SPRING 2019 | VOL. 71, NO.1

CHEMICAL INFORMATION BULLETIN

A Publication of the Division of Chemical Information of the ACS
ISSN: 0364-1910

Spring 2019 National Meeting:

Chemistry for New Frontiers

Orlando, Florida
March 31st - April 4th

Table of Contents:

Cover image of Orlando Convention Center courtesy of the American Chemical Society National Meeting Webpage.

Message from the Division Chair

Welcome to the first issue of the 2019 *Chemical Information Bulletin!*

My purpose in writing is to brief you on some of the activities that CINF has been doing for the betterment of the chemical information community. I would like to thank Erin Davis, Past-chair, and Jeremy Garritano, Chair-elect, for their close collaboration since the beginning of my tenure.

National Meeting Program. CINF's program at national meetings provides an excellent opportunity to learn about the most recent developments in chemical information and cheminformatics. Our program is diverse and rich, and benefits from the leadership of our Program Chair, Sue Cardinal, and the dedication of many symposium organizers. As our area of chemistry keeps growing in impact, we have been working collaboratively with other divisions to develop joint programming and explore new topics.

We would like to encourage the participation of younger chemists. With the support of ACS Publications, we award up to three $1,000 awards to graduate students and postdoctoral researchers per meeting to present in the CINF Scholarships for Scientific Excellence poster session. The call for applications is now open for the fall 2019 session in San Diego. It is a wonderful opportunity to increase the visibility of your work and network with the diverse group of professionals that CINF compromises. In addition, we are also considering developing a CINF early-career award, and I would like to hear any feedback that you may have on that.

Gender Disparity in CINF Awards. As our Awards Chair, Rajarshi Guha, described in the Winter 2018 *Chemical Information Bulletin* feature "Tackling Gender Disparity in the Skolnik Award", only one woman has won CINF's flagship award since 1976 (Yvonne C. Martin, 2009). The Skolnik Award selection committee does not submit nominations, but relies on the chemical information community at-large to identify and nominate candidates. We would like to encourage nominations of candidates who have made outstanding contributions to, and achievements in the theory and practice of chemical information science; nominators are not required to be ACS or CINF members. CINF serves a diverse and inclusive community of professionals from all over the world, and our awards are granted regardless of race, religion, country or ethnic origin, sex, gender identity and expression, sexual orientation, age, and disability.

Strategic Planning. Our last strategic plan dates from 2006, so it's time to re-energize CINF and invest in the future. We would like to grow the impact of our division and our domain, and increase the value that we provide to our members. We believe that strategic planning will provide a strong foundation for that.

CINF has been awarded an Innovative Project Grant from the ACS Divisional Activities Committee to conduct a strategic planning retreat. In addition, we have been able to secure supplemental funding from the Leadership Advisory Board and the retreat will be held during the second half of this year. We are now in the early planning stages, and will keep you informed of our progress. Our goal is to represent all the voices of the chemical information community in this process.

In the meantime, let's connect! Feel free to send me an email at elsa.alvaro@northwestern.edu, if you have comments or questions, or if you'd like to participate more actively in shaping the division.

Sincerely,
Elsa Alvaro

ACS Technical Division
Chemical Information (CINF)

Report from the ACS Leadership Institute

I wanted to take the opportunity to highlight the ACS Leadership Institute for our members. Along with Elsa Alvaro, I recently attended the 2019 ACS Leadership Institute in Atlanta, GA, on behalf of CINF. Held in January each year, the Institute brings together representatives from local sections, national committees, technical divisions, and Younger Chemists Committee and student leaders for a weekend of leadership training and networking.

Overall, depending on which role you have (local section, technical division, etc.), you have a predetermined track so you can meet other volunteers, leaders, and ACS staff that relate most to your role within the society. As Chair-elect of CINF, I attended the Technical Division track.

To allow everyone to get their bearings, a Networking Lunch started off the Institute. Also during the lunch, we heard about some of the Society's plans for celebrating the International Year of the Periodic Table. We were encouraged to share our plans and ideas at our tables, and allowed to share the ideas with the entire room before adjourning to our sessions.

I first went to the division session, hosted by the Committee on Divisional Activities (DAC) and found myself assigned to a table of attendees that represented a wide range of technical divisions (BMGT, CATL, MEDI, and POLY). And if you don't know what all of those stand for…well that was just the beginning of the parade of acronyms throughout the institute. (To save you a search – Business Development and Management, Catalysis Science and Technology, Medicinal Chemistry, and Polymer Chemistry.)

For the rest of the afternoon there was a variety of presentations from ACS staff about the services the society can provide to technical divisions. Mixed in were some general topical updates and opportunities for us to discuss issues at our tables and report out. Some of the topics included membership growth (challenges and opportunities), connecting with student members, recruiting members, and engaging and retaining members. After dinner, we returned to our division session for a few more topics and then there was a networking reception.

The next day was more customized to each individual. Before arriving, we each had the opportunity to choose two specific leadership courses of interest (or for experienced ACS leaders there was a full-day session on becoming an "Extraordinary Leader"). I chose "Engaging and Motivating Volunteers" and "Leading Change."

In these sessions, all attendees were mixed. So there was an opportunity to meet leaders of Local Sections and National Committees as well as student leaders and ACS staff. For the session on volunteers, I was lucky enough to have my table agree to work on a CINF volunteer opportunity: CINF Fundraising Chair. Throughout the session we worked through four steps of volunteer involvement as it relates to the CINF position: 1) scoping the responsibilities, 2) recruiting the right volunteers, 3) promoting a satisfying experience, and 4) fostering long-term commitment. Throughout the morning, the process gave me additional ideas for enhancing our volunteer network beyond that single position: we can recognize our members more (functionaries, those with membership anniversaries, etc.) , strive for better succession planning, make sure new committee chairs and functionaries start with as much confidence and support as possible, and create better communication and feedback mechanisms through which to interact with our members.

The "Leading Change" session was very useful in not only delving into the reasons and barriers for change, but also outlining one particular model (Kotter International's eight-Step Process for Leading Change) that could be used to implement change within an organization. The worksheets and group work at our tables were useful for understanding the process, but the time went by too quickly. If we had each brought our own challenge, I could see working through this model over an entire day instead of four hours. As CINF embarks on a new strategic plan, I gleaned some insight into a potential process for implementing the plan after it has been developed.

The day ended with a reception and ACS Resource Fair: an opportunity to talk to representatives from various ACS departments and services such as Membership, Web Strategy, ACS Student Chapters, and ACS Meetings.

The final day was a wrap-up within our division session, focusing primarily on making connections with other technical divisions and discussing the broader ACS strategic planning process.

Many of the leadership development courses are also available at each National Meeting and most are free for ACS members. So you do not need to be an elected official or committee chair to benefit from the ACS Leadership Development System. There are even a few that are available on-demand as eLearning Courses. Please check out the web site to learn more:
https://www.acs.org/content/acs/en/careers/leadership.html.

If you have any other questions about the Leadership Institute, please feel free to contact me.

Jeremy Garritano
jg9jh@virginia.edu

Twenty-five Years Ago

Last summer I promised another trip down memory lane before the spring 2019 national ACS meeting, and here it is. Diving into my archives, I found that in 1994 the spring meeting was held in San Diego. One innovation was the change of day for the start of technical programming: it began on the Sunday of the meeting for the first time. (Previously it had started on the Monday.)

CAS had apparently not heard from ACS about this change, so the CAS User Meeting clashed with prime technical programming time. The open meeting of the Committee on CAS formed part of the CINF technical program, as usual. CASLINK was demonstrated at the user meeting: it was an innovation for searching five CAS files with one command. At the open meeting it was reported that Derwent Chemical Codes could automatically be generated, and images from patents could be displayed using STN Express version 3.2. Reading my report 25 years later, I am amused to see that someone complained that CAS document delivery by U.S. Mail was not fast enough!

The biggest CAS news, though, was the out-of-court settlement of the lawsuit with Dialog, the specific terms of which could not be divulged. The case had dragged on for more than three years, and had cost both parties a great deal of money. ACS was insured for such legal disputes but, eventually, financial coverage was almost exhausted, and $2 million a year from ACS reserves would have been needed to continue the law suit. By then, Tierney and Massie were newly-appointed as heads of Dialog and CAS. The lawsuit was not an issue that they themselves owned, yet their relationship seemed likely to be affected by it. Almost equally important to the decision to settle out of court were user comments. It was in the interests of Massie and Tierney to solve the problem.

On the cheminformatics front, MDL announced that they were going to port ISIS/Host to Silicon Graphics (SGI) workstations. MDL and SGI hosted a joint reception, together with a panel discussion entitled “How Combined Technologies Impact the Drug Discovery Process." MDL, BIOSYM, and SGI were proposing a three-party drug discovery solution. (There was no love lost between MDL and TRIPOS, incidentally.) At the MDL/SGI reception, after an initial dive for the drinks and the excellent buffet, invitees were submitted to a video of Steve Goldby, CEO of MDL, talking to someone very senior in SGI. Unfortunately there was severe “ghosting” on the video and the multiple Goldby images may have made some guests wonder whether they had indulged too heavily in the free drinks. The rather woodenly-staged panel discussion concerned combinatorial chemistry, molecular diversity, interdisciplinary drug discovery teams, 3D searches, and docking.

Amusingly, someone said: “Chemists all use electronic mail and file sharing nowadays." Fast forward 25 years and it is a given that everyone uses email and collaborates: there would be no need to draw attention to it.

Elsewhere, the launch of Accord for Excel by Synopsys Scientific Systems attracted a lot of attention. Industry observers gossiped that if Oxford Molecular were to survive, it had to stop trying to support such a large and very varied range of minor products. The company had a capitalization of £29.5 million, yet the previous year’s revenues were only £1.4 million, and losses were £1.2 million. Chemical Design Limited was advertising “New Chem-X." Shortly after the ACS meeting, Evans and Sutherland divested itself of TRIPOS, and Molecular Simulations (MSI) acquired BioCAD.

CD-ROMs were still very popular. SilverPlatter Information (remember them?) took a whole page advertisement in the *Chemical Information Bulletin*. Beilstein Information Systems were about to ship the Windows version of Current Facts in Chemistry on CD-ROM, with the same look and feel as CROSSFIRE. Springer was showing InfoChem’s CD-Select. I myself reviewed CD-Select in the June issue of DATABASE magazine. I also reviewed Derwent World Drug Index. (Incidentally, at the end of the ACS meeting, Derwent Publications announced that it was changing its name to Derwent Information) . The Pharmaprojects CD-ROM was an unfriendly DOS product with no chemical structure search. Richard Love of ACS gave a technical talk about the work leading up to the publication of *J. Am. Chem. Soc.* and *Biochemistry* on CD-ROM from 1994; he began with a section on the advantages of SGML.

In spring 1994, I estimated that ChemDraw had a 90% hold on the Macintosh market for chemical structure editors, while ChemWindow had almost completely cornered the PC market. Windows ChemDraw, unfortunately, did not have OLE and DDE support at that time. ISIS/Draw had not made big inroads into the chemical structure drawing market.

Pool, Heller, and Milne were selling SciWords for Chemistry, a product which contained 75,000 words for merging with your normal word processor dictionary. The words included chemical names, element symbols, individuals’ surnames, and common abbreviations. SciWords for Agriculture and SciWords for Environment were also on offer.

In the technical program, there were symposia on chemical similarity and superposition, and on similarity searching. A number of talks, in various sessions, mentioned file conversion. ConSystant 2.0, a program written by Richard Hong of Hawk Scientific Systems for converting connection table formats, ran on three platforms, and handled seventeen chemical formats, and six graphical output formats.

Richard pointed out that in the course of converting structures, you lose information at each step, something that was particularly true of modeling formats. James Bentley of Burroughs Wellcome described some work involving Macromodel and DOCK. In one of his procedures, elevent programs were used, six of them file format conversion programs.

Speakers from the University of Arizona described version 2 of Babel which handled thirty-six input formats and twenty-eight output formats. Plus ça change, plus c'est la même chose. For my report on the fall 2018 meeting (see http://warr.com/), I transcribed a talk by Noel O'Boyle of NextMove Software on the vagaries of exchanging SMILES strings between programs.

There was also a CINF symposium on uses of the Internet. I cannot imagine a session having that title nowadays. Gary Wiggins spoke about LISTSERVs (such as chminf-l, which began in May 1991) . He remarked that postings have included an announcement of the ACS-Dialog settlement, and a discussion on whether or not to wash your bananas before peeling them. *PC Magazine* had described a LISTSERV as "a crude but effective cross between a chain letter and a shouting match."

Incidentally, the computational chemistry list also began in 1991; a brief history has been published recently in the *Journal of Cheminformatics* (https://jcheminf.biomedcentral.com/track/pdf/10.1186/s13321-018-0322-7).

As a final thought, I looked at the list of buyers of my report on the 1994 meeting in San Diego, and I noted the following long-lost names: American Cyanamid, Boots, Burroughs Wellcome Co., Ciba-Geigy, Fisons, Hoechst, Parke-Davis, Rhône Poulenc Rorer, Sandoz Agro, Schering Plough, Wellcome Research Labs, Wyeth-Ayerst, and Zeneca Agrochemicals. Fortunately for me, new licensees have appeared as the years passed.

Thanks for reading. Maybe I will be allowed to make another trip down memory lane before this year's fall meeting in San Diego. If so, I will be chuckling over things that were exciting in Washington, DC, in fall 1994.

Wendy Warr
March 2, 2019

CINF Social Networking & Business Events at the Spring 2019 ACS Meeting

Please Join Us at these Division of Chemical Information Events!

Division of Chemical Information

The ACS Division of Chemical Information is pleased to host the following social networking events at the Spring 2019 ACS National Meeting in Orlando, FL.

Sunday Welcoming Reception & Scholarships for Scientific Excellence Awards
6:30-8:30 pm, Sunday, March 31st – West Hall C, Orange County Convention Center.
Sponsored by: **InfoChem and Springer Nature**

Scholarships for Scientific Excellence
Sponsored exclusively by: **ACS Publications**

Tuesday Luncheon (ticketed event: contact Division Chair, Elsa Alvaro)
12:00-1:30 pm Tuesday, April 2nd – Room W315B, Orange County Convention Center.
Sponsored exclusively by: **Royal Society of Chemistry**
Speaker: Dr. Wendy A. Warr
Principal Consultant
Wendy Warr & Associates
Presentation: "How to right a scientific paper"

CINF Business Meetings

Saturday, March 30th, 12:30-2:30 PM

CINF Education Committee: Orange County Convention Center, **Room W221E**

CINF Program Committee: Orange County Convention Center, **Room W224B**

Saturday, March 30th, 3-6 PM

CINF Executive Committee: Orange County Convention Center, **Room W224B**

Sunday, March 31st, 12-2 PM

CSA Trust Meeting: Orange County Convention Center, **Room W330B**

CINF Meeting Announcements

Dear CINF Members,

Looking forward to seeing you at the ACS meeting in Orlando taking place March 31st- April 4th, 2019 at the Orange County Convention Center. The CINF program consists of nine symposia and includes one-hundred fifteen talks and eight posters. We will celebrate the hundredth year of IUPAC with a symposium “Creating a Common Language for Chemistry: IUPAC’s Past, Present & Future Roles." For cheminformaticians, there are symposia on Drug Discovery, Applications of Cheminformatics to Environmental Science, Deep Learning, and Web-Based Chemoinformatics Platforms. Information professionals will be interested in Libraries & External Partners Working Together and Assessing Chemistry Outreach. Our Careers Symposium will describe careers in chemical information.

If you can’t make this conference, perhaps the next will pique your interest? See the Call for Papers https://callforpapers.acs.org/sandiego2019/CINF for San Diego to get a preview. Abstracts are due on March 18th, 2019.

We are always looking for ideas for rich programs, for organizers, and for speakers. Feel free to share ideas or volunteer to help by emailing me, Sue Cardinal at scardinal@library.rochester.edu.

Regards,

Sue Cardinal

CINF Program

Sunday, March 31st

Drug Discovery: Informatics Approaches
West Hall B4 - Theater 10
Organizer E. Davis
Cosponsored by MEDI

Partnering Up in the New Frontier: Libraries & External Partners Working Together
West Hall B4 - Theater 11
Organizer S.K. Cardinal & M. Qiu

In this time of limited resources, learn how libraries and external partners collaborate synergistically to mutually benefit patrons, local communities, businesses, government agencies and beyond. The libraries also benefit as they are seen in a new light, as more than a provider of information. This symposium highlights pairs of speakers that can share both perspectives of their partnership, relating their experiences and describing what they have learned.

Collaborations & Data Sharing in Rare & Orphan Disease Drug Discovery
West Hall B4 - Theater 10
Organizer R. J. Bienstock
Cosponsored by MEDI

A rare or orphan disease is one which affects a small percentage of the population, (defined in the EU as a life-threatening or debilitating disease occurring in no more than 1 in 2,000 people; in the US as any disease that affects fewer than 200,000 (1 in 1,500 people) or in Japan as a disease that affects 1 in 2,500 people) . There are about 6,000-8,000 rare diseases affecting 8% of the world's people. Most are genetic in origin. Due to their low prevalence, it is essential that genetic and other data be shared between organizations across the globe to have a chance of impacting treatment or drug discovery. This has motivated collaborative research efforts including the Global Rare Diseases Registry Repository, sponsored by the National Center for Advancing Translational Science in the US, and on an international level, the International Rare Disease Research Consortium and ERA-Net for Research Programmes on Rare Diseases. Registry and tissue repository databases provide valuable data resources to a large community of researchers. Examples include the Matchmaker Exchange, MME (http://www.matchmakerexchange.org) a federated network of genotype and rare phenotype databases which facilitates the interaction between multiple disconnected projects. This symposium discusses ways in which sharing of rare disease data can be facilitated, how to organize data their types, and formats, and how to improve sharing and collaboration efforts.

Careers in Chemical Information
West Hall B4 - Theater 11
Organizer N. Bharti
Cosponsored by SCHB

The aim of this symposium is to introduce students and interested colleagues to various practices in chemical information careers, and to provide guidance to chemists and chemical engineers who may be considering a career change. We aim to give a complete view of the chemical information profession, covering many different types of information careers in which chemists or chemistry majors can apply their training, experience, talent, and skills.

Monday, April 1st

Web-based Chemoinformatics Platforms
West Hall B4 - Theater 10
Organizer J. L. Medina-Franco

The goal of the symposium is to present recent advances on developments and applications of open or commercial web platforms for chemoinformatic applications.

Creating a Common Language for Chemistry: IUPAC's Past, Present & Future Roles
West Hall B4 - Theater 11
Organizer L. R. McEwen & B. Lawlor
Cosponsored by HIST
Financially supported by IUPAC

Founded in July 1919 to create a common language for chemistry to enable organization of chemical information at a time when chemists routinely named compounds according to their own personal preferences, the International Union of Pure and Applied Chemistry (IUPAC) is evolving with the times. The language that IUPAC initially created was based upon the mode of communication one hundred years ago, verbal or written, and for the past century that language and its related standards have successfully facilitated the rapid advancement of science. Notwithstanding vast changes in chemistry and information technology over the past century, the communication challenges that chemists face remain surprisingly similar. IUPAC standards have played a critical role in ensuring innovative approaches to information management are consistent, and backwards compatible to ensure continuing access to the rich historic record of the discipline.

Fast forward to the twenty-first century, chemists are now involved in global collaborative research, sharing digital information over international networks, and using computers as an integral part of research and communication. The written and spoken word are no longer sufficient in an era of digital information and interoperable data exchange world-wide. IUPAC has taking a leading role in the development of a new language for chemistry: one that is understandable by both humans *and* their computers! From the advancement of the JCAMP-DX standard in NMR spectroscopy to the development of the InChI code for chemical structure representation in partnership with the InChI Trust, IUPAC is aggressively working to ensure that chemical information can be seamlessly distributed and unequivocally understood around the globe. And several new initiatives have emerged in recent years as IUPAC has reached out to like-minded organizations such as CODATA, the Research Data Alliance, the Royal Society of Chemistry, the Beilstein Institute, and, of course, the Division of Chemical Information of the American Chemical Society. In celebration of IUPAC's 100th Anniversary in 2019, this symposium will provide a glimpse of the past, but more importantly take a look at the promising future of the worldwide practice, sharing, and communication of digital chemical information. In addition, a new initiative established by IUPAC to even more broadly promote the essential value of the chemical sciences, entitled the "Top Ten Emerging Technologies in Chemistry," will be discussed and the first results announced.

Tuesday, April 2nd

Web-Based Chemoinformatics Platforms
West Hall B4 - Theater 10
Organizer J. L. Medina-Franco

See description above.

Deep Learning
West Hall B4 - Theater 11
Organizer J. L. Medina-Franco
Cosponsored by COMP

The computational chemistry community has developed and uses machine learning (ML) algorithms for decades. However, a new wave of high expectations has risen around deep learning (DL) . This symposium will bring together DL practitioners from different areas. It will start with an introduction to DL technical details, followed by current applications in chemistry, and to finalize with a critical overview of the perspectives of the field. Contributions to any of these areas were welcomed, particularly those that critically compare traditional ML and DL methods.

Assessing Chemistry Outreach
West Hall B4 - Theater 10
Organizer M. R. Hartings
Cosponsored by YCC

The future success of the chemical enterprise is, in part, dependent upon the quality of outreach to nonprofessional scientists. The importance of outreach activities has recently been highlighted by organizations such as the American Chemical Society, the National Academies, and the National Science Foundation. While many chemists have been working on the practice of outreach, the ability to assess audience participation, engagement, and effectiveness has not kept pace. These data, and how chemists share them among themselves, is crucial towards the attempt to measure and optimize outreach. In this symposium, specialists in science communication and outreach will detail their efforts to assess their efforts using qualitative and quantitative methodologies.

Wednesday, April 3rd

Applications of Cheminformatics to Environmental Science
West Hall B4 - Theater 10
Organizer A. J. Williams
Cosponsored by ENVR

Cheminformatics and computational chemistry have had an enormous impact in regard to providing environmental chemists and toxicologists access to data, information, and knowledge. With an overwhelming array of online resources and an increasingly-rich collection of software tools, the ability to source information continues to expand. Scientists typically seek chemical data in the form of chemical properties, their function and use, as well as information regarding their exposure potential, persistence in the environment and their transformation in environmental and biological systems. Commonly, the most pressing concern regarding chemicals is their potential as environmental toxicants. The increasing rate of production and release of new chemicals into commerce requires improved access to historical data and information to assist in hazard and risk assessment. High-throughput *in vitro* and *in silico* analyses increasingly are being brought to bear to screen chemicals rapidly for their potential impacts and interweaving this information with more traditional *in vivo* toxicity data and exposure estimation to provide integrated insight into chemical risk, is a burgeoning frontier on the cusp of cheminformatics and environmental sciences. This symposium has speakers that will provide an overview of the present state of data, tools, databases and approaches available to environmental chemists.

The session will hopefully cover environmental chemistry databases, data modeling and delivery, computational hazard and risk assessment, prioritizing environmental chemicals, data exchange and integration.

Deep Learning
West Hall B4 - Theater 10
Organizer J. L. Medina-Franco
Cosponsored by COMP

See description above.

Chemical Structure Association Trust

Applications Invited for CSA Trust Grants for 2019 and 2020

The Chemical Structure Association (CSA) Trust is an internationally recognized organization established to promote the critical importance of chemical information to advances in chemical research. In support of its charter, the trust has created a unique grant program and is now inviting the submission of grant applications for 2019. The deadline for receipt of proposals for the 2020 Grant is also being announced at this time.

Purpose of the Grants: The grant program has been created to provide funding for the career development of young researchers who have demonstrated excellence in their education, research or development activities that are related to the systems and methods used to store, process and retrieve information about chemical structures, reactions and compounds. One or more grants will be awarded annually up to a total combined maximum of ten thousand U.S. dollars ($10,000). Grantees have the option of payments being made in U.S. dollars or in pounds sterling equivalent to the U.S. dollar amount. Grants are awarded for specific purposes, and within one year each grantee is required to submit a brief written report detailing how the grant funds were allocated. Grantees are also requested to recognize the support of the trust in any paper or presentation that is given as a result of that support.

Who is Eligible? Applicant(s), age 35 or younger, who have demonstrated excellence in their chemical information related research and who are developing careers that have the potential to have a positive impact on the utility of chemical information relevant to chemical structures, reactions and compounds, are invited to submit applications. Proposals from those who have not received a Grant in the past will be given preference. While the primary focus of the grant program is the career development of young researchers, additional bursaries may be made available at the discretion of the trust. All requests must follow the application procedures noted below and will be weighed against the same criteria.

Which Activities are Eligible? Grants may be awarded to acquire the experience and education necessary to support research activities, for example, for travel to collaborate with research groups, to attend a conference relevant to one's area of research (including the presentation of an already-accepted research paper), to gain access to special computational facilities, or to acquire unique research techniques in support of one's research. Grants will not be given for activities completed prior to the grant award date.

Application Requirements:

Applications must include the following documentation:

1. A letter that details the work upon which the grant application is to be evaluated as well as details on research recently completed by the applicant;

2. The amount of grant funds being requested and the details regarding the purpose for which the grant will be used (e.g., cost of equipment, travel expenses if the request is for financial support of meeting attendance, etc.). The relevance of the above-stated purpose to the trust's objectives and the clarity of this statement are essential in the evaluation of the application;

3. A brief biographical sketch, including a statement of academic qualifications and a recent photograph;

4. Two reference letters in support of the application.

Additional materials may be supplied at the discretion of the applicant only if relevant to the application and if such materials provide information not already included in items 1-4. A copy of the completed application document must be supplied for distribution to the grants Committee and can be submitted via regular mail or e-mail to the committee chair (see contact information below).

Deadline for Applications:

Application deadline for the 2019 Grant is **April 19th, 2019**. Successful applicants will be notified no later than **May 17th, 2019**. The deadline for 2020 is **March 28th, 2020** and successful applicants will be notified by **May 7th, 2020**.

Address for Submission of Applications:

The application documentation can be mailed via post or emailed to: Bonnie Lawlor, CSA Trust Grant Committee Chair, 276 Upper Gulph Road, Radnor, PA 19087, USA. If you wish to enter your application by e-mail, please contact Bonnie Lawlor at chescot@aol.com prior to submission so that she can contact you if the e-mail does not arrive.

Chemical Structure Associate Trust: Recent Grant Awardees

2018

Stephen Capuzzi: Division of Chemical Biology and Medicinal Chemistry at the University of North Carolina Eshelman School of Pharmacy, Chapel Hill (USA), was awarded a Grant to attend the 31st ICAR in Porto, Portugal from 06/11/2018 to 06/15/2018, where he presented his research entitled "Computer-aided Discovery and Characterization of Novel Ebola Virus Inhibitors."

Christopher Cooper: Cavendish Laboratory, University of Cambridge, UK, was awarded a grant to present his current research on systematic, high-throughput screening of organic dyes for co-sensitized dye-sensitized solar cells. He presented his work at the Solar Energy Conversion Gordon Research Conference and Seminar held June 16-22, 2018 in Hong Kong.

Mark Driver: Chemistry Department, University of Cambridge, UK, was awarded a grant to offset costs to attend the 7th EUCheMS conference where he will present a poster on his research that focuses on the development and applications of a theoretical approach to model hydrogen bonding.

