and CCS. We describe the development, optimization, and validation of the CCS calculation module of ISiCLE. We have architected ISiCLE for use on super-computing resources, including a refactoring of the popular MOBCAL[51–53] code for trajectory-based mobility calculations, and validated the calculation of 1983 CCS values against experimental data. Calculation accuracy is compared to similar first-principles approaches[54,55] as well as the property-based machine learning tool, MetCCS.[39] Finally, we provide a growing database of calculated CCS values, available at metabolomics.pnnl.gov, and a demonstration of the utility of calculated CCS in three examples.

MATERIALS AND METHODS

Validation Set Molecules.

Lists of molecule standards and their measured CCS were collected from in-house data and from the literature.[56–62] Values were tabulated (see the Supporting Information) along with their associated CCS relative standard deviation, observed mass, IUPAC International Chemical Identifiers (InChI),[63] SMILES string, formula, chemical name, source citation DOI, and chemical class information. For details on how InChI were obtained and processed, please see the Supporting Information Methods section. Only CCS collected on drift tube IM (DTIM) instruments with nitrogen (N_2) buffer gas were included, and only protonated, $[M + H]^+$, deprotonated, $[M - H]^-$, and sodiated $[M + Na]^+$ molecules were considered. If the CCS of the same compound and adduct ion (herein simply referred to as “adduct”) was measured by two different sources, their CCSs were included as two separate entries. All CCS values that were obtained in-house were collected using an Agilent 6560 Ion Mobility Q-TOF MS (Agilent Technologies, Santa Clara) with seven stepped electric field voltages, as described by Zheng et al.[64]

After analysis by ClassyFire,[1] our validation set was found to have 14 chemical superclasses and 76 classes. Chemical class composition of the validation set is shown in Figure 1. A total of 1308 unique compounds are included with masses ranging from 68.0374 to 1072.3806 Da.

ISiCLE CCS Calculation Module.

CCS values were calculated using ISiCLE, a high-throughput, automated computational pipeline built using the Python Snakemake framework,[65] enabling scalability, portability, provenance, fault tolerance, and automatic job restarting. Snakemake is a workflow management system that provides a readable Python-based workflow definition language and execution environment that scales, without modification, from single-core workstations to compute clusters through as-available job queuing based on a task dependency graph. See the SI Discussion for a more thorough discussion on the benefits and justification for using a workflow engine. ISiCLE offers three calculation methods, Lite, Standard, and AIMD (ab initio molecular dynamics), each increasing in calculation accuracy and computational complexity. This work focuses on the Standard method, though Lite and AIMD methods are introduced and discussed. ISiCLE source code is available in the SI, and up-to-date versions are available for download on GitHub (github.com/pnnl/isicle).

The ISiCLE module for calculating IM CCS started with the generation of 3D structures of ionized compounds (in .mol and .mol2 file formats) from a chemical structure identifier, such as the InChl of neutral parent compounds, and ended with the calculation of CCS values for various conformers of the ionized compounds using the trajectory method.[51] A conformer is any unique 3D arrangement of atoms for a molecule with the same bonding connectivity; i.e., a conformer is one of many constitutional stereoisomers of a molecule. For this work, protonated, deprotonated, and sodiated forms were considered for each molecule, but ISiCLE can be used to process other adducts as well (e.g., $[M + K]^+$ and $[M + 2Na]^{2+}$).[41]

The Standard pipeline, depicted in Figure 2, involves a series of intermediate steps for conformer generation using molecular dynamics (MD) simulations and geometry optimization using quantum chemical calculations, via density functional theory (DFT), on PNNL supercomputing resources. Each step of the CCS calculation pipeline was executed using a series of Python and shell scripts developed in-house, all coordinated through the Snakemake workflow. The details of each step in the Snakemake workflow are described below.

InChl to 3D Structure Creation.

Each processed (desalted, neutral, and major tautomer) InChI was converted into a two-dimensional (2D) representation of the compound using OpenBabel.[66,67] Three-dimensional structures were then generated and subsequently optimized using the generalized amber force field (GAFF) in OpenBabel. Ionized forms of the neutral structures were generated by identifying ionization sites in each parent 3D structure based on pK_a values, which were automatically calculated inline using the ChemAxon command line tool, cxcalc.[3] The strongest acidic atom (lowest pK_a) was assigned as the deprotonation site, and the strongest basic atom (highest pK_b) was assigned as the protonation and sodiation sites. All ionized structures were saved in the .mol2 file format.

Conformer Generation and Geometry Optimization.

During experimental analysis of authentic chemical reference standards, a continuous distribution of CCS values, or even multiple CCS values, could be observed for a single ionized molecule or complex. This necessitated the use of in silico conformer sampling methods to capture the CCS distribution, lest in silico predictions failed to achieve the required levels of accuracy.[69] Our approach ultimately selected a set of 30 total conformers, three each from a series of 10 simulated annealing MD steps. Twenty of the 30 conformers were chosen such that they sampled the extremities of geometrical space, whereas the remaining 10 were chosen to represent the most common regions of geometrical space. Briefly, ionized structures were used to seed conformer generation by in vacuo MD simulations, using SANDER (simulated annealing with NMR-derived energy restraints) from AmberTools17,[70] which raised the temperature such that energy barriers between conformation populations could be overcome, and subsequently cooled the system as a means of producing low-energy, stable conformers. For MD and simulated annealing details, please see the SI Material and Methods. Ten conformers from the low-temperature equilibration stage (300 K) of each simulated annealing cycle were randomly selected and then down-selected to three by identifying the two most dissimilar conformers and the single most similar conformer, leading to a total of 30 conformers. The dissimilar conformers were determined as the two conformers with the largest pairwise root-mean-square deviation (RMSD) of their atomic positions, while the most similar conformer had the lowest pairwise RMSD sum among the 10 conformers. The three selected conformers were sufficiently representative of the ten conformers in a single simulated annealing step.[40] Thus, a total of 30 conformer geometries were used for subsequent geometry optimization with DFT using NWChem.

Density Functional Theory Calculation.

To further optimize the resulting molecular geometries, quantum chemical DFT calculations were performed using NWChem, an open-source, high-performance computational chemistry software developed at PNNL, similar to the methods described in previous studies.[41,71] The B3LYP exchange-correlation functional was used for all energy and geometry optimization calculations.[72–75] All basis sets were obtained from the Environmental Molecular Sciences Laboratory (EMSL) Basis Set Exchange,[76,77] which included the Pople basis set at the 6-31+G** level (a double-ζ valence potential basis set having a single polarization function).[78–80]

CCS Calculation via MOBCAL-SHM.

CCS values of the geometry-optimized conformers were calculated using the trajectory method, as implemented in our new version of MOBCAL, called MOBCAL-SHM (shared memory; see the Results section). MOBCAL-SHM source code and binaries are available in the SI, and up-to-date versions are available for download on GitHub (github.com/pnnl/mobcal-shm). MOBCAL[51–53] was selected among comparable alternatives as its implementation of the trajectory method is generally accepted to be the "gold standard" for computational CCS calculation.[81–85] The original version of MOBCAL is computationally intensive; therefore, to improve computational efficiency, we developed and optimized the parallel MOBCAL-SHM, written in C. For details on the modifications made to speed up MOBCAL, please see the SI Material and Methods.

Averaging Calculated CCS Values of Conformers for Comparison to Experimental Values.

