Metabolomics Association of North America
Summer Symposium 2019
hosted by West Coast Metabolomics Center, UC Davis

Quantum Chemistry and Computational Methods for Compound Identification

September 04, 2019
9 a.m. – 7 p.m.
UC Davis Genome Center Auditorium

Keynote speaker
Stefan Grimme,
University of Bonn, Germany
“Quantum chemistry prediction of NMR and EI mass spectra”

Speakers:
Yannick Djoumbou Feunang,
Corteva Agriscience, USA
CFM-ID 3.0, ClassyFire, Biotransformer

Ryan Renslow, PNNL, USA
ISiCL ion mobility prediction

Stephen Stein, NIST, USA
MS and MS/MS libraries, hybrid searches

Jennifer Wei, Google Brain, USA
neural networks for EI mass spectra

Lloyd W. Sumner, University of Missouri, USA
combining MS/MS with NMR, and CCS

Xiuxia Du, University of North Carolina, USA
MS data processing, data deconvolution

Pieter Dorrestein, UC San Diego, USA
GNPS, MS/MS similarity propagation

Zheng-Jiang Zhu, CAS, Shanghai, China
MetDNA, ion mobility, DIA MS

David Grant, University of Connecticut, USA
Ecom50, predicting IR spectra, retention index

Ivana Blazenovic, DiscernDX, USA
compound ID confidence scoring

Supported by: RCN 1743594
Supported by: U2C ES030158
### MANA Summer Symposium 2019

#### Quantum Chemistry and Computational Methods for Compound Identification

**September 04, 2019**

**Program**

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00-9:00</td>
<td>breakfast and registration</td>
<td></td>
</tr>
<tr>
<td>9:00 - 9:10</td>
<td>welcome - Fiehn, Oliver</td>
<td>US</td>
</tr>
<tr>
<td>9:10 - 10:00</td>
<td>Grimme, Stefan</td>
<td>Germany</td>
</tr>
<tr>
<td>10:00 - 10:30</td>
<td>Djoumbou Feunang, Yannick</td>
<td>US</td>
</tr>
<tr>
<td></td>
<td>break</td>
<td></td>
</tr>
<tr>
<td>11:00 -11:30</td>
<td>Renslow, Ryan</td>
<td>US</td>
</tr>
<tr>
<td>11:30-12:00</td>
<td>Stein, Stephen</td>
<td>US</td>
</tr>
<tr>
<td>12:00-12:30</td>
<td>Wei, Jennifer</td>
<td>US</td>
</tr>
<tr>
<td></td>
<td>lunch break</td>
<td></td>
</tr>
<tr>
<td>1:30-2:00</td>
<td>Sumner, Lloyd</td>
<td>US</td>
</tr>
<tr>
<td>2:00-2:30</td>
<td>Du, Xiuxia</td>
<td>US</td>
</tr>
<tr>
<td>2:30-3:00</td>
<td>Dorrestein, Pieter</td>
<td>US</td>
</tr>
<tr>
<td></td>
<td>break</td>
<td></td>
</tr>
<tr>
<td>3:30-4:00</td>
<td>Zhu, Zheng-Jiang</td>
<td>China</td>
</tr>
<tr>
<td>4:00-4:30</td>
<td>Grant, David</td>
<td>US</td>
</tr>
<tr>
<td>4:30-5:00</td>
<td>Ivana Blazenovic</td>
<td>US</td>
</tr>
<tr>
<td>5:00-7:00</td>
<td>reception, posters</td>
<td></td>
</tr>
</tbody>
</table>
“Quantum chemistry prediction of NMR and EI mass spectra”

The GFN-xtB family of semi-empirical tight-binding methods is introduced[1]. The methods follow a global and element-specific parameters only strategy and are consistently parameterized for all elements through radon. Their original purpose and main target for the parameter optimization has been the computation of molecular geometries, vibrational frequencies, and non-covalent interactions. They are effectively used in the framework of meta-dynamics (MTD) to globally explore chemical compound, conformer, and reaction space.[2] For typical conformational search problems of organic drug molecules, the new MTD(RMSD) algorithm yields lower energy structures and more complete rotatmer/conformer ensembles at reduced computational effort. They are used in a fully automated procedure to compute high-resolution 1H-NMR spectra[3]. TB methods combined with the Fermi-smearing technique can also describe difficult electronic structures and covalent bond breaking at least qualitatively correct. This enables the 'first-principles' automated quantum chemistry computation of electron ionization mass spectra even for transition metal complexes[4]. The scope and limitations of various 'low-cost' quantum chemistry methods in typical chemistry applications like pKa or CD spectra calculation[5] are discussed.


Stefan Grimme, PhD studied Chemistry and finished his Ph.D. in 1991 in Physical Chemistry on laser spectroscopy. He habilitated in Theoretical Chemistry in the group of Sigrid Peyerimhoff. In 2000, he got the C4 chair for Theoretical Organic Chemistry at the University of Muenster. In 2011, he accepted an offer as the head of the founded Mulliken Center for Theoretical Chemistry at the University of Bonn. He has published more than 490 research articles and is the recipient of the 2013 Schrödinger medal of the World Organization of Theoretically Oriented Chemists (WATOC). In 2015, he was further awarded the ,Gottfried Wilhelm Leibniz Preis, of the DFG. His main research interests are the development and application of quantum chemical methods for large molecules, density functional theory, noncovalent interactions, and their impact in chemistry.

grimme@thch.uni-bonn.de
Endogenous (metabolic) and exogenous (environmental) transformations are mainly responsible for an estimated 95% of the human chemical exposome (~3 million compounds) that remains largely uncharacterized or unknown. The identification and characterization of these compounds is crucial in understanding the short- and long-term effects that they may have on human and animal health, as well as the environment. Yet, it is a daunting task. In this presentation, I will describe several computational tools that I contributed to developing to enable Metabolomics. In particular, I will present: 1) ClassyFire and ChemOnt, freely available resources for the automated hierarchical structure-based classification of chemicals; 2) BioTransformer, a software tool for predicting metabolic biotransformation products arising from human metabolism and environmental microbial degradation; and 3) CFMID 3.0, the latest release of CFM-ID, a software tool designed to accurately predict mass spectra for rapid compound identification. These tools were developed as part of my research at the University of Alberta, Canada.