Geqing Wang: La Trobe Institute for Molecular Sciences, La Trobe University, Australia, was awarded a grant to present his work at the Fragment-based Lead Discovery Conference (FBLD2018) in San Diego, USA in October 2018. The current focus of his work is the development of novel anti-virulence drugs which potentially overcome the problems of antibiotic resistance of Gram-negative bacteria.

Roshan Singh: University of Oxford, UK, was awarded a grant to conduct research within Dr. Marcus Lundberg's Group at Uppsala University, Sweden, as part of a collaboration that he has set up between them and Professor Edward Solomon's Group at Stanford University, California. He conducts research within Professor John McGrady's group at the University of Oxford. The collaboration will look to consolidate the experiments on heme Fe(IV)=O complexes currently being studied by Solomon's Group with future multireference calculations to be conducted within Lundberg's Group.

2017

Jesus Calvo-Castro: University of Hertfordshire, England, was awarded a grant to cover travel to present his worked at the Fifth International Conference on Novel Psychoactive Substances to be held in Vienna, Austria from August 23-24, 2017. He works on the development of novel methodologies for the in-the-field detection of novel psychoactive substances (NPS), where chemical structure and information play a crucial role.

Jessica Holien: St. Vincent's Institute of Medical Research, Fitzroy, Victoria, Australia, was awarded a grant to cover travel to present her work at the 2017 Computer-aided Drug Design (CADD) Gordon Research Conference scheduled to take place July 16-21, 2017 in Mount Snow, VT, USA. She is a postdoctoral researcher at St. Vincent's and is responsible for a range of computational molecular modeling including; compound database development, virtual screening, docking, homology modeling, dynamic simulations, and drug design.

Chemical Structure Associate Trust: Recent Grant Awardees

2016

Thomas Coudrat: Monash University, Australia, was awarded a grant to cover travel to present his work at three meetings in the United States: the Open Eye Scientific CUP XVI, the American Chemical Society Spring Meeting, and the Molsoft ICM User Group Meeting. His work is in ligand directed modeling.

Clarisse Pean: Chimie Paris Tech, France, was awarded a grant to cover travel to give an invited presentation at the 2016 Pacific Rim Meeting on Electrochemical and Solid State Science later this year.

Qian Peng: University of Oxford, England, was awarded a grant to attend the 23rd IUPAC Conference on Physical Organic Chemistry. His research is in the development of new ligands for asymmetric catalysis.

Petteri Vainikka: University of Turku, Finland, was awarded a grant to spend the summer developing and testing new methods for modeling organic solvents in organic solutions with Dr. David Palmer and his group at the University of Strathclyde, Glasgow, Scotland.

Qi Zhang: Fudan University, China, was awarded a grant to attend a Gordon Conference on enzymes, coenzymes and metabolic pathways. His research is in enzymatic reactions.

2015

Dr. Marta Encisco: Molecular Modeling Group, Department of Chemistry, La Trobe Institute for Molecular Science, La Trobe University, Australia was awarded a grant to cover travel costs to visit collaborators at universities in Spain and Germany and to present her work at the European Biophysical Societies Association Conference in Dresden, Germany in July 2015.

Jack Evans: School of Physical Science, University of Adelaide, Australia was awarded a grant to spend two weeks collaborating with the research group of Dr. Francois-Xavier Coudert (CNRS, Chimie Paris Tech) .

Dr. Oxelandr Isayev: Division of Chemical Biology and Medicinal Chemistry, UNC Eshelman School of Pharmaacy, University of North Carolina at Chapel Hill, was awarded a grant to attend summer classes at the Deep Learning Summer School 2015 (University of Montreal) to expand his knowledge of machine learning to include deep learning (DL) . His goal is to apply DL to chemical systems to improve predictive models of chemical bioactivity.

Aleix Gimeno Vives: Cheminformatics and Nutrition Research Group, Biochemistry and Biotechnology Dept., Universitat Rovira I Virgili was awarded a grant to attend the Cresset European User Group Meeting in June 2015 in order to improve his knowledge of the software that he is using to determine what makes an inhibitor selective for PTP1B.

2014

Dr. Adam Madarasz: Institute of Organic Chemistry, Research Centre for Natural Sciences, Hungarian Academy of Sciences was awarded a grant for travel to study at the University of Oxford with Dr. Robert S. Paton, a 2013 CSA Trust Grant winner, in order to increase his experience in the development of computational methodology which is able to accurately model realistic and flexible transition states in chemical and biochemical reactions.

Maria José Ojeda Montes: Department of Biochemistry and Biotechnology, University Rovira i Virgili, Spain was awarded a grant for travel expenses to study for four months at the Freie University of Berlin to enhance her experience and knowledge regarding virtual screening workflows for predicting therapeutic uses of natural molecules in the field if functional food design.

Dr. David Palmer: Department of Chemistry, University of Strathclyde, Scotland was awarded a grant to present a paper at the fall 2014 meeting of the American Chemical Society on a new approach to representing molecular structures in computers based upon on ideas from the Integral Equation Theory of Molecular Liquids.

Sona B. Warrier: Departments of Pharmaceutical Chemistry, Pharmaceutical Biotechnology, and Pharmaceutical Analysis, NMIMS University, Mumbai was awarded a grant to attend the International Conference on Pure and Applied Chemistry to present a poster on her research on inverse virtual screening in drug repositioning.

2013

Dr. Johannes Hachmann: Department of Chemistry and Chemical Biology at Harvard University, Cambridge, MA was awarded a grant for travel to speak on “Structure-property relationships of molecular precursors to organic electronics" at a workshop sponsored by the Centre Européen de Calcul Atomique et Moléculaire (CECAM) that will take place October 22 – 25, 2013 in Lausanne, Switzerland.

Dr. Robert S. Paton: University of Oxford, UK was awarded a grant to speak at the Sixth Asian Pacific Conference of Theoretical and Computational Chemistry in Korea on July 11, 2013. Receiving the invitation for this meeting has provided Dr. Paton with an opportunity to further his career as a Principal Investigator.

Dr. Aaron Thornton: Material Science and Engineering at CSIRO in Victoria, Australia was awarded a grant to attend the 2014 International Conference on Molecular and Materials Informatics at Iowa State University with the objective of expanding his knowledge of web semantics, chemical mark-up language, resource description frameworks, and other online sharing tools. He will also visit Dr. Maciej Haranczyk, a prior CSA Trust Grant recipient, who is one of the world leaders in virtual screening.

2012

Tu C. Le: CSIRO Division of Materials Science & Engineering, Clayton, VIV, Australia was awarded a grant for travel to attend a cheminformatics course at Sheffield University and to visit the Membrane Biophysics group of the Department of Chemistry at imperial College London.

2011

J. B. Brown: Kyoto University, Kyoto, Japan. J.B. was awarded a grant for travel to work with Professor Ernst Walter-Knappat the Freie University of Berlin and Professor Jean-Phillipe Vert of the Paris MinesTech to continue his work on the development of atomic partial charge kernels

2010

Noel O'Boyle: University College Cork, Ireland was awarded a grant to both network and present his work on open source software for pharmacophore discovery and searching at the 2010 German Conference on Cheminformatics.

2009

Laura Guasch Pamies: University Rovira & Virgili, Catalonia, Spain was awarded a grant to do three months of research at the University of Innsbruck, Austria.

2008

Maciej Haranczyk: University of Gdansk, Poland was awarded the Grant to travel to Sheffield University, Sheffield, UK, for a six-week visit for research purposes.

2007

Rajarshi Guha: Indiana University, Bloomington, IN, USA was awarded a grant to attend the Gordon Research Conference on Computer Aided Design in August 2007.

2006

Krisztina Boda: University of Erlangen, Erlangen, Germany was awarded a grant to attend the 2006 Spring National Meeting of the American Chemical Society in Atlanta, GA, USA.

2005

Dr. Val Gillet & Professor Peter Willett: University of Sheffield, Sheffield, UK were awarded a grant for student travel costs to the 2005 Chemical Structures Conference held in Noordwijkerhout, the Netherlands.

2004

Dr. Sandra Saunders: University of Western Australia, Perth, Australia was awarded a grant to purchase equipment needed for her research.

2003

Prashant S. Kharkar: Institute of Chemical Technology, University of Mumbai, Matunga, Mumbai was awarded a grant to attend the conference, Bioactive Discovery in the New Millennium, in, Lorne, Victoria, Australia (February 2003) to present a paper, The Docking Analysis of 5-Deazapteridine Inhibitors of Mycobacterium avium complex (MAC) Dihydrofolate reductase (DHFR).

2001

Georgios Gkoutos: Imperial College of Science, Technology and Medicine, Department of Chemistry. London, U.K. was awarded a grant to attend the conference, Computational Methods in Toxicology and Pharmacology Integrating Internet Resources, (CMTPI-2001) in Bordeaux, France, to present part of his work on internet-based molecular resource discovery tools.

CINF Technical Program

SUNDAY MORNING

Section A

Drug Discovery: Informatics Approaches
West Hall B4 - Theater 10
Organizer E. Davis
Cosponsored by MEDI

8:30 CINF 1. Analysis of billions of Synthetically Accessible Virtual Inventory (SAVI) compounds as to their drug potential. **H. Patel**, W. Ihlenfeldt, M. Nicklaus

8:55 CINF 2. Drug repurposing is a common phenomenon: Bibliometric and cheminformatics evidence based on PubMed data. **N.C. Baker**, S. Ekins, A.J. Williams, A. Tropsha

9:20 CINF 3. Biology scale modeling in chemical-proteomics: Data management and analytics. **H. Wang**

9:45 CINF 4. Monomer.org. **D.J. Milton**

10:10 Intermission.

10:20 CINF 5. Extensive data-driven modeling of food-derived bioactive peptides that inhibit the angiotensin I-converting enzyme. **D.P. Russo**, Y. Zhang, H. Zhu

10:45 CINF 6. BIOFACQUIM: A compound database of natural products from Mexico. **B.A. Pilon-Jimenez**, J. Medina-Franco

11:10 CINF 7. Analysis of tautomeric transforms in chemical databases in the context of redesign of handling of tautomerism for InChI V2. **D. Dhaked**, M. Nicklaus

11:35 CINF 8. Kinase inhibitor selectivity data analysis. **Z. Luo**, V. Ulshoefer

Section B

Partnering Up in the New Frontier: Libraries & External Partners Working Together
West Hall B4 - Theater 11
Organizer S.K. Cardinal & M. Qiu

8:30 Introductory Remarks.

8:40 CINF 9. Universities and scholarly publishers collaborating to help students and postdocs advance their research and get published. **G. Baysinger,** S. O'Reilly

9:10 CINF 10. Partnership between librarians and non-profit stakeholders in research information ecosystem: WikiEdu and carpentries. **Y. Li**

9:40 CINF 11. FAIR chemical data for health and safety: Connecting the dots with cheminformatics and librarianship. **L.R. McEwen**, E. Bolton

10:10 Intermission.

10:20 CINF 12. 30 years of Reaxys: Chemical information for the chemists. **J.N. Currano**, J. Dolenc, O. Renn, J. Swienty-Busch

10:50 CINF 13. PubChem as a resource for chemical information training. **S. Kim**, E. Bolton

11:20 Panel Discussion.

11:50 Concluding Remarks.

SUNDAY AFTERNOON

Section A

Drug Discovery: Informatics Approaches
West Hall B4 - Theater 10
Organizer E. Davis
Cosponsored by MEDI

1:30 CINF 14. SuCOS: A pharmacophoric-shape overlap metric for comparing binding modes. **S. Leung**, M. Bodkin, F. von Delft, P. Brennan, G.M. Morris

1:55 CINF 15. LigandNet: A machine-learning-based toolkit for predicting ligand activity to proteins. **M. Hassan**, D. Castaneda, D. Shrestha, I. Salama, S. Sirimulla

2:20 CINF 16. Machine learning-based prediction of compound profiling matrices. **R. Rodríguez Pérez**, J. Bajorath

Collaborations & Data Sharing in Rare & Orphan Disease Drug Discovery
West Hall B4 - Theater 10
Organizer R. J. Bienstock
Cosponsored by MEDI

3:20 CINF 17. Collaborations and data sharing in rare disease. **R.J. Bienstock**

3:40 CINF 18. Genetic and Rare Diseases (GARD) information center treatment profiles. **Q. Zhu**, D. Nguyen, N. Southall, A. Chen, E. Sid, A. Pariser

4:05 CINF 19. Biomedical data translator: Supporting data integration and rare disease research. **N. Southall**, C. Colvis

4:30 CINF 20. Data-driven drug discovery for rare diseases: tales from the trenches. **F. van den Broek**

Section B

Careers in Chemical Information
West Hall B4 - Theater 11
Organizer N. Bharti
Cosponsored by SCHB

1:30 Introductory Remarks.

1:35 CINF 21. Computational chemistry and chemoinformatics career opportunities at the NIH (NIEHS). **R.J. Bienstock**

1:55 CINF 22. Careers in publishing chemical information: From the lab bench to the editorial office to the database. **G. Jones**

2:15 CINF 23. Water-quality data and publications for careers in chemistry information. **E.C. Wild**

2:35 CINF 24. Scientist in EH&S: Changing the tradition in laboratory safety. **S. Singh**, N. Bharti

2:55 Intermission.

3:05 CINF 25. Antony Williams, the ChemConnector: A career path through a diverse series of roles and responsibilities. **A.J. Williams**

3:25 CINF 26. Careers in science: Science policy and general advice. **E. Dunlea**

3:45 CINF 27. How interests and experience led to a career in chemical literature informatics. **N.C. Baker**

4:05 CINF 28. Lab to library: A career in chemistry librarianship. **N. Ruhs**

MONDAY MORNING

Section A

Web-Based Chemoinformatics Platforms
West Hall B4 - Theater 10
Organizer J. L. Medina-Franco

8:00 Introductory Remarks.

8:05 CINF 37. Designing drug candidates and chemical probes in cyberspace. **B. Villoutreix**

8:35 CINF 38. Cheminformatics tools and applications on the web: Challenges, examples, and the future. **D. Fourches**

9:05 CINF 39. SynSpace: A user-friendly web- and cloud-based design platform to expand synthetically-enabled scaffold and lead analogue space for medicinal chemistry and AI-assisted drug discovery. **G. Makara**, G. Pocze, L. Kovacs, O. Demeter, I. Szabo

9:35 CINF 40. Exploring an expanded chemical universe using www.chemmaps.com. **A. Borrel**, D. Fourches, N. Kleinstreuer

10:05 Intermission.

10:20 CINF 41. Exploring chemical space at gdb.unibe.ch. D. Probst, M. Awale, A. Thakkar, J. Reymond

10:50 CINF 42. Developing an integrated model management solution to assure quality of predicted data at the US EPA's National Center of Computational Toxicology. **C. Grulke**, A.J. Williams, A. Singh, J. Edwards

11:20 CINF 43. US-EPA CompTox chemicals dashboard: A web-based data integration hub for environmental chemistry data. **A.J. Williams**, C. Grulke, R. Judson, J. Wambaugh, J. Dunne, J. Edwards

Section B

Creating a Common Language for Chemistry: IUPAC's Past, Present & Future Roles
West Hall B4 - Theater 11
Organizer L. R. McEwen & B. Lawlor
Cosponsored by HIST
Financially supported by IUPAC

8:30 Introductory Remarks.

8:35 CINF 44. IUPAC Commission on Isotopic Abundances and Atomic Weights: Its history, role, and work. **J. Meija**

9:00 CINF 45. Archives of the International Union of Pure and Applied Chemistry at the Science History Institute. **R.S. Brashear**

9:25 CINF 46. "A" in IUPAC: Applying the common language for chemistry to meet world needs. **M.C. Cesa**

9:50 Intermission.

10:05 CINF 47. Accidental nomenclaturist: A journey from bench chemist to ACS-NTS and IUPAC member. **M.M. Rogers**

10:30 CINF 48. iGROW: IUPAC global recognition opportunities for women. **F. Meyers**, C. Ribes, A.K. Wilson

10:55 CINF 49. Role of IUPAC Committee on Chemistry Education in communicating chemistry. **M.H. Towns**

11:20 CINF 50. Short history of IUPAC InChI algorithm. **S.R. Heller**

11:45 Concluding Remarks.

MONDAY AFTERNOON

Section A

Web-Based Chemoinformatics Platforms
West Hall B4 - Theater 10
Organizer J. L. Medina-Franco

1:10 Introductory Remarks.

1:15 CINF 51. Web Force-Field (WebFF) repository: Molecular dynamics force-field data for soft materials at multiple levels of granularity. **F.R. Phelan**, H. Sun

1:45 CINF 52. CavityPlus: A web server for protein cavity detection with pharmacophore modelling, allosteric site identification, and covalent ligand-binding ability prediction. **J. Pei**

2:15 CINF 53. iSpiEFP: Automating the computational workbench. **Y. Bui**, L.V. Slipchenko

2:45 CINF 54. ProteinsPlus and SMARTSviewer: Two web applications for the modeling and cheminformatics community. **R. Fährrolfes**, R. Schmidt, M. Rarey

3:15 Intermission.

3:30 CINF 55. D-Peptide Builder: A web-based application to enumerate the chemical space of peptides. **B. Diaz Eufracio**, J. Medina-Franco, O. Palomino-Hernández, A. Arredondo-Sanchez

4:00 CINF 56. Freely available online resource for prediction of novel multitarget anti-HIV agents. **D. Druzhilovskiy**, D. Filimonov, L. Stolbov, P. Savosina, V. Poroikov, M.C. Nicklaus

4:30 CINF 57. ZINC15.docking.org: Over 1.5 billion compounds you can search and buy; 550 million lead-like you can dock. **J.J. Irwin**

Section B

Creating a Common Language for Chemistry: IUPAC's Past, Present & Future Roles
West Hall B4 - Theater 11
Organizer L. R. McEwen & B. Lawlor
Cosponsored by HIST
Financially supported by IUPAC

1:30 Introductory Remarks.

1:35 CINF 58. Towards a "Digital IUPAC:" Coordinating community needs for digital data standards. **L.R. McEwen,** D. Martinsen, H.A. Lawlor

2:00 CINF 59. Renovating the IUPAC *Gold Book* for the digital era and the next 100 years. **S.J. Chalk**

2:25 CINF 60. ISMC: IUPACs interdivisional Sub-committee on Materials Chemistry. **C.K. Ober,** V. Gubala

2:50 Intermission.

3:05 CINF 61. FAIR data in the 21st century: The role of scientific unions in facilitating interdisciplinary data science in Chemistry and the Earth Sciences. **S. Stall,** L.R. McEwen

3:30 CINF 62. Top Ten Emerging technologies in chemistry: A new initiative from IUPAC and Chemistry International. **F. Gomollon-Bel,** J. Garcia Martinez, B. Lawlor

3:55 CINF 63. IUPAC and its next century: A Secretary General's perspective. **R. Hartshorn**

4:20 Panel Discussion.

4:50 Concluding Remarks.

MONDAY EVENING

Sci-Mix
West Hall C
Organizer R.J. Bienstock

8:00 CINF 774 Analyzing the effectiveness of a pilot community service learning project in the undergraduate chemistry laboratory. **H.H. Grewal,** J. Khalil, C.C. Lovallo, K. Ho

Section A

Web-Based Chemoinformatics Platforms
West Hall B4 - Theater 10
Organizer J. L. Medina-Franco

8:00 Introductory Remarks.

8:05 CINF 64. 3decision®: Bringing structural data analytics to the masses. **G. Jonasson**

8:35 CINF 65. Leveling the playing field: Illuminating understudied targets with Pharos. **T. Sheils**, D. Nguyen, N. Southall, T.I. Oprea, V. Siramshetty

9:05 CINF 66. Chembench: A publicly-accessible, integrated cheminformatics portal. **E. Muratov**, S. Capuzzi, V.M. Alves, V. tkachenko, A. Korotcov, D. Korn, W. Lam, T. Thornton, D. Pozefsky, A. Tropsha

9:35 CINF 67. K4DD database: Ligand binding kinetics at its best. **G.F. Ecker**, L. Richter

10:05 Intermission.

10:20 CINF 68. MOEsaic: The application of matched molecular pair analysis to SAR exploration. **G. Fortin**

10:50 CINF 69. ARENA360: An integrated informatics solution for drug discovery. **C. Betton**, Z. Luo, V. Ulshoefer

11:20 CINF 70. Delivering computational chemistry to cheminformatics: collaborative drug discovery with LiveDesign. **E. Davis**

11:50 Concluding Remarks.

Section B

Deep Learning
West Hall B4 - Theater 11
Organizer J. L. Medina-Franco
Cosponsored by COMP

8:00 Introductory Remarks.

8:05 CINF 71. Advances in deep learning and their applied utility toward chemical informatics & drug discovery. **E. Clark**, W.E. Hahn, R. St Clair, P. Morris, M. Teti

8:35 CINF 72. How much can we learn from SMILES as text? **H. Sun**

9:05 CINF 73. Novel, active learning approach for deep learning of chemical data: Extracting more chemical insights by choosing less. **M. Haghighatlari**, J. Hachmann

9:35 CINF 74. Application of machine learning to skin cancer detection and classification. **A.C. Terentis**, J. Strasswimmer

10:05 Intermission.

10:20 CINF 75. Deep learning for the characterization and identification of small molecules. **S. Colby**, J. Nunez, N. Hodas, C. Corley, R. Renslow

10:50 CINF 76. Virtual high-throughput screening: A combined deep-learning approach. **P. Morris**, R. St Clair, M. Teti, E. Clark, W.E. Hahn

11:20 CINF 77. Learn deep before deep learning. **K. Martinez Mayorga**, G. Gómez Jiménez, A. Madariaga-Mazon

Section A

Assessing Chemistry Outreach
West Hall B4 - Theater 10
Organizer M. R. Hartings
Cosponsored by YCC

1:15 CINF 78. Going beyond popular: Assessing SciPop Talks! **R.M. Burks**, K. Deards, E. DeFrain

1:35 CINF 79. Understanding interest, relevance, & self-efficacy: Chemistry at the museum and beyond. **E.L. Howell**, S. Yang, D.A. Scheufele

1:55 CINF 80. Collecting, understanding, and utilizing audience feedback to increase interest, relevance, and self-efficacy related to hands-on chemistry activities in a museum. **G.M. Haupt**

2:15 CINF 81. Advancing inclusive excellence in academic chemistry departments from the top down through a discipline-based evidenced-based approach. **R. Hernandez**, D. Stallings, S.K. Iyer

2:35 CINF 82. Science outreach: What does it mean to be successful, and how do we know? **J. Garbarino**

2:55 Intermission.

3:10 CINF 83. Amplifying your social impact: A collaborative approach to chemistry outreach. **M.T. Gallardo-Williams**, G. Van Den Driessche, A. Malico

3:30 CINF 84. Evaluating impact. S. Kundu

3:50 CINF 85. How can I measure the success of my online outreach? **D. Reese**r, S. Hadden, M. Ruhl, A.T. Yarnell

4:10 CINF 86. Mapping the chemistry Twitter community: A reproduction of academic power structures or an opportunity to empower marginalized voices? **P. Vincent-Ruz**, D. Reeser, M.R. Hartings

Section B

Deep Learning
West Hall B4 - Theater 11
Organizer J. L. Medina-Franco
Cosponsored by COMP

1:30 Introductory Remarks.

1:35 CINF 87. Interpretable molecular design based on layer-wise relevance propagation. **Y. Kwon**, K. Kim, I. Kim, J. Yoo, W. Son, Y. Choi, H. Lee, J. Shin

2:05 CINF 88. Machine-learned model for molecular simulations of liquid and water vapor. **T. Loeffler**, T. Patra, H. Chan, S. Sankaranarayanan

2:35 CINF 89. Prediction of chemical reactivity with a graph-convolutional neural network model. **C.W. Coley**, W. Jin, L. Rogers, T.F. Jamison, T. Jaakkola, W.H. Green, R. Barzilay, K.F. Jensen

3:05 Intermission.

3:20 CINF 90. Predicting bond dissociation energies through deep learning. **Y. Guan**, Y. Kim, P. St. John, S. Kim, R.S. Paton

3:50 CINF 91. Multitask prediction of site selectivity in aromatic C-H functionalization reactions. **T.J. Struble**, C.W. Coley, K.F. Jensen

4:20 CINF 92. Molecular transformer for chemical reaction prediction and uncertainty estimation. **P. Schwaller**, T. Laino, T. Gaudin, C. Bekas, A.A. Lee

WEDNESDAY MORNING

Section A

Applications of Cheminformatics to Environmental Science
West Hall B4 - Theater 10
Organizer A. J. Williams
Cosponsored by ENVR

8:00 Introductory Remarks.

8:05 CINF 93. Environmental chemical information in PubChem. **J. Zhang**, E. Bolton

8:25 CINF 94. EPA CompTox chemicals dashboard: An online resource for environmental chemists. **A.J. Williams**, C. Grulke, J. Dunne, J. Edwards

8:45 CINF 95. Mapping of chemical identifiers to DSSTox to enable data integration in the US-EPA CompTox Chemicals Dashboard. **C. Grulke**, I. Thillainadarajah, P. Browne, A.J. Williams, A. Richard

9:05 CINF 96. Consistency checking the experimental data available from the USEPA NCCT CompTox database. S.J. Chalk, A.J. Williams, C. Grulke

9:25 Intermission.

9:40 CINF 97. Literature-based cheminformatics for research in chemical toxicity. **N.C. Baker**, A.J. Williams, T. Knudsen

10:00 CINF 98. Green chemistry and open data. **J. Zhang**, E. Bolton

10:20 CINF 99. Development of the alternatives assessment dashboard webtool. **L. Vegosen**, T. Martin

10:40 CINF 100. Application of chemical informatics to alternatives assessment. **W. Barrett, S.R.** Takkellapati, K. Tadele, L. Vegosen, M.A. Gonzalez

11:00 CINF 101. Prediction of toxicity using WebTEST (Web-services Toxicity Estimation Software Tool). **T. Martin, A.J.** Williams, V. Tkachenko

11:20 CINF 102. Case study in quantitative GenRA predictions using repeated dose toxicity studies from ToxRefDB v2.0. **G. Helman**, G. Patlewicz, I. Shah, K. Paul Friedman, L. Pham, S. Watford

11:40 CINF 103. Enhancement of acute toxicity prediction by multi-task learning. S. Sosnin, D. Karlov, I.V. Tetko, M.V. Fedorov

Section B

Deep Learning
West Hall B4 - Theater 11
Organizer J. L. Medina-Franco
Cosponsored by COMP

8:00 Introductory Remarks.

8:05 CINF 104. Prediction of toxicity: Deep learning with small and imbalanced datasets. **G.F. Ecke**r, J. Hemmerich, E. Asilar

8:35 CINF 105. Imputing compound activities based on sparse and noisy data. **T. Whitehead**, B. Irwin, P.A. Hunt, M.D. Segall, G. Conduit

9:05 CINF 106. Machine learning in the context of bioactivity. **J. Sieg**, M. Rarey

9:35 CINF 107. ML and AI in the design of new drug lead compounds. **S. Keinan**, W.J. Shipman, E.H. Frush, E. Addison

10:05 Intermission.

10:20 CINF 108. Influence of compound profiling matrix density on the performance of multi-task deep neural networks and random forest models. **R. Rodríguez Pérez**, J. Bajorath

10:50 CINF 109. Many possible roles of deep learning in drug discovery: Separating truth from hype. **R. Abel**, K. Leswing, K. Marshall, J. Staker, C. McQuaw, S. Jerome, S. Mondal, S. Bhat

11:20 CINF 110. Industry perspective: Deep learning for QSAR models. **J. Shen**

11:50 Concluding Remarks.