Reported CCS values are normally a single value per adduct and often chosen based on experimental signal strength, centroid analysis, and relative CCS peak location (e.g., to avoid selecting the CCS of a multimer). Thus, to calculate a single CCS value for each ionized structure from a set of conformers, a number of methods were evaluated, including methods similar to those implemented by Paglia et al.[54] and Bowers et al.[55] Putative methods resulted from the Cartesian product of three sets, (i) optimization scheme, (ii) number/type of conformers used in the average, and (iii) averaging method.

Optimization Scheme.—The optimization scheme explored whether the final geometry optimization by DFT was necessary to achieve lowest error, as it was the most computationally intensive step in the pipeline. Thus, we performed DFT calculations for each conformer to two levels of efficacy, optimization for energy only and optimization for energy and structure.

Number/Type of Conformers.—Methods of sampling conformers from the MD step, which produces two least-similar conformers and a single most-representative conformer for each of the 10 simulated annealing steps, were also evaluated. This set therefore includes the use of all sampled conformers (30), only the least similar conformers (20), and the most representative conformers (10).

Averaging Method.—Averaging methods included the mean and median CCS of conformers for each ionized structure, as well as three energy-based methods, (a) CCS of the lowest energy conformer, (b) the mean CCS of those conformers with relative energy less than 5 kcal/mol, and (c) the sum of each conformer's CCS contribution, Boltzmann-weighted by relative energy. We hypothesized that Boltzmann weighting, based on calculated DFT energies, would shift the overall CCS distribution toward higher probability conformers, thus creating CCS distributions that are characteristic of IM experiments.

Combined, the two optimization schemes, three conformer sampling methods, and five averaging methods resulted in 30 putative approaches for reducing a distribution of CCS values across conformers to a single CCS value per ionized structure. A comparison of these approaches with respect to mean absolute error (MAE) is summarized in Table S1, which includes results of similar approaches. The method introduced by Paglia et al. is similar to taking the lowest-energy conformer among all sampled conformers, DFT optimized for energy only. The method introduced by Bowers et al. is similar to averaging all energy- and structure-optimized conformers with relative delta energy less than 5 kcal/mol. Additional steps were taken to account for inaccuracies in the various components of the pipeline, including calibration. Please see the details in the SI Material and Methods.

Lite and ab Initio Molecular Dynamics (AIMD)-Based Method.

For applications that do not require as high CCS accuracy, a Lite method was created for rapid calculation. For applications that require higher CCS accuracy, at the cost of additional computational time, an AIMD-based method was created. Details on both of these methods are provided in the SI Material and Methods.

In Silico Library and Online Database.

In addition to the validation set, ISiCLE was used to calculate CCS values for the HMDB, Universal Natural Product Database (UNPD),[86] and the Distributed Structure-Searchable Toxicity (DSSTox) Database.[87] CCS values (protonated, deprotonated, and sodiated forms) were calculated using the ISiCLE Lite method for compounds from several databases that fell within the 50–1100 Da mass range (~80k from the HMDB, ~205k from the UNPD, and 720k from the DSSTox). Additionally, some compounds from the HMDB were run through ISiCLE Standard CCS calculations. All of these values are available at metabolomics.pnnl.gov, which is being regularly updated to expand the number of compounds and to replace Lite CCS values with Standard CCS values as they become available.

RESULTS

The efforts detailed in this work produced ISiCLE to address long-standing challenges hindering identification of the vast set of features in complex biological samples for which standards do not exist. Identification of small molecules requires accurate libraries of chemical properties that can be reliably measured experimentally, such as CCS. The essential tool for identification is an accurate library of properties for matching, not the authentic reference material itself. Authentic reference materials have been the preferred approach for obtaining libraries because the error in the features in the library are limited to relatively small experimental errors. Where computational tools can produce libraries with known errors, those libraries can be of value for providing evidence for the presence of a molecule in a sample, similar to a library made from experimental analysis of authentic compounds with known instrument error.

In silico methods must have validated error ranges and are also fast enough to make scientific contributions on a meaningful time scale, especially when cultivating libraries of in silico properties large enough for robust and comprehensive compound identification. Moreover, when possible, methods should attempt to reduce reliance on reference standards or training sets, as these impose limitations on novel molecule identification and discovery. The following results demonstrate ISiCLE's success in terms of accuracy, achieving 3.2% unsigned error; throughput, processing molecules in a matter of hours; and out-of-sample generalization in cases where other approaches have failed.

Mobility Calculation Improvements.

Increasing the speed of MOBCAL was one of the key factors for improving the throughput of CCS calculations. Briefly, our new MOBCAL-SHM reduced average CCS computation time from 10.8 to 0.08 node-hours, amounting to a 135-fold increase in efficiency. For comparison, Zanotto and co-workers recently reported a 48-fold efficiency increase in a refactored version of MOBCAL.[88] Full results regarding the mobility calculation improvements and computational efficiency can be found in the SI Results.

Validation.

Among our explored approaches for averaging calculated CCS values of conformers (Table S1), the lowest error was similar for the top several methods. These included (i) DFT optimization of energy and structure, (ii) averaging over either 20 or 30 conformers, and (iii) averaging by taking the minimum-energy conformer or by Boltzmann weighting. Because Boltzmann weighting by energy offers theoretical improvements over minimum-energy methods,[38] it was selected for the Standard method of ISiCLE. Additionally, following linear calibration, the Boltzmann method yields the lowest error. Figure 3 shows calculated CCS results for the validation set, plotted against *m/z*.

ISiCLE achieves 3.2% MAE when evaluated against experimental CCS values. Compared to other methods of CCS calculations on the same set of molecules, ISiCLE performs significantly better. Methods developed by Paglia et al.[54] and Bower et al.[55] achieve errors of 5.3 and 5.2%, respectively. The MetCCS approach achieved a MAE of 3.3%.

Applications.

To demonstrate the utility of ISiCLE, we used calculated CCS, mass, and other properties in three example applications involving real samples.

Application 1: Degradation Products in Sediment.—Environmental samples of New York/New Jersey Waterway Sediment (NIST SRM 1944[89]) were analyzed by DTIM-MS, with determination of accurate mass and CCS features for multiple compounds. CCS was calculated in silico using the Lite method of ISiCLE for 21 possible degradation products (e.g., 2-hydroxyfluorene, 3-hydroxyflourene, and 4,5-pyrenediol[90–92]) of 9 polycyclic aromatic hydrocarbons (e.g., fluorene, pyrene, and 1,6-dimethylphenanthrene) present in the sediment. Evidence for the presence of the parent compounds and predicted degradation products was built by comparing measured and calculated accurate mass and CCS. For example, experimental data for 4,5-pyrenediol[90–92] matched the predicted values within 1.1% (Table 1, representative data shown in Figure S4).

Application 2: U.S. Environmental Protection Agency (EPA) Non-Targeted Analysis Collaborative Trial (ENTACT) Challenge.—We participated in the ENTACT interlaboratory challenge,[93] designed for the objective testing of nontargeted analytical chemistry methods using a consistent set of blinded synthetic mixtures. Each mixture contained an unknown number of chemicals (later revealed to be 95–365 compounds) in dimethyl sulfoxide. All compounds were selected from the EPA ToxCast chemical library.[94] Further details on ENTACT are outlined in Sobus et al.[95] and Ulrich et al.[96] We used calculated CCS in part to provide evidence for the presences of compounds in each synthetic mixture, along with high resolution mass and isotopic signature.[93] For CCS calculations, the ISiCLE Standard method was used for 16% of molecules in the ToxCast chemical library, and the Lite method was used for the remaining molecules. In the end, our ToxCast CCS library had values for 11 633 adducts. The addition of calculated CCS to either the combination of mass and isotopic signature information or to mass or isotopic signature alone increased the confidence in correctly determining that a molecule was present in a sample for 84% of molecules. The increase in confidence was determined by finding the percent of true positives that had higher confidence scores due to at least one feature with a measured CCS within 5% of the predicted CCS, showcasing the importance of CCS in this multiattribute approach. Compared to the true positive experimental standards spiked in these samples, calculated CCS errors for Standard and Lite methods of ISiCLE were 3.1 and 5.4%, respectively (Table 1). This out-of-sample test demonstrates consistent CCS error values compared to the initial validation set. Experimental and calculated CCS values from this study are available in the library introduced below.