Yannick Djoumbou Feunang joined Corteva Agriscience in 2018 as a Data scientist/Cheminformatician, after a PhD and Post-Doc journey at the University of Alberta, Canada. Over the last 10 years, his research interests have focused on exploring the interface between Biology and Chemistry, using Computing Science. In 2018, he earned the Scholarship for Excellence from the Chemical Information division of the American Chemical Society, for his work on in silico metabolism prediction. Moreover, he has worked on in silico MS-spectra prediction and compound identification, Cheminformatics database development, and Biochemical/Biomedical ontology development projects.

Currently, Yannick is a member of the Chemistry Data Science and Informatics (ChemDSI) group at Corteva Agriscience, where he contributes to the development of predictive models, as well as the development of the Cheminformatics workbench, cutting-edge Cheminformatics scientific computing platform.
We are advancing the use of computationally calculated chemical properties to enable the identification of molecules with reduced reliance on the use of data from analysis of authentic reference materials. Our molecular identification pipeline, ISiCLE (In Silico Chemical Library Engine), uses a large-scale computational chemistry platform that exploits PNNL’s high-performance computational quantum chemistry software, NWChem, to calculate small molecule chemical properties, such as collision cross section (CCS) and nuclear magnetic resonance (NMR) chemical shift. In this symposium I will discuss the details of the ISiCLE workflow, as well as report on recent changes on how adduct ionization sites are predicted. I will present several successful application examples, including identifying environmental degradation products, separating molecular isomers, and correcting mislabeled isobaric isosteres. Finally, the challenges and future plans of this approach will be discussed, including a comparisons to recent deep learning approaches.

**Ryan Renslow** is an expert in mathematical modeling, with a focus in biological systems and molecular modeling and quantum chemical calculations. Dr. Renslow has over 50 publications, and expertise in quantitative biology, density functional theory, and biological data analysis automation. A chemical engineer by training, his research has been routinely multi-disciplinary, working alongside microbiologists, ecologists, physicians, dentists, human health researchers, and researchers specializing in metabolomics and exposomics. Ryan Renslow received his B.S. and M.S. and his PhD in Chemical Engineering at the Washington State University, Pullman, WA. In 2010 he was the NIH Protein Biotechnology Fellow, Washington State University, Pullman, WA. Ryan Renslow received 2015 the Linus Pauling Distinguished Postdoctoral Fellowship at PNNL, Richland, WA.
There is wide agreement that the inability to reliably identify a large fraction of the ions generated in mass-spectral-based metabolomics studies greatly hinders progress in this field. While identification through libraries can, in principle, provide a solution, libraries are, and are likely to remain too limited in scope to provide a solution to this problem. Currently, perhaps 10% of the ions observed in urine or plasma can be identified by libraries. However, current libraries contain an enormous amount of information linking chemical structures to spectra that can be exploited identify compounds absent from the library. This is done by, in effect, making use of fragmentation pathways of for spectra present in the library to tentatively compounds not present in the library. We discuss two such methods: 1) the ‘hybrid search’ which finds compounds that differ from a library compound by an ‘inert’ chemical group and 2) our MS Interpreter program which validates the correctness of a spectrum/structure pair using basic ion fragmentation principles and thermochemistry, both derived from library spectra.

Stephen E. Stein, PhD is NIST Fellow, and Director of the NIST Mass Spectrometry Data Center in the Biomolecular Measurement Division, Material Measurement Laboratory, National Institute of Standards and Technology. Dr. Stein conceives, plans, directs and personally conducts advanced research and development in chemical data analysis, with emphasis on mass spectral data and search algorithms, deconvolution algorithms, quality measurement and new approaches for chemical identification in proteomics and metabolomics, methods for automated chemical data analysis, algorithms for spectra and property prediction from chemical structural data.

Dr. Stein received his BS in Chemistry from U Rochester, NY, in 1969, and his PhD in Physical Chemistry from U Washington in 1974. He was Associate Professor in Chemistry at West Virginia U until 1982 and serves as director of the NIST MS data center since 1988.
We propose to improve the library’s coverage by augmenting it with synthetic spectra that are predicted from candidate molecules using machine learning. We contribute a lightweight neural network model that quickly predicts mass spectra for small molecules, averaging 5 ms per molecule with a recall-at-10 accuracy of 91.8%. Achieving high-accuracy predictions requires a novel neural network architecture that is designed to capture typical fragmentation patterns from electron ionization. We analyze the effects of our modeling innovations on library matching performance and compare our models to prior machine-learning-based work on spectrum prediction.

Jennifer Wei is a software engineer with the Brain Research team in Cambridge, MA. She received her PhD in Chemical Physics at Harvard University. Her primary research interests are the applications of machine learning for small molecules towards predicting molecular properties and chemical phenomena. She has published research on applications of machine learning to reaction prediction, inverse design of molecules, and mass spectrometry prediction.

weijennifer@google.com
Lloyd W. Sumner, U Missouri, USA

“Development of Integrated Computational and Empirical Tools to Address the Metabolomics Grand Challenges of Confident Metabolite Identification and Increased Depth-of-Coverage”

The vast utility of metabolomics is well documented in the literature; however, its full scientific promise has not yet been realized due to multiple technical challenges. These grand challenges include large-scale confident metabolite identifications and greater depth of coverage. We have developed sophisticated spectral, computational and integrated experimental metabolomics tools for the systematic and biologically directed annotation of plant metabolomes and for greater metabolome depth of coverage. This presentation will describe a UHPLC-QTOF-MS/MS mass spectral library, custom software entitled Plant Metabolite Annotation Toolbox (PlantMAT), software for CCS prediction, sophisticated UHPLC-timsTOF-MS/MS and UHPLC-QTOF-MS-SPE-NMR instrumental ensembles that are being used for ‘sequencing’ the first plant metabolomes.