WEDNESDAY AFTERNOON

Section A

Applications of Cheminformatics to Environmental Science
West Hall B4 - Theater 10
Organizer A. J. Williams
Cosponsored by ENVR

1:15 CINF 111. OPERA models for physicochemical properties, environmental fate and toxicological endpoints to support regulatory purposes. **K. Mansouri**, R. Judson, A.J. Williams, N. Kleinstreuer

1:35 CINF 112. Applications of a chemotype-enrichment approach to the ToxCast data landscape and beyond: Inverting the SAR paradigm. **A. Richard**, R. Lougee, C. Grulke, N.C. Baker, J. Wang, A.J. Williams

1:55 CINF 113. Framing chemical safety and risk management: Ontological perspectives from laboratory procedures and incident reports. **C.M. Shimizu**, L.R. McEwen

2:15 CINF 114. Evaluation of the chemotype-enrichment workflow as a tool for independent evaluation biological activity thresholds and a comparison with traditional QSAR methods. **R. Lougee**, A. Richard, C. Grulke

2:35 CINF 115. Case study in quantitative GenRA predictions using acute oral toxicity. **G. Helman**, I. Shah, G. Patlewicz

2:55 CINF 116. Comprehensive computational approach for predicting human skin sensitization as suggested alternative to animal testing. **E. Muratov**, V.M. Alves, J. Borba, R. Braga, S. Capuzzi, A.C. Silva, N. Kleinstreuer, C.H. Andrade, A. Tropsha

3:15 Intermission.

3:25 CINF 117. Predicting chemical-assay interference using Tox21 qHTS data. **A. Borrel**, R. Huang, M. Xia, K. Houck, R. Judson, N. Kleinstreuer

3:45 CINF 118. Methods for in silico screening of use and exposure data in authority databases. **S. Fischer**

4:05 CINF 119. Novel nanodescriptors applied in QNAR: Combination of virtual nanomaterial library and geometrical structure of nanomaterial. **X. Yan**, A. Sedykh, W. Wang, B. Yan, H. Zhu

4:25 CINF 120. Reaction library for predicting direct phototransformation products of aquatic organic contaminants. **C Yuan**, C.T. Stevens, E.J. Weber

4:45 CINF 121. Cheminformatics and non-targeted analysis: A two-way street. **E.M. Ulrich**, J. Sobus, S. Newton, C. Grulke, A. Richard, R. Singh, A. McEachran, K. Phillips, K. Mansouri, J. Wambaugh, K.K. Isaacs, A.J. Williams

5:05 CINF 122. Elucidation of chemical dark matter using 'standards-free' small molecule identification. **R. Renslow**, S. Colby, D. Thomas, J. Nunez, Y. Yesiltepe, N. Govind, J.R. Cort, J. Teeguarden

5:25 Concluding Remarks

Notes From Our Sponsors

Division of Chemical Information Sponsors
Spring 2019

The American Chemical Society Division of Chemical Information is very fortunate to receive generous financial support from our sponsors to maintain the high quality of the Division's programming and to promote communication between members at social functions at the ACS Spring 2019 National Meeting in Orlando, FL, and to support other divisional activities during the year, including scholarships to graduate students in chemical information.

The Division gratefully acknowledges contribution from the following sponsors:

Silver	**Royal Society of Chemistry** **ACS Publications**
Bronze	**Springer Nature**
Contributor	**Bio-Rad Laboratories** **InfoChem**

Opportunities are available to sponsor Division of Chemical Information events, speakers, and material. Our sponsors are acknowledged on the CINF web site, in the Chemical Information Bulletin, on printed meeting materials, and at any events for which we use your contribution. For more information please review the Sponsorship Brochure at http://www.acscinf.org/PDF/CINF_Sponsorship_Brochure.pdf.

Please feel free to contact me if you would like more information about supporting the ACS Division of Chemical Information.

Graham Douglas
Chair pro tem, Fundraising Committee 2019
Email: **Sponsorship@acscinf.org**
Tel: 510-407-0769

The ACS CINF Division is a non-profit tax-exempt organization with taxpayer ID no. 52-6054220.

2019 CINF Officers & Functionaries

Chair

Elsa Alvaro
Northwestern University
elsa.alvaro@northwestern.edu

Chair-Elect

Jeremy Garritano
University of Virginia
jg9jh@virginia.edu

Past-Chair

Erin Davis
Schrödinger
erinsdavis@gmail.com

Secretary

Tina Qin
Harvard University
qinnamsu@gmail.com

Treasurer

Stuart Chalk
University of North Florida
schalk@unf.edu

CINF Councilors

Bonnie Lawlor
chescot@aol.com

Andrea Twiss-Brooks
University of Chicago
atbrooks@uchicago

Svetlana N. Korolev
University of Wisconsin, Milwaukee
skorolev@uwm.edu

CINF Alternate Councilors

Rachelle Bienstock
RJB Computational Modeling LLC
rachelleb1@gmail.com

Chuck Huber
UC Santa Barbara
huber@library.ucsb.edu

Jeremy Garritano
University of Virginia
jg9jh@virginia.edu

Archivist/Historian

Bonnie Lawlor
See Councilor

Awards Committee Chair

Rajarshi Guha
Vertex Pharmaceuticals
rajarshi.guha@gmail.com

Careers Committee Chair

Neelam Bharti
Carnegie Mellon University
nbharti@andrew.cmu.edu

Communications and Publications Committee Chair

Graham Douglas
graham_c_douglas@hotmail.com

Education Committee Chair

Grace Baysinger
Stanford University
graceb@stanford.edu

Finance Committee Chair

Stuart Chalk
See Treasurer

Fundraising Interim Committee Chair

Graham Douglas
graham_c_douglas@hotmail.com

Membership Committee Chair

Donna Wrublewski
Caltech Library
dtwrub@caltech.edu

Nominating Committee Chair

Erin Davis
See Past Chair

Program Committee Chair

Sue Cardinal
University of Rochester
scardinal@library.rochester.edu

CIB Editor Spring

Kortney Rupp
Lawrence Livermore National Laboratory
rupp5@llnl.gov

CIB Editor Summer

David Shobe
avidshobe@yahoo.com

CIB Editor Fall

Teri Vogel
UC San Diego
tmvogel@ucsd.edu

CIB Editor Winter

Judith Currano
The University of Pennsylvania
currano@pobox.upenn.edu

Webmasters

Stuart Chalk
University of North Florida
schalk@unf.edu

Rachelle Bienstock
RJB Computational Modeling LLC
rachelleb1@gmail.com

Spring 2019 CIB Contributors

Articles & Features

Elsa Alvaro
Wendy A. Warr
Jeremy Garritano
Sue Cardinal
Bonnie Lawlor

Sponsor Information

Graham Douglas

Production

Elsa Alvaro
Jeremy Garritano
Bonnie Lawlor
Sue Cardinal
Kortney K. Rupp
Wendy A. Warr

Schedule of Future ACS National Meetings

258th	Aug. 25-29	2019	San Diego, CA	Chemistry of Water
259th	Mar. 22-26	2020	Philadelphia, PA	Macromolecular Chemistry: The Second Century
260th	Aug. 23-27	2020	San Francisco, CA	Chemistry from Bench to Market
261st	Mar. 21-25	2021	San Antonio, TX	TBA
262nd	Aug. 22-26	2021	Atlanta, GA	TBA
263rd	Mar. 20-24	2022	San Diego, CA	TBA
264th	Aug. 21-25	2022	Chicago, IL	TBA
265th	Mar. 26-30	2022	Indianapolis, IN	TBA
266th	Aug.13-17	2023	San Francisco, CA	TBA

CINF 1

Analysis of billions of Synthetically Accessible Virtual Inventory (SAVI) compounds as to their drug potential

Hitesh Patel[1], *hitesh.patel@nih.gov, Wolf-Dietrich Ihlenfeldt*[2], *Marc Nicklaus*[1]. *(1) Computer-Aided Drug Design Group, Chemical Biology Laboratory, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Frederick, Maryland, United States (2) Xemistry GmbH, Konigstein, Germany*

The Synthetically Accessible Virtual Inventory (SAVI) project aims at computationally generating a very large number of easily and inexpensively synthesizable novel screening compounds for the purpose of drug discovery. In SAVI, products are not generated by just applying SMARTS transform patterns to building block structures of unknown availability. We instead combine a set of transforms richly annotated with chemical context, coming from, or being newly developed in the mold of, the original LHASA project knowledgebase, with a set of highly annotated, reliably available, purchasable starting materials. These are reacted together *in silico* with custom scripts and capabilities in the chemoinformatics toolkit CACTVS. Each product is annotated with a number of computed properties seen as important in current drug design, including rules for identifying potentially reactive or promiscuous compounds. After having produced and made publicly available the first (beta) set of 283 million SAVI products annotated with proposed one-step syntheses, we are here reporting on the second full production run aimed at creating a database of more than one billion high-quality, easily synthesizable screening samples using approximately 1 million building blocks and more than 50 transforms, including modern popular chemistry such as Suzuki coupling and Buchwald-Hartwig reaction etc. We will present an analysis of the thus generated molecules, as to properties such as structural diversity, novel rings formed, and drug-likeness. We will report on the current status, ongoing developments, as well as scientific and technical challenges of the project.

CINF 2

Drug repurposing is a common phenomenon: Bibliometric and cheminformatics evidence based on PubMed data

Nancy C. Baker[1], *baker.nancy@epa.gov, Sean Ekins*[2], *Antony J. Williams*[3], *Alexander Tropsha*[4]. *(1) ParlezChem, Hillsborough, North Carolina, United States (2) Collaborations Pharmaceuticals, Fuquay Varina, North Carolina, United States (3) ChemConnector, Wake Forest, North Carolina, United States (4) Univ of North Carolina, Chapel Hill, North Carolina, United States*

Cheminformatics includes literature informatics, a field of endeavor encompassing document retrieval and literature mining. Herein, we mined biomedical literature

annotated in PubMed to assess the scope of drug repurposing, the phenomenon of using known drugs to treat conditions that the drugs were not prescribed for. We have analyzed the evolution of all drug–disease relationships mined from PubMed taking the date of the first report on specific drug-disease connectivity as an indication of the initial indication vs. subsequent use for another indication. The analysis extends back to drug treatments from the 1940's and provides a historical overview up to the present day. We find that most drugs have been tested as treatments against more than one disease and some compounds have been tried in hundreds of diseases. Caffeine, for instance, has been tried clinically or experimentally on over 350 diseases. Specific examples of drugs that have been tested and used in a variety of diseases will be discussed, examining what motivated researchers to try the drug on a new disease. While in the majority of cases these drugs were tried on diseases in therapeutic areas close to their original use, there are striking, and perhaps instructive repurposing attempts where drugs have been tried in unexpectedly novel therapeutic areas.

CINF 3

Biology scale modeling in chemical-proteomics: Data management and analytics

***Huijun wang**, huijun.wang@merck.com. Merck, Kenilworth, New Jersey, United States*

Understanding drug-target interactions at both protein and pathway signaling network level in healthy vs. disease states is critical for the success of drug discovery and development. In the post-genomic era, quantitative chemical proteomics is emerging as a powerful tool to identify and validate novel druggable targets by means of (i) deconvolution of the molecular mechanism of action (MMoA); (ii) proteome selectivity assessment of bioactive molecules and (iii) druggability assessment for proteins of therapeutic interest with unclear MMoA identified through diverse approaches. Internally, we utilized in biochemical assays, transcriptomics and chemoproteomics experiments to elucidate drug-target interaction from various angles. At its core, chemoproteomics has the unprecedented power to unbiasedly discover, and unambiguously quantify, hundreds to thousands of protein interactions in a disease-relevant biological system perturbed with a controlled chemical insult, such as a drug or investigational molecule. Yet, the effective means to systematically mine, integrate and derive relevant information out of such complex big data sets remains a scientific frontier today. To maximize the value of various chemical biology data, to fuel in the hypothesis generation-testing cycle in Target Identification and Validation (TIDVal), we invested in a Chemical Biology Data Management System (CBDMS) as the infrastructure foundation to capture systems-biology perspectives of drug-proteome and transcriptome dynamics, on-target engagement, off-target effects and polypharmacology. Herein we report current progress of this endeavor, particularly in chemical proteomics data handling, analysis and visualization in the context of several exemplary chemical proteomics experiments to identifying novel targets.

CINF 4

Monomer.org

***Donald J. Milton**, jmilton@ionisph.com. Ionis Pharmaceuticals, La Jolla, California, United States*

Digital representation of complex chemical entities like biopolymers and large macromolecules have pushed past the practical use of traditional cheminformatics tools. Compounds like antibody-drug conjugates and antisense oligonucleotides require require both chemical and biological (sequence) management. HELM (Hierarchical Editing Language for Macromolecules) developed at Pfizer and currently managed by the Pistoia Alliance provides a path forward for these complex entities. The lack of reference monomer library, however, is highly limiting and prevents the global adoption of this specification. monomer.org provides this system-of-record of monomers and allows HELM and other macromolecule notations to exist as pointers to a stable and permanent cloud-based library. Here we will highlight the initial dataset and discuss the curation/review process. We will give real-world examples and provide a vision of how complex entities will be managed in the future.

CINF 5

Extensive data-driven modeling of food-derived bioactive peptides that inhibit the angiotensin I-converting enzyme

Daniel P. Russo[1], *danrusso@scarletmail.rutgers.edu, Ying-Hua Zhang*[2], *Hao Zhu*[3,1]. *(1) Center for Computational and Integrative Biology, Rutgers University, Camden, New Jersey, United States (2) Key Laboratory of Dairy Science, Ministry of Education, Northeast Agricultural University, Harbin, Heilongjiang, China (3) Chemistry Department, Rutgers Univesity, Camden, New Jersey, United States*

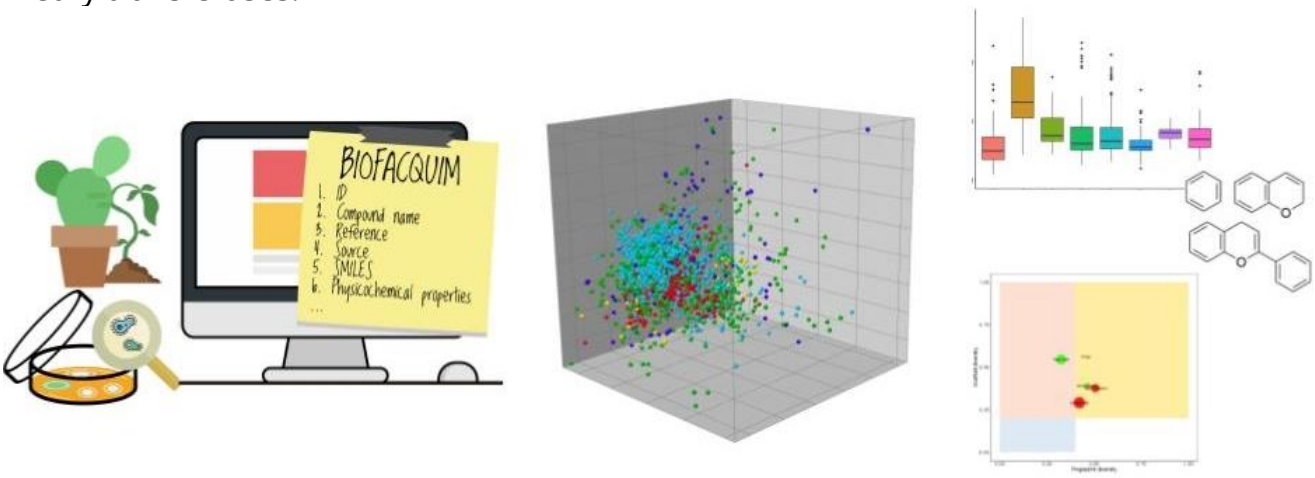
Approximately a third of all adults over the age of 20 have high blood pressure, a precursor to a variety of heart and kidney diseases and a risk factor for heart attacks and strokes. In humans, blood pressure is regulated by the renin-angiotensin hormone system. The angiotensin I-converting enzyme (ACE), as a key component of this system, catalyzes the conversion of angiotensin I to angiotensin II which acts a signaling molecule to narrow blood vessels resulting in blood pressure increase. Compounds that inhibit ACE activity have successfully been developed as treatments for controlling blood pressure and rank among the most widely prescribed drugs on the market. Furthermore, small peptides from a variety of food origins such as milk, soy, or fish, have been delineated as ACE inhibitors. These food-originating ACE-inhibiting peptides have gained remarkable interest over the years due to their therapeutic potential, safe toxicity profile and little side effects. Unlike small molecules, there are no curated bioactivity data repositories for peptides, hindering further modeling studies. In this work, we present the results of a data-driven modeling study to investigate the ACE inhibition of small peptides. First, a large database of peptides with ACE inhibitions was compiled from a variety of sources. This database consisted of 4,529 peptide sequences with IC_{50} data for ACE inhibition and various lengths ranging from 2-50 amino acids. To our knowledge, this is the largest database characterization for ACE inhibiting peptides to date. These peptides were grouped by the number of residues and used as the basis for several quantitative structure-activity relationship model developments using a variety of machine learning algorithms combined with a variety of descriptors. Several models showed good correlation with the experimentally-derived activities through cross-validation ($r^2 > 0.5$). Additionally, predictions of peptides, which were not included in the current database, showed cleared evidence for amino acid preference to strongly increase/decrease ACE-inhibitions, which varies based on peptide length. In summary, we show how data-driven informatics modeling studies can be an applicable method to perform peptide virtual screening to select new ACE-inhibiting peptides which have potentially therapeutic effects.

CINF 6

BIOFACQUIM: A compound database of natural products from Mexico

Beatriz A. Pilon-Jimenez, *angiepilon96@gmail.com, José L. Medina-Franco. Department of Pharmacy, School of Chemistry, National Autonomous University of Mexico (UNAM), Ciudad de México, Mexico*

Natural product databases have a major impact on drug discovery projects and other research areas. The number of public compound databases with molecules from natural origin is increasing. Thus far, several countries have started initiatives to build databases that are representative of their diversity. Examples include France, Brazil, and Panama. Herein, we discuss the advances to build "BIOFACQUIM", a compound database with natural products isolated in Mexico. We will present the assembly, curation, and implementation on a public platform. We will also discuss advances in the chemoinformatic characterization of the content and coverage in chemical space. Specifically, it will be presented the profile of physicochemical properties, scaffold content, and diversity, as well as structural diversity based on molecular fingerprints. Finally, we will also discuss the progress in the profiling of the natural product database for drug discovery, identifying novel compounds with activity as inhibitors of epigenetic targets with emphasis on DNA methyltransferases.



Assembly, curation and cheminformatic analysis of a compound database of natural products from Mexico for drug discovery.

CINF 7

Analysis of tautomeric transforms in chemical databases in the context of redesign of handling of tautomerism for InChI V2

Devendra Kumar Dhaked*, devendrakumar.dhaked@nih.gov, Marc Nicklaus. Computer-Aided Drug Design Group, Chemical Biology Laboratory, Center for Cancer Research, National Cancer Institute, National Institutes of Health, NCI-Frederick, Frederick, Maryland, United States*

The International Chemical Identifier (InChI), even though in principle intended to be tautomer-invariant, handles tautomerism only to a limited extent in its current version (1.05). For example, keto-enol and 1,5 tautomerism are not turned on by default in Standard InChI and this propagates to the Nonstandard InChI. Expanding the set of the

20 prototropic tautomerism rules implemented as the standard rule set in the chemoinformatics toolkit CACTVS, we have added more than 40 additional new transforms (expressed as SMIRKS) derived from various literature sources that document experimental evidence of specific types of tautomeric interconversions. We have compiled 21 prototropic, 14 ring-chain, and 8 valence tautomeric transforms. We have applied these transforms to several large databases, such as PubChem, Aldrich Market Select (AMS), ChEMBL, and the organic part of the Cambridge Structural Database (CSD), to determine (a) the occurrence rate at which each of the transforms is applicable to molecules in these databases, (b) the success rate of current standard InChI[Key] for handling these types of tautomerism and (c) the success rate of nonstandard InChI[Key] for the same goal upon activation of keto-enol and 1,5 tautomerism. Based on these transforms, a tautomer prediction web tool ("Tautomerizer") has been developed that enumerates all possible tautomers of a given molecule using one, a selectable subset, or all of the rules. These transforms will form the basis of the final decision on the Redesign of the Handling of Tautomerism for InChI V2 and this will have a significant impact on the future design and behavior of InChI.

CINF 8

Kinase inhibitor selectivity data analysis

Zhaowen Luo[1]*, zhaowen.luo@emdserono.com, Veit Ulshoefer*[2]*. (1) EMD Serono Research and Development Institute, Inc, Wayland, Massachusetts, United States (2) Global Early Development, , Merck KGaA, Darmstadt, Germany*

Kinases are very attractive small molecular drug targets; over 40 kinase inhibitors are already approved by the FDA. One critical aspect of kinase inhibitors is the selectivity profile among kinases – there are over 700 kinases (protein, lipid, even carbohydrate), all have very similar small molecule binding pocket - ATP binding pocket. As a result, small molecule kinase inhibitors usually hit multiple kinase inhibitors, with very few exceptions of allosteric inhibitors. To address the issue, a set of tools were developed at EMD Serono to analyze, visualize and evaluate kinase selectivity profiles. Among them, there are a web base application to generate kinase inhibitory plots on kinome, a heatmap application to analyze multiple compounds kinase inhibitory profile for hit selection, as well as a set of Spotfire templates to address difference between kinase assays. By profiling kinase selectivity of approved small molecule kinase inhibitors, we identify a few metrics to better assess kinase selectivity and their effect on clinical outcomes. In addition, we also found that selectivity is not as important for oncology indications as initially thought, promiscuous inhibitors actually have a higher chance to make it to the market. Beside selectivity profile analysis, we also compared assay panel results from different CROs for the same kinases. While for majority of kinases, results are agreed with each other among suppliers, there are also a quite a few exceptions.

All these tools and results not only provide deep insight into kinase profiles, but also help project teams to evaluate kinase drug candidates and make critical decisions.

CINF 9

Universities and scholarly publishers collaborating to help students and postdocs advance their research and get published

Grace Baysinger[1], *graceb@stanford.edu*, *Shannon O'Reilly*[2], *S_OReilly@acs.org. (1) Stanford University Libraries, Stanford, California, United States (2) American Chemical Society, Washington, District of Columbia, United States*

Universities and scholarly publishers share a mission – to help students publish high-quality research and advance the field. Through collaboration, universities and publishers can work together to accomplish their shared goals and at the same time advance the research careers of students. This presentation will look at the 2018 collaboration between Stanford University's Science and Engineering Libraries and ACS Publications. In the spring, Stanford helps graduate students and postdocs "GearUp" by holding a half-day event. ACS Publications, through the ACS on Campus program, hosts professional development events on scholarly publishing and science communication to help students and researchers advance their careers. In April 2018, Stanford hosted an event with the theme "GearUp for Scientific and Technical Publishing" that included a session on Graphical Abstracts, an important part of any publication as they help users understand the research and play a key role in helping users find and choose which articles to read. A presentation did not exist prior to the event. Stanford University and ACS Publications, through the ACS on Campus program, worked together to develop a 30 minute presentation and identify a speaker, a Senior Editor of *The Journal of Physical Chemistry*. Through this collaboration, Stanford University hosted a successful event with a session on graphical abstracts and ACS Publications was able to add the presentation to its program portfolio. Now the presentation developed by Stanford is shown to universities across the US and the world. Working together, Stanford and ACS Publications were able to accomplish their goals and offer a needed resource to young researchers.

CINF 10

Partnership between librarians and non-profit stakeholders in research information ecosystem: WikiEdu and carpentries

Ye Li, *yeli@mines.edu. Arthur Lakes Library, Colorado School of Mines, Golden, Colorado, United States*

As scholarly and research information evolves to be more open in the digital age, new types of non-profit organizations beyond professional societies are emerging and

starting to take important roles in the ecosystem. Among them, Carpentries (https://carpentries.org) focus on teaching foundational coding and data science skills and Wiki Education (https://wikiedu.org) connects higher education to Wikipedia with the goal to enhance the most popular reference. Both organizations consider libraries, especially academic libraries, as their close partners to reach a broad audience globally. In this presentation, we will share the experiences of building and sustaining partnership with Wiki Education and Carpentries from the library perspective. Through these partnerships, the library was able to take advantage of the expertise and resources shared by the external partners to engage faculty and students for in-depth collaboration in curriculum development, research, and publishing. Meanwhile, Wiki Education and Carpentries relied on librarians' knowledge and network of the campus communities to accomplish their missions. Through the collaborative effort, the campus research and learning communities are better connected with the research information ecosystem as both information consumers and creators.

CINF 11

FAIR chemical data for health and safety: Connecting the dots with cheminformatics and librarianship

Leah R. McEwen[1], lrm1@cornell.edu, Evan Bolton[2], bolton@ncbi.nlm.nih.gov. (1) Clark Library, Cornell University, Ithaca, New York, United States (2) National Center for Biotechnology Information, Bethesda, Maryland, United States

Chemical data have been described for over 100 million characterized compounds and many data points are openly accessible in some form. Leveraging current technologies in data architecture and semantics, it is possible to programmatically compile and transpose these data at large scales for further analysis concerning a variety of chemical problems. Analyzing chemical data to support decision making in health and safety presents a range of additional complicating socio-technical factors to find, access, integrate and review (FAIR) appropriate data. Navigating these challenges becomes a classic information literacy scenario involving a diversity of data sources and community expertise in health and safety. Connecting the needs of specific use cases with available data and automated processing surfaces many opportunities for librarians to build bridges in partnership with different user communities and further address broader issues with data. This presentation will discuss how library expertise can help to maximize potential for cheminformatics in laboratory health and safety through the ongoing collaboration of a chemistry librarian with the PubChem chemical knowledge-base.

CINF 12

30 years of Reaxys: Chemical information for the chemists

Judith N. Currano[1], currano@pobox.upenn.edu, Jozica Dolenc[3], dolenc@chem.ethz.ch, Oliver Renn[2], oliver.renn@aol.de, Jürgen Swienty-Busch[4], J.Swienty-Busch@elsevier.com. (1) Chemistry Library, University of Pennsylvania, Jenkintown, Pennsylvania, United States (2) Chemistry & Applied Biosciences, ETH Zurich, Mittelbiberach, Germany (3) ETH Zurich, Zurich, Germany (4) Elsevier Information Systems GmbH, Frankfurt, Germany

When Beilstein and Gmelin decided to convert their handbooks into an online database, nobody could foresee the remarkable success of the chemistry information solution that, 30 years later, continues to deliver high quality chemistry data on substances and reactions to chemists. Over the past three decades, the system and the data have been developed to address the dynamic changes and needs of a modern information infrastructure, with the help of numerous chemists in industry and academia. Librarians have played a key role in advising the Reaxys product team on building a solution that fits to the needs of a diverse collection of chemists and have connected the team with a variety of different end-users to provide first-hand feedback on product ideas and release versions. This talk will present the history of Reaxys and its current state, will highlight the ways in which librarians have provided input into the development process, and will discuss Reaxys implementations at universities.