Application 3: High Accuracy CCS for Positional and cis/trans Isomers of Chlorogenic Acids.—We recently reported the ability of SLIM-MS to provide ultrahigh resolution IM separations[40] of positional and *cis/trans* isomers of dicaffeoylquinic acids (diCQAs), chlorogenic acids with reported anti-HIV and anti-inflammatory benefits.[40] Experimental CCS and CCS calculated using the ISiCLE AIMD-based method were compared for 3,5-diCQA isomers. To further evaluate the accuracy of ISiCLE, we expanded the calculations to encompass all eight reported diCQA isomers, including 1-*trans*,3-*trans*; 1-*trans*,5-*trans*; 3-*trans*,4-*trans*; 3-*cis*,5-*cis*; 3-*cis*,5-*trans*; 3-*trans*,5-*cis*; 3-*trans*,5-*trans*; and 4-*trans*,5-*trans*-diCQA. Resulting MAE were 4.8, 2.6, and 0.8% for Lite, Standard, and AIMD-based methods of ISiCLE (see Figure S5), respectively, compared to 6.4% for MetCCS. This out-of-sample set. i.e., set of compounds not present during model training, clearly demonstrates the performance-accuracy trade-off and reveals sub-1% error when the AIMD-based method is used. All CCS calculations in the near future could be performed with the AIMD-based method as computational power increases. This example also reveals one of the drawbacks of machine learning approaches that do not consider 2D or 3D molecular structures in their CCS calculation, such as MetCCS. The training parameters for these methods do not sufficiently differ between isomers to accurately distinguish their CCS values. Conformer consideration and 3D electron structure calculations alleviate this issue and can more accurately reflect the experimentally observed CCS values.

In Silico Library and Online Database.

CCS values for $[M + H]^+$, $[M - H]^-$, and $[M + Na]^+$ adducts are made available at metabolomics.pnnl.gov, currently totaling 1455 and over 1 million entries for experimental and calculated values, respectively. This community resource will be updated as more values become available. The website provides additional information, including chemical name, SMILES, InChI, 2D structure, formula, and mass.

DISCUSSION

In Silico Libraries Contributing to Molecular Identification.

We have developed ISiCLE as an additional tool[97–100] for the in silico generation of chemical property libraries. These tools are facilitating the departure from complete reliance on experimentally derived chemical properties for complex mixture characterization by potentially offering evidence for the presence of comprised molecules. Experimental characterization of authentic reference materials is an expensive, time-consuming practice that cannot accommodate candidate molecules that are (i) without a form available for purchase, (ii) without a protocol to synthesize, or (iii) as of yet undiscovered. ISiCLE therefore enables expansion of chemical property libraries through calculation, and although initially dependent on experimental standards for calibration and validation, it will ultimately see use as a generative approach for creating significantly larger chemical property libraries than are currently possible.

Through the incorporation of computationally derived libraries in small-molecule identification pipelines, the comprehensive characterization of complex samples becomes tractable with a sufficiently representative library. As ISiCLE evolves toward greater accuracy and diversity of calculated properties (CCS, NMR chemical shifts,[50] and beyond), and more molecules are added to the in silico library, compound identification may be confidently made without reliance upon data from experimental analysis of authentic compounds. It will be the work of institutions, such as the Metabolomics Standards Initiative and metabolomics societies, to establish frameworks and criteria for assessing the confidence of “identifications” made with in silico libraries. As an estimated >99% of metabolites are currently undiscovered,[4,101–103] accommodation of computational methods with confidence is imperative for the advancement of our fields.

Library for the Molecular Universe?

The over a million compound library reported here is a transformational increase over existing reference libraries. Nonetheless, as a minor fraction of chemical space and of chemical properties, libraries of its size and composition alone are not sufficient for identifying all reported features in complex biological samples. We recognize the likelihood that features will emerge from untargeted analyses that cannot be resolved based on a combination of CCS and mass and that match no library entries because they represent a currently unknown chemical structure. For these latter cases, measured attributes, such as high accuracy mass and isotopic signature, can be used to generate plausible molecular formulas that can then be correlated to possible chemical structures, for example, through in silico metabolism simulators,[104] deep learning-based neural networks,[105] or with more effort, combinatorial searching of a given formula.[106] Additional attributes of these new molecular structures, CCS, NMR chemical shifts,[50] retention times, and MS/MS spectra, can then be calculated and, where sufficient data exist, be used to identify the subset most likely to represent the feature. Thus, ISiCLE can be used to generate attributes of probable chemical structures, with errors small enough to support down selection and provisional identification. A growing library of these new compounds would eventually come to represent an increasing portion of molecular space.

Rapid and extensive growth of in silico libraries is also an attractive approach to reduce the number of unidentifiable features in complex samples. Processing all molecules available in databases, such as HMDB, UNPD, PubChem, and others,[22,24,25] using tools like ISiCLE would establish libraries of our known or recorded molecular universe. This level of library expansion would eventually cover the majority of known biologically relevant molecules to include those for which authentic chemical standards are not available.

Comparison with Other CCS Calculation Methods.

Structure-based approaches that utilize first-principles of quantum chemical calculations leverage our understanding of the underlying physics to predict chemical properties directly. Compared to approaches that predict chemical properties without first-principles simulation, such as the machine learning-based MetCCS, ISiCLE performs comparably with decreased CCS error but with increased computation time. However, ISiCLE offers promise in that it will theoretically generalize more effectively to out-of-sample characterizations, a critical factor in growing an in silico chemical property library. Machine learning methods, like MetCCS, are limited by the size and scope of the initial training set, and thus ultimately limited to the number of authentic chemical standards available for purchase. Furthermore, machine learning is challenged by chemicals with similar properties and similar structures, such as constitutional and configurational isomers (e.g., *cis/trans* isomers), as demonstrated above in application 3 with diCQA. The input properties required for MetCCS were nearly identical for all 8 isomers, despite CCS values for this set spanning a range of over 43 Å^2, leading to predicted CCS errors as high as 9.5% (1-*trans*,5-*trans*-diCQA). We have demonstrated that our approach can surmount this challenge and with high accuracy (MAE of 0.8% for this set). In addition, ISiCLE offers scalability across HPC resources, portability, provenance, and fault tolerance.

It is important to note, that CCS and *m/z* are highly correlated in most cases (Pearson product-moment correlation coefficient of 0.96 for our validation set of molecules, 0.92 in Marklund et al.[107]). Because of this, even a simple linear regression model can achieve an MAE of 4.4% compared to experimental values. However, this type of model predicts the same CCS for a given formula, preventing its use entirely when trying to distinguish between molecules with the same formula (such as the diCQA isomers), and this model has a very long tail of high error predictions (e.g., 56% larger variance compared to the ISiCLE Standard method). To that point, for molecules that fall far from the linearity zone (regression residual greater than 10%), the MAE using the linear model reaches 12.9%, compared to 6.2% for ISiCLE Standard. Thus, the use of molecular structure is useful in reducing CCS error generated from machine learning methods, which include linear regression, that only consider molecular properties.