Lloyd W. Sumner acquired his Ph.D. in analytical chemistry from Oklahoma State University in 1993. He joined the Samuel Roberts Noble Foundation in 1999 and rose to the rank of Professor within the Plant Biology Division. He then moved to the University of Missouri, Columbia in January 2016 as a Professor in the Biochemistry Department and Director of the MU Metabolomics Center. Dr. Sumner’s research program focuses upon the development, integration, and application of large-scale biochemical profiling technologies to better understand plant specialized metabolism. Dr. Sumner’s research is or has been graciously supported by the University of Missouri, Noble Foundation, NSF 2010, NSF MCB, NSF MRI, NSF-JST, NSF-IOS, NSF-PGRP and The Oklahoma Commission for the Advancement of Science and Technology. Dr. Sumner is currently an AAAS Fellow; Former Treasurer and President of the Metabolomics Society, Lifetime Honorary Member of the Metabolomics Society, Distinguished Alumni of Cameron University, and President Emeritus of the Phytochemical Society of North America.

sumnerlw@missouri.edu
Spectral deconvolution is an essential step in preprocessing untargeted GC-MS metabolomics data. It computationally separates ions that are in the same mass spectrum but belong to coeluting compounds that are not resolved completely by chromatography. As a result of this computational separation, spectral deconvolution produces pure fragmentation mass spectra. Each mass spectrum consists of ions from a single compound, allowing identification or annotation of compounds through searching spectral libraries.

Traditionally, spectral deconvolution has been achieved by using a two-step model peak approach. The first step detects the presence of components and selects a model peak for each perceived component. At this stage, a component generally consists of multiple chromatographic peaks that have a very similar peak shape. From these peaks, a model peak is selected that can best represent the elution profile of the component. The second step decomposes each detected peak into a linear combination of the model peaks and constructs a pure fragmentation spectrum for each perceived component. This two-step approach is computationally efficient and provides model peaks that are similar to real chromatographic peaks in shape. However, there is always the risk that a selected model peak has actually been produced by two or more co-eluting components and therefore is inappropriate to serve as a model peak. This inappropriate model peak selection would cause incorrect pure fragmentation spectra for all the involved co-eluting compounds and eventually errors in library matching. To address the limitations of the traditional approach, we have developed a multivariate curve resolution (MCR)-based method. At this symposium, I will describe the fundamental differences between the two approaches and how we have implemented the MCR-based method while overcoming the inherent computational complexity.

**Xiuxia Du** is an expert in metabolomics informatics, statistics, and machine learning. In particular, she is focused in developing computational algorithms and software tools for preprocessing mass spectrometry-based metabolomics data. The suite of ADAP algorithms and the associated software tool that her team have developed provide metabolomics researchers an alternative informatics approach to preprocess their untargeted LC-MS and GC-MS metabolomics data. The source code of all of the algorithms and the related software tools is publicly available on Github for close scrutiny by the metabolomics community.

Prof Xiuxia Du received her B.S. and M.S. in Electrical Engineering at the Hefei University of Technology, China. In 2005 she finished her PhD at Washington University in St. Louis and started her postdoc at PNNL in Computational Proteomics. Professor Du is since 2008 a Professor at the Department of Bioinformatics & Genomics University of North Carolina at Charlotte.

xiuxia.du@uncc.edu
Pieter Dorrestein, UC San Diego, USA

“Mass Spectrometry Annotation and Analysis at the Repository Scale”

This lecture will describe the emerging approaches that are being developed by the GNPS user community to annotate molecules not just at the individual compound level but also using different levels of information at the repository scale. It will describe the logistical challenges and informatics solutions to enable comparison of data from different labs, often collected on different instruments, but also computational strategies for discovery and annotation of biologically relevant molecular ion partners of small molecules, another level of annotation. Finally it will describe the importance of FAIR principles to improve the amount of information that can be obtained using mass spectrometry.

Pieter Dorrestein is an Associate Professor at the University of California - San Diego. He is the Director of the Therapeutic Discovery Mass Spectrometry Center and a Co-Director, Institute for Metabolomics Medicine in the Skaggs School of Pharmacy & Pharmaceutical Sciences, Departments of Pharmacology, Chemistry and Biochemistry. Dr. Dorrestein was trained by Tadgh Begley in the chemical biology of enzymes involved in vitamin biosynthesis and by Neil Kelleher and Christopher Walsh, whom were co-sponsors of his NRSA postdoctoral fellowship, in Top and Middle down mass spectrometry on proteins that made small molecules of therapeutic value. Since his arrival to UCSD in 2006, Dr. Dorrestein has been pioneering the development of mass spectrometry methods to study the chemical and ecological crosstalk between population of organisms for agricultural, diagnostic and therapeutic applications.

pdorrestein@ucsd.edu
Zheng-Jiang Zhu, CAS, Shanghai, China

Interdisciplinary Research Center on Biology and Chemistry, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Shanghai, 200032 P. R. China

“MetDNA: Metabolic Reaction Network-based Recursive Metabolite Annotation for Untargeted Metabolomics”

Large-scale metabolite annotation is a challenge in liquid chromatogram-mass spectrometry (LC-MS)-based untargeted metabolomics. Here, we develop a metabolic reaction network (MRN)-based recursive algorithm (MetDNA) that substantially expands metabolite annotations without the need for a comprehensive standard spectral library. Metabolites and their reaction-paired neighbor metabolites tend to share similar MS2 spectra due to structural similarity. Based on this rationale, MetDNA characterizes initial seed metabolites using a small tandem spectral library, and utilize their experimental MS2 spectra as surrogate spectra to annotate their reaction-paired neighbor metabolites which are subsequently served as the basis for recursive analysis. We further showcase the utility and versatility of MetDNA using different LC-MS instrumentations, data acquisition methods, and biological sample types, and demonstrate that about 2,000 metabolites can cumulatively be annotated from one experiment. MetDNA largely expands the annotation of metabolites, thereby allowing quantitative assessment for not just metabolic pathways but also multi-omic studies, such as integrative analysis between metabolomics and transcriptomics.

Zheng-Jian Zhu received his B.S. degree in Chemistry from Nanjing U in 2006, and his Ph.D. degree in Chemistry from U Massachusetts at Amherst in 2011. He worked as postdoctoral researcher with Gary Siuzdak at Scripps from 2011 to 2013. From 2013, he was appointed as Professor in Shanghai Institute of Organic Chemistry (SIOC) and Interdisciplinary Research Center on Biology and Chemistry (IRCBC), Chinese Academy of Sciences (CAS).