CINF 13

PubChem as a resource for chemical information training

Sunghwan Kim, kimsungh@ncbi.nlm.nih.gov, Evan Bolton. National Library of Medicine, National Institutes of Health, Bethesda, Maryland, United States

Libraries at many large academic institutions provide chemical information training programs for students. However, these programs are based on commercial chemical information resources, which come with non-trivial subscription fees. These fees are often too expensive for small organizations, including many primarily undergraduate institutions (PUIs) and community colleges (CCs). It leads to disparity in access to chemical information as well as learning opportunities among students. This issue may be addressed at least in part by developing free online training programs based on public chemical databases, such as PubChem (https://pubchem.ncbi.nlm.nih.gov). PubChem has a great potential as an online resource for chemical education, but it also has important issues that students and teachers should keep in mind, such as data accuracy, data provenance, structure standardization, terminologies and so on. In this presentation, we will discuss various aspects of PubChem as a resource for chemical information training.

CINF 14

SuCOS: A pharmacophoric-shape overlap metric for comparing binding modes

Susan Leung[1,2,5]*, susan.leung@st-hildas.ox.ac.uk, Michael Bodkin*[3]*, Frank von Delft*[4,5]*, Paul Brennan*[2,5]*, Garrett M. Morris*[1]*. (1) Statistics, University of Oxford, Oxford, United Kingdom (2) TDI NDMRB, University of Oxford, Oxford, United Kingdom (3) Evotec, Oxford, United Kingdom (4) Diamond Light Source, Didcot, United Kingdom (5) Structural Genomics Consortium, Oxford, United Kingdom*

One of the fundamental assumptions of hit-to-lead fragment-based drug discovery is that the binding mode of the fragment will be structurally conserved upon synthetic elaboration. Indeed, this was borne out by a recent survey of the X-ray crystal structures of fragments and elaborated-fragments by Malhotra and Karanicolas. Hence, during virtual screening of elaborated molecules, it is reasonable to keep only those screened molecules that retain the crystallographically observed binding mode. One of the most common ways of quantifying binding mode similarity is Root Mean Square Deviation (RMSD) of the positions of corresponding atoms. Protein-Ligand Interaction Fingerprints (PLIFs) are an increasingly used alternative way to compare binding modes, and in particular, explicit interactions made between the ligand and the protein. We present *SuCOS,* an open-source RDKit-based implementation of Malhotra and Karanicolas' Combined Overlap Score (COS). SuCOS has a Pearson correlation coefficient with COS of 0.92. We compared the performance of RMSD, PLIF-Tversky/PLIF-Tanimoto, and SuCOS on (i) Malhotra and Karanicolas' dataset of paired larger and smaller molecules bound to the same protein; (ii) redocking of the larger and smaller molecules into their respective proteins using AutoDock Vina; and (iii) cross-docking of the larger molecule into the smaller molecule's cognate protein structure using AutoDock Vina. We show that combined volumetric and 3D-pharmacophoric-based metrics like SuCOS are superior to RMSD when comparing an elaborated fragment (larger molecule) with its original fragment hit counterpart (smaller molecule). When the molecules are identical, such as in redocking, the threshold of 2 Å RMSD is often used. However, this often disregards the size of the molecules being compared. The SuCOS score ranges from 0 to 1, regardless of molecular size, and is therefore suitable for defining a more universal threshold. SuCOS also has potential applications in ligand-based and implicit structure-based virtual screening.

CINF 15

LigandNet: A machine-learning-based toolkit for predicting ligand activity to proteins

Md Mahmudulla Hassan, Daniel castaneda, Dewan Shrestha, Ibrahim Salama, ***Suman Sirimulla****, ssirimulla@utep.edu. School of Pharmacy, The University of Texas at El Paso, El Paso, Texas, United States*

LigandNet is a Machine Learning (ML) toolbox that combines different ligand-based models into an open source platform that can predict if a ligand may have an activity to a specific protein. Finding a ligand that will bind to a human protein and have a

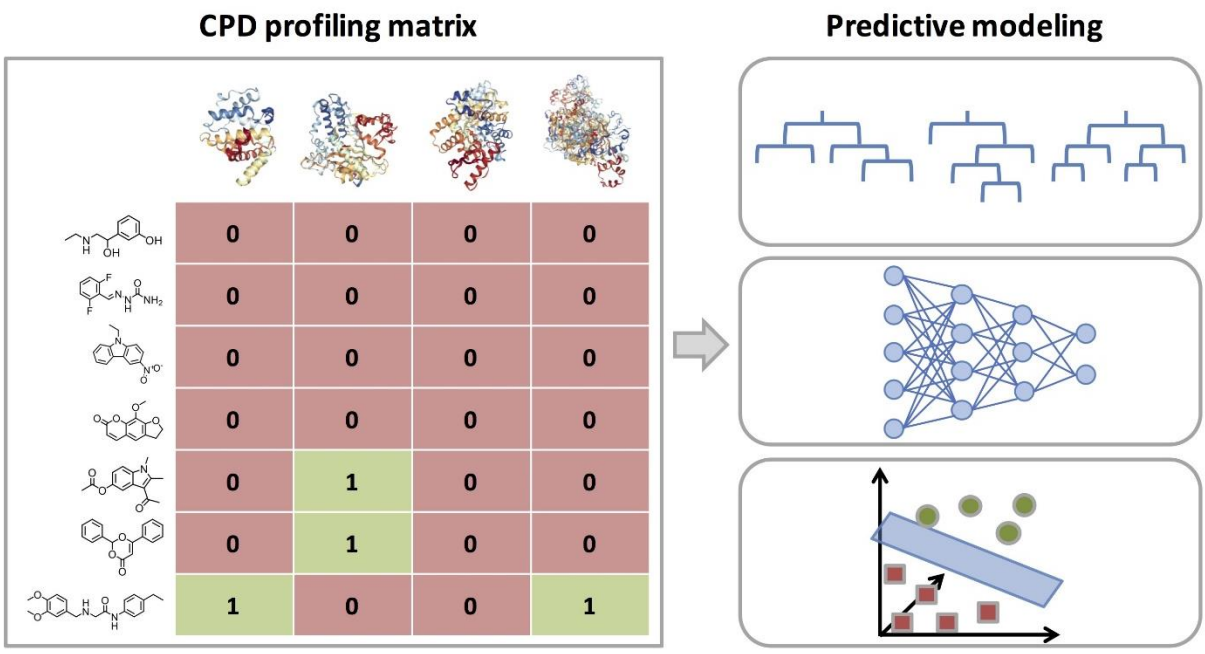
significant signaling effect through the cell can be an expensive task. Nowadays, ML models are being employed throughout many scientific fields and are being successfully applied to drug discovery research. In this project, we have applied advanced ML approaches such as Random Forest (RF), Support Vector Machine (SVM), Linear Regression (LR), Extra Tree Classifier (ETC) and Deep Learning (DL) to classify the ligands as active/binder or inactive/nonbinder. We obtained the known active ligands for each of 1704 proteins from Pharos (pharos.nih.gov) database. For each of the known active ligands, decoys were generated using DecoyFinder (http://urvnutrigenomica-ctns.github.io/DecoyFinder/) and Zinc database (http://zinc.docking.org/). ML models were developed for each of the protein-ligand sets by using known actives and generated decoys. ECFP6 fingerprints and Topological Pharmacophore Atomic Triplet Fingerprint (TPATF) from MayaChemTools (http://www.mayachemtools.org/) were employed as feature generators in developing the models. Models were validated using highest positive predictive value (PPV), sensitivity, and area under the curve of the receiver operating characteristic plots (ROC-AUC) to determine which model works best with each dataset of the proteins, determining the accuracy and precision of each ML approach. The developed models are available on GitHub (https://github.com/sirimullalab/LigandNet).

CINF 16

Machine learning-based prediction of compound profiling matrices

Raquel Rodríguez Pérez[2,1], *raquel.rp13@gmail.com, Jurgen Bajorath*[2]. *(1) Medicinal Chemistry, Boehringer Ingelheim, Biberach an der Riss, Germany (2) Life Science Informatics, University of Bonn, B-IT, Bonn, Germany*

Machine learning (ML) including deep learning (DL) is increasingly applied in pharmaceutical research for a variety of applications including predictive modeling of high-throughput screening (HTS) data. However, HTS data display a large imbalance between active and inactive compounds and contain experimental noise, which complicates ML-based predictions. Compound profiling matrices represent a special form of HTS data. Such matrices are obtained by screening a compound collection against multiple targets and present challenging prediction tasks for ML (Figure 1). We have compared a variety of ML and DL methods for the prediction of large compound profiling matrices. Different molecular representations and prediction strategies were explored. Although predictive performance varied across assays comprising matrices, promising results were obtained for many targets. Perhaps surprisingly, DL did not further increase the predictive performance of standard ML methods such as random forests (RF). RF models correctly prioritized active compounds for most assays of test matrices indicating the ability of standard ML techniques to identify novel hits under rather challenging experimental conditions.



CINF 17

Collaborations and data sharing in rare disease

***Rachelle J. Bienstock**, rachelleb1@gmail.com. RJB Computational Modeling LLC, Chapel Hill, North Carolina, United States*

A rare or orphan disease is one which affects a small percentage of the population, (defined in the EU as a life-threatening or debilitating disease occurring in no more than 1 in 2000 people; in the US as any disease that affects fewer than 200,000 (1 in 1500 people) or in Japan as a disease that affects 1 in 2500 people). There are about 6000-8000 rare diseases affecting 8% of the worlds people. Most are genetic in origin. Due to their low prevalence, it is essential that genetic and other data be shared between organizations across the globe to have a chance of impacting treatment and/or drug discovery. This has motivated collaborative research efforts including the Global Rare Diseases Registry Repository, sponsored by the National Center for Advancing Translational Science in the US, and on an international level, the International Rare Disease Research Consortium and ERA-Net for Research Programmes on Rare Diseases. Registry and tissue repository databases provide valuable data resources to a large community of researchers. Examples also include the Matchmaker Exchange, MME (http://www.matchmakerexchange.org), a federated network of genotype and rare phenotype databases which facilitate the interaction between multiple disconnected projects. This presentation with introduce the symposium and discuss ways in which sharing of rare disease data can be facilitated, and innovations in collaboration methods.

CINF 18

Genetic and Rare Diseases (GARD) information center treatment profiles

Qian Zhu, qian.zhu@nih.gov, Dac-Trung Nguyen, Noel Southall, Alice Chen, Eric Sid, Anne Pariser. NCATS, NIH, Potomac, Maryland, United States

The Genetic and Rare Diseases (GARD) Information Center was established in 2002 to provide up-to-date information for approximately 7,000 genetic and rare diseases. This resource, which is currently managed by the Office of Rare Diseases Research (ORDR) within the National Center for Advancing Translational Sciences (NCATS), has remained an important portal for patients, health-care professionals, and researchers seeking to understand the current state of genetic and rare diseases. Recent advances in our understanding of the genetic and molecular basis of disease have resulted in great progress toward effective management and treatment of rare diseases. To provide better insight into this progress, herein we describe our ongoing efforts to further develop GARD as a resource to encourage scientific research. Specifically, we are working to integrate GARD data with other data sources to identify treatments approved or in development for each rare disease. FDA approved drugs from FDA Orphan Drug Designations will be integrated along with mappings between GARD and ClinicalTrials.gov to highlight completed and ongoing trials for each disease. Treatments not identified through direct data mining can also be inferred via genetic factors, which are the underlying cause of most rare diseases. Ultimately, these treatment development profiles will be integrated into the GARD resource for public use.

CINF 19

Biomedical data translator: Supporting data integration and rare disease research

Noel Southall, noel.southall@gmail.com, Christine Colvis. National Center for Advancing Translational Sciences (NCATS), National Institutes of Health, Bethesda, Maryland, United States

The National Center for Advancing Translational Sciences (NCATS) — one of 27 Institutes and Centers at the National Institutes of Health (NIH) — was established to transform the process of delivering treatments and cures for disease to patients. One strategic focus for the Center is on rare diseases. With 7000 rare diseases, a one-disease-at-a-time approach to clinical and translational science will not work. Instead, NCATS seeks to find what is common among diseases and demonstrate how advances for one disease can seed innovation in other areas. Today there is a tremendous amount of biomedical research data and data available from disease classifications, health records, clinical trials and adverse event reports that could be useful for understanding health and disease and for developing and identifying

treatments for all diseases. Ideally, these data would be mined collectively to provide insights into the relationship between molecular and cellular processes (the targets of rational drug design) and the signs and symptoms of diseases. But currently, these very rich yet different data sources are housed in various locations, often in forms that are not compatible or interoperable with each other. All of these factors limit the ability to get more treatments to more patients more quickly. To address these problems, NCATS launched the Biomedical Data Translator program, called “Translator” for short. Through this program, NCATS is supporting research to develop a groundbreaking computational tool that enables connections among conventionally siloed data types. Once completed, Translator will be able to draw on data sources ranging from air quality measurements to electronic health records and molecular mechanisms of disease. Initial demonstration projects focus on data sharing, novel integrations of data, and novel analytic capabilities, leading to a better understanding of the relationships between rare and common diseases and helping NCATS to realize its rare disease strategic vision.

CINF 20

Data-driven drug discovery for rare diseases: tales from the trenches

Frederik van den Broek, f.broek@elsevier.com. Elsevier, Amsterdam, Netherlands

For the pharmaceutical industry as a whole, addressing the challenge of rare or orphan diseases is high on the agenda. But for the patients and their families, rare diseases can be very isolating and it can often feel like the potential for new treatments is low. One avenue for potential treatments is to identify drug repurposing candidates for the rare disease in question. This talk will give an overview of various collaborative projects undertaken in the last few years, which involved the combination, normalisation and analysis of data from various disparate sources, including some valuable lessons learnt along the way.

CINF 21

Computational chemistry and chemoinformatics career opportunities at the NIH (NIEHS)

Rachelle J. Bienstock, rachelleb1@gmail.com. RJB Computational Modeling LLC, Chapel Hill, North Carolina, United States

Dr. Bienstock will discuss career opportunities available in the areas of computational chemistry and chemoinformatics at the NIH specifically, and in federal government research laboratories in general. Dr. Bienstock worked for many years as a computational chemist (contractor) at The National Institute of Environmental Sciences (NIEHS), one of the National Institute of Health (NIH) laboratories. She will discuss the collaboratory and

interactive research environment within government labs in general, and specifically some of the types of projects on which she worked. These projects involved protein modeling, protein-protein interaction modeling, docking and ligand design, and working closely with experimental groups. She will contrast the type of projects, opportunities and the work environment for computationally oriented chemists within the NIH, with those available within industry, academic and other employment sectors.

CINF 22

Careers in publishing chemical information: From the lab bench to the editoral office to the database

Guy Jones, jonesg@rsc.org. Royal Society of Chemistry, Cambridge, United Kingdom

The systems for communicating chemical information are evolving as swiftly as the discoveries that underpin the science from peer-reviewed academic journals and curated databases, to new outlets such as pre-print servers and data repositories. There are many opportunities within this environment for chemists looking to start a career in scientific publishing. This talk will provide an overview of some of these roles, providing insights into life working in journals, peer review, scientific ethics, journal management, chemical databases, people management, open science and research data, from the perspective of a reformed organometallic chemist working at the Royal Society of Chemistry – a not-for-profit society publisher dedicated to advancing excellence in the chemical sciences.

CINF 23

Water-quality data and publications for careers in chemistry information

Emily C. Wild, ewild@princeton.edu. Library, Scholarly Collections & Research Services, Princeton University, Princeton, New Jersey, United States

Library and information materials have become more available online and within databases for water quality. Nevertheless, finding all the relevant historical and current information resources have become more challenging for research chemists as core source materials evolve from various access portals and information providers change missions and areas of study. As libraries downsize and library collections move to offsite facilities, chemistry librarians have a vital role in showing chemists and other professionals how to find and use water-quality information through time within many library and information sources. Online content is indexed and available as full-text content; however, there are many databases to navigate and many journals, conference proceedings, government reports and other series that are only partially available as full-text content, while some information sources remain available only in print and(or) may

be on a library shelf and "in-line" to be indexed. This presentation will be an overview of the print, digital, and online chemistry information sources used in hydrology for water-quality research inquiries and how one can find careers in geochemistry information.

CINF 24

Scientist in EH&S: Changing the tradition in laboratory safety

shailendra singh[1]*, shailen2@andrew.cmu.edu, Neelam Bharti*[2]*. (1) Environmental Health and safety, Carnegie Mellon University, Pittsburgh, Pennsylvania, United States (2) Mellon Institute Library, Carnegie Mellon University, Pittsburgh, Pennsylvania, United States*

Laboratory environments can be hazardous places to work, as lab personnel are exposed to numerous potential hazards including chemical, biological, physical, and radioactive hazards. The academic research community can benefit from an environmental, health, and safety (EH&S) professional, who has subject matter expertize, has the scientific background to understand laboratory processes, knows technical language and the vocabulary of regulatory nomenclature, and has hazard assessment skills. The theme of this presentation will be how a laboratory scientist can emerge as an EH&S professional and how research lab experience can be a given advantage in EH&S. At the same time, I will also highlight some of the professional challenges it bring to a scientist turned EH&S professional when dealing with the research community and traditional EH&S system.

CINF 25

Antony Williams, the ChemConnector: A career path through a diverse series of roles and responsibilities

Antony J. Williams*, tony27587@gmail.com. National Center for Computational Toxicology, Environmental Protection Agency, Wake Forest, North Carolina, United States*

Antony Williams is a Computational Chemist at the US Environmental Protection Agency in the National Center for Computational Toxicology. He has been involved in cheminformatics and the dissemination of chemical information for over twenty-five years. He has worked for a Fortune 500 company (Eastman Kodak), in two successful start-ups (ACD/Labs and ChemSpider), for the Royal Society of Chemistry (in publishing) and, now, at the EPA. Throughout his career path he has experienced multiple diverse work cultures and focused his efforts on understanding the needs of his employers and the often unrecognized needs of a larger community. Antony will provide a short overview of his career path and discuss the various decisions that helped motivate his change in career from professional spectroscopist to website host and

innovator, to working for one of the world's foremost scientific societies and now for one of the most impactful government organizations in the world. *This abstract does not necessarily represent the views or policies of the U.S. Environmental Protection Agency.*

CINF 26

Careers in science: Science policy and general advice

***Edward Dunlea**, edunlea@andrew.cmu.edu. Mellon College of Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, United States*

Considering a career path outside academia? It can be confusing to figure out what else is out there, how to look for jobs, and how to decide what's right for you. As a Ph.D. Chemist, I have been down this road myself and can offer some insights gleaned from my experiences in program management within the government, science policy both at a non-profit and within the private sector, and research administration in academia. My talk will focus on the field of science policy – what it is and how to find jobs – and I will also offer some general advice on how to translate the skills learned as a chemist into careers outside of academia.

CINF 27

How interests and experience led to a career in chemical literature informatics

***Nancy C. Baker**, baker.nancy@epa.gov. Leidos, Hillsborough, North Carolina, United States*

Nancy Baker is a literature informatics researcher at Leidos and works currently at the EPA supporting research in chemical toxicity through innovative methods in literature informatics. After an education in liberal arts and languages, she started a career in computing and eventually landed at the drug company GSK. During her 16 years at the pharmaceutical company building software, she worked with chemists and biologists whose enthusiasm sparked her interest in science. She returned part-time to the classroom to study chemistry, biology, and genetics and eventually enrolled in a PhD program Information Science and the Program in Bioinformatics and Computational Biology at the University of North Carolina in Chapel Hill. Nancy will provide an overview of her career path and discuss the merits of nontraditional paths to a nontraditional career.

CINF 28

Lab to library: A career in chemistry librarianship

***Nicholas Ruhs**, nruhs@fsu.edu. University Libraries , Florida State University, Tallahassee, Florida, United States*

With the advent of the digital age we have seen a major shift in how chemistry scholars process, store, and retrieve information. Scholars often call upon academic librarians to assist them in navigating this ever-increasing availability of information. Since much of this information is highly specialized and subject-specific, there exists a growing need for librarians with chemistry skills, training, and expertise. However, a career in librarianship is often overlooked or not considered by chemistry students or those looking to transition from a professional research or teaching position. Students and professionals have often acquired expertise in areas that transfer well to a career in librarianship, including skills in information and reference retrieval, data management, data analysis, citation management, and instruction. As a former graduate research chemist turned librarian, I can attest to the skills overlap between the two roles. In this talk, I will describe my career journey from Ph.D. chemistry student to academic librarian, my considerations for making such a move, and how the responsibilities of my current position compare with those of different chemical professions. Strategies for those considering a career in chemical librarianship will also be discussed.

CINF 29

Computational-aided design of diversity: Chemical libraries based on natural products

Fernanda Saldivar[1]*, fer.saldivarg@gmail.com, Elena Lenci*[2]*, Andrea Trabocchi*[2]*, José L. Medina-Franco*[1]*. (1) School of Chemistry, Departament of Pharmacy, National Autonomous University of Mexico, Mexico City, Mexico City, Mexico (2) Departament of Chemistry Ugo Schiff, University of Florence, Florence, Italy*

Compounds of natural origin are among the most favorable source to find drug candidates. However, research in this field has been more complex, expensive and inefficient compared to the investigation of small molecules from synthesis. Therefore, the use of computational methods to find new hits structures from natural products (NP) represents an alternative to overcome these problems and perform a more rational and economic search. Compounds from NP also offer innovative ring systems with geometries suitable for the spatial positioning of the side chains. Therefore, much of the NP have been widely used for the design of chemical libraries and the design based on fragments. In this context, diversity-oriented synthesis (DOS) and biology-oriented synthesis (BIOS) represent two main approaches for the synthesis of collections inspired by NP. Among the different NP classes, bicyclic acetals and spiroacetals are important starting points in both approaches due to their intrinsic biological value and unique structural characteristics. For this reason, this class of compounds was selected and, through the use of chemoinformatic tools, workflows were generated to automate the search and analysis of scaffolds of biological importance, for the subsequent design of DOS libraries.

The generated workflows can be used for the analysis of other scaffolds of biological importance and the subsequent design of chemical libraries.

CINF 30

Classification models of pesticides by mode of action

*Adriana Osnaya-Hernandez[1], **Gabriela Gómez-Jiménez**[1], gabbgomez@hotmail.com, Daniel Chavez[1], Fernando Cortes-Guzman[1], Abraham Madariaga[2], Karina Martinez Mayorga[1]. (1) Physical Chemistry, Instituto de Quimica, UNAM, Scottsdale, Arizona, United States (2) Physical Chemistry, Universidad Nacional Autonoma de Mexico, Mexico City, Mexico*

Pesticides are substances widely used in agriculture to control pests, including weeds. One of the most informative and commonly used endpoint in regulatory settings is the acute oral toxicity in rat. This endpoint can be predicted using QSAR models on different computational platforms. These are typically global or similarity-based models. Classifications by modes of action (MOA) could generate local models with higher accuracy. In this study, we present a classification model by MOA of pesticides to estimate acute oral toxicity in rat. The IUPAC - Pesticides Property DataBase (PPDB) was used as a reference. IUPAC-PPDB contains 2288 pesticides including insecticides, fungicides and herbicides and are classified by MOA. Two databases were assigned to the most common MOA: the EPA T.E.S.T. database (v. 4.2.1), which contains 7413 pesticides, and pesticides used in Mexico (Cofepris Pesticides Database: CPD, which contains 150 compounds). Depending on the MOA, assignments were performed using Random Forest or KNN algorithms. An initial exploration of the databases consisted of the analysis of the chemical space coverage of the three databases (PPDB, T.E.S.T and CPD). Predictive models, using these data, are in progress. As an example, we present a model of acute oral toxicity in rat, developed using 121 similar pesticides. Descriptors were calculated with Dragon 6. Predictive models (MLR), developed in QSARINS, consisted of multiple linear regressions, estimated with ordinary least squares, and genetic algorithms for variable selection. Leave-one-out cross validation, Y-scrambling and applicability domain was analyzed. The statistics of the best model are: Model development and internal validation: R2=0.816, Q2LOO=0.777. External validation: Q2F1=0.647 Retx=0.644, CCC=0.789.

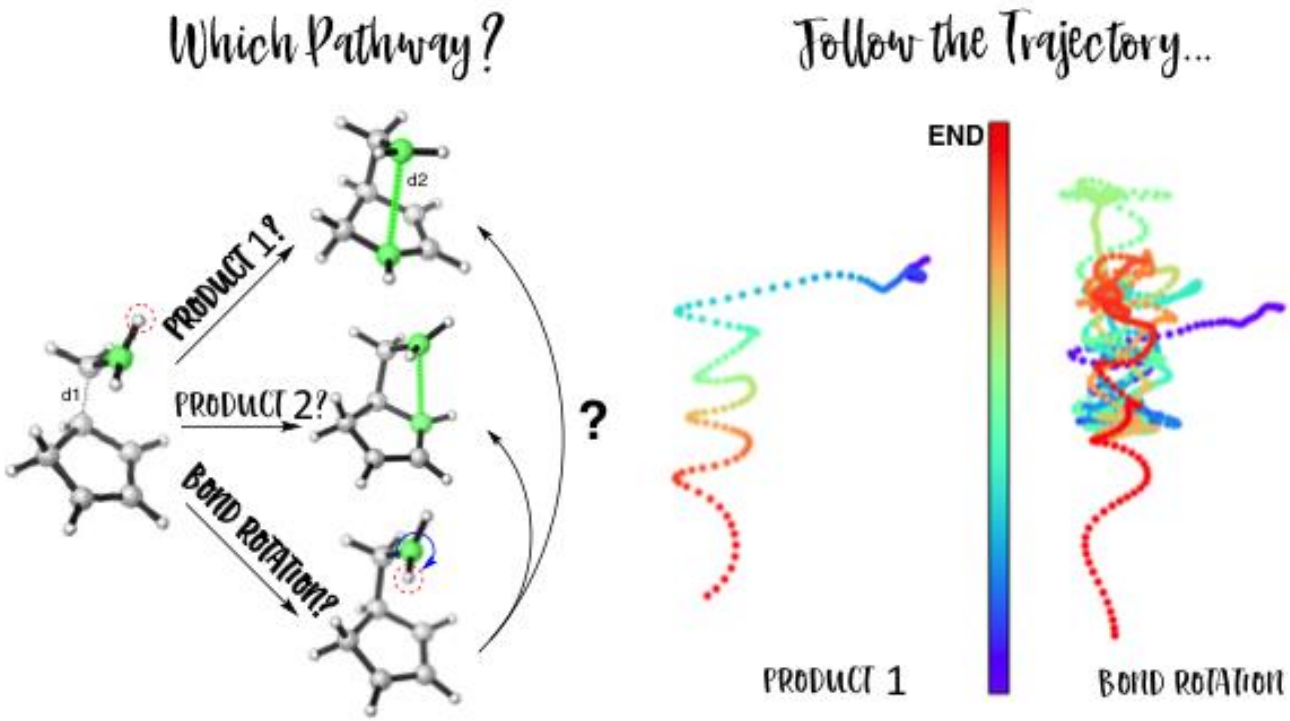
CINF 31

Understanding stereoselectivity in radical cation diels-alder reactions using quasi-classical dynamics

Jacqueline Tan[1], *Jacqueline.tan@chem.ox.ac.uk, Robert S. Paton[2,3]. (1) Oxford, Oxford, United Kingdom (2) Chemistry Research Laboratory, University of Oxford, Oxford,*

United Kingdom (3) Chemistry, Colorado State University, Colorado, Colorado, United States

The Diels-Alder (DA) reaction is among the most important and versatile methods in creating ring molecules, and factors governing stereoselectivity and rate have been widely studied. However, the chemoselectivity in products are often restricted, as electronic natures of the reactants (typically electron-rich diene and electron-poor dienophile) have to be matched. On the other hand, in radical chemistry, the removal of a free election helps promote an entirely different chemical environment within the reactants. This is experimentally shown to be imperative in forming a different set of products not predicted in neutral DA reactions using photochemistry or redox chemistry. Using an emerging technique known as quasi-classical dynamics, coupled with Density Functional Theory, we investigated the radical cation DA reactions of anethole with isoprene and cyclopentadiene to provide quantitative understanding to this observed phenomena. We also aim to demonstrate the applicability of this understanding in reactions where minor products are actually desired. This has never been applied to the study of radical cycloadditions before.