Additional discussions are located in the SI Discussion section, including a discussion regarding the different ISiCLE methods (i.e., Lite, Standard, and AIMD-based methods) as well as their computational efficiency.

CONCLUSION

In this article we present the development of ISiCLE, a computational tool for accurate and supercomputing-enabled prediction of chemical properties using quantum chemical methods. This work offers (1) the first open-source, scalable (from desktop to HPC resources), and portable quantum chemistry-based collision cross section calculation workflow for the community; (2) an advanced conformer sampling method for higher accuracy property prediction, based on Boltzmann weighting to ensure that highly probable conformers are more represented; (3) a refactoring of the gold standard mobility calculation method (MOBCAL) with a speedup of over 2 orders of magnitude; (4) a validation of the whole pipeline on the largest experimental data set in the literature to date (unique values); (5) a comparison of our approach with those in the literature, including competing machine learning approaches; and (6) a public library of over 1 million entries, covering the Human Metabolome Database, the EPA DSSTox exposure database, and the Universal Natural Product Database.

The transformation of the field of metabolomics toward global comprehensive identification of compounds in complex samples is underway. Among the many innovations that are necessary to reach this goal, e.g., ultrahigh resolution separation and higher throughput NMR, the development of in silico libraries of chemical properties to provide evidence for the presence of the multitude of compounds for which authentic standards do not exist is a critical step. Development of ISiCLE, including MOBCAL optimization, is an important first step toward meeting the goal of establishing large scale in silico libraries. ISiCLE has an easy to use software package for calculating chemical properties, including CCS, incentivizing adoption.

ISiCLE’s AIMD-based method produced CCS values with absolute errors of 0.8%, approaching measurement error where the less computationally intensive implementations each had absolute errors less than current methods. Looking forward, ISiCLE’s reliance on first-principles and full 3D chemical structures may provide advantages over machine learning approaches derived from 2D structural information, particularly for positional isomers. Our successful use of ISiCLE for identification of the diCQA positional isomers highlights this import point for the field.

Recent funding of Compound Identification Development Cores by the National Institutes of Health reflects growing recognition of the challenge to the community that our limited libraries and availability of chemical standards pose. As momentum in the development of new methodologies for chemical identification through innovations in computational and experimental methods grows, a parallel need to consider best practices for use of these methods for chemical identification will also have to be fostered. In addition, it is clear to us, that calculation of additional attributes, e.g., NMR chemical shifts, IR spectra, and others, will be necessary to increase the dimensionality of the array of attributes for chemical identification. Together, these improvements will help bring about the required paradigm shift away from the reference-material-based library building and, as a consequence, a rapid advancement in compound identification and biomedical discovery.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This research was supported by the Genomic Science Program (GSP), Office of Biological and Environmental Research (BER), U.S. Department of Energy (DOE), and is a contribution of the Metabolic and Spatial Interactions in Communities Scientific Focus Area. Additional support was provided by the Pacific Northwest National Laboratory (PNNL); Laboratory Directed Research and Development program via the Global Forensic Chemical Exposure Assessment for the Environmental Exposome project and the Microbiomes in Transition Initiative; and the National Institutes of Health, National Institute of Environmental Health Sciences grant U2CES030170 and National Cancer Institute grant R03CA222443. Ion mobility spectrometry-mass spectrometry analyses were performed in the Environmental Molecular Sciences Laboratory (EMSL), a national scientific user facility sponsored by BER and located at PNNL. Simulations were performed on PNNL Institutional Computing and EMSL high-performance computing resources. PNNL is operated for DOE by Battelle Memorial Institute under contract DE-AC05-76RL01830. The authors thank Dr. Neeraj Kumar (PNNL) for valuable discussions on the MD simulations, and we thank Dr. Erin Baker (NCSU) and Dr. Samuel Payne (BYU) for discussions regarding the experimental data and statistics.

REFERENCES

(1). Feunang YD; Eisner R; Knox C; Chepelev L; Hastings J; Owen G; Fahy E; Steinbeck C; Subramanian S; Bolton E; Greiner R; Wishart DS J. Cheminf 2016, 8, 61.

(2). Wishart DS; Feunang YD; Marcu A; Guo AC; Liang K; Vazquez-Fresno R; Sajed T; Johnson D; Li C; Karu N; Sayeeda Z; Lo E; Assempour N; Berjanskii M; Singhal S; Arndt D; Liang Y; Badran H; Grant J; Serra-Cayuela A; Liu Y; Mandal R; Neveu V; Pon A; Knox C; Wilson M; Manach C; Scalbert A Nucleic Acids Res. 2018, 46 (D1), D608–D617. [PubMed: 29140435]

(3). cxcalc, 16.11; ChemAxon, 2017.

(4). Dobson CM Nature 2004, 432 (7019), 824–828. [PubMed: 15602547]

(5). Rosenberg RN; Robinson AB; Partridge D Clin. Biochem 1975, 8 (6), 365–368. [PubMed: 1204210]

(6). Dirren H; Robinson AB; Pauling L Clin Chem. 1975, 21 (13), 1970–1975. [PubMed: 1192592]

(7). Robinson AB; Dirren H; Sheets A; Miquel J; Lundgren PR Exp. Gerontol 1976, 11 (1–2), 11–16. [PubMed: 1278265]

(8). Gates SC; Sweeley CC; Krivit W; Dewitt D; Blaisdell BE Clin Chem. 1978, 24 (10), 1680–1689. [PubMed: 699273]

(9). Politzer IR; Githens S; Dowty BJ; Laseter JL J. Chromatogr. Sci 1975, 13 (8), 378–9. [PubMed: 1159029]

(10). Nicholson JK; Buckingham MJ; Sadler PJ Biochem. J 1983, 211 (3), 605–15. [PubMed: 6411064]

(11). Bales JR; Higham DP; Howe I; Nicholson JK; Sadler PJ Clin Chem. 1984, 30 (3), 426–432. [PubMed: 6321058]

(12). Flint HJ; Porteous DJ; Kacser H Biochem. J 1980, 190 (1), 1–15. [PubMed: 6449928]

(13). Middleton RJ; Kacser H Genetics 1983, 105 (3), 633–650. [PubMed: 6416922]

(14). Rashed MS J. Chromatogr., Biomed. Appl 2001, 758 (1), 27–48.

(15). Clayton PT J. Inherited Metab. Dis 2001, 24 (2), 139–50. [PubMed: 11405336]

(16). Kuhara T Mass Spectrom. Rev 2005, 24 (6), 814–27. [PubMed: 15376278]

(17). Castle AL; Fiehn O; Kaddurah-Daouk R; Lindon JC Briefings Bioinf. 2006, 7 (2), 159–65.

(18). Sumner LW; Amberg A; Barrett D; Beale MH; Beger R; Daykin CA; Fan TW; Fiehn O; Goodacre R; Griffin JL; Hankemeier T; Hardy N; Harnly J; Higashi R; Kopka J; Lane AN; Lindon JC; Marriott P; Nicholls AW; Reily MD; Thaden JJ; Viant MR Metabolomics 2007, 3 (3), 211–221. [PubMed: 24039616]

(19). Beisken S; Eiden M; Salek RM Expert Rev. Mol. Diagn 2015, 15 (1), 97–109. [PubMed: 25354566]

(20). Tulp M; Bohlin L Trends Pharmacol. Sci 2002, 23 (5), 225–231. [PubMed: 12008000]

(21). Fiehn O Plant Mol. Biol 2002, 48 (1–2), 155–71. [PubMed: 11860207]

(22). Pence HE; Williams AJ Chem. Educ 2010, 87 (11), 1123–1124.