Dr. Zhu develops mass spectrometry based metabolomics technology with applications in health and disease related research. Dr. Zhu has published >50 peer-reviewed papers on the prestigious journals including Nature, Nature Chemistry, Nature Biotechnology, Nature Communications, JACS, eLife, Analytical Chemistry and Bioinformatics with >3000 times in citation. Dr. Zhu is supported by the Thousand Youth Talents Program. See http://www.zhulab.cn for more information.

jiangzhu@sioc.ac.cn
David Grant, U Connecticut, USA

“Predicting retention times, Ecom50, CCS, IR and MSMS spectra”

The long-term objective of the research in my lab is to develop analytical, computational and database tools that can be used to rapidly identify the chemical structures of metabolites in human biofluids. Our computational algorithms predict physical/chemical properties of compounds contained in large chemical databases. The physical/chemical properties chosen are those that can be experimentally determined for any unknown compound by HPLC-mass spectrometry. These include retention indices, collision induced dissociation fragmentation spectra, Ecom50, collision-cross section and infrared spectra. Compounds in databases (for example PubChem) whose predicted properties most closely match experimental properties are prioritized as likely candidates for the unknown. By facilitating the structural identification of unknown chemical compounds these tools will enhance the ability of metabolomics to compliment and synergize other areas of biomedical research.

David Grant received BS degrees in Biology and Chemistry from Idaho State University in 1979 and a PhD in Environmental Toxicology from Michigan State University in 1987. After postdoctoral training in molecular toxicology at The University of California-Davis, he began a tenure-track position in 1995 in the Department of Pharmacology and Toxicology at the University of Arkansas for Medical Sciences. He moved to the University of Connecticut in 2001 and is currently a Professor in the Department of Pharmaceutical Sciences. His research is focused on developing analytical and computational approaches to determine the structure of unknown metabolites in human biofluids. Dr. Grant has served as a member of The American Society for Pharmacology and Experimental Therapeutics, The International Society for the Study of Xenobiotics, The American Society for Mass Spectrometry and The Metabolomics Society. He serves on NIH study sections in the areas of metabolomics and analytical biochemistry.

david.grant@uconn.edu
The annotation of small molecules remains a major challenge in untargeted mass spectrometry-based metabolomics. Critical discussion is needed on structured elucidation approaches and software that are designed to help during the annotation of unknown compounds. Only by elucidating unknown metabolites first is it possible to biologically interpret complex systems, to map compounds to pathways and to create reliable predictive metabolic models for translational and clinical research. We present in silico fragmentation tools such as MS-FINDER, CFM-ID, MetFrag, ChemDistiller and CSI:FingerID that can annotate compounds from existing structure databases and that have been used in the CASMI (critical assessment of small molecule identification) contests. The use of retention time models from liquid chromatography and the utility of collision cross-section modelling from ion mobility experiments are discussed. Furthermore, we present annotation of all spectra in a sample using variety of freely available m/z-RT databases, mass spectral libraries (MoNA, NIST17, LipidBlast, CarniBlast etc).
Generating in silico electron ionization GC-MS spectra of biologically important purine and pyrimidine derivatives using QCEIMS

Jesi Lee (1, 2), Tobias Kind (1), Lee-Ping Wang (2), Oliver Fiehn (1)

(1) West Coast Metabolomics Center, University of California, Davis, 451 Health Sciences Drive, Davis, California 95616, United States
(2) Chemistry Department, University of California, Davis, One Shields Avenue, Davis, California 95616, United States

Mass spectrometry is one of the most commonly used methods for compound identification in metabolomics. However, many metabolites do not have reference mass spectra available for correct annotation. In the attempt to solve this problem, theoretically predicted mass spectra can be used as a substitute for experimental spectra. The Quantum Chemistry Electron Ionization Mass Spectra (QCEIMS) program developed by the Grimme group can predict 70 eV electron ionization mass spectra from any given input structure.

QCEIMS produces ensemble of conformers from the ground state molecular dynamics and computes the molecular orbital spectrum by performing a single-point calculation and Mullikan population analysis. In this work, we generated in silico electron ionization GC-MS spectra of a diverse collection of biologically important purine and pyrimidine derivatives using QCEIMS. Purines and pyrimidines are building blocks for RNA and DNA synthesis and members of several important metabolic pathways. We investigated the accuracy of predicted spectra by comparing the in silico spectra to the experimental spectra in the NIST17 database.

The cosine and dot product similarity scores were calculated to assess the prediction accuracy as well as rank ordering. We were able to obtain a number of molecules with dot product similarity scores above 850, which is usually sufficient for initial compound annotations. Other metabolites had cosine similarity score above 600. For some compounds the main fragmentation patterns and overall m/z distribution were accurately predicted by the computational spectra; however the predicted intensities were less accurate. Despite of the complexity of structures, QCEIMS was able to predict good quality spectra. Overall, the result showed that QCEIMS is a promising method to assist in compound identification.
Generation of in-silico electron ionization mass spectra using quantum chemistry and conformational flexibility investigation

Shunyang Wang (1, 2), Tobias Kind (1), Oliver Fiehn (1)

(1) West Coast Metabolomics Center, University of California, Davis, 451 Health Sciences Drive, Davis, California 95616, United States
(2) Chemistry Department, University of California, Davis, One Shields Avenue, Davis, California 95616, United States

Background: Mass spectrometry plays an important role in metabolomics and electron ionization (EI) is one of the standard ionization methods in GC-MS. It is important to understand the mechanism behind the associated gas phase reactions and improve ways to simulate EI mass spectra. The Quantum Chemical Electron Ionization Mass Spectrometry (QCEIMS) method developed by Grimme allows for the investigation of gas phase reaction trajectories and can simulate 70 eV mass spectra. In order to improve the accuracy of these in-silico mass spectra, it is important to understand the factors that influence the final results. We choose the conformational flexibility of molecules as an initial starting point for our investigation. Using a-priori knowledge in our simulation will allow us to improve the quality of in-silico mass spectra by modulating specific calculational parameters.