CINF 32

Application of *ab initio* molecular dynamic simulation in 4D fingerprints

Yi-Shu Tu[1], *georgetu@gmail.com, Y. Jane Tseng*[1], *Michael Appel*[2]. *(1) Biomedical Informatics, National Taiwan University, Taipei, Taiwan (2) USDA-ARS, Dunlap, Illinois, United States*

The invention of 4D Fingerprints (4DFP) is a breakthrough achievement of 3D and 4D Quantitative Structure-Activity Relationship (QSAR) methods since it eliminates the requirement of manual alignments through the chemical structures. In addition, it adds the dynamic consideration into 3D QSAR instead of considering only conformation in the minimum energetic state. However, the 4DFP only use traditional molecular dynamic simulation with predefined forcefields. Applying traditional molecular dynamic simulation methods to small molecules may lead to some inaccurate bonding dynamic problems, and limited applicable structures. Here we present the use of PM3 semi-empirical methods with *ab initio* molecular dynamics simulation to generate the 4DFP of trichothecene molecules. With contemporary computer technologies, there is no significant computational time increase by replacing traditional molecular dynamics with PM3 molecular dynamics. However, the PM3 methods can improve the accuracy of studying and applying dynamics of small molecules.

CINF 33

ASKCOS: Data-driven synthetic route design and validation for small organic molecules

Connor W. Coley[1], *ccoley@mit.edu, Hanyu Gao*[1], *William H. Green*[2], *Klavs F. Jensen*[3]. *(1) Chemical Engineering, MIT, Cambridge, Massachusetts, United States (2) Rm E17-504, MIT, Cambridge, Massachusetts, United States (3) Dept of Chem Eng Rm 66 350, MIT, Cambridge, Massachusetts, United States*

In this poster, I will summarize our development of the open access ASKCOS synthesis planning suite (askcos.mit.edu). Advances in laboratory automation promise to decrease the manual effort of synthesis, but determining *how* to synthesize a compound currently requires time and effort investment from expert chemists. Our overall synthesis planning workflow contains a number of interconnected modules. First, we address the problem of retrosynthetic planning and how the recursive expansion and search strategy are both conducive to machine learning approaches. Second, we address the challenge of *in silico* reaction validation, which can be addressed by solving the inverse problem of forward reaction prediction. This poster will describe how our techniques for retrosynthesis and forward prediction—leveraging advances in data science and machine learning techniques—are integrated into an overall workflow that, for a given molecular target, predicts a rank-ordered list of reaction paths that connect the target to purchasable starting materials via a series of plausible reaction steps. The software is able to find robust synthetic pathways in

seconds for simple organic molecules and in less than a minute even for modern small molecule APIs. I will also discuss how our software has been adopted by a medicinal and process chemistry groups in several pharmaceutical and chemical companies and opportunities for extension.

CINF 34

Hierarchical H-QSAR modeling method that integrates binary/multi classification and regression models for predicting acute oral systemic toxicity

XINHAO LI, xli74@ncsu.edu, Denis Fourches. Chemistry, North Carolina State University, Raleigh, North Carolina, United States

Acute toxicity is an important endpoint for chemical risk assessment. As animal testing is still the major avenue for assessing chemical's adverse effects, *in silico* screening represents perhaps the most promising alternative approach. Quantitative structure-activity relationships (QSAR) models can be used to estimating chemical induced toxicity with continuous (*regression*) or discrete (*classification*) predictions. It is still quite unclear how those different types of models can complement and help each other to afford the best prediction accuracy for a given chemical. Herein, we will discuss a novel, dual-layer hierarchical modeling method to build QSAR models for predicting categorical (binary toxic/ nontoxic and four EPA-defined categories) and continuous (LD_{50}) endpoints for assessing rat acute oral toxicity. The training and external test sets were compiled by NICEATM/NCCT and contained 8,213 and 2,843 compounds, respectively. Each molecule has at least one label of 'toxic/nontoxic', EPA category (class I, II, III, and IV) and LD_{50} values. The first layer of independent models (*base* models) was solely built with computed chemical descriptors. Then, a second layer of models (*hierarchical* models) were built by stacking all the cross-validated predictions from the *base* models. In addition, the applicability domain of these models was defined according to the '*model prediction zone*', which is based on the prediction probabilities of all binary and multiclass models, resulting in four prediction zones: '*in-zone*' (low confidence), '*half-zone-binary*', '*half-zone-multiclass*', and '*out-zone*' (high confidence). All models were then evaluated using the external test set. The results showed that *(i)* hierarchical models take advantage of the knowledge learned by the base models, resulting in improved prediction performances (R^2 = 0.47, RMSE = 0.62 for base model; R^2 = 0.54, RMSE = 0.58 for hierarchical regression model on external test set); *(ii)* for each individual model, 'out-zone' predictions afforded the best prediction performances (up to R^2 = 0.65, RMSE = 0.51 for hierarchical regression model with a coverage of 64% on the external test set); and *(iii)* hierarchical models have larger coverage of '*out-zone*' predictions than those of the corresponding base models. Overall, this hierarchical H-QSAR modeling method relying on the full stacking of binary, multiclass, and regression base models represents a promising approach for chemical risk assessment.

CINF 35

In silico platform as an alternative to animal testing for acute toxicity

Joyce Borba[1,2]*, joycevillaverde@gmail.com, Vinicius M. Alves*[1]*, Arhur C. Silva*[2]*, Kirsten Overdahl*[3]*, Stephen Capuzzi*[1]*, Erik Overdahl*[1]*, Daniel Korn*[1]*, Ruither Silva*[1]*, Steven Hall*[2]*, Rodolpho Braga*[2]*, Nicole Kleinstreuer*[4]*, Carolina H. Andrade*[2]*, Eugene Muratov*[1]*, Alexander Tropsha*[1]*. (1) Laboratory for Molecular Modeling, University of North Carolina, Chapel Hill, North Carolina, United States (2) Laboratory for Molecular Modeling and Drug design, Federal University of Goias, Goiania, Goias, Brazil (3) Nicholas School of the Environment, Duke University, Durham, North Carolina, United States (4) NTP-ICVAM, RTP, North Carolina, United States*

Acute toxicity tests are used to identify hazard potentials appearing after very short exposure times.Since 2009, animal testing for cosmetic products was prohibited in Europe, and in 2016, US EPA published a guideline for waiving the so-called “6 pack” battery tests (acute oral toxicity, acute dermal toxicity, acute inhalation toxicity, skin irritation and corrosion, eye irritation and corrosion, skin sensitization) to reduce animal testing of pesticides. Computational models that can accurately predict potential hazards for cosmetics, drugs and pesticides find growing use in both laboratory research and regulatory decision support. As part of our overarching program on the development of the 6 pack virtual screening platform, we have developed a new tool for the rapid identification of potential skin sensitizing compounds as well as compounds causing acute inhalation toxicity. We have compiled, curated, and integrated the largest publicly available dataset for the endpoints listed in Table 1. These data were used to develop an ensemble of QSAR models using several approaches such as MuDRA, Random Forest, and Deep Learning. All models were developed and validated according to the OECD QSAR principles. Different metrics of external model accuracy are shown in Table 1. Model interpretation revealed several SAR rules, which can guide structural optimization of toxicants toward making them into non-toxicants. Virtual screening of the REACH database using the developed QSAR models identified several potential toxicants. Models have been made available within our publicly accessible PredSkin web portal (http://www.labmol.com.br/predskin/) and will be shown as part of the presentation. These models can be employed by users to identify both putative toxicants and non-toxicants in chemical libraries of their interest.

Table 1. Computational models for four of the 6-pack animal tests

	Skin sensitization	Eye irritation and corrosion	Acute inhalation toxicity	Acute oral toxicity
Compiled data	10861 records	6387 records	4647 records	8994 records
Curated data	1000 compounds	3547 compounds	528 compounds	8506 compounds
Model metrics	CCR=0.89 Se = 0.94/Sp = 0.84/ PPV = 0.91/NPV = 0.89	CCR=0.88 Se = 0.87/Sp = 0.88/ PPV = 0.83/NPV = 0.91	CCR=0.73 Se = 0.70/Sp = 0.76/ PPV = 0.73/NPV = 0.73	CCR=0.72 Se = 0.74/Sp = 0.78/ PPV = 0.73/NPV = 0.85

CINF 36

BDEDB: A bond-dissociation energy database and instant prediction

Yanfei Guan[1], yanfei.guan@chem.tamu.edu, Robert S. Paton[1,2]. (1) Chemsitry, Colorado State University, Fort Collins, Colorado, United States (2) Chemistry Research Laboratory, University of Oxford, Oxford, United Kingdom

The homolytic cleavage of covalent bonds is of great importance to many chemistry areas, especially in metabolism, combustion and catalysis. An instant predictor for bond dissociation energies (BDEs) of a large number of molecules that doesn't require specific devices (super-cluster) is therefore of great value to a broad cross-section of chemists. However, due to the absence of databases for BDEs, it is quite difficult to develop and evaluate such tools. Herein, we summarize our progress in building a BDE database, which significantly facilitates accessing, querying and analysis for a large amount of BDE data. This BDE database is a SQL database and is built on the Atomic Simulation Environment Data Base. A workflow was built to automatically grab molecules from PubChem dataset, optimize molecules/radicals, calculate BDEs, and populate the BDE database. Furthermore, the constructed database is used as training set for several different machine learning models to predict BDEs. The best model, modified Schnet (a continuous-filter convolutional neural network), provides comparable accuracy to DFT calculations at significantly lower costs. Using the selected model, a web-based instant BDE predictor will be built.

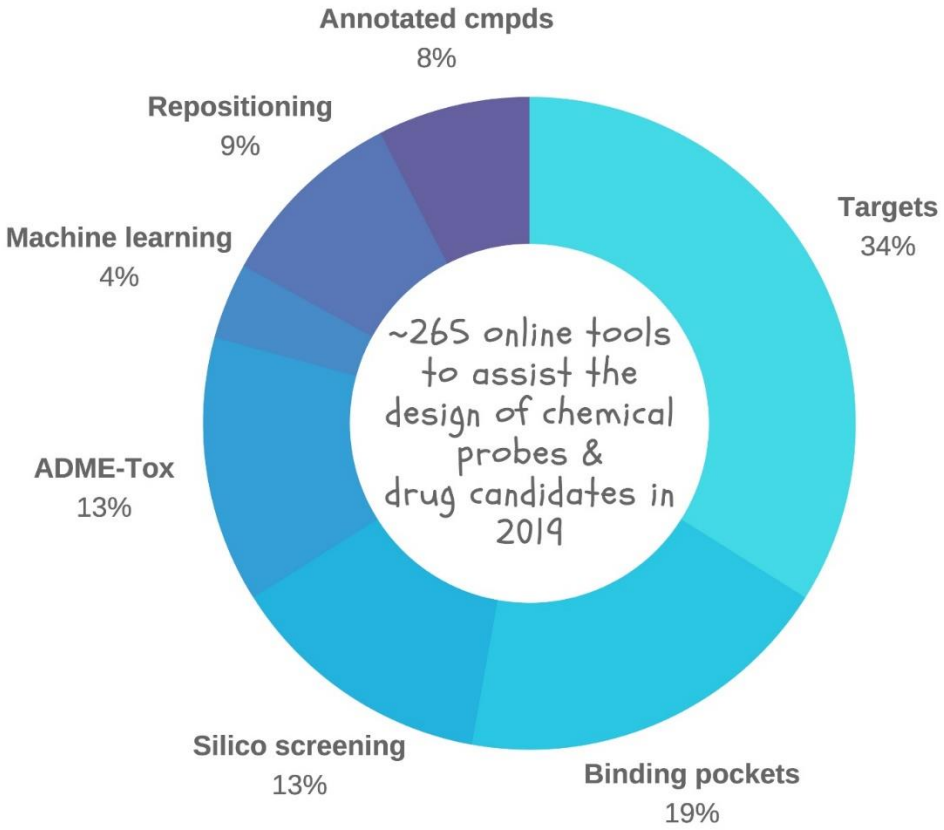
CINF 37

Designing drug candidates and chemical probes in cyberspace

Bruno Villoutreix[1,2,3], bruno.villoutreix@gmail.com. (1) Inserm U973, Paris, France (2) Lille University, Lille, France (3) Pasteur Institute, Lille, France

In silico resources enabling and supporting chemical biology and the first steps of drug discovery have blossomed during these past twenty years. For example, in 2005, I found about 500 databases and in silico tools (online and standalone packages) that could be

of particular interest to medicinal chemistry. These included some databases with annotated small molecules and targets, structural predictions and analysis of (putative) therapeutic targets, ligand-based and structure-based virtual screening tools, QSAR models and ADME-Tox prediction engines… In 2019, I now have a compilation of about 3200 URLs (www.vls3d.com), dedicated to structural bioinformatics and chemoinformatics. Among these, about 265 online tools and databases can be used to assist the design of chemical probes including short peptides and drug candidates. Clearly, if fully exploited, these inspiring tools and databases help to identify bioactive molecules, to improve the properties of small compounds and to investigate targets, off-targets and anti-targets. I will highlight some recently reported web services in the field of chemo- and pharmaco-informatics, discuss approaches that are usually needed during the course of a project and give some some examples of tools developed in our laboratory such as FAF-Drugs, AMMOS and MTiOpenScreen.



CINF 38

Cheminformatics tools and applications on the web: Challenges, examples, and the future

Denis Fourches, *dfourch@ncsu.edu. Chemistry, North Carolina State University, Raleigh, North Carolina, United States*

It has never been easier to find easy-to-use cheminformatics tools and applications on the web. With cheminformatics-powered websites such as PubChem, PDB, CompTox Dashboard, or Chemspider, researchers without cheminformatics skills can mine large sets of chemical structures, visualize three-dimensional structures of proteins, and even explore protein-ligand interfaces. This is a true game-changer for the whole community, but those new tools also come with challenges and still untapped opportunities. Herein, we will discuss the potential challenges the current generation of web services can face (*e.g.,* oversimplification of chemical biological data, black boxes). Then, we will present our two most recent webservers: *(i)* RealityConvert.com, which generates chemical input files for 3D printers and augmented/virtual reality devices, and *(ii)* ChemMaps.com, which allows for a user-friendly, browser-based exploration of the chemical space. Finally, we will discuss several avenues that could likely shape the future of web-based, connected cheminformatics platforms.

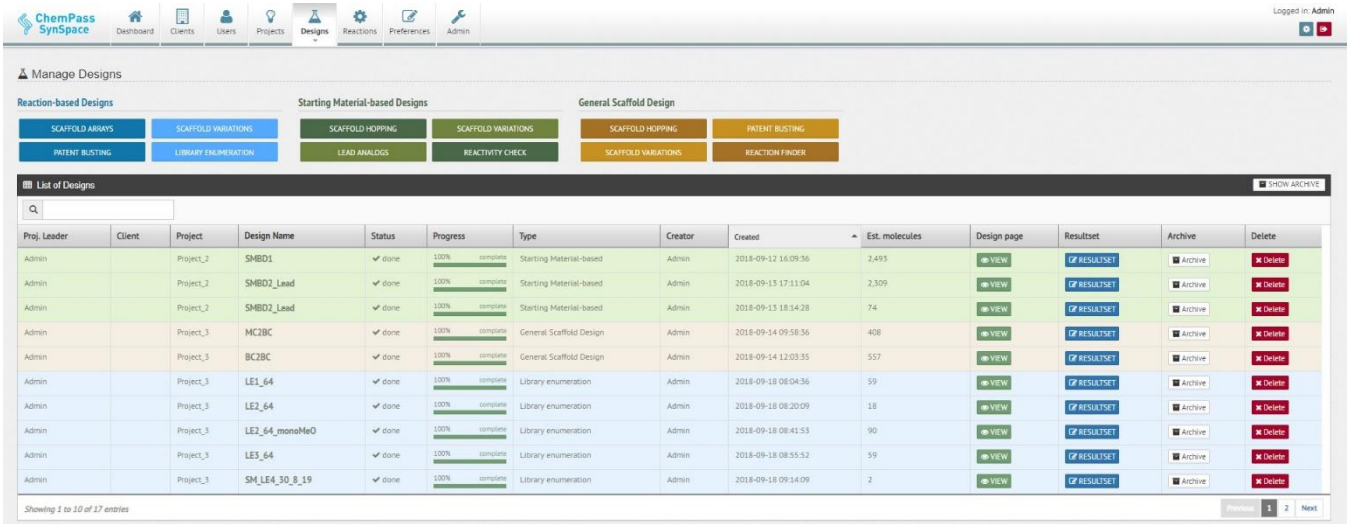
CINF 39

SynSpace: A user-friendly web- and cloud-based design platform to expand synthetically-enabled scaffold and lead analogue space for medicinal chemistry and AI-assisted drug discovery

Gergely Makara, *gergely.makara@chempassltd.com, Gabor Pocze, Laszlo Kovacs, Orsolya Demeter, Istvan Szabo. ChemPass Ltd., Budapest, Hungary*

Throughout the world mankind is facing an ageing and growing population that requires more effective and safer medicines in all therapeutic areas. Despite significant advances in our understanding of the biological basis of diseases, pharmaceutical R&D is struggling to sustain the level of productivity and efficiency it reached in the second half of the 20th century. High rates of failure and the increasing cost of drug discovery as well as extended research and development timelines hinder the development of medicines. Due to these challenges there has been an increasing need for substantial innovations in the pharmaceutical sector. It has been shown that if the selection of the synthetic targets in lead optimization cycles is supported by QSAR or deep learning methods, the number of compounds synthesized as well as the cycle time for each iteration can be significantly reduced. We have developed a rule-based artificial intelligence technology that can produce a large number of novel and synthetically-enabled lead analogues and scaffold hopping designs around lead structures. Since its introduction, the cloud-based SynSpace software has been found by multiple organizations to generate a larger number of relevant novel ideas around leads than medicinal chemist teams can do. Thus, SynSpace is a valuable addition to the medicinal chemistry toolbox.
We have also been developing automated lead analysis tools and a synthesis-based

library enumerator that - in conjunction with SynSpace - can automatically carry out scaffold hopping and lead analogue idea generation and thereby offer large sets of novel and project specific lead-like structures to advanced AI platforms for selection. These platforms have the biggest impact on a number of key parameters in drug discovery: cycle time, number of discovery cycles, the number of compounds to be synthesized and coverage of IP space. Improvements in these factors can be converted into higher success rates and major resource savings towards a more economical and productive candidate development phase.



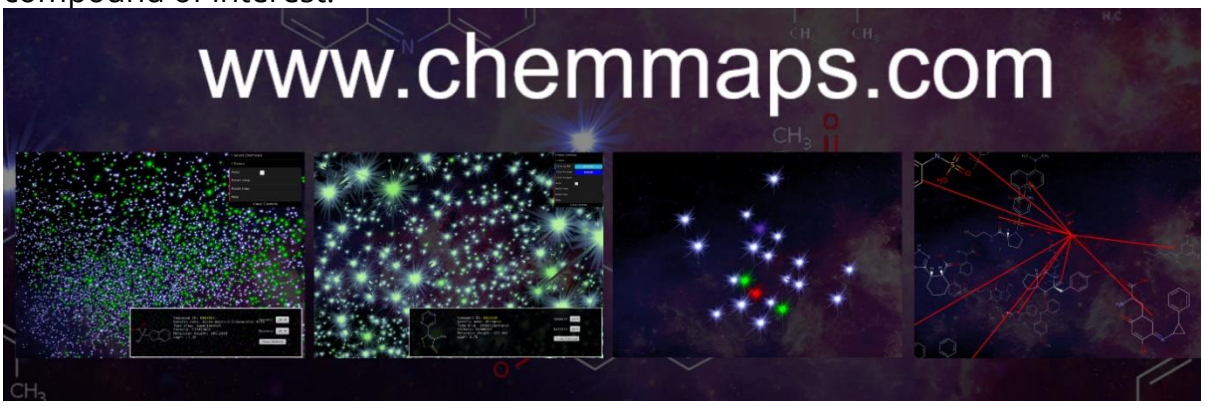
CINF 40

Exploring an expanded chemical universe using www.*chemmaps.com*

Alexandre Borrel[1]*, alexandre.borrel@univ-paris-diderot.fr, Denis Fourches*[2]*, Nicole Kleinstreuer*[1]*. (1) Biostatistics and Computational Biology Branch, NIEHS, Morrisville, North Carolina, United States (2) Chemistry, North Carolina State University, Raleigh, North Carolina, United States*

Easily navigating chemical space has become more important due to the increasing size and diversity of publicly-accessible databases such as DrugBank, ChEMBL, or DSSTox, and associated high-throughput screening (HTS) and other datasets. Modelers typically rely on complex projection techniques using molecular descriptors computed for all the chemicals to be visualized. However, the multiple cheminformatics steps required to prepare, characterize, compute and explore those molecules, are technical, typically necessitate scripting skills, and thus represent a real obstacle for non-specialists. Inspired by the popular Google Maps application, we developed the ChemMaps.com webserver to easily navigate chemical spaces. The first version of ChemMaps.com was developed to browse and visualize the space of

2,000 FDA-approved drugs and over 6,000 drug candidates based on the DrugBank database (https://www.drugbank.ca/) and was extended on ~47,000 environmental chemicals. In this new version, the chemical coverage has been extended to include the full DSSTox inventory (>700,000 chemicals and additional browsing, searching, and exporting/importing options were updated and developed. Users can now upload their own set of chemicals and visualize them on the available maps and/or define a new map from them. All computed data, e.g. coordinates, chemical descriptors, etc. can now be downloaded. Different navigation options have been also developed, including a distance computing on the fly for two selected chemicals and a faster and more responsive environment.). Users can search for specific compounds, overlay regulatory classification and labeling based on animal toxicity data, explore and export nearest neighbor space, refine the projections based on physicochemical properties, and link out to the EPA's CompTox Dashboard (https://comptox.epa.gov/dashboard) for detailed information on a chemical. Work is ongoing to embed ChemMaps.com on the EPA's CompTox Dashboard to provide real-time chemical space visualization specific to the compound of interest.



Screenshots of www.chemmaps.com, from left to right: initial DrugMap view, zoom in on EnvMap, selection of few chemicals, and representation of chemicals on the space with the closest neighborhoods connected.

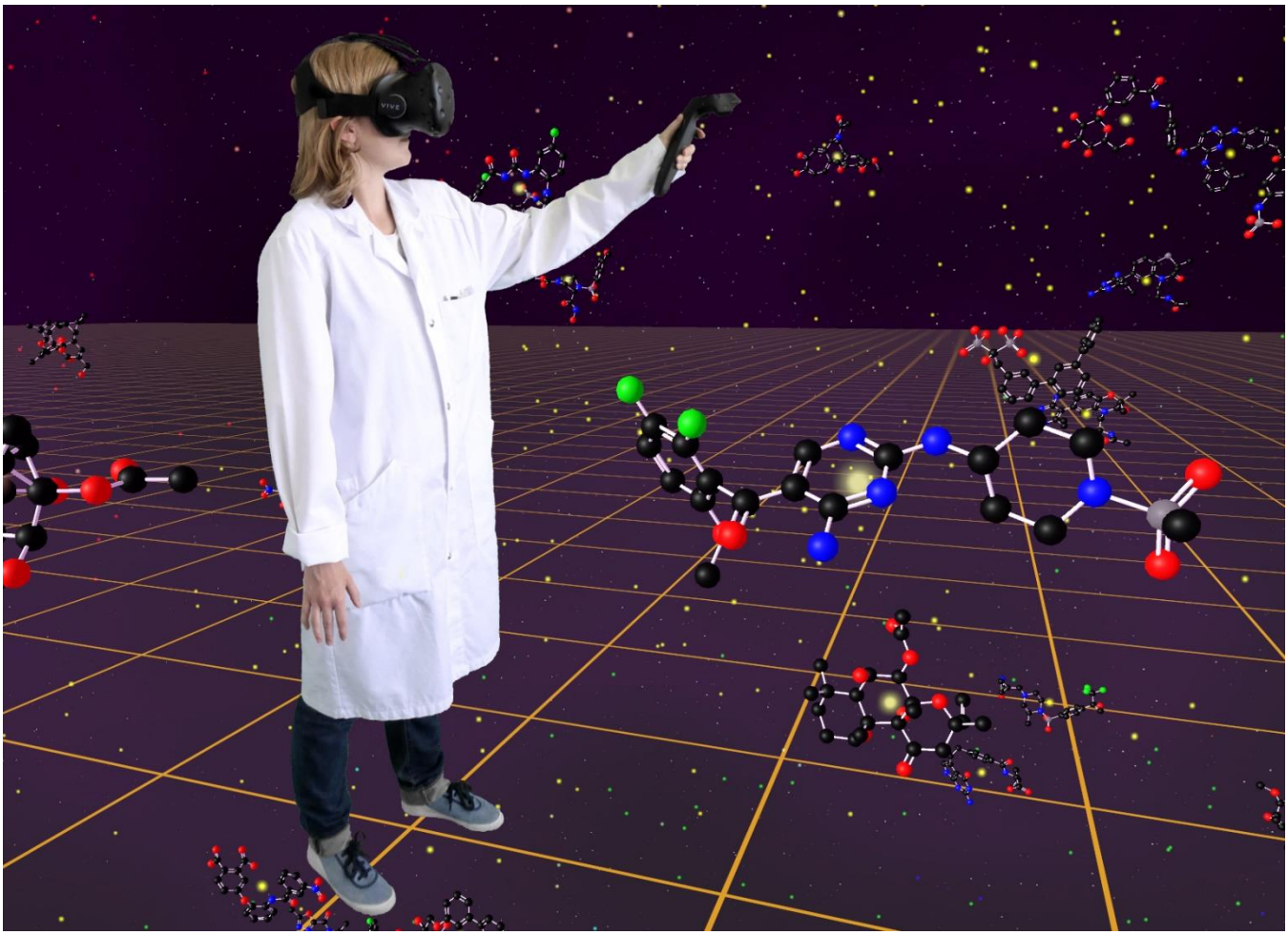
CINF 41

Exploring chemical space at gdb.unibe.ch

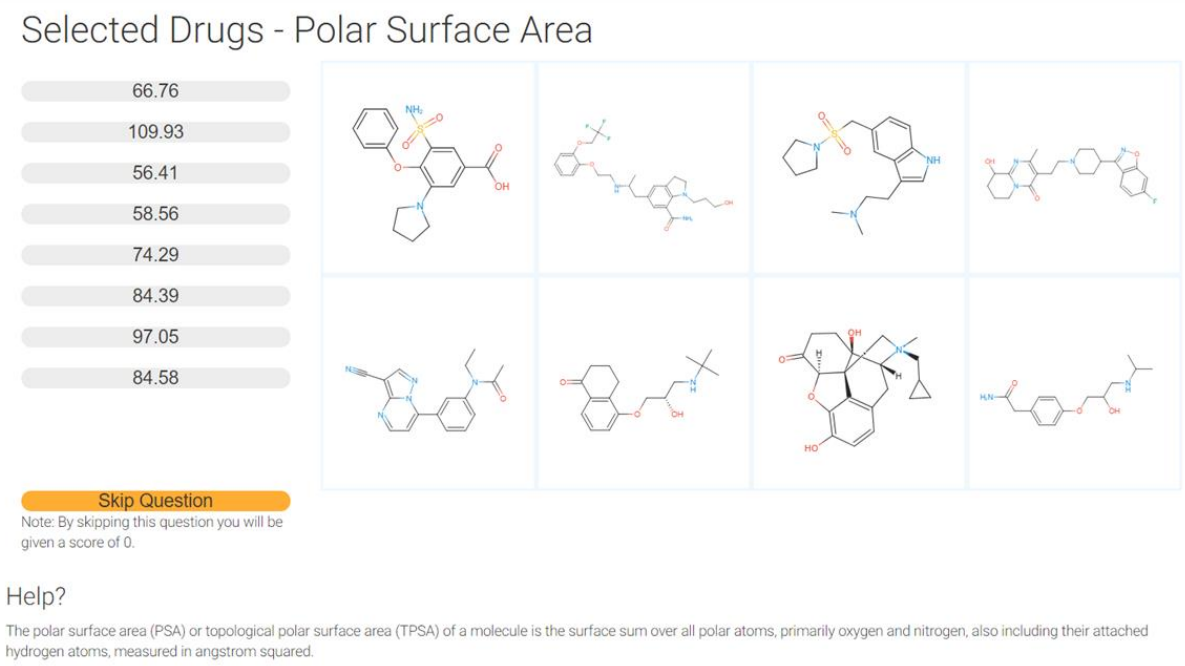
Daniel Probst, Mahendra Awale, Amol Thakkar, ***Jean-Louis Reymond****, jean-louis.reymond@ioc.unibe.ch. Chem Dept Univ of Bern, Bern, Switzerland*

Our website www.gdb.unibe.ch offers a variety of tools to assist medicinal chemists in designing and evaluating drug molecules. At the design stage, our chemical universe databases GDB11, GDB13, GDB17, FDB17, and GDB4c offer exhaustive lists of possible molecules. These have been produced by systematic enumeration following simple rules of chemical stability and synthetic feasibility, whilst offering vast potential for

innovation. These databases can be downloaded, or directly searched online by fingerprint similarity using our MQN and SMIfp fingerprints. We also offer multi-fingerprint similarity browsers to search ZINC (purchasable compounds) and ChEMBL (bioactive compounds). At the stage of evaluating drug molecules, our polypharmacology browsers PPB and PPB2 can be used to predict off-targets of any hit molecule based on ChEMBL data. We also propose general tools to support the exploration of large chemical databases such as: Faerun, capable of mapping large data sets, such as all patented molecules from SureChEMBL in a 3D-chemical space; SmilesDrawer, to draw molecular structures from SMILES in a web-browser without server communication; WebMolCS, to create 3D-chemical space maps of any data set defined by the user; MHFP6, a new fingerprint to compare molecules by circular substructures, outperforming ECFP4 in drug analog searches and suitable for big data settings; a Virtual Reality Chemical Space to explore DrugBank; and the chemistry learning game, to acquire better understanding of structural formulas.