(23). Kim S; Thiessen PA; Bolton EE; Chen J; Fu G; Gindulyte A; Han L; He J; He S; Shoemaker BA; Wang J; Yu B; Zhang J; Bryant SH Nucleic Acids Res. 2016, 44 (D1), D1202–D1213. [PubMed: 26400175]

(24). Erthal LCS; Marques AF; Almeida FCL; Melo GLM; Carvalho CM; Palmieri LC; Cabral KMS; Fontes GN; Lima LMTR. Biophys. Chem 2016, 218, 58–70. [PubMed: 27693831]

(25). Lemmon E; McLinden M; Friend D; Linstrom P; Mallard W Nist standard reference database number 69. NIST Chemistry WebBook; National Institute of Standards and Technology: Gaithersburg, 2011.

(26). Schymanski EL; Singer HP; Slobodnik J; Ipolyi IM; Oswald P; Krauss M; Schulze T; Haglund P; Letzel T; Grosse S; Thomaidis NS; Bletsou A; Zwiener C; Ibanez M; Portoles T; de Boer R; Reid MJ; Onghena M; Kunkel U; Schulz W; Guillon A; Noyon N; Leroy G; Bados P; Bogialli S; Stipanicev D; Rostkowski P; Hollender J Anal. Bioanal Chem 2015, 407 (21), 6237–6255. [PubMed: 25976391]

(27). Randazzo GM; Tonoli D; Strajhar P; Xenarios I; Odermatt A; Boccard J; Rudaz SJ Chromatogr. B: Anal. Technol. Biomed. Life Sci 2017, 1071, 11–18.

(28). Vinaixa M; Schymanski EL; Neumann S; Navarro M; Salek RM; Yanes O TrAC, Trends Anal. Chem 2016, 78, 23–35.

(29). D’Atri V; Causon T; Hernandez-Alba O; Mutabazi A; Veuthey JL; Cianferani S; Guillarme D J. Sep Sci 2018, 41 (1), 20–67. [PubMed: 29024509]

(30). Bocker S Curr. Opin. Chem. Biol 2017, 36, 1–6. [PubMed: 28025165]

(31). Allen F; Greiner R; Wishart D Metabolomics 2015, 11 (1), 98–110.

(32). Wolf S; Schmidt S; Muller-Hannemann M; Neumann S BMC Bioinf. 2010, 11, 148.

(33). Gerlich M; Neumann S J. Mass Spectrom 2013, 48 (3), 291–8. [PubMed: 23494783]

(34). Kangas LJ; Metz TO; Isaac G; Schrom BT; Ginovska-Pangovska B; Wang L; Tan L; Lewis RR; Miller JH Bioinformatics 2012, 28 (13), 1705–13. [PubMed: 22592377]

(35). Duhrkop K; Shen H; Meusel M; Rousu J; Bocker S Proc. Natl. Acad. Sci. U. S. A 2015, 112 (41), 12580–5. [PubMed: 26392543]

(36). Allen F; Pon A; Wilson M; Greiner R; Wishart D Nucleic Acids Res. 2014, 42, W94. [PubMed: 24895432]

(37). Bouteiller Y; Gillet J-C; Grégoire G; Schermann JP J. Phys. Chem. A 2008, 112 (46), 11656–11660. [PubMed: 18942809]

(38). Willoughby PH; Jansma MJ; Hoye TR Nat. Protoc 2014, 9 (3), 643–660. [PubMed: 24556787]

(39). Zhou ZW; Shen XT; Tu J; Zhu ZJ Anal. Chem 2016, 88 (22), 11084–11091. [PubMed: 27768289]

(40). 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 J. Phys. Chem. Lett 2017, 8 (7), 1381–1388. [PubMed: 28267339]

(41). Zheng X; Zhang X; Schocker NS; Renslow RS; Orton DJ; Khamsi J; Ashmus RA; Almeida IC; Tang K; Costello CE; Smith RD; Michael K; Baker ES Anal. Bioanal. Chem 2017, 409 (2), 467–476. [PubMed: 27604268]

(42). Hill HH Jr; Siems WF; St Louis RH. Anal. Chem 1990, 62 (23), 1201A–1209A.

(43). Deng L; Ibrahim YM; Garimella SVB; Webb IK; Hamid AM; Norheim RV; Prost SA; Sandoval JA; Baker ES; Smith RD Anal. Chem 2016, 88 (20), 10143–10150. [PubMed: 27715008]

(44). Deng LL; Ibrahim YM; Hamid AM; Garimella SVB; Webb IK; Zheng XY; Prost SA; Sandoval JA; Norheim RV; Anderson GA; Tolmachev AV; Baker ES; Smith RD Anal. Chem 2016, 88 (18), 8957–8964. [PubMed: 27531027]

(45). Garimella SVB; Hamid AM; Deng L; Ibrahim YM; Webb IK; Baker ES; Prost SA; Norheim RV; Anderson GA; Smith RD Anal. Chem 2016, 88 (23), 11877–11885. [PubMed: 27934097]

(46). Hamid AM; Ibrahim YM; Garimella SVB; Webb IK; Deng LL; Chen TC; Anderson GA; Prost SA; Norheim RV; Tolmachev AV; Smith RD Anal. Chem 2015, 87 (22), 11301–11308. [PubMed: 26510005]

(47). Ibrahim YM; Hamid AM; Cox JT; Garimella SVB; Smith RD Anal. Chem 2017, 89 (3), 1972–1977. [PubMed: 28208272]

(48). Ibrahim YM; Hamid AM; Deng LL; Garimella SVB; Webb IK; Baker ES; Smith RD Analyst 2017, 142 (7), 1010–1021. [PubMed: 28262893]

(49). Stow SM; Causon TJ; Zheng XY; Kurulugama RT; Mairinger T; May JC; Rennie EE; Baker ES; Smith RD; McLean JA; Hann S; Fjeldsted JC Anal. Chem 2017, 89 (17), 9048–9055. [PubMed: 28763190]

(50). Yesiltepe Y; Nuñez JR; Colby SM; Thomas DG; Borkum MI; Reardon PN; Washton NM; Metz TO; Teeguarden JG; Govind N; Renslow RS J. Cheminf 2018, 10 (1), 52.

(51). Mesleh MF; Hunter JM; Shvartsburg AA; Schatz GC; Jarrold MF J. Phys. Chem 1996, 100 (40), 16082–16086.

(52). Campuzano I; Bush MF; Robinson CV; Beaumont C; Richardson K; Kim H; Kim HI Anal. Chem 2012, 84 (2), 1026–33. [PubMed: 22141445]

(53). Shvartsburg AA; Jarrold MF Chem. Phys. Lett 1996, 261 (1–2), 86–91.

(54). Paglia G; Williams JP; Menikarachchi L; Thompson JW; Tyldesley-Worster R; Halldorsson S; Rolfsson O; Moseley A; Grant D; Langridge J; Palsson BO; Astarita G Anal. Chem 2014, 86 (8), 3985–3993. [PubMed: 24640936]

(55). Theoretical Collision Cross Sections, https://labs.chem.ucsb.edu/bowers/michael/theory_analysis/cross-sections/.

(56). Stephan S; Hippler J; Kohler T; Deeb AA; Schmidt TC; Schmitz OJ Anal. Bioanal. Chem 2016, 408 (24), 6545–55. [PubMed: 27497965]

(57). Henderson SC; Li J; Counterman AE; Clemmer DE J. Phys. Chem. B 1999, 103 (41), 8780–8785.