Methods: We selected 464 molecules based on chemical descriptors including the Kier flexibility index (PHI), the number of rotatable bonds (RBN) and rotatable bond fraction descriptor (RBF). Descriptors were calculated using the AlvaDesc program. We initial 3-D structures using the MMFF94 force field within the Avogadro software. We passed molecule structures to QCEIMS and used semi-empirical method (OM2) to conduct molecular dynamics and density functional theory (DFT PBE0/SV(p)) to calculate the ionization potential. After simulation of several hundred molecular dynamics trajectories, the ions are used to calculate m/z peaks based on molecular weight calculations of the fragment formulas and peak abundances are generated by counting the occurrence of specific fragments across several hundred independent trajectories. This information is then used by QCEIMS to calculate a simulated mass spectrum. In order to improve our simulations, we also used several starting conformers. The obtained in-silico spectra were compared against the NIST17 library using cosine and dot-product similarities.

Results: The average simulation time per molecule is 1.5 hours for molecules less than 130 Da using 44 CPU cores. For 214 simulated compounds the mass spectral similarity score was larger than 700, while 151 spectra had a score of less than 500. Similarity scores ranging from 850-1000 are usually considered sufficient enough for compound identification.; Therefore, a-priori knowledge that can be directly calculated from a molecular structure would be useful to generate a predictive quality estimate. The conformational flexibility of molecules containing the elements carbon, hydrogen and oxygen, including caged inflexible structures such as adamantane and very flexible alkanes had no influence on the quality of spectra. Future investigations will probe additional QCEIMS parameters to allow for simulation of high quality in-silico mass spectra.
Probabilistic annotation of carbohydrate structures facilitates systematic organization of information on carbohydrates as inflammatory biomarkers

Hesam Dashti\textsuperscript{1,2}, Olga Demler\textsuperscript{1}, Jonathan R. Wedell\textsuperscript{2}, William M. Westler\textsuperscript{2}, Hamid R. Eghbalnia\textsuperscript{2}, John L. Markley\textsuperscript{2}, and Samia Mora\textsuperscript{1}

\textsuperscript{1} Department of Medicine, Brigham and Women’s Hospital, Harvard Medical School, Boston, MA
\textsuperscript{2} Department of Biochemistry, University of Wisconsin, Madison, WI

Introduction. Glycans and composite carbohydrates are routinely studied as markers of systemic inflammation and cardiometabolic diseases. Many of these are attached to macromolecules, and the chemical composition and stereochemistry of these carbohydrate moieties underlie their biomedical activities. Oligosaccharides may undergo major structural modifications that make it challenging to identify their parent carbohydrate (polyhydroxy aldehydes or ketones) fragments. We hypothesize that advanced cheminformatics methods can provide a robust platform for identification of carbohydrates and annotating their stereochemistry.

Methods. We consider the parent monosaccharides as the basis and calculate a pair of penalties for every possible molecular fragment in a given molecule. These penalties compensate for two factors: (1) differences between a molecular fragment and its corresponding parent, and (2) the number of atoms in a given molecule that constitute non-saccharide ligands. The overall penalty scores are then used to calculate a probability of “carbohydrate similarity” for the molecular fragments.

Results. We evaluated our approach through analysis of the structures in the Ligand Expo dataset, which is used by the Worldwide Protein Data Bank to annotate ligands bound to macromolecules. We compared the results of our method with the manually annotated contents of the Ligand Expo, which indicated our method could identify nearly an order of magnitude more ligands as of carbohydrate origin than were annotated as such in the Ligand Expo.

Conclusion. We demonstrate a novel method for high throughput structural annotation of carbohydrates. The more robust annotations achieved by our approach can be used to correlate information concerning biomarkers of interest with metadata (e.g., ontology, and bioactivities) and experimental data (e.g., from X-ray crystallography, nuclear magnetic resonance spectroscopy, and mass spectrometry).
Automatic molecular annotation of mass spectrometry imaging data

Jan H. Kobarg1; Nikolas Kessler2; Wiebke Timm2; Janina Oetjen2; Klaus Steinhorst1; Stefan Schiffler1; Shannon Cornett3; Aiko Barsch2; Heiko Neuweger2; Alice Ly2; Dennis Trede1; Na Parra4; Xuejun Peng1

1SciLS, Bremen, Germany; 2Bruker Daltonik GmbH, Bremen, Germany; 3Bruker Daltonics Inc., Billerica MA, USA; 4Bruker Daltonics Inc., San Jose CA, USA

Mass spectrometry imaging (MSI) has emerged as a promising analytical technique to spatially resolve components of metabolic processes. However, bioinformatics remains a challenging bottleneck for metabolomics imaging studies, particularly in untargeted metabolomics imaging where a gap between extracting the m/z of an image and identifying the underlying metabolite and projecting it into a biological context.

A novel approach to automatically annotate metabolites and lipids of MALDI-imaging data was established based on accurate mass, isotopic pattern fit, retention time (if available), mass spectral similarity and ion mobility: (i) spectra are statistically compared and discriminative m/z-markers; (ii) peaks are annotated by molecular formula generation or targeted methods; (iii) spatial patterns of annotated peaks are compared. Images of metabolites and lipids from fresh-frozen rat tissues of brain, kidney and testis were collected in positive ionization mode at spatial step sizes ranging from 20-50 μm² pitch.

First, MALDI MSI data was visualized and regions of interest (ROIs) were defined both manually, using co-registered stained microscopy data, and automatically, spatial segmentation by means of spectral clustering. For the ROI’s, multiple ions were identified as discriminative between different ROIs by means of Receiver Operating Characteristic (ROC) analysis. For the rat brain lipids data set, 62 peaks were selected for further analysis; Second, subsamples of the spectra in each ROI were deisotoped incl. fine structure and adducts were detected yielding only putative monoisotopic metabolite peaks; Third, peaks were annotated using untargeted (molecular formula generation) or targeted (e.g. HMDB, LipidMaps) annotation methods. For the rat brain lipids data set, 23 peaks were annotated, five from molecular formula generation, four from targeted matching against a list of lipids reportedly in brain tissue and 14 from targeted matching against the LipidMaps database. Among these were the following previously reported in literature: SM(36:1) – 731.60616 m/z, PC(34:1) – 760.58495 m/z and PC(36:1) – 788.61574 m/z. This annotation step can be integrated with measurements from chromatographic retention times and ion mobility cross section and MS/MS to improve annotation confidences.
Integrating 4D peak picking of LC-TIMS-MS/MS data into GNPS feature based molecular networking for metabolomics and lipidomics analysis.