Exploring DrugBank in Virtual Reality, available at http://viz.gdb.tools.



The Chemistry Learning Game challenges you to associate various properties (PSA shown) and molecule names with drugs and related molecules

CINF 42

Developing an integrated model management solution to assure quality of predicted data at the US EPA's National Center of Computational Toxicology

***Christopher Grulke**, grulke.chris@epa.gov, Antony J. Williams, Amar Singh, Jeff Edwards. National Center for Computational Toxicology, Environmental Protection Agency, Wake Forest, North Carolina, United States*

The Computational Toxicology Program within the U.S. Environmental Protection Agency (EPA) integrates advances in biology, chemistry, and computer science to help prioritize chemicals for further research based on potential human and environmental health risks. A key component of this prioritization effort is the use of New Approach Methods (NAMs) which include both *in vitro* and *in silico* methods of estimating more traditional risk-assessment inputs (e.g. *in vivo* toxicity). Without NAMs, the number of chemicals is too large and the available data too sparse to effectively prioritize. However, the application of NAM generated data within EPA to support a regulatory context requires a high degree of quality assurance. While this need is apparent for *in silico* NAMs (e.g. QSAR models), it is also vital for most *in vitro* NAMs which use computational tools to process screening results into informative scores. To meet minimal requirements, well-defined workflows must be applied to versioned and well-

documented computational methods to yield reproducible results. In addition, the management solution must be flexible enough to efficiently integrate newly developed models from our cheminformatics researchers that could inform a prioritization task without limiting the options afforded to our researchers. To meet these requirements, we have designed a series of loosely coupled microservices that are genericized to enable many of our use-cases but specialized enough to efficiently support our most pressing need of supplying predictions for the Comptox Chemicals Dashboard (https://comptox.epa.gov/dashboard). *This abstract does not necessarily represent the views or policies of the U.S. Environmental Protection Agency.*

CINF 43

US-EPA CompTox chemicals dashboard: A web-based data integration hub for environmental chemistry data

***Antony J. Williams**, tony27587@gmail.com, Christopher Grulke, Richard Judson, John Wambaugh, Jeremy Dunne, Jeff Edwards. National Center for Computational Toxicology, Environmental Protection Agency, Wake Forest, North Carolina, United States*

The National Center for Computational Toxicology (NCCT) at the US Environmental Protection Agency has measured, assembled and delivered to the community an enormous quantity and diversity of data for the environmental sciences, including high-throughput *in vitro* screening data, *in vivo* and functional use data, exposure models and chemical databases with associated properties. The CompTox Chemicals Dashboard is a web-based application providing access to data associated with ~770,000 chemical substances but is not limited only as a data delivery hub. The application also provides the ability to perform real-time prediction for a series of physicochemical and toxicological endpoints as well as conduct read across using both chemical and biological data to support the read-across process. A batch search capability provides a facile approach to accessing data for thousands of chemicals at a time in user consumable formats such as Excel and SDF files. This presentation will provide an overview of the CompTox Chemicals Dashboard, how it has developed into an integrated data hub for environmental data and real-time predictions for physicochemical and toxicological endpoints of interest to environmental scientists. *This abstract does not necessarily represent the views or policies of the U.S. Environmental Protection Agency.*

CINF 44

IUPAC Commission on Isotopic Abundances and Atomic Weights: Its history, role, and work

Juris Meija[1,2], *Juris.Meija@nrc-cnrc.gc.ca. (1) Metrology Research Centre, National Research Council Canada, Ottawa, Ontario, Canada (2) Commission on Isotopic abundances and Atomic Weights, International Union of Pure and Applied Chemistry, Research Triangle Park, North Carolina, United States*

It is hard to imagine IUPAC without the Periodic Table and, in turn, without atomic weights. As IUPAC celebrates its centennial, its oldest body, the Commission on Isotopic abundances and Atomic Weights (CIAAW) turns 120. Atomic weights lay the foundations for many scientific measurements many of which go largely unnoticed. The work of the CIAAW relies on the volunteers who are willing to engage in evaluation of isotope ratio measurements for the benefit of broader goals. Since 1902, the International Committee has been shaped by 120+ expert volunteers. To complicate matters, many view the Periodic Table and changes therein as a part of larger cultural fabric of science so any changes are likely to be debated for a long time. This presentation will outline the impact and the role of IUPAC in setting the international standards for atomic weights along with the challenges facing the CIAAW over the last century.

CINF 45

Archives of the international union of pure and applied chemistry at the Science History Institute

Ronald S. Brashear, *rbrashear@sciencehistory.org. Science History Institute, Philadelphia, Pennsylvania, United States*

In the mid-1990s, the archives of IUPAC's Commission on Atomic Weights and Isotopic Abundances were given to the Science History Institute (at that time the Chemical Heritage Foundation), thanks to the efforts of Steffen Peiser. In 1997, the impending move of IUPAC's headquarters from Oxford to Research Triangle Park, North Carolina, provided an opportunity for the Institute to become the permanent home of the remaining IUPAC records that were no longer needed by the current administration. At the present time the total IUPAC archive consists of 388 boxes or 214 linear feet (65 linear meters) and 150 photographs. It is an important source for historians on the development and organization of science. It is one of the most heavily used collections in the Institute's Othmer Library, primarily because of how important IUPAC is in the overall organization and disciplinary identity of chemistry in the twentieth century as well as the central role it plays in the management of chemical nomenclature and in the standardization of atomic weights, physical constants, and formats of publications. This talk will discuss the circumstances surrounding the acquisition of the archive, its contents, and how scholars can best gain access to the material. I will also discuss the possibility of finding and adding additional material to the archive.

CINF 46

"A" in IUPAC: Applying the common language for chemistry to meet world needs

Mark C. Cesa, markcesa@comcast.net. INEOS Nitriles, Wheaton, Illinois, United States

The International Union of Pure and Applied Chemistry, IUPAC, is the global organization that provides objective scientific expertise and develops the essential tools for the application and communication of chemical knowledge for the benefit of humankind and the world. In addition to the invaluable work that IUPAC does to support and enable advances in "pure" chemistry research, IUPAC also reaches out through its interdisciplinary Divisions and Standing Committees to the global chemistry enterprise to enhance chemistry's role in the world and to help ground public policy in sound science. IUPAC partners with international organizations including UNESCO, the UN Strategic Approach to International Chemicals Management, the Organization for the Prohibition of Chemical Weapons, and the U. S. National Academies on a range of projects and programs. These efforts include collaborations on guidelines for codes of conduct for chemists, safety and security training for chemists in developing countries, sustainability, environmental monitoring and protection, and resources for education and outreach on dual use chemicals and for nomenclature standards for medicinal chemistry, among many others. As IUPAC moves into its second century, the Union aims both to react to rapid advances in science and to lead the continuing development of global chemistry.

CINF 47

Accidental nomenclaturest: A journey from bench chemist to ACS-NTS and IUPAC member

Michelle M. Rogers, michelle.m.rogers@gmail.com. Product Safety and Compliance, The Lubrizol Corporation, Chagrin Falls, Ohio, United States

In the talk I will share my journey from a bench chemist at a specialty chemical company to becoming President of the ACS Committee on Nomenclature, Terminology and Symbols (NTS) and my ongoing involvement with IUPAC and how that was pivotal to my journey. When I started in industry, I thought my days of IUPAC nomenclature were behind me. Little did I know that IUPAC and a chance meeting with their nomenclature committee would be part of the catalyst to change the course of my career. In 2009, I was selected to attend the IUPAC general assembly in Glasgow, Scotland as a young observer. It was at this conference that I had the opportunity to learn the diverse areas that IUPAC was involved in and attended my first meeting Division VIII – Chemical Nomenclature and Structure Representation. After that conference I got a taste for international collaboration and volunteered to serve on an ACS committee and became a member of the US National Committee to IUPAC. My role within my company

changed and I joined the regulatory team, where part of my role was naming chemicals to support EU REACH registrations. It was through this that I identified the need for more robust nomenclature guidance for complex industrial chemicals. To help address this area, I decided to get involved with Division VIII of IUPAC and the ACS-NTS committee. Through these experiences I have developed a deep understanding of continuing need to develop a common language for chemistry both for people and computers.

CINF 48

iGROW: IUPAC global recognition opportunities for women

*Fabienne Meyers[1], Carolyn Ribes[2], **Angela K. Wilson**[3], wilson@chemistry.msu.edu. (1) IUPAC, Boston, Massachusetts, United States (2) The Dow Chemical Company, Terneuzen, Netherlands (3) Division of Chemistry, Michigan State University, East Lansing, Michigan, United States*

Globally, fewer women than men are recognized for their accomplishments in science and engineering. In 2011, initiated as part of the International Year of Chemistry celebration, an international awards program was launched to acknowledge, honor, and celebrate the accomplishments of women in chemistry and chemical engineering. Seventy-one outstanding scientists from 6 continents have received this award based on excellence in basic or applied research, distinguished accomplishments in teaching or education, or demonstrated leadership or managerial excellence in the chemical sciences. At each IUPAC World Congress, a symposium is also held to discuss challenges, solutions, and opportunities for women in science. Highlights of previous award symposia and related activities will be presented. The 2019 awards will be presented in Paris in July during the IUPAC World Chemistry Congress and General Assembly. To celebrate the fifth edition of the award, a special issue of Pure and Applied Chemistry will feature papers by over twenty recipients of the award.

CINF 49

Role of IUPAC Committee on Chemistry Education in communicating chemistry

***Marcy H. Towns**[1,2], mtowns@purdue.edu. (1) Purdue University, West Lafayette, Indiana, United States (2) Committee on Chemical Education, International Union of Pure and Applied Chemistry, Research Triangle Park, North Carolina, United States*

The IUPAC Committee on Chemistry Education (CCE) aims to develop global collaborative working relationships relevant to the learning and teaching of chemistry. It emphasizes the importance of high quality student-centered learning practices as well as identifying and discussing the learning outcomes in chemistry education. The activities that support these aims are numerous and quite broad. The CCE supports and is engaged in numerous activities that develop high quality materials for learning

chemistry and engages with other IUPAC divisions in developing technical reports that inform the global community about important matters such as the proposed redefinition of the quantity “amount of substance” and its unit, the mole. The CCE supports initiatives that increase the engagement of under represented groups in chemistry. Additionally it supports initiatives that raise awareness, social responsibility, and the understanding of the nature of science as well as environmental and ethical issues that are related to chemistry. This presentation will highlight some of the initiatives that the CCE has carried out and has been engaged in where the sharing of ideas and information has impacted chemistry education.

CINF 50

Short history of IUPAC InChI algorithm

***Stephen R. Heller**, steve@hellers.com. Retired, Silver Spring, Maryland, United States*

This presentation will describe the origins of the InChI project at both IUPAC and NIST. The need for a computer based open source freely available mechanism for describing defined chemical structures had been examined and discussed for more than a decade prior to the start of this project, but never took hold until the right time, right place, right outside resources. and right people all came together to create the perfect “good storm”. The success of InChI can be seen in its uncoerced adoption and support by the chemical community. Lastly, the current state and current and planned enhancements for InChI and the InChIKey will be presented.

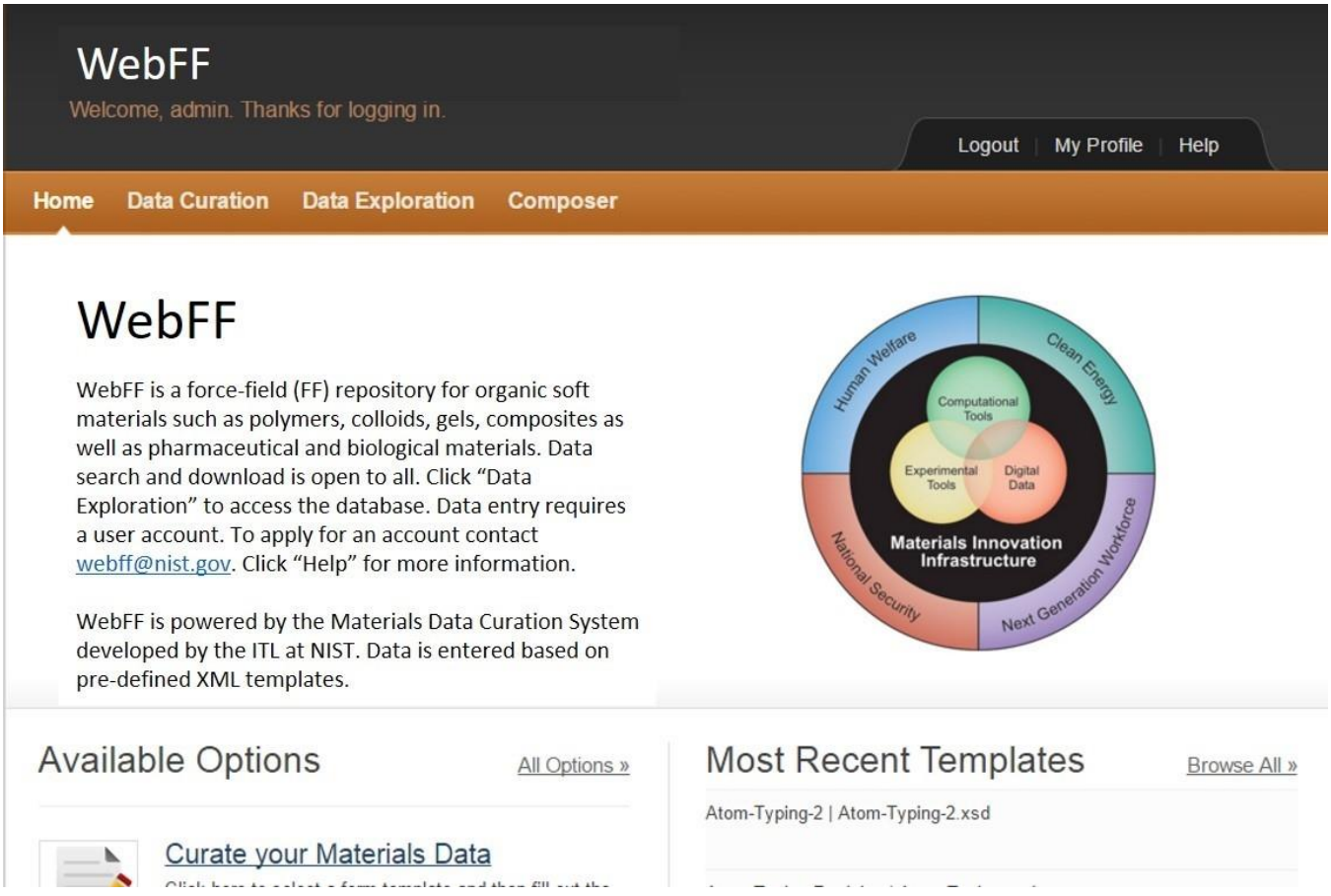
CINF 51

Web Force-Field (WebFF) repository: Molecular dynamics force-field data for soft materials at multiple levels of granularity

***Frederick R. Phelan**[1], frederick.phelan@nist.gov, Huai Sun[2]. (1) STOP 8542, NIST, Gaithersburg, Maryland, United States (2) Department of Chemistry, Shanghai Jiao Tong University, Shanghai, China*

This talk will describe WebFF, an open and extensible force-field repository, designed to support the Materials Genome Initiative (MGI) for soft materials. The repository is built using the NIST Materials Data Curation System (MDCS). The MDCS has a web-based interface built on top of the RESTful API (Python) and a NoSQL database system to support ontology based database descriptions using XML schema. Users interact with the repository through two main portals. The *Data Curation Portal* supports upload of published force-field data with appropriate metadata descriptors to support provenance based data sharing. The *User Portal* supports search for curated force-field data based on the metadata descriptors and download in a number for common formats. The initial release of the repository features three integrated XML schemas: the first two are for

Class I and Class II organic, atomistic force-fields, respectively, and the third accommodates a wide range of coarse-grained force-field data. WebFF offers authors and developers a means to make their data publicly shareable in a manner that should play well with publishers. New datasets may be curated interactively or using our Python based toolset to upload large datasets *en masse*. New data entered into the repository will be assigned a DOI in order to facilitate direct access, and the metadata descriptors enable direct links to published articles to provide data provenance. The datasets may also include attachments, provided that they do not violate copyright agreements. User search, download and data curation to the site will be described and demonstrated. Further details of the repository and descriptions of the XML schemas may be found in the WebFF Reference Manual (URL: https://webff-documentation.readthedocs.io/en/latest/index.html) and on GitHub (URL: https://github.com/usnistgov/WebFF-Documentation).



CINF 52

CavityPlus: A web server for protein cavity detection with pharmacophore modelling, allosteric site identification, and covalent ligand-binding ability prediction

Jianfeng Pei, jfpei@pku.edu.cn. Academy for Advanced Interdisciplinary Studie, Peking University, Beijing, China

CavityPlus is a web server that offers protein cavity detection and various functional analyses. Using protein three-dimensional structural information as the input, CavityPlus applies CAVITY to detect potential binding sites on the surface of a given protein structure and rank them based on ligandability and druggability scores. These potential binding sites can be further analysed using three submodules, CavPharmer, CorrSite, and CovCys. CavPharmer uses a receptor-based pharmacophore modelling program, Pocket, to automatically extract pharmacophore features within cavities. CorrSite identifies potential allosteric ligand-binding sites based on motion correlation analyses between cavities. CovCys automatically detects druggable cysteine residues, which is especially useful to identify novel binding sites for designing covalent allosteric ligands. Overall, CavityPlus provides an integrated platform for analysing comprehensive properties of protein binding cavities. Such analyses are useful for many aspects of drug design and discovery, including target selection and identification, virtual screening, *de novo* drug design, and allosteric and covalent-binding drug design. The CavityPlus web server is freely available at http://repharma.pku.edu.cn/cavityplus or http://www.pkumdl.cn/cavityplus.

CINF 53

iSpiEFP: Automating the computational workbench

Yen Bui, ybui@purdue.edu, Lyudmila V. Slipchenko. Chemistry, Purdue University, Lafayette, Indiana, United States

Computational molecular modeling has made great strides in providing support for experimental studies in chemistry, physics and biology following continual advancements in hardware, networking, and data management. However, utilizing computational methods itself is daunting for the novice user as molecular simulations require not only thorough theoretical background, but also technical experience in data parsing between various programs, data visualization of large data sets, and terminal-based programming. Moreover, working with increasingly larger datasets and various data formats becomes a serious time sink and error source even for the most experienced users. To address questions of data compatibility, analysis and visualization, we introduce iSpiEFP – a local graphical user interface (GUI) that streamlines multi-scale calculations with Effective Fragment Potential (EFP) – a sophisticated ab initio based method for modeling non-covalent interactions. iSpiEFP serves as a workflow manager for system visualization, access to a cloud-based amazon

web server (AWS) database of EFP parameters, high-performance simulations, and data analysis. The talk will begin with a brief overview of modern methods for describing non-covalent interactions. It will proceed with validation of EFP on biological datasets and conclude with a demonstration of iSpiEFP as the tool enabling users to perform state-of-the-art simulations of ligand-protein binding, molecular crystal structures and optical properties of photoactive materials.

CINF 54

ProteinsPlus and SMARTSviewer: Two web applications for the modeling and cheminformatics community

*Rainer Fährrolfes, Robert Schmidt, **Matthias Rarey**, rarey@zbh.uni-hamburg.de. University of Hamburg, Hamburg, Germany*

Structure-based design is characterized by graphic-intense, interactive applications. Not surprisingly, up to day monolithic desktop applications dominate the software landscape in this area. Using modern JavaScript-based graphic plugins, several tasks in structure-based modeling can be transferred to the web. ProteinsPlus (http://proteins.plus) is an easy to use web service unifying access, assessment, and preprocessing of protein structures. Important processes like exploring ligands, predicting hydrogen positions and optimizing the hydrogen bond network, estimating electron density support, detecting binding pockets and estimating their druggability are fully integrated. Highly efficient database searches for active sites allow to quickly create structure ensembles, getting an overview of known bound ligands, active site mutations and conformational flexibility. A RESTful service enables the full integration into other web resources or even workflow systems like KNIME. As a second example, we present a cheminformatics web service named SMARTSview (http://smartsview.zbh.uni-hamburg.de). The design of SMARTS expressions usually applied for filtering compound collections remains a very technical procedure. To support it, we developed three software components: SMARTSview to visualize SMARTS expressions, SMARTSminer to create them from compound sets, and SMARTScompare to calculate the pattern similarity. All three of them are quite unique and available as a web service. With SMARTScompare, a method for searching chemical patterns in pattern collections like LINT or PAINS is made available for the first time.

CINF 55

D-Peptide Builder: A web-based application to enumerate the chemical space of peptides

Barbara Diaz Eufracio, debi_1223@hotmail.com, José L. Medina-Franco, Oscar Palomino-Hernández, Aaron Arredondo-Sanchez. DIFACQUIM, UNAM, Ciudad de Mexico, Ciudad de Mexico, Mexico

Peptides are an important source molecules in drug discovery. Peptides conform aunique class of pharmaceutical compounds, molecularly poised between small molecules and proteins. Due to this feature peptides may act as modulators of protein – protein interactions in addition to some other targets.
This work introduces D-Peptide Builder, a free online server for the rapid enumeration of combinatorial peptide libraries and explore their chemical space. The server enumerate linear and cyclic-peptide libraries from different sizes and structural modifications from a pool of naturally occurring amino acids.
The chemical space of the newly generated peptides can be visualized through machine learning techniques such as PCA and t-SNE. Finally, the current version of D-Peptide Builder allows the analysis of the intra-set similarity by computing pair-wise similarity values with differents fingerprints of different design and the Tanimoto coefficient.

CINF 56

Freely available online resource for prediction of novel multitarget anti-HIV agents

Dmitry Druzhilovskiy[1], explorermf@yandex.ru, Dmitry Filimonov[1], Leonid Stolbov[1], Polina Savosina[1], Vladimir Poroikov[1], Marc C. Nicklaus[2]. (1) Department of Bioinformatics, Institute of Biomedical Chemistry, Moscow, Russian Federation (2) NCI-Frederick Bldg 376 RM 207, Natl Inst Health NCI Ft Detrick, Frederick, Maryland, United States

AIDS is one of the multifaceted diseases, and this underlying complexity hampers its complete cure. In addition to that, HIV-infected patients have an increased risk of comorbidities, such as infectious diseases as well as noncommunicable disorders. Multitarget drug therapy is considered a better treatment option for HIV and infections as compared to targeted drug therapies and helps to overcome drug resistance, reduction in drug dosage and toxicity. In such a scenario, machine learning approaches can be used to alleviate the process of drugs discovery and development.
We attempt to address the existing challenges by developing a computational approach to predict multitarget agents with anti-HIV properties effectively and as a prediction platform for the exploration of the Synthetically Accessible Virtual Inventory (SAVI) library. SAVI includes about 283 million molecules annotated with a proposed one-step synthetic route from commercially available starting materials. The SAVI database is well suited for ligand-based methods of virtual screening to select the most promising hits for experimental testing.
Among the various therapeutic targets for HIV treatment, protease, reverse transcriptase, integrase, fusion, viral protein R, tumor susceptibility gene 101, viral infectivity factor, Rev, Tat, GP120, toll-likes receptors, P24 capsid protein, HIV-LTR and

other are the primary focus. Besides, it should be noted that there are no less than 100 different HIV comorbidities diseases. We have developed freely available web service to select new potential anti-HIV agents based on selected targets from SAVI. We applied the computer program PASS (Prediction of Activity Spectra for Substances) and similarity estimates based on MNA (Multilevel Neighborhoods of Atoms) and QNA (Quantitative Neighborhoods of Atoms) descriptors. PASS has been developed and applied for drug design & discovery during the past 25 years and has been validated in more than 500 independent papers reporting on the application of PASS predictions to the discovery of new pharmaceutical agents. The combined multitarget approach of machine learning and similarity assessment allows estimating the anti-HIV potential of new compounds. This approach led to the selection of dozens of molecules recommended for synthesis and testing for antiretroviral activity, we have identified a few promising hits against the identified targets which may give new drugs to combat HIV infection after wet lab validation.