(58). Hoaglund CS; Valentine SJ; Sporleder CR; Reilly JP; Clemmer DE Anal. Chem 1998, 70 (11), 2236–2242. [PubMed: 9624897]

(59). May JC; Goodwin CR; Lareau NM; Leaptrot KL; Morris CB; Kurulugama RT; Mordehai A; Klein C; Barry W; Darland E; et al. Anal. Chem 2014, 86 (4), 2107–2116. [PubMed: 24446877]

(60). Wyttenbach T; Bushnell JE; Bowers MT J. Am. Chem. Soc 1998, 120 (20), 5098–5103.

(61). Hines KM; May JC; McLean JA; Xu L Anal. Chem 2016, 88 (14), 7329–7336. [PubMed: 27321977]

(62). Henderson SC; Valentine SJ; Counterman AE; Clemmer DE Anal. Chem 1999, 71 (2), 291–301. [PubMed: 9949724]

(63). Heller S; McNaught A; Stein S; Tchekhovskoi D; Pletnev I J. Cheminf 2013, 5, 7–7.

(64). Zheng XY; Aly NA; Zhou YX; Dupuis KT; Bilbao A; Paurus VL; Orton DJ; Wilson R; Payne SH; Smith RD; Baker ES Chem. Sci 2017, 8 (11), 7724–7736. [PubMed: 29568436]

(65). Köster J; Rahmann S Bioinformatics 2012, 28 (19), 2520–2522. [PubMed: 22908215]

(66). OpenBabel, v2.4; 2017.

(67). O'Boyle NM; Banck M; James CA; Morley C; Vandermeersch T; Hutchison GR J. Cheminf 2011, 3 (1), 33.

(68). Wang J; Wolf RM; Caldwell JW; Kollman PA; Case DA J. Comput. Chem 2004, 25 (9), 1157–74. [PubMed: 15116359]

(69). Lee JW; Lee HHL; Davidson KL; Bush MF; Kim HI Analyst 2018, 143 (8), 1786–1796. [PubMed: 29561029]

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

(71). Graham TR; Renslow R; Govind N; Saunders SR J. Phys. Chem. C 2016, 120 (35), 19837–19847.

(72). Becke AD J. Chem. Phys 1993, 98 (7), 5648–5652.

(73). Lee C; Yang W; Parr RG Phys. Rev. B: Condens. Matter Mater. Phys 1988, 37 (2), 785–789.

(74). Vosko SH; Wilk L; Nusair M Can. J. Phys 1980, 58 (8), 1200–1211.

(75). Devlin FJ; Finley JW; Stephens PJ; Frisch MJ J. Phys. Chem 1995, 99 (46), 16883–16902.

(76). Feller D J. Comput. Chem 1996, 17 (13), 1571–1586.

(77). Schuchardt KL; Didier BT; Elsethagen T; Sun L; Gurumoorthi V; Chase J; Li J; Windus TL J. Chem. Inf. Model 2007, 47 (3), 1045–1052. [PubMed: 17428029]

(78). Francl MM; Pietro WJ; Hehre WJ; Binkley JS; Gordon MS; DeFrees DJ; Pople JA J. Chem. Phys 1982, 77 (7), 3654–3665.

(79). Hariharan PC; Pople JA Theoretica chimica acta 1973, 28 (3), 213–222.

(80). Rassolov VA; Ratner MA; Pople JA; Redfern PC; Curtiss LA J. Comput. Chem 2001, 22 (9), 976–984.

(81). Ruotolo BT; Benesch JLP; Sandercock AM; Hyung S-J ; Robinson CV Nat. Protoc 2008, 3, 1139. [PubMed: 18600219]

(82). Laganowsky A; Reading E; Allison TM; Ulmschneider MB; Degiacomi MT; Baldwin AJ; Robinson CV Nature 2014, 510 (7503), 172–175. [PubMed: 24899312]

(83). Forsythe JG; Petrov AS; Walker CA; Allen SJ; Pellissier JS; Bush MF; Hud NV; Fernandez FM Analyst 2015, 140 (20), 6853–6861. [PubMed: 26148962]

(84). Lapthorn C; Pullen FS; Chowdhry BZ; Wright P; Perkins GL; Heredia Y Analyst 2015, 140 (20), 6814–6823. [PubMed: 26131453]

(85). Boschmans J; Jacobs S; Williams JP; Palmer M; Richardson K; Giles K; Lapthorn C; Herrebout WA; Lemiere F; Sobott F Analyst 2016, 141 (13), 4044–4054. [PubMed: 27264846]

(86). Gu J; Gui Y; Chen L; Yuan G; Lu HZ; Xu X PLoS One 2013, 8 (4), No. e62839. [PubMed: 23638153]

(87). Richard AM; Williams CR Mutat. Res. Fundam. Mol. Mech. Mutagen 2002, 499 (1), 27–52.

(88). Zanotto L; Heerdt G; Souza PCT; Araujo G; Skaf MS J. Comput. Chem 2018, 39 (21), 1675–1681. [PubMed: 29498071]

(89). Wise SA; Poster DL; Schantz MM; Kucklick JR; Sander LC; Lopez de Alda M; Schubert P; Parris RM; Porter BJ Anal. Bioanal. Chem 2004, 378 (5), 1251–1264. [PubMed: 14745475]

(90). Dean-Ross D; Moody J; Cerniglia CE FEMS Microbiol. Ecol 2002, 41 (1), 1–7. [PubMed: 19709233]

(91). Reichert WL; Le Eberhart B-T; Varanasi U Aquat. Toxicol 1985, 6 (1), 45–56.

(92). Luan TG; Yu KS; Zhong Y; Zhou HW; Lan CY; Tam NF Chemosphere 2006, 65 (11), 2289–96. [PubMed: 16806399]

(93). Nuñez JR; Colby SM; Thomas DG; Tfaily MM; Tolic N; Ulrich EM; Sobus JR; Metz TO; Teeguarden JG; Renslow RS, Advancing Standards-Free Methods for the Identification of Small Molecules in Complex Samples. 2018, arXiv:1810.07367 [q-bio.BM]. arXiv.org e-Print archive. https://arxiv.org/abs/1810.07367.

(94). Richard AM; Judson RS; Houck KA; Grulke CM; Volarath P; Thillainadarajah I; Yang C; Rathman J; Martin MT; Wambaugh JF; Knudsen TB; Kancherla J; Mansouri K; Patlewicz G; Williams AJ; Little SB; Crofton KM; Thomas RS Chem. Res. Toxicol 2016, 29 (8), 1225–51. [PubMed: 27367298]

(95). Sobus JR; Wambaugh JF; Isaacs KK; Williams AJ; McEachran AD; Richard AM; Grulke CM; Ulrich EM; Rager JE; Strynar MJ; Newton SR J. Exposure Sci. Environ. Epidemiol 2018, 28, 411.

(96). Ulrich EM; Sobus JR; Grulke C; Richard A; Newton S; Mansouri K; Williams A; Strynar M. J. Anal. Bioanal. Chem 2019, 411, 853. [PubMed: 30519961]

(97). Allen F; Pon A; Wilson M; Greiner R; Wishart D Nucleic Acids Res. 2014, 42 (W1), W94–W99. [PubMed: 24895432]

(98). Ridder L; van der Hooft JJJ; Verhoeven S Mass Spectrom. 2014, 3, S0033–S0033.

(99). Ruttkies C; Schymanski EL; Wolf S; Hollender J; Neumann S J. Cheminf 2016, 8 (1), 3.