Florian Zubeil\textsuperscript{1}; Nikolas Kessler\textsuperscript{1}; Heiko Neuweger\textsuperscript{1}; Sven Meyer\textsuperscript{1}; Ulrike Schweiger-Hufnagel\textsuperscript{1}; Aiko Barsch\textsuperscript{1}; Xuejun Peng\textsuperscript{2}; Na Parra\textsuperscript{2}

\textsuperscript{1}Bruker Daltonik GmbH, 28359 Bremen, Germany; \textsuperscript{2}Bruker Daltonics Inc., San Jose CA, USA

As throughput of metabolomic and lipidomic analyses continuously expands, effective workflows for analyzing the resulting datasets are of increasing importance. Recently molecular networking has become a vital tool in the metabolomics community as it quickly allows the identification of compounds with similar fragmentation patterns which are often structurally related. It also allows propagation of annotations from known compounds to related derivatives. While this approach mainly focusses on the fragment spectra, important information can be deduced from the precursor spectra, i.e. intensity, accurate mass and isotopic pattern of the analytes. Herein, a workflow to integrate analyte information for untargeted profiling from the software MetaboScape into GNPS feature based molecular networking was proposed.

A lipid extract was generated from SRM1950 reference plasma and metabolite extracts of 15 hop pellets were obtained by ethanolic extraction. Samples were analyzed by timsTOF Pro (Bruker) operated in PASEF mode (4D LC-TIMS-QTOF-MS/MS-dataset) and impact II (Bruker) (3D LC-QTOF-MS/MS-dataset). Initial data processing (peak picking, deisotoping, de-adducting and consensus spectra creation) was performed in a novel version of MetaboScape (Bruker). Lipid annotation was performed in SimLipid (Premier Biosoft). A mgf file of consensus spectra was created using a dedicated GNPS file export in MetaboScape which was subjected to feature based molecular networking on the GNPS website (https://gnps.ucsd.edu). The resulting networks were analyzed in CytoScape 3.6.1 with additional attributes exported from MetaboScape.

Two sample sets of ethanolic extracts of different hop pellets (3D workflow) and lipid extract of SRM1950 reference plasma (4D workflow) were analyzed. This unique workflow allows the integration of three- and four-dimensional Time aligned Region complete eXtraction (T-ReX) peak picking results as well as annotation workflows of MetaboScape with GNPS feature based molecular networking. The nodes in the resulting molecular network are enriched by useful information about the precursor ions like the intensity in individual samples, molecular formula, annotation, CCS values, group mean and maximum intensity. The latter information are important indicators to assess distribution of a specific analyte between sample groups (by group mean). Additionally, the maximum intensity can help to determine if a purification of the analyte is feasible. Likewise, interpretation of the resulting molecular network is greatly simplified by displaying generated molecular formulas instead of precursor masses as node labels. For hop extracts, this approach allowed for the detection of several previously unidentified compounds that are predominant in only some hop types and share the same cluster with Xanthohumol, indicating their chemical similarity. Additionally, this workflow enables the integration of 4D peak picking results from PASEF-MS/MS data into GNPS molecular networking which is demonstrated on a lipid sample. This enables a straightforward processing of TIMS-MS data including the benefit of cleaner MS/MS spectra of co-eluting analytes for molecular networking which are generated by ion mobility separation.
Small Molecule Accurate Recognition Technology (SMART) to accelerate structural determination of metabolomics

Chen Zhang
UC San Diego
chz023@ucsd.edu

We present the Small Molecule Accurate Recognition Technology (SMART) 2.0, a system that integrates the benefits of 2D HSQC NMR with advances in deep learning to enhance and improve the efficiency of metabolomic research.

This tool is highly effective in both assisting metabolomic dereplication efforts as well as the automatic identification of a new compound as belonging to a particular compound family. To effectively accomplish this goal, we, a deep Convolutional Neural Network (CNN) with autoencoder was trained on a constructed database containing over 30,000 HSQC spectra as the training set. This resulted in a remarkable and highly accurate 3D clustering of different classes of molecules based on their respective HSQC spectra.

To demonstrate the utility of SMART, multiple newly isolated compounds were automatically located with their known analogues in the embedded clustering space, thereby streamlining the discovery pipeline for new natural products. The result shows that we can greatly accelerate the rate of known compound identification as well as rapidly generating hypotheses about the relationship of new molecules to those used for the training - based entirely on their NMR properties. In addition, because white Gaussian noise, impurities, or solvent effects are often seen in experimental HSQC spectra, we investigated the robustness of the SMART 2.0 to recognize HSQC spectra in the presence of significant noise or artifacts.
Re-use of Public Data:  
Repository Assisted Analysis of Untargeted LC-MS Data

Christine Aceves  
UC San Diego  
caceves@ucsd.edu

One of the current challenges facing the field of untargeted liquid-chromatography mass spectrometry (LC-MS) is the ability to determine the significance of a signal with regards to a specific biological question. Ideally, one could use the vast amount of data currently sitting in public repositories to help make informed decisions about the relevance of a peak, however several factors prevent this from occurring.

First, chromatographic differences such as gradient and column impede samples being compared across datasets. Second, although many samples are available publically, a lack of metadata standardization and availability leaves much of them unworkable. Using curated metadata, over 32,000 samples were summed over time on a per file basis using MS1 data to form representative m/z vectors. Once samples were standardized and stripped of retention time, we developed a pipeline in which data can be pulled from the repository, using the metadata, to assist experiment scale analysis in a statistics driven way. By picking a target sample group (such as a disease state) and comparing it iteratively to subsets of the repository we have developed a method for returning a putative list of masses unique to said target group.