CINF 57

ZINC15.docking.org: Over 1.5 billion compounds you can search and buy; 550 million lead-like you can dock

John J. Irwin, jji@cgl.ucsf.edu. Pharmaceutical Chemistry, University of California San Francisco, San Rafael, California, United States

ZINC is a database of commercially available compounds for virtual screening. Recently, we have been adding many new molecules and preparing them in biologically relevant forms and 3D formats for docking. Some uses of ZINC include:

a) acquire a database for docking. At the time of abstract submission, this was 350M. We expect it will be closer to 550M by the time this is presented. We plan to be over 1 billion by summer 2020.

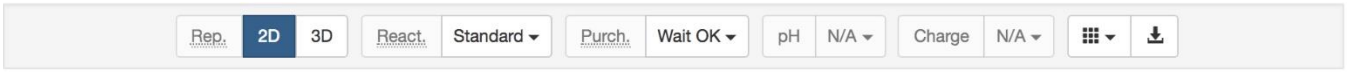
b) analog by catalog. At the time of abstract submission, this was 1.1 billion. By presentation time, we expect it to be above 1.5 billion. We plan to be above 2 billion by summer 2020. ZINC categorizes purchasable molecules into broad bins by delivery time and price to help you find what you need.

c) Predicted biological activity of purchasable chemical space. We use the SEA+Tc method to predict biological targets for 1.5 billion molecules and make these freely available for download.

d) Nearest bioactive, metabolite, natural product, drug. ZINC can identify the most similar purchasable compounds to precedented actives.

e) Similarity to actives reported for a target. ZINC can sort in decreasing order of

similarity the actives for any gene product represented in ChEMBL to your query molecule. In the past 2 years, we have purchased over 1200 make-on-demands. 92% of these arrived and were tested, over half of them in 4 weeks and nearly all within 6.



		Molecular Weight (up to, Daltons)											
		200	250	300	325	350	375	400	425	450	500	>500	LogP
LogP (up to)	-1	24K	199K	1.1M	1.6M	2M	844K	210K	51K	37K	16K	5.3K	6.1M
	0	147K	1.2M	5.3M	7.6M	9.9M	3.3M	1.4M	409K	296K	105K	3.6K	30M
	1	446K	4.2M	18M	24M	39M	12M	7.1M	2.3M	1.7M	634K	7.4K	109M
	2	629K	7.2M	30M	35M	88M	31M	22M	7.9M	6.2M	2.8M	20K	230M
	2.5	245K	3.6M	21M	22M	52M	23M	18M	7.5M	6.2M	2.8M	22K	156M
	3	141K	2.8M	18M	25M	48M	25M	22M	10M	8.7M	3.8M	38K	164M
	3.5	60K	1.8M	14M	19M	34M	24M	23M	12M	11M	5M	63K	143M
	4	16K	725K	7.9M	11M	15M	17M	20M	12M	11M	5.7M	94K	101M
	4.5	2K	172K	3.4M	5.9M	8.7M	11M	14M	10M	10M	5.7M	131K	70M
	5	96	20K	927K	2.3M	3.7M	5.2M	8.6M	6.4M	7.6M	4.7M	159K	40M
	>5	28	862	45K	179K	556K	1.2M	2.1M	2.6M	3M	2.5M	805K	13M
Totals, by Weight		1.7M	22M	119M	153M	300M	155M	139M	71M	66M	34M	1.3M	**1063M** Substances **1.9K** Tranches

CINF 58

Towards a “Digital IUPAC”: Coordinating community needs for digital data standards

Leah R. McEwen[1,2], *lrm1@cornell.edu, David Martinsen*[3,2], *Helen A. Lawlor*[4,2]*. (1) Clark Library, Cornell University, Ithaca, New York, United States (2) Committee on Publications and Cheminformatics Data Standards, International Union of Pure and Applied Chemistry, Research Triangle Park, North Carolina, United States (3) Retired,*

Washington, District of Columbia, United States (4) Retired, Radnor, Pennsylvania, United States

Anticipating the need for computer readable chemistry standards in the digital era, in 2014 the International Union of Pure and Applied Chemistry (IUPAC) expanded the remit of its publications to address a framework for “Digital IUPAC.”[1] A new Subcommittee on Cheminformatics Data Standards (SCDS) was established to explore the needs of the chemistry community in this area. With the breadth of activity internationally around data, SCDS primarily takes a coordinating role, striving to “prioritize and efficiently meet those needs through the collaborative efforts.”[2] Towards this end, SCDS has been involved in organizing a number of community workshops in collaboration with other chemistry and data organizations worldwide, including the Research Data Alliance and the Committee on Data of the International Science Council (CODATA), among others. In addition to enabling cross-connection in the community, SCDS looks for challenges and opportunities where specific work on digital chemical standards can enhance the exchange of chemical data and information. Recent workshops have focused on chemical representation, spectroscopic data formats, and metadata for Findable, Accessible, Interoperable, and Re-usable (FAIR) data for humans and machines. This presentation will summarize the scope of these efforts to date and outline future directions and needs for developing data standards for chemistry in the digital era.

CINF 59

Renovating the IUPAC gold book for the digital era and the next 100 years

***Stuart J. Chalk**, schalk@unf.edu. Department of Chemistry, University of North Florida, Jacksonville, Florida, United States*

In the past 100 years IUAPC has become well known for the development of nomenclature standards for chemicals and terminology for communication of chemically related concepts. Initially published in IUPAC's Pure and Applied Chemistry (PAC), terminology recommendations have been incorporated into the IUPAC Color Books published by the divisions. Subsequently, many terms from the color books have been incorporated into the Gold Book - the Compendium of Chemical Terminology. While the work to date as been focused on the standardization of concept (term) definitions for human use, the aggregated set of all PAC recommendations on terminology constitutes a corpus of high quality definitions for entries into an ontology for use in computer representation of chemical concepts. This is sorely needed at a time when there is a significant move toward machine learning approaches to understand science both within and outside chemistry. With it chemical data scientists can envision and apply this 'common language' into their cheminformatics work, promoting interoperability in chemical data.

This paper will review the history of the PAC recommendations, the color books, and highlight how the existing guidelines for the development of terms supports the renovation of the terms for computer use. The current update to the Gold Book website will be discussed (including machine processability) as well as future ontological representation of the terms. Finally, this development will be highlighted as one of the most important in the next 100 years of IUPAC.

CINF 60

ISMC: IUPACs interdivisional sub-committee on materials chemistry

Christopher K. Ober[1], cko3@cornell.edu, Vladimir Gubala[2]. (1) Cornell Univ, Ithaca, New York, United States (2) University of Kent, Canterbury, United Kingdom

The International Union of Pure and Applied Chemistry will have its 100thanniversary this year. Founded by members of the academic and industrial chemistry community it has led the creation of the language of chemistry so these different groups could better communicate and chemistry could become a truly international community. Like any organization of this age it continues to evolve and reflect the changing face of chemistry. One area of recent growth has been in the subject of materials chemistry which now touches many aspects of chemical science. Examples include nanoscience where chemistry is a critical discipline involving many different aspects of chemistry ranging from inorganic chemistry to polymer chemistry to medicine. To reflect these changes and the broad interest in materials chemistry, the Interdivisional Sub-committee on Materials Chemistry was created to explore common interests between the divisions of Inorganic Chemistry, Physical and Biophysical Chemistry and Polymer Chemistry. More recently the divisions of Chemistry and Human Health and Chemistry and The Environment have become involved. This presentation will discuss how IUPAC recognizes and deals with changing aspects of chemical science to stay relevant while delivering on its original mission.

CINF 61

FAIR data in the 21st century: The role of scientific unions in facilitating interdisciplinary data science in Chemistry and the Earth Sciences

Shelley Stall[2], SStall@agu.org, Leah R. McEwen[1], lrm1@cornell.edu. (1) Clark Library, Cornell University, Ithaca, New York, United States (2) American Geophysical Union, Washington DC, District of Columbia, United States

The International Union of Pure and Applied Chemistry (IUPAC) and the American Geophysical Union (AGU) are both celebrating centennials in 2019 and looking towards the digital future of scientific communication in Chemistry and Earth Sciences,

respectively. Scientific unions have played key roles in facilitating scientific communication in their respective disciplines through development of standard definitions, procedures, publication practices and other processes of exchange across sectors. Data science presents many opportunities for innovative analysis across massive and diverse types of data from multiple disciplines and contexts. The scale of these efforts is surfacing many previously hidden challenges in data exchange and has prompted the formulation of the FAIR principles of Findability, Accessibility, Interoperability, and Re-usability of data by humans and machines [1]. These principles reflect the underlying core values of scientific data integrity in the long-standing work and expertise of the unions. Scientific unions have the opportunity to enable the incorporation of FAIR principles into the workflows of stakeholder communities of practice, to facilitate accurate data exchange and to enrich flow of data into interdisciplinary data projects emerging around the world. This talk will cover the FAIR data efforts and initiatives getting underway in IUPAC and AGU, and explore opportunities for synergy and collaboration.

CINF 62

Top ten emerging technologies in chemistry: A new initiative from IUPAC and *Chemistry International*

Fernando Gomollon-Bel[3,4]*, gomobel@gmail.com, Javier Garcia Martinez*[2]*, Helen A. Lawlor*[1]*. (1) Retired, Radnor, Pennsylvania, United States (2) Department of Inorganic Chemistry, University of Alicante, Alicante, Spain (3) Department of Engineering, University of Cambridge, Cambridge, United Kingdom (4) European Young Chemists' Network, EuChemS, Brussels, Belgium*

Throughout its first one hundred years, IUPAC has contributed to the worldwide understanding and application of the chemical sciences for the well-being of our world and humankind. This paper discusses The 'Top Ten Emerging Technologies in Chemistry' - a new initiative established by IUPAC to even more broadly promote the essential value of the chemical sciences. Beginning in 2019, the year that commemorates IUPAC's 100th anniversary, this initiative will provide a review of the most relevant technologies emerging in the field of Chemistry. The resulting paper will be published annually in IUPAC's news journal, *Chemistry International*. Every year, the 'Top Ten Emerging Technologies in Chemistry' review article will showcase innovative, emerging technological advances in chemistry, materials, and engineering, highlighting how these advances contribute to society's well-being and to a more sustainable future for our planet. Note that for this initiative an emerging technology is defined one that hovers between a new scientific discovery and a fully-commercialized technology. In 2018, IUPAC solicited nominations for outstanding emerging technologies to chemists

all around the globe. Then, IUPAC selected a panel of prestigious researchers who served as judges and who reviewed all the nominations and, based upon their collective knowledge and experience, selected their own top ten. The results, along with a compelling article reviewing the potential of the chosen emerging technologies, will be published in the April 2019 issue of *Chemistry International* and presented for the first time at the 2019 ACS Spring National Meeting in Orlando.

CINF 63

IUPAC and its next century: A secretary general's perspective

Richard Hartshorn[1,2]*, richard.hartshorn@canterbury.ac.nz. (1) University of Canterbury, Christchurch, New Zealand (2) Secretary General, International Union of Pure and Applied Chemistry, Research Triangle Park, North Carolina, United States*

IUPAC (www.iupac.org) has a proud history in development of nomenclature, curation of the periodic table, chemical education, and critical evaluation of data, among many other things. A century of achievement in these areas means that it can rightly claim to have provided the essential tools for the application and communication of modern chemical knowledge. The challenge for IUPAC, now, is to look forward to the next century, to identify critical needs for the discipline and to build international consensus around the standards and activities that are required to meet those needs. As Secretary General of IUPAC, I have significant input into the future direction of IUPAC and its work. I will present my perspective on what the future holds for the organisation, and highlight the importance of promoting cheminformatics as a critical field for growth, of facilitating cross-disciplinary and inter-organisational activities, and developing ways to involve more young chemists in IUPAC work. There will be some history, some explanations, some crystal ball-gazing, and a few long names (merely because I have to play up to expectations of someone still involved in nomenclature development – at least a little bit).

CINF 64

3decision®: Bringing structural data analytics to the masses

Gabriella Jonasson, gabriella.jonasson@discngine.com. Discngine, Paris, France

Over half of today's drug discovery projects rely on rational structure-based drug design (SBDD) techniques. Unfortunately, the structural information that these projects have available is still not used to its full potential.

The major hurdles lie in the complexity of structural data analytics and in an inconsistent persistence of the data. Considering the exponential growth of the production of new protein structures, these obstacles will become increasingly significant in the drug discovery industry. 3decision®, a new structural knowledge management solution, addresses these issues by transforming the massive amount of data, coming from 3D structures and metadata, into persistent structural knowledge. The solution's unique web-based user interface puts the focus on the collaboration between different types of users. Structural biologists, medicinal chemists, and molecular modelers can thus easily access complex analyses and generate, test, connect ideas with each other and the rest of the community.

CINF 65

Leveling the playing field: Illuminating understudied targets with Pharos

Timothy Sheils[1]*, timothy.sheils@nih.gov, Dac-Trung Nguyen*[1]*, Noel Southall*[1]*, Tudor I. Oprea*[2]*, Vishal Siramshetty*[1]*. (1) NCATS, NIH, Potomac, Maryland, United States (2) University of New Mexico, Albuquerque, New Mexico, United States*

Pharos is a web resource that originated from the NIH "Illuminating the Druggable Genome" (IDG) program to characterize the understudied regions of the druggable genome. These so-called “dark” regions are important because they have potential druggable opportunities toward a wide range of disease areas. While displaying data on well known targets is fairly simple, it is much more difficult to display data, or the lack thereof, on understudied targets. To better facilitate analysis and prioritization of understudied targets, we recently introduced two new features within Pharos. First, Pharos utilizes an “illumination graph” to display the strengths and weakness of the knowledge landscape for a target. Using a standardized visualization, it is possible to show what areas of the target have the most potential for illumination, which can direct funding and research into those realms. Second, Pharos has the capability to generate

topics of interest for users. Topics can range from disease types to protein classes, and are an aggregation of the data available on each group of targets, diseases and small molecules. There are several benefits from this grouping. Users are able to view a target as part of a larger subset, one that frequently includes dark targets. In this context, we are also able to highlight a target's potential by combining the IDG target level and knowledge score. Using these criteria we are able to illuminate targets with the most and least knowledge available, as well as targets that have druggable potential, or an easily bridged knowledge gap. These features can be valuable resources in the quest to increase the amount of knowledge available about dark targets.

CINF 66

Chembench: A publicly-accessible, integrated cheminformatics portal

Eugene Muratov[1]*, murik@email.unc.edu, Stephen Capuzzi*[1]*, Vinicius M. Alves*[1]*, Valery tkachenko*[2]*, Alexander Korotcov*[2]*, Daniel Korn*[3]*, Wai In Lam*[3]*, Thomas Thornton*[3]*, Diane Pozefsky*[3]*, Alexander Tropsha*[1]*. (1) Chemical Biology & Medicinal Chemistry, University of North Carolina - Chapel Hill, Chapel Hill, North Carolina, United States (2) Science Data Software, LLC, Rockville, Maryland, United States (3) Computer Science, University of North Carolina, Chapel Hill, North Carolina, United States*

The enormous increase in the amount of publicly available chemical genomics data and growing emphasis on data sharing and open science mandates cheminformaticians to make their models publicly available for broad use by the scientific community. Chembench is one of the first publicly-accessible, integrated cheminformatics Web portals. It has been extensively used by researchers from different fields for curation, visualization, analysis, and modeling of chemogenomics data. Since its launch in 2008, Chembench has been accessed more than 1 million times by ~15K users from a total of 98 countries. We report on the recent updates and improvements that increase the simplicity of use, computational efficiency, accuracy, and accessibility of a broad range of tools and services for computer-assisted drug design and computational toxicology available on Chembench. Current Chembench is based on Open Data Science Platform, which is fully compliant with FAIR and 5 V principles. Chembench remains freely accessible at https://chembench.mml.unc.edu.

CINF 67

K4DD database: Ligand binding kinetics at its best

Gerhard F. Ecker*, gerhard.f.ecker@univie.ac.at, Lars Richter. Dept. Pharmaceutical Chemistry, University of Vienna, Wien, Austria*

Residence time - and more recently - the association rate constant kon are increasingly acknowledged as important parameters for in vivo efficacy and safety of drugs.

However, their broader consideration in drug development is limited by a lack of knowledge how to optimize these parameters. In the light of the IMI funded collaborative project K4DD - Kinetics for drug discovery, a data warehouse for hosting all data derived within the project has been developed. With the end of the project in 2017, an aggregated version of the data has been donated to ChEMBL. In the meanwhile, we performed an extensive literature search and enriched the original K4DD data with more than 900 additional kinetic triples (kon, koff, KD). This allows complex queries and analyses across multiple targets, especially exploiting matched molecular pairs.

CINF 68

MOEsaic: The application of matched molecular pair analysis to SAR exploration

Guillaume Fortin, gfortin@chemcomp.com. Chemical Computing Group, Montreal, Quebec, Canada

With the increasing size of data sets and the parallel development of multiple structural series in medicinal chemistry projects, managing and analyzing structure activity/property relationship data is becoming ever more challenging. Tools and methods for the efficient visualization, analysis and profiling of compounds therefore remain of deep interest. Here, we describe a new web-based application, MOEsaic, which enhances typical medicinal chemistry workflows aimed at interrogating the SAR/SPR data through the application of interactive matched molecular pairs (MMP) analysis and R-group profiling.

CINF 69

ARENA360: An integrated informatics solution for drug discovery

Christophe Betton[1], christophe.betton@emdserono.com, Zhaowen Luo[2], Veit Ulshoefer[3]. (1) Research and Bioinformatics, EMD Serono, Inc., Billerica, Massachusetts, United States (2) EMD Serono Research and Development Institute, Inc, Wayland, Massachusetts, United States (3) Research and Bioinformatics, Merck KGaA, Darmstadt, Germany

EMD Serono has a centralized data warehouse for all research data. The datastore captures about two million registered entities, hundreds of millions of biological endpoints as well as half a million documents. To better share data among the internal R&D community, a web-based platform is developed to provide an integrated solution for data retrieval, mining, and analysis.
ARENA360 provides easy browsing functionality for each compound, the data are presented in specific sections: basic chemical structure and related data; compound batches and inventory; and biological assay data grouped by categories – in vitro, in vivo, ADME etc. In addition, by using Chemaxon technology, ARENA360 also provides

structure searching functions, as well as advanced data mining options.
In addition to internal data, reference compounds are annotated using Drugbank and Cortellis databases, so data from external databases are also available along with the internal ones. Furthermore, several design tools are integrated, such as nearest neighbor searching and internally developed chemistry design applications. Compounds are also annotated by their target profiles based on biological assay results, therefore allowing users to drill into biology via the linked target database. Within this target database, users can find target profiles, disease indications, and even competitive intelligence information.
ARENA360 is not limited to NCEs, but also has NBE capabilities. Small molecule and protein drug candidates can be searched and browsed side by side in one web interface. The platform is built based on many emerging technologies for big data and provides an integrated view for the internal and external data for the drug candidates; it is used daily by over a thousand scientists.

CINF 70

Delivering computational chemistry to cheminformatics: collaborative drug discovery with LiveDesign

Erin Davis, erinsdavis@gmail.com. Schrödinger Inc, Seattle, Washington, United States

Drug Discovery has inarguably become dependent upon a plethora of computational tools and data, requiring increasing collaboration across traditionally siloed areas of computational modeling and medicinal chemistry. These areas have become more and more critical to reducing attrition and improving predictions earlier in the R&D process, approaching a nearly quantitative level. With this comes substantial challenges of tracking project progression, accessing data across various tool sets, and sharing ideas across teams. Too often files are lost, ideas are not traceable from inception to synthesis, or scientists just can't form the ad-hoc queries they want across various datasets.
Herein we present LiveDesign, a highly-collaborative and intuitive web-based platform for fostering creativity by bringing computational modeling alongside experimental and predictive data. LiveDesign also recognizes the necessity of heavy extensibility, with easy plug and play gadgets through well-established web technologies to reach into other tools as needed. This talk will cover how LiveDesign is helping streamline early phase drug discovery by demonstrating the democratization of modeling and data with several use cases.

CINF 71

Advances in deep learning and their applied utility toward chemical informatics & drug discovery

Evan Clark[1,2], *eclark28@fau.edu*, *William E. Hahn*[3], *williamedwardhahn@gmail.com*, *Rachel St Clair*[4], *Paul Morris*[4], *Mike Teti*[4]. *(1) MedBios, Lake Worth, Florida, United States (2) Biomedical Science, Florida Atlantic University, Boca Raton, Florida, United States (3) College of Science, Florida Atlantic University, Boca Raton, Florida, United States (4) Complex Systems and Brain Sciences, Florida Atlantic University, Boca Raton, Florida, United States*

With new approaches to gradient descent based on a novel branch of computational calculus known as automatic differentiation, deep neural networks have proven in the last half decade to outperform several machine learning methods across multiple divisions of data science. Their utility has already been demonstrated in the development of fully-autonomous vehicle navigation, speech recognition, object detection and that are in production today. With the advent of deep convolutional networks with automatic differentiation, it is now commonplace to build what are known as 'end-to-end models'. These classes of models map directly from inputs to output (i.e. SMILES string to toxicity, structure to functionality, and inversely, mapping from desired function to possible structure). We discuss the three most popular deep learning models, convolutional neural networks, long-short term memory networks, and generative adversarial networks (GANs) and their utility to several branches of computational chemistry. GAN's are of great particular interest to computational chemistry; previous models were limited to classification or regression tasks whereas the latest of generative models can generate high resolution multi dimensional (2-dimensional & 3-dimensional) structures and sequences, providing new compound candidates for virtual high throughput screening. Furthermore, these deep networks are transparent in that the model parameters the networks learn can be visualized, interpreted, and most importantly transferred to new applications.

CINF 72

How much can we learn from smiles as text?

Hongmao Sun, *hongmao.sun@nih.gov. NIH, Rockville, Maryland, United States*

Deep learning (DL) has been achieving unprecedented successes in many fields, including medical diagnosis and medical image analysis. Learning capability of DL is challenged in this study by deciphering structural features embedded in the text of Smiles strings without aid of any parsers. Using preprocessed canonical Smiles strings as input, recurrent neural network (RNN) was employed to decipher the patterns buried in the Smiles strings, and the output was fed to deep neural network (DNN) to construct regression logP models for the 10,851 compounds in the Starlist and classification models for aqueous solubility of 11,208 compounds. The predictive performance of the logP model is comparable to the current benchmark – clogP, as measured by root mean square (RMS) of errors. Excellent predictive power was also achieved for classification models of aqueous solubility.The areas under the receiver operating characteristic

curve (AUC-ROC) were above 0.90 for both the test and validation sets (10% each). The results not only uncover the tremendous structural information embedded in the Smiles strings, but also open a new field of DL-based parser-free and descriptor-free QSAR approaches.

CINF 73

Novel, active learning approach for deep learning of chemical data: Extracting more chemical insights by choosing less

Mojtaba Haghighatlari[1], *mojtabah@buffalo.edu, Johannes Hachmann*[2]*. (1) Chemical and Biological Engineering, University at Buffalo, Buffalo, New York, United States (2) Dept of Chemical and Biological Engineering, University at Buffalo, SUNY, Buffalo, New York, United States*

We can easily tailor deep learning (DL) architectures to learn the underlying physics of the problems even from simple chemical representations. However, training a deep artificial neural network typically requires large data sets from experimental characterization or even physics-based models, which are dramatically demanding and correspondingly limited in scope. Here, we present a novel active learning sampling method for the training of the DL models to leverage small size data sets. This approach establishes accuracy and speed up in a closed loop of data modeling and molecular design. As a proof of concept, we report capabilities of this method for accurately predicting high refractive index (RI) of organic molecules in a virtual high-throughput screening approach. The data generation protocol requires calculation of electronic and optical properties of molecules using the density functional theory (DFT) and molecular dynamics (MD) simulation. We demonstrate that the proposed approach can reduce the size of training set by %72 to achieve the same accuracy in comparison to the model trained on a larger set. This is a significant speed up in terms of data acquisition (i.e., DFT and MD calculations). We further apply this method to several DL architectures and discuss the involved challenges to extract insights from DL models as a guide for our active learning approach.

CINF 74

Application of machine learning to skin cancer detection and classification

Andrew C. Terentis, *terentis@fau.edu, John Strasswimmer. Florida Atlantic Univ, Boca Raton, Florida, United States*

In the field of machine learning, the goal of statistical classification is to use an object's characteristics to identify which class it belongs to. We are seeking to utilize the techniques of machine learning for rapid, automated detection and removal of skin cancers in a clinical setting using Raman spectroscopy for the detection and laser

ablation for the removal. In a preliminary study we obtained twenty-five tissue samples from eleven patients undergoing Mohs surgery to remove squamous cell carcinomas (SCC). Raman spectra were collected from both untreated and ablated normal and SCC tissue samples. Spectra were then subjected to principal component analysis (PCA) to reduce the dimensionality of the data. The first five PCs, which collectively encompassed more than 65% of the data variance, were then used as the predictors for a Binary Logistic Regression (BLR) that classified individual spectra into the two categories (normal vs. SCC). For non-ablated samples, cancers were detected with 92% sensitivity and 60% specificity, while for ablated samples SCC were identified from the Raman spectra with 95% sensitivity and 100% specificity. We are currently generating a much larger training data set and exploring different machine and deep learning approaches to Raman spectroscopic data analysis to determine the best way of achieving our goal of rapid and accurate detection of cancers using Raman spectroscopy.

CINF 75

Deep learning for the characterization and identification of small molecules

Sean Colby, sean.colby@pnnl.gov, Jamie Nunez, Nathan Hodas, Courtney Corley, Ryan Renslow. Pacific Northwest National Laboratory, Richland, Washington, United States

Robust and comprehensive identification of small metabolites in complex samples will revolutionize our understanding of metabolic interactions in biological systems. Existing and emerging technologies have enabled measurement of chemical properties of molecules in complex mixtures and, in concert, are sensitive enough to resolve even stereoisomers. Despite these experimental advances, small molecule identification is inhibited by (i) a deficiency in reference properties (e.g. mass spectra, collisional cross section, and other measurable properties), limiting the number of possible identifications, and (ii) the lack of a method to generate candidate matches from experimental features. To this end, we developed a variational autoencoder (VAE) to learn a continuous numerical, or latent, representation of molecular structure, to simultaneously characterize and expand reference libraries for small molecule identification. We extended the VAE to include a chemical property decoder, trained as a multitask network, in order to shape the latent representation such that it assembles according to desired chemical properties. The approach is unique in its application to metabolomics and small molecule identification, its focus on properties that can be obtained from experimental instruments (mass, CCS), and its training paradigm, which involved a cascade of transfer learning iterations. First, molecular representation is learned from a large dataset of unlabeled structures. Next, properties calculated in silico are used to continue training with property prediction, as experimental property data is limited. Finally, the network is trained with the limited experimental data. This allows the network to learn as much as possible at each stage, enabling success with progressively smaller datasets without overfitting. Once trained, the network can be used to predict

chemical properties directly from structure, as well as generate candidate structures with chemical properties similar to some arbitrary input. Our approach is orders of magnitude faster than first-principles simulation for property prediction. Additionally, the ability to generate molecules along manifolds defined by chemical property analogues positions this work as highly useful in a number of application areas, including metabolomics and small molecule identification, drug discovery and design, chemical forensics, and beyond.