(100). Zhou ZW; Xiong X; Zhu ZJ Bioinformatics 2017, 33 (14) , 2235–2237. [PubMed: 28334295]

(101). Reymond J-L; Awale M ACS Chem. Neurosci 2012, 3 (9), 649–657. [PubMed: 23019491]

(102). Reymond JL Acc. Chem. Res 2015, 48 (3), 722–730. [PubMed: 25687211]

(103). Ruddigkeit L; van Deursen R; Blum LC; Reymond JL J. Chem. Inf. Model 2012, 52 (11), 2864–2875. [PubMed: 23088335]

(104). Jeffryes JG; Colastani RL; Elbadawi-Sidhu M; Kind T; Niehaus TD; Broadbelt LJ; Hanson AD; Fiehn O; Tyo KE; Henry CS J. Cheminf 2015, 7, 44.

(105). Gómez-Bombarelli R; Wei JN; Duvenaud D; Hernández-Lobato JM; Sánchez-Lengeling B; Sheberla D; Aguilera-Iparraguirre J; Hirzel TD; Adams RP; Aspuru-Guzik A ACS Cent. Sci 2018, 4 (2), 268–276. [PubMed: 29532027]

(106). Gugisch R; Kerber A; Kohnert A; Laue R; Meringer M; Rücker C; Wassermann A MOLGEN 5.0, A Molecular Structure Generator In Advances in Mathematical Chemistry and Applications; Basak SC, Restrepo G, Villaveces JL, Eds.; Bentham Science Publishers, 2015; pp 113–138.

(107). Marklund EG; Degiacomi MT; Robinson CV; Baldwin AJ; Benesch JL Structure 2015, 23 (4), 791–9. [PubMed: 25800554]

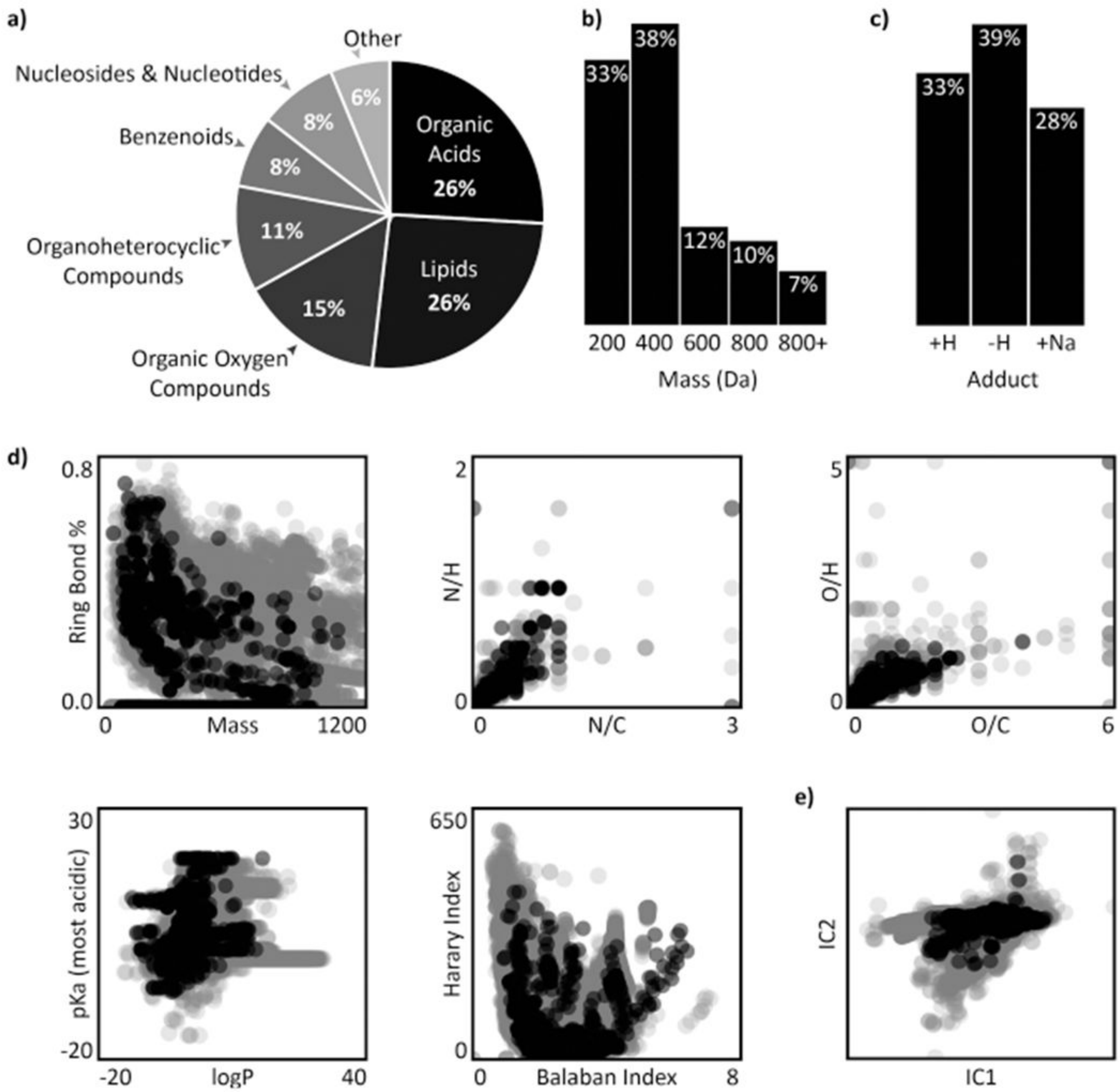


Figure 1.

Validation set property distribution and chemical space coverage. (a) Superclass distribution of compounds, as determined by ClassyFire.[1] (b) Mass distribution with mass labels corresponding to [X-200, X]. (c) Adduct distribution. (d, e) Comparison of the validation set to the Human Metabolome Database (HMDB),[2] with black points corresponding to compounds found in the validation set, and gray points corresponding to compounds found in the HMDB (v4.1, only those with masses 50–1200). (d) Distribution of predicted properties, with the ring bond percentage (number of bonds in rings divided by the total number of bonds), log *P*, pK_a, Balaban index, and Harary index calculated using cxcalc.[3] (e)

Independent component analysis performed on the properties plotted in (d), with properties normalized to have a mean of 0 and standard deviation of 1.