Applying MS2 analytical techniques such as molecular networking and MASST to this focused list of precursor masses can then begin to allow us to answer important biological questions in an efficient and reproducible manner.
Identifying oxidized triacylglycerides and inflammatory lipids in plasma of tuberculosis patients

Manoj Khadka
Emory University
manoj.khadka@emory.edu

Tuberculosis is one of the top ten causes of death worldwide and is the top cause of mortality by a single infectious agent. Approximately 1.6 million people died of TB in 2017 among which roughly 18.75 % were HIV infected. Despite the global decline in incidence, the decline needs to be increased to 4-5% to reach the 2020 milestones of the End TB Strategy. It is thus vital to understand TB pathogenesis and disease progression mechanisms to develop new diagnostics for early detection and drugs to combat multidrug resistant TB.

Here we utilized an untargeted lipidomics approach to study changes in plasma lipidome in negative, latent and active tuberculosis cohorts and attempted to gain insight into the disease pathogenesis based on lipid modulation among different cohorts. Lipids were extracted and analyzed via liquid chromatography mass spectrometry and the data were statistically treated to identify the difference in lipid profiles among different TB cohorts using univariate and multivariate statistical approaches. Principal component analysis revealed clustering of data by negative, latent and active tuberculosis groups.

Major differences were observed in phospholipids, di- and triglycerides. Most of the lysophospholipid species and oxidized triglycerides were significantly more abundant in active tuberculosis and phosphatidylcholine and triglyceride species were found to be significantly higher in latent tuberculosis cohorts. Bile acids, eicosanoids and docosanoids among others were observed to be significantly abundant in active tuberculosis. Based on the findings, we suggest that bacterial fatty acid utilization pattern can be observed in active TB at the plasma level. The observation of oxidized and inflammatory lipids modulation implies their role in tuberculosis immunopathogenesis. This study adds additional understanding of immunopathogenesis between latent and active tuberculosis and furthers our understanding of potential biomarkers of disease progression.
Associations between metabolites in aqueous humor and plasma in mice investigated by non-targeted GC/MS

Michael J. Muehlbauer¹, James R. Bain¹, Peter Buchwald², Midhat Abdulreda²

¹ Duke Molecular Physiology Institute, Durham, NC
² Diabetes Research Institute, University of Miami Miller School of Medicine, Miami, FL

The aqueous humor (Aq.H) in the anterior chamber of the eye (ACE) offers a unique microenvironment for the assessment of locally enriched metabolites, not all of which detectible in plasma, and may be advantageous for the implantation of pancreatic islets (Abdulreda et al., Diabetologia 2019).

As a first step towards establishing this novel concept, we used non-targeted metabolomics by gas chromatography-mass spectrometry (GC-MS) to evaluate metabolite levels in mouse Aq.H compared against simultaneously collected plasma. Samples were obtained from 1) C57BL/6 mice (three groups of males and three groups of females, 5 per group), and 2) NOD mice (five groups, all males, 3-7 per group, and generally nondiabetic), with Aq.H from each group pooled for analysis. All samples were analyzed using an Agilent 5977B GC/MSD with data deconvoluted and annotated by AMDIS using a spectral retention time-locked library modified from Fiehn (Anal Chem 2009).

Features were aligned across samples with intensities evaluated as the log-2 transformed integrated peak area. Other features that could not be annotated from our library were aligned by spectral similarity between samples using SpectConnect (Styczynski et al., Anal Chem 2007). Results indicated significant positive correlations between the Aq.H and plasma in the average intensities of many metabolites, several of which have known relevance to islet function and diabetes. The study further highlights the potential utility of metabolomics analyses for identifying specific biomarkers in Aq.H that could have possible diagnostic and prognostic value in diabetes and other diseases.
Metabolomics Approach to identify the changes in Aroma Volatiles in Climacteric and Non-Climacteric Melons Grown seven locations

Rita Metrani1,9, G.K. Jayaprakasha1,9, Kevin Crosby1,9, John Jifon3,9, Paul Brierley3,9, Sadhana Ravishankar3,9, Daniel Leskovar4,9, Tom Turini5,9, Jonathan Schultheis6,9, Tim Coolong7,9, Wenjing Guan8,9, and Bhimanagouda S. Patil1,9

1Vegetable & Fruit Improvement Center, Texas A&M AgriLife, College Station, TX 77845; 2Texas A&M AgriLife Research and Extension Center, Weslaco, TX 78596; 3University of Arizona, Tucson, AZ 85721; 4Texas A&M AgriLife Research and Extension Center, Uvalde, TX 78801; 5University of California, Fresno, CA 93710; 6North Carolina State University, Raleigh, NC 27695; 7University of Georgia, Tifton, GA 31793; 8Purdue University, Vincennes, IN 47591; 9National Center of Excellence for Melon at the Vegetable and Fruit Improvement Center, Texas A&M University, College Station, TX 77845.