CINF 76

Virtual high-throughput screening: A combined deep-learning approach

Paul Morris[3,1]*, pmorris2012@fau.edu,* ***Rachel St Clair***[3,1]*, rstclair2012@fau.edu,* ***Mike Teti***[3,1]*, mteti@fau.edu, Evan Clark*[3,1,2]*, William E. Hahn*[3]*. (1) MedBios, Lake Worth, Florida, United States (2) Biomedical Science, Florida Atlantic University, Boca Raton, Florida, United States (3) Center For Complex Systems & Brain Sciences, Florida Atlantic University, Boca Raton, Florida, United States*

Chemoinformatics aims to assist in atomistic system applications that depend on molecular interactions, structural characteristics, and functional properties. The arrival of deep learning and the abundance of easily accessible chemical data from repositories like Pubchem and ChemSpider have enabled advancements in developmental fields such as drug discovery. For example, virtual High-Throughput Screening (vHTS) is one such technique that integrates deep learning and chemical processing to provide in silico biomolecular simulation. In this technique, a deep learning approach that combines several computational models and architectures was developed to define an in silico screening method for target-binding molecules. To accomplish this, the inherent embedding of an end-to-end deep learning model trained to encode the structural characteristics and functional properties of a molecule via a corresponding SMILES string was repurposed through transfer learning to predict target binding chemicals and proteins. Integrative approaches in computational chemistry, such as the method applied here, can provide essential advancements in in silico chemical research when compared to traditional in vitro and in vivo experiments.

CINF 77

Learn deep before deep learning

Karina Martinez Mayorga[1]*, kmtzm@unam.mx, Gabriela Gómez Jiménez*[2]*, Abraham Madariaga-Mazon*[1]*. (1) Physical Chemistry, Instituto de Quimica, UNAM, Scottsdale, Arizona, United States (2) Instituto de Química, UNAM, Mexico City, Mexico*

Machine learning algorithms have been developed and used in chemoinformatics and computational chemistry, for decades. However, a new wave of high expectations has

risen around the use of deep learning. In this talk, we will discuss current applications, suggest areas of opportunity, and point out limitations and perspectives of the field.

CINF 78

Going beyond popular: Assessing *SciPop Talks!*

Raychelle M. Burks[1], *rburks@stedwards.edu, Kiyomi Deards*[2], *Erica DeFrain*[2]. *(1) Chemistry, St. Edward's University, Austin, Texas, United States (2) Libraries, University of Nebraska, Lincoln, Nebraska, United States*

Now entering its sixth year, *SciPop Talks!* is a successfull speaker series at the intersection of pop culture and science. Its continued operation, now spanning two unaffliated campuses, speaks to its popularity. But what is behind high attendance numbers? Why do participants attend and what do they take away? In this talk, *SciPop Talks!* organizers will share what their survey data shows thus far and what future assessments will be done.

CINF 79

Understanding interest, relevance, & self-efficacy: Chemistry at the museum and beyond

Emily L. Howell[1,2], *eleahyhowell@gmail.com, Shiyu Yang*[2], *Dietram A. Scheufele*[2,3]. *(1) The Nelson Institute for Environmental Studies, University of Wisconsin-Madison, Madison, Wisconsin, United States (2) Department of Life Sciences Communication, University of Wisconsin-Madison, Madison, Wisconsin, United States (3) Morgridge Institute for Research, Madison, Wisconsin, United States*

As part of a larger project to understand public engagement around chemistry, this study describes a unique research collaboration between the Museum of Science

Boston (MOS), the Science Museum of Minnesota (SMM), and science communication researchers at the University of Wisconsin-Madison (UW) to better understand people's views of chemistry and what factors shape those views. This work was part of an NSF-funded research study (ChemAttitudes, award #1612482) and this presentation focuses on UW's role assessing views of chemistry in a more general public, to better understand and complement the findings MOS and SMM captured on the museum floor. It will outline: 1) questions that shaped initial research design, 2) the surveying process and analysis methods we used, and 3) how each step of data collection informed subsequent collections and analysis, as well as some of the strengths and challenges of the collaboration. The goal is to illustrate how collaborative research can inform study design through multiple angles as well as present some research design and analysis that might be useful to engagement practitioners. We touch briefly on each of these areas in more detail below. First, we will outline how our collaboration began, what the goals were, and what research questions we formed from those goals. Second, we will describe the survey design, which used two waves of surveying through the online platform Amazon Mechanical Turk. The platform is an attractive option for surveying larger groups and running survey-embedded experiments and is inexpensive enough that it could be useful for running pilot tests or collecting data to supplement evaluation data. In this section we will focus in particular on how we translated research questions into features of the survey and analysis and what some of the strengths and challenges were with our methods of data collection and analysis. We will also discuss some of the results as examples of how they complemented findings from MOS and SMM. Finally, we will close with a summary of the challenges and successes of this particular collaboration and tips for creating such collaborations and using a variety of methods to understand science outreach outcomes. This section will include the steps we took to create regular and effective communication across the research team, taking into account challenges of communicating across geographical distances, professional experiences, and sometimes distinct research goals.

CINF 80

Collecting, understanding, and utilizing audience feedback to increase interest, relevance, and self-efficacy related to hands-on chemistry activities in a museum

Gretchen M. Haupt, ghaupt@smm.org. Research and Evaluation, Science Museum of Minnesota, Saint Paul, Minnesota, United States

Understanding the impact of programming and outreach relies on engaging with your audience about their experiences. This can occur through informal dialogue and/or systematic data collection processes. The Museum of Science Boston (MOS) and the Science Museum of Minnesota (SMM) have been collecting visitor feedback on facilitated chemistry activities in museums as part of a design based research (DBR)

study that aims to increase public interest and understanding of chemistry as well as increase public perception of chemistry's relevance and increase the public's self-efficacy with respect to chemistry. The DBR approach involves applying a theoretical framework to the development of hands-on educational activities about chemistry, while also testing and modifying the framework itself. This talk focuses on the methodologies researchers at MOS and SMM used to engage public audiences in order to gather and understand data about activities' impacts on visitors in regards to increasing interest, relevant, and self-efficacy. These data were also used to inform further iterations of the activities to achieve close alignment between the projects' theoretical framework and actual impacts. The presenter will discuss the methods used to gather data (surveys and interviews, observations, and video capture), the challenges and opportunities related to collecting data in a free-choice learning environment, sense-making activities related to the data, along with brief findings and their utilization.

CINF 81

Advancing inclusive excellence in academic chemistry departments from the top down through a discipline-based evidenced-based approach

Rigoberto Hernandez, r.hernandez@jhu.edu, Dontarie Stallings, Srikant K. Iyer. Chemistry, Johns Hopkins University, Baltimore, Maryland, United States

The Open Chemistry Collaborative in Diversity Equity (OXIDE) is aimed at institutional reform that lowers inequitable barriers hindering the success of faculty from diverse groups. The collaborative itself is a partnership with the department heads of research-active chemistry departments, social scientists and other key stakeholders. The lowering of these barriers increases the likelihood that individuals already in the tenure pipeline will have equitable chances of success and thereby leads to changes in faculty demographics that move closer to those of the broader U.S. population. The creation of a more equitable climate is also expected to encourage more disadvantaged students to enter academic careers in the chemical sciences. Through biennial National Diversity Equity Workshops (NDEWs) since 2011, OXIDE convenes the chairs of chemistry departments to collectively identify effective practices in collaboration with leaders from the social and behavioral sciences, and to enable them to hold each other accountable for advancing inclusive excellence from the top down.

CINF 82

Science outreach: What does it mean to be successful, and how do we know?

Jeanne Garbarino, jgarbarino@rockefeller.edu. RockEDU Science Outreach, The Rockefeller University, New York, New York, United States

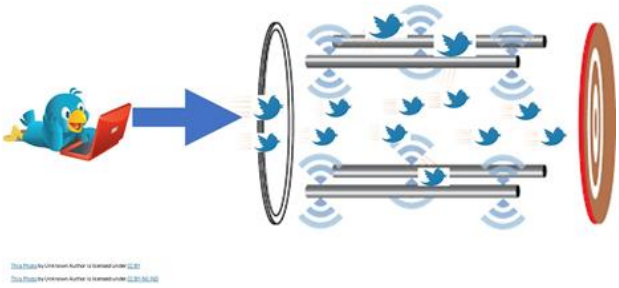
As the scientific enterprise evolves to meet the needs of modern researchers, we are seeing a greater emphasis on the value of science outreach efforts, both as a practice and as a profession. This, in turn, has placed a greater emphasis on the requirements for assessment and evaluation strategies related to science outreach. However, there are virtually unlimited ways to design programs and activities related to science outreach making it incredibly difficult to establish standard tools to measure impact. Drawing from direct experiences at RockEDU Science Outreach at The Rockefeller University, as well as findings from SciOut18 -- the first national meeting dedicated to the field -- this talk will touch on the core elements for determining science outreach success in the context of individuals and communities.

CINF 83

Amplifying your social impact: A collaborative approach to chemistry outreach

Maria T. Gallardo-Williams*, Maria_Gallardo@ncsu.edu,* ***George Van Den Driessche****, georgevdd@gmail.com, Alexandra Malico. Chemistry, North Carolina State University, Raleigh, Cary, North Carolina, United States*

In 2017, an estimated 2.34 billion people world-wide reported having an active social media account, with approximately 336 million of these users having an active Twitter account. Social media platforms, like Twitter, offer obvious benefits for promoting and sharing scientific research. Each user becomes an instant news source for their work, and are capable to report lab updates, news, and discoveries in near real time, allowing for direct and public engagement with the scientific community. This allows research to be promoted for increased visibility and citations. However, amidst the noise of a billion profiles, how can you be heard and measure your impact? The NC State chemistry communication team has adapted an amplification method that relies on echoing content through faculty-student account collaborations. This approach relies on promoting departmental news using faculty (@NCStateChem) and student (@NCStateChemGSA) run accounts, which are then amplified through research lab or personal community member profiles. Conversely, we then amplify research-related content from lab and personal accounts. Our approach relies on the belief that growing the departmental brand is dependent upon the brand of our researchers, teaching faculty, and students, and vice-versa. In this presentation, we will share insights into our social media strategy for Twitter (#ChemPack, student hosts), as well as an emerging faculty-student led Instagram account, and provide an overview of our in-house amplification network monitoring impressions and engagement rate as a metric to assess our impact.



CINF 84

Evaluating impact

***Suze Kundu**, s.kundu@digital-science.com. Digital Science, London, United Kingdom*

As Professor Sir Mark Walport, former Chief Scientific Advisor of the UK Government, once said, "science isn't finished until it is communicated". Given that research is carried out to help people, be that through clean water, cheap energy, or advancements in medicine, researchers have a duty to engage with the public on the aims, processes and outcomes of their research. Yet even today, despite the rise in popularity and deeper understanding of the need for societal engagement, there is relatively little emphasis on or acknowledgement of this part of the research cycle. In this discussion, tools to help researchers, institutions and funders monitor, manage and share best practice of impact through evaluation will be showcased, along with some suggestions on how to nudge the academic world into better supporting this important and indeed fundamental stage of research.

CINF 85

How can I measure the success of my online outreach?

***Dorea Reeser**, D_Reeser@acs.org, Sondra Hadden, Michael Ruhl, Amanda T. Yarnell. C&EN, American Chemical Society, Washington, District of Columbia, United States*

Online engagement can be key to the success of STEM outreach. But how do you decide where to spend your content and social media resources? Learn how C&EN, a leader in chemistry news, measures the impact and success of its online campaigns. C&EN Audience Engagement Editor Dorea Reeser will share the lessons we've learned, and key takeaways that you can apply to your own online outreach as soon as you step out the door.

CINF 86

Mapping the chemistry Twitter community: A reproduction of academic power structures or an opportunity to empower marginalized voices?

Paulette Vincent-Ruz[1], *pvincentruz@gmail.com, Dorea Reeser*[2], *Matthew R. Hartings*[3]. *(1) University of Pittsburgh, Pittsburgh, Pennsylvania, United States (2) Chemical & Engineering News, Washington, District of Columbia, United States (3) Chemistry, American University, Gaithersburg, Maryland, United States*

Social media is increasingly the place where chemists engage in knowledge sharing, public discussions, and debates. However, at first glance, it may be difficult not only to examine how strong are the relationships between users but also if there is anything to gain from engaging in the twitter chemistry community. Furthermore, despite being an easy application to access, Twitter can easily reproduce existing power structures within academia. Therefore, it is essential that we not only understand the structure of the chemistry Twitter community but also whether its development has allowed for the expression and dissemination of marginalized voices. Social network analysis is a technique that explicitly takes social structure into account and allows us to investigate intra- and inter-group behavior. Based on a network analysis

of over 6,000 chemistry twitter users we were able not only identify “who is following who?” but also highlight the structure of the network’s relationships, and identify users whose position is particularly dominant. Once we established the basic network structure, we then proceeded to analyze whether there were important sub-communities based on field, location, language or topics of interest. The analysis allowed us to identify not only community clusters but also users that acted as broadcasters of information between groups. Finally, an analysis of influential users was conducted to identify whether their characteristics reflected the current hierarchical structures of academia, or whether Twitter allowed marginalized voices to thrive and empowered them to become central figures within the chemistry Twitter community. This presentation will discuss the results and implications on how to use Twitter to identify and empower marginalized voices within chemistry, as well as its potential to create collaborations and support to chemists at various stages of their careers.

CINF 87

Interpretable molecular design based on layer-wise relevance propagation

Youngchun Kwon, kown10@gmail.com, Kyungdoc Kim, Inkoo Kim, Jiho Yoo, Won-Joon Son, Youn-Suk Choi, Hyo Sug Lee, Jaikwang Shin. Platform Technology Laboratory, Samsung advanced institute of technology, Suwon, Korea (the Republic of)

Many machine learning-based researches have been carried out to accelerate the materials discovery and deep neural networks (DNNs) demonstrated impressive performance in terms of materials property prediction and inverse design. However, current technologies remain at a statistical level due to the limitation in analyzing and unraveling the deep learning models. Thus, we couldn't figure out the logic behind their decision and couldn't do rational molecular design based on clear evidence like human researchers. To overcome this, we present an interpretable molecular design method that extracts crucial sub-structures in molecules for expressing specific properties using layer-wise relevance propagation (LRP) and then use this information to design new molecules. In the proposed workflow, extended connectivity fingerprint (ECFP) is used as an input representation of molecules and DNN models for property prediction are constructed. Following the prediction of property of a seed molecule, LRP identifies important bits in the input ECFP by running a backward pass in the DNN model where the nodes that contribute significantly to the output-layer receive higher relevance. In order to design new molecules for target, the selected bits are randomly modified and then the resulting fingerprints are reconstructed into actual molecular structures through a recurrent neural network model. Successive iterations of this procedure gradually enhance the property and automatically evolve the molecular structures to meet the target. The effectiveness of our *de novo* design method was demonstrated successfully by evolving organic molecules toward the increase and decrease of their maximum light-absorbing wavelengths by adopting PubChem as a training library. As a result, the ratio of newly designed molecules which met the target condition was 86.5%, which was about twice as high as the previous random approach type evolutionary design. And the speed of property improvement was about 1.5 times higher. This research provided an innovative molecular design methodology to alleviate the painstaking effort of chemists and reduce the turnaround time for materials development. In this respect, our interpretable molecular design is expected to be one of the promising tools to explore enormous chemical space efficiently.

CINF 88

Machine-learned model for molecular simulations of liquid and water vapor

Troy Loeffler[2], Tarak Patra[1], Henry Chan[2], Subramanian Sankaranarayanan[1], skrssank@anl.gov. (1) Argonne National Lab, Naperville, Illinois, United States (2) Argonne National Laboratory, Schaumburg, Illinois, United States

An accurate and computationally efficient molecular level description of mesoscopic behavior of ice-water systems remains a grand challenge. There are a number of different water models that have been developed in an attempt to adequately capture the various thermodynamic anomalies exhibited by water in its various phases. While the solid and liquid phases have received much attention, studies on the vapor phase remains scarce. As an alternative,machine-learning techniques can be used to interpolate atomistic simulations or first principles calculations. Such machine-learning potentials (MLPs) enable linear-scaling atomistic simulations with an accuracy that is close to the reference method at a fraction of the computational cost. Here, we train an artificial neural network interfaced directly with Monte-Carlo based sampling to develop an ML model for water vapor. We evaluate its predictive power and make systematic comparisons with other existing water models including the best available polarizable and non-polarizable models. The limitations and strengths of our ML model in the context of water vapor simulations will be discussed.

CINF 89

Prediction of chemical reactivity with a graph-convolutional neural network model

Connor W. Coley[1], ccoley@mit.edu, Wengong Jin[2], Luke Rogers[1], Timothy F. Jamison[3], Tommi Jaakkola[2], William H. Green[4], Regina Barzilay[2], Klavs F. Jensen[5]. (1) Chemical Engineering, MIT, Cambridge, Massachusetts, United States (2) CSAIL, MIT, Cambridge, Massachusetts, United States (3) Chemistry/18-590, MIT, Cambridge, Massachusetts, United States (4) Rm E17-504, MIT, Cambridge, Massachusetts, United States (5) Dept of Chem Eng Rm 66 350, MIT, Cambridge, Massachusetts, United States

We present a supervised learning approach to predict the products of organic reactions given their reactants, reagents, and solvent(s). By training a graph convolutional neural network model on hundreds of thousands of reaction precedents from the patent literature, the neural model makes informed predictions of chemical reactivity. It predicts the major product correctly over 85% of the time (requiring around 100 ms per example) using a benchmark dataset of 470k reactions from the USPTO. This accuracy improves upon the state of the art by over 5% and performs on par with expert chemists with years of formal training in a small-scale human benchmarking study. We gain additional insight into predictions via the design of the neural model, specifically the inclusion of a global attention mechanism, revealing an understanding of chemistry qualitatively consistent with manual approaches. This is an initial step toward model interpretability for the prediction of organic reactivity.

CINF 90

Predicting bond dissociation energies through deep learning

Yanfei Guan[1], yanfei.guan@chem.tamu.edu, Yeonjoon Kim[2], Peter St. John[2], Seonah Kim[3], Robert S. Paton[1,4]. (1) Chemsitry, Colorado State University, Fort Collins, Colorado, United States (2) National Renewable Energy Laboratory, Golden, Colorado, United States (3) National Bioenergy Center, National Renewable Energy Laboratory, Lakewood, Colorado, United States (4) Chemistry Research Laboratory, University of Oxford, Oxford, United Kingdom

The homolytic cleavage of covalent bonds underpins much of the chemistry involved in combustion, metabolism and catalysis, among many other areas. The ability to routinely and rapidly predict bond dissociation energy (BDE) values is therefore of great value to a broad cross-section of chemists. However, predicting BDE values for all breakable bonds in molecules using conventional applications of quantum chemistry, while accurate, can be computationally expensive and tedious. Herein, we summarize our progress in building a deep learning model and BDE database, which serve as training set of the network. This approach is applied to thousands of molecules in the PubChem dataset, where it is demonstrated to provide comparable accuracy to DFT calculations at significantly lower costs. Model training could be carried out in several hours on GPUs. We developed an automated workflow built on the Atomic Simulation Environment Data Base to obtain DFT reference BDE values for all breakable bonds in a large collection of organic molecules. We modified a continuous-filter convolutional neural network (Schnet), which refines atom representations considering interactions with neighboring atoms, to predict BDE for molecules.

CINF 91

Multitask prediction of site selectivity in aromatic C-H functionalization reactions

Thomas J. Struble[1], tstruble@mit.edu, Connor W. Coley[1], Klavs F. Jensen[2]. (1) Chemical Engineering, MIT, Cambridge, Massachusetts, United States (2) Dept of Chem Eng Rm 66 350, MIT, Cambridge, Massachusetts, United States

Accurate predictions of site selectivity for aromatic functionalization reactions is crucial for prioritizing target compounds in drug discovery and routes in process chemistry. Selectivity of aromatic C-H functionalization reactions is highly dependent on subtle electronic plus steric features of the substrate and can be controlled by reaction conditions such as catalysts. Prediction of reaction outcomes using machine learning is possible using both template and template free approaches and is generalizable to a large corpus of reactions but comes with a tradeoff of poor accuracy with site and regio-specificity. Current approaches to site selectivity prediction achieves good accuracy based on electronic calculations but does not address steric or reaction conditions and generally takes 1-10 min per molecule. We report a generalizable approach to prediction of site selectivity that is accomplished using a shared graph-convolutional neural network fed into a multitask predictor trained on ~58,000 examples from Reaxys. Overall mean reciprocal rank for all C-H functionalization tasks is 92% and predictions

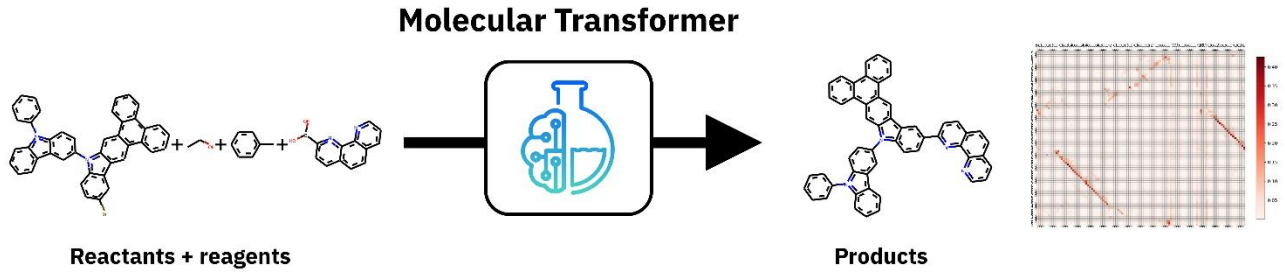
are made in approximately 200 ms. Likelihoods are given for each individual task allowing a chemist to quickly determine which C-H functionalization reactions might proceed with high selectivity and to prioritize those targets or routes.

CINF 92

Molecular transformer for chemical reaction prediction and uncertainty estimation

Philippe Schwaller[1,2], *phischwaller@gmail.com, Teodoro Laino*[1], *Théophile Gaudin*[1,3], *Costas Bekas*[1], *Alpha A. Lee*[2]*. (1) IBM Research GmbH, Rueschlikon, Switzerland (2) University of Cambridge, Cambridge, United Kingdom (3) University of Toronto, Toronto, Ontario, Canada*

Organic synthesis is one of the key stumbling blocks in medicinal chemistry. The necessary yet unsolved step in planning synthesis is solving the forward problem: given reactants and reagents, predict the products. We treat reaction prediction as a machine translation problem between SMILES strings of reactants-reagents and the products. We show that a multi-head attention Molecular Transformer model outperforms all algorithms in the literature, achieving a top-1 accuracy above 90% on a common benchmark dataset. Our algorithm requires no handcrafted rules, and accurately predicts subtle chemical transformations. Crucially, our model can accurately estimate its own uncertainty, with an uncertainty score that is 89\% correct in terms of classifying whether a prediction is correct. Furthermore, we show that model is able to handle inputs without reactant-reagent split and including stereochemistry, which makes our method universally applicable across existing datasets.



CINF 93

Environmental chemical information in PubChem

Jian Zhang*, jiazhang@ncbi.nlm.nih.gov, Evan Bolton. National Institutes of Health, Bethesda, Maryland, United States*

Many environmental science data resources exist for chemical hazard and environment related information. PubChem integrates a number of these, including data from EPA, USGS, OSHA, CDC, ILO, and ECHA. The results is a public, open access platform that allows you to readily locate for chemicals information on the chemical and physical properties, toxicology, ecology, and safety and hazard information, among others. The presentation will give the overview of the environment related chemical data in PubChem, ways to access it, and other useful tips and tricks.

CINF 94

EPA CompTox chemicals dashboard: An online resource for environmental chemists

***Antony J. Williams**, tony27587@gmail.com, Christopher Grulke, Jeremy Dunne, Jeff Edwards. National Center for Computational Toxicology, Environmental Protection Agency, Wake Forest, North Carolina, United States*

The U.S. Environmental Protection Agency (EPA) Computational Toxicology Program integrates advances in biology, chemistry, and computer science to help prioritize chemicals for further research based on potential human health risks. This work involves computational and data driven approaches that integrate chemistry, exposure and biological data. As an outcome of these efforts the National Center for Computational Toxicology (NCCT) has measured, assembled and delivered an enormous quantity and diversity of data for the environmental sciences including high-throughput *in vitro* screening data, *in vivo* and functional use data, exposure models and chemical databases with associated properties. The CompTox Chemicals Dashboard is a freely accessible user-friendly web application that provides access to data associated with ~770,000 chemical substances. These data include human and ecological hazard data for tens of thousands of chemicals from multiple databases and peer-reviewed publications. Experimental and predicted physicochemical properties, product and functional use information and bioassay screening data associated with the ToxCast and Tox21 high throughput screening programs are also available. The application provides batch searching capability to source data for thousands of chemicals at a time as well as real-time QSAR predictions for almost two dozen physicochemical and toxicity endpoints. This presentation will provide an overview of the CompTox Chemicals Dashboard and its value to the community as an informational hub. *This abstract does not necessarily represent the views or policies of the U.S. Environmental Protection Agency.*

CINF 95

Mapping of chemical identifiers to DSSTox to enable data integration in the US-EPA CompTox Chemicals Dashboard

Christopher Grulke[1], *grulke.chris@epa.gov, Inthirany Thillainadarajah*[1], *Patience Browne*[2], *Antony J. Williams*[1], *Ann Richard*[1]. *(1) National Center for Computational Toxicology, Environmental Protection Agency, Wake Forest, North Carolina, United States (2) Organisation for Economic Co-operation and Development (OECD), Paris, France*

The Computational Toxicology Program within the U.S. Environmental Protection Agency (EPA) integrates advances in biology, chemistry, and computer science to help prioritize chemicals for further research based on potential human and environmental health risks. A key component of this integration effort is the mapping of chemical identifiers from a broad range of data sources to Distributed Structure-Searchable Toxicity (DSSTox) substances using our chemical list curation protocol (CLCP). Source identifiers typically consist of chemical names, sometimes with CAS-RN, and less often with SMILES or mol files. Structure-centric databases, such as PubChem and ChemSpider, use purely automated approaches to resolve Source substance identifiers (SSIDs) with the goal of mapping to unique chemical structures (CIDs). In contrast, the goal of the DSSTox CLCP is to accurately map SSIDs to a unique DSSTox substance (DTXSID) and, if possible, a unique structure (DTXCID). The CLCP uses a combination of automated mappings to DTXSIDs and expert manual curation review to resolve conflicts in list identifiers (e.g., the chemical common name maps to one DTXSID, CASRN maps to another). The CLCP has resulted in the mapping of nearly 350 data sources to DSSTox substances, and in the process identified and resolved conflicting source/structure information for the chemistry associated with each list member. Three specific applications of the CLCP will be described: (1) mapping of chemicals identified as “active” under the EPA Toxic Substances Control Act (TSCA), (2) collection and mapping of a set of endocrine reference chemicals, and (3) mapping chemical substances to animal toxicity values stored in our Toxicity Value database. Lists processed through the CLCP that are approved for public-release are published on the List page (https://comptox.epa.gov/dashboard/chemical_lists) of the CompTox Chemicals Dashboard and support modeling efforts within the National Center of Computational Toxicology. *This abstract does not reflect U.S. EPA or OECD policy.*

CINF 