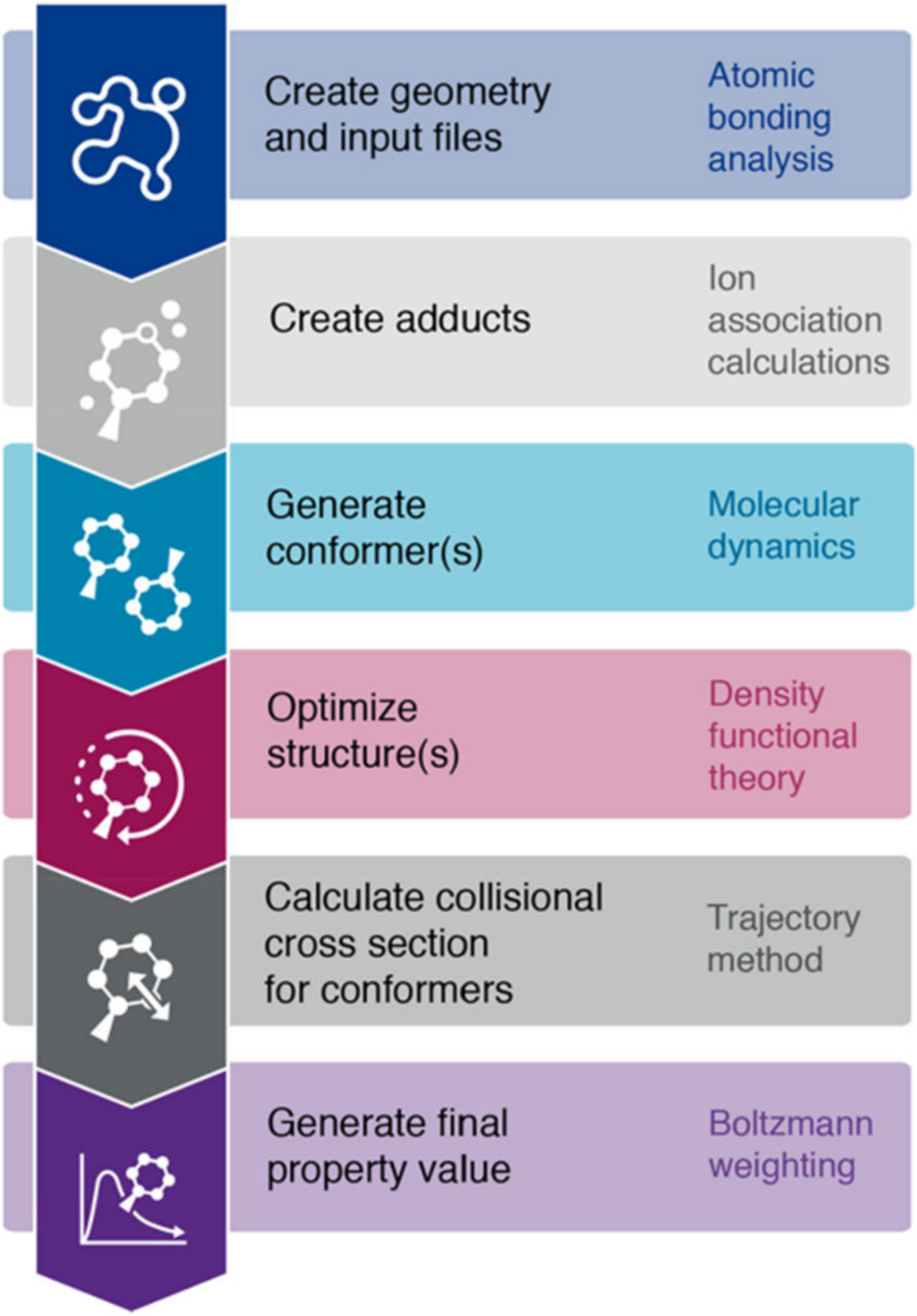

Create geometry and input files
Atomic bonding analysis
Create adducts
Ion association calculations
Generate conformer(s)
Molecular dynamics
Optimize structure(s)
Density functional theory
Calculate collisional cross section for conformers
Trajectory method
Generate final property value
Boltzmann weighting

Figure 2.

Schematic overview of the ISiCLE module for CCS calculation. Major computational tasks are listed for the Standard method and, where appropriate, the associated method used. Tasks include preparation of input geometry from InChI, adduct formation, conformer generation by molecular dynamics, structure optimization by density functional theory, CCS calculation by the trajectory method, and finally, final CCS prediction by Boltzmann weighting across conformers.

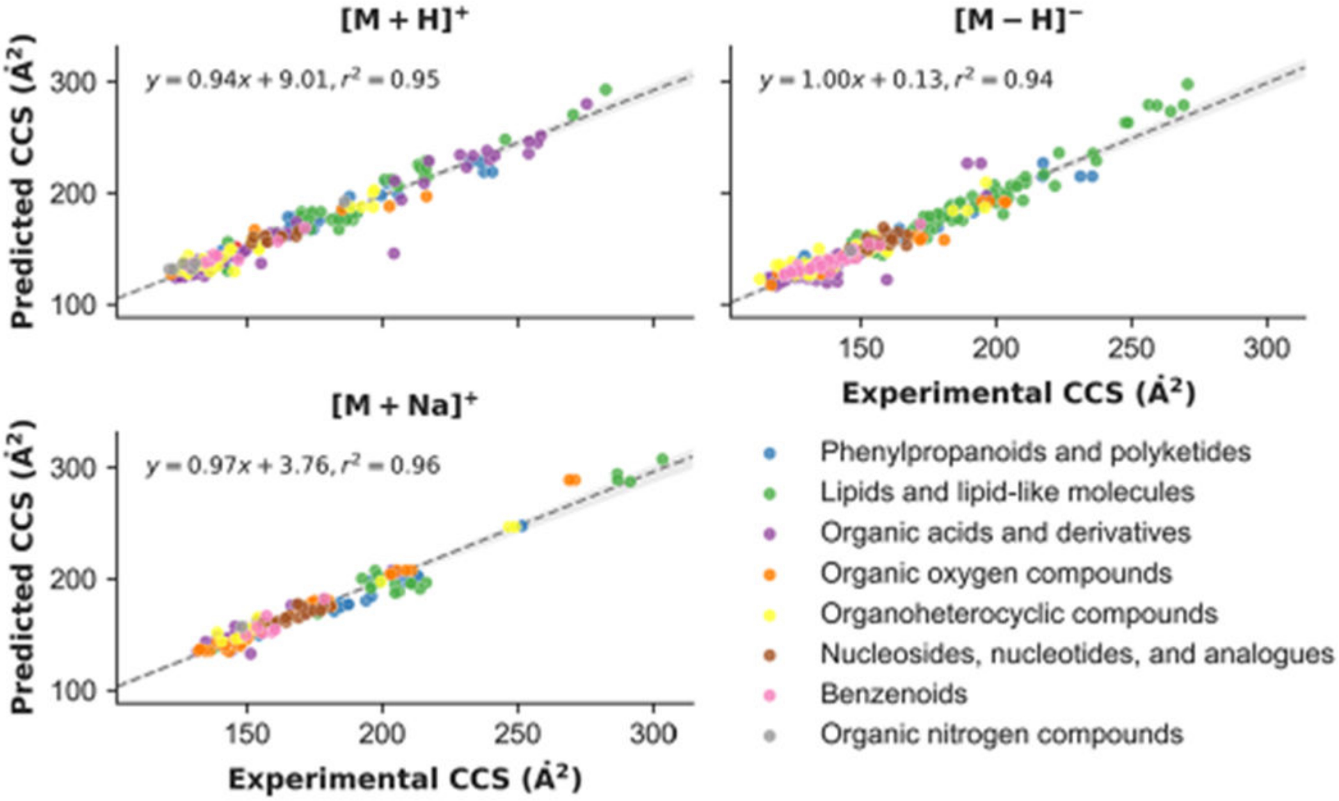


Figure 3.

Calculated CCS versus *m/z*. Visual representation of CCS values calculated by ISiCLE Standard for the validation set, plotted against *m/z* by adduct ion, colored by chemical class as determined by ClassyFire.[1]

Table 1.

Performance Comparison[a]

method	validation set (%)	PAH degradation product (application 1) (%)	EPA ENTACT mixtures (application 2) (%)	diCQA isomers (application 3) (%)
ISiCLE (Lite)	6.4	1.1	5.4	4.8
ISiCLE (Standard)	3.2	–	3.1	2.6
ISiCLE (AIMD-based)	–	–	–	0.8
MetCCS[39]	3.3	–	–	6.4
Paglia et al.[54]	5.3	–	–	3.1
Bowers et al.[55]	5.2	–	–	3.7

[a] MAE is shown for each method and dataset, where applied. The hierarchy of ISiCLE methods (Lite, Standard, and AIMD-based) is captured, as well as ISiCLE’s performance relative to similar methods (Paglia et al.[54] and Bowers et al.[55]) and the machine learning-based MetCCS.[39].