Cucumis melo fruits are highly valued for its sweet and refreshing flesh; however, the flavor is highly influenced by aroma as dictated by volatile organic compounds. These compounds are the most important factors of fruit quality as perceived by consumers. Using the optimized GC-MS-SPME method, the flavor profiles of four cantaloupes and three honeydew varieties from seven locations were investigated. In the present study, 208 volatile compounds were identified and quantified, which includes esters, alcohols, aldehydes, hydrocarbons, ketones, sulfur containing compounds and acids including fatty acids. Principal component analysis was conducted to understand the multi-location effect on melon varieties and difference between cantaloupes and honeydews. Ethyl butanoate, ethyl 2-methylbutanoate, 2-methylbutyl acetate, ethyl hexanoate, hexyl acetate, benzyl acetate are the main esters associated with fresh, fruity, sweet aroma-flavor attributes which are high in Western Shipper, Tuscan type-Da Vinci, Harper type, and commercial preferred varieties. Aldehydes and alcohols dominate the volatile profile of honeydew melon due to green grassy aroma and flavor. (E)-2-Nonenal (odorant in cucumber), (E,Z)-2,6-nonadienal (cucumber aldehyde) and (E,Z)-3,6-nonadien-1-ol imparts green cucumber-like flavor and aroma to the melon which is high in honeydew and certain cantaloupe varieties. Several antimicrobial compounds α-terpinene, limonene, 1,8-cineole, β-phellandrene, citronellal, decanal, α-copaene, benzaldehyde, (E)-2-decenal, δ-cadinene, α-farnesene, 1-decanol, eugenol, cadinol, thymol, α-eudesmol, β-eudesmol, geranylacetone and β-ionone have been identified and quantified which could help in preventing the bacterial outbreaks in melons. The levels of the antimicrobial compounds were high in certain cantaloupes. Overall Tuscan type-Da Vinci and Western Shipper ranked highest among cantaloupe varieties. This study would help to determine the influence of growing conditions and melon cultivars to understand the levels of amino acids. This project based upon the work supported by NIFA-SCRI- 2017-51181-26834 through National Center of Excellence for Melon at the Vegetable and Fruit Improvement Center of Texas A&M University.
Variation in amino acids profiles in different varieties of cantaloupe grown in seven locations at different states was investigated. Results demonstrated that melons are excellent sources of essential amino acids and neurotransmitters. The levels of amino acids varied among the four varieties grown in different environmental conditions. Tuscan type (TT-DV) harvested from Weslaco-TX and North Carolina had significantly higher levels of γ-aminobutyric acid (GABA), inhibitory neurotransmitters which plays a significant role to prevent excitotoxicity. It content varies from 1006±43.1 to 3187±50.8 μg/g depending on the varieties. Similarly, other essential amino acids also showed significant variation in their levels. Valine, essential amino acid important for muscle synthesis, and function of nervous system. It levels varies from 303.9±16.4 to 541.5±27.0 μg/g and significantly higher in TT-DV variety from Arizona and North Carolina followed by Weslaco-TX. In addition, Western Shipper (F-39) from Indiana had significantly higher followed by Weslaco-TX. Likewise, phenylalanine content was significantly higher in TT-DV melon harvested from Uvalde-TX and Indiana. There was no significant difference in tryptophan among the different locations. This study would help to determine the influence of growing conditions and melon cultivars to understand the levels of amino acids. This project based upon the work supported by NIFA-SCRI- 2017-51181-26834 through National Center of Excellence for Melon at the Vegetable and Fruit Improvement Center of Texas A&M University.
**Purpose**: To determine whether lysophospholipid profiles are altered in the optic nerve (ON) between human control and glaucoma samples, and whether conversion enzymes in lysophospholipid pathways are also altered in glaucomatous ON compared to controls.

**Methods**: Lipid extraction of control (n = 11) and glaucomatous (n = 12) ON samples was conducted using Bligh and Dyer method. High-resolution mass spectrometry was performed using a Q-exactive mass spectrometer coupled with an Accela 600 HPLC system. Analysis was performed for lysophospholipids (lyso-phosphatidylcholines, lyso-phosphatidylserines, lysophosphatidylethanolamines, lysophosphatidylinositol, lyso-sphingomyelins) using LipidSearch v.4.1 and MzMine v. 2.0. Further bioinformatics analyses were performed using MetaboAnalyst 4.0 and Graphpad Prism 8. Lysophospholipid synthesis and degradation pathway maps were created using Kyoto Encyclopedia of Genes and Genomes (KEGG) based tools.

The enzymes’ specifics were found in UniProt and BRENDA databases. The mRNA expression level in normal and glaucomatous human ON was determined from Gene Expression Omnibus (GEO) entry GSE45570 [3] and analyzed in Graphpad Prism 8. Protein amounts were determined using Phastgel and Dot Blot and were used for normalization of protein amounts across samples. Western blot and ELISA were performed using established protocols.
Using NMR and FT-ICR for bacterial and fungal metabolomics

Teresa Palazzo¹
Postdoctoral scholar
Pacific Northwest National Laboratory
teresa.palazzo@pnnl.gov

Metabolomics research at EMSL at Pacific Northwest National Lab (PNNL) encompasses a wide range of environmental topics. Active areas of research span from metabolic flux analysis of antibacterial resistant E. coli, all the way to fungal lignin catabolism with the long-term goal of valorizing this abundant and renewable, yet recalcitrant material.

We utilize high field Nuclear Magnetic Resonance (NMR) and Fourier Transform Ion Cyclotron Resonance Mass Spectrometry (FTICR MS) to study these complex systems. Our current efforts are focused on updating and improving our metabolomics workflows, to create a more efficient, reproducible, and high-throughput framework in order to keep up with increasing demand. Several current exemplar projects are highlighted, and each is tagged with a ‘call-to-arms’ of sorts – we need you to help us determine how we might use quantum mechanics and other computational tools to enhance our workflows and help us address these complex questions.
A combined NMR and LC-MSE metabolomics approach to investigate mechanism of action of antitubercular compounds

Isin Tuna Sakallioglu1, Denise Zinniel2, Boone Evans3, Patrick Dussault3, Raul Barletta2, Robert Powers1

1,3Department of Chemistry, University of Nebraska-Lincoln, Lincoln, NE 68588, USA
2School of Veterinary Medicine and Biomedical Sciences, University of Nebraska-Lincoln, Lincoln, NE 68588, USA

Tuberculosis is one of the most common life-threatening diseases in developing countries, still ranking among the top ten causes of death worldwide. In 2017, about 10 million of the global population was diagnosed with tuberculosis and over 1.6 million succumbed to the disease. Although several therapeutic agents have been developed over the years, most of these treatments require the patient to take multiple drugs over an extended period of time, which can lead to significant side effects such as emergence of antibiotic resistance. Therefore, there is an urgent need to develop new anti-mycobacterial agents possessing a novel mechanism of action, one that would provide a better activity against latent and drug-resistant strains of tuberculosis. In 2014, Sittiwong et al. have identified a new set of molecules, fatty acid analogues, which have shown promising results in inhibiting Mycobacterium tuberculosis growth. Several of these fatty acid analogues have displayed low micromolar minimum inhibitory concentration (MIC50) values for inhibition.

Our goal is to examine the metabolic effect of fatty acid analogues on Mycobacterium Smegmatis (M. smegmatis), a wild-type laboratory strain analogous to Mycobacterium tuberculosis, and subsequently determine their mechanism of action by using MS and NMR based metabolomics. The objective of finding the mechanism of action of chemical leads using a metabolomics approach is based on the premise that drugs with similar therapeutic targets will have similar impact on metabolome, which will be observable by statistical methods, e.g. clustering analysis. It is hypothesized that if the chemical lead can induce metabolic separation from known drugs then this result would potentially indicate a new antibiotic of medical significance which can then be further explored as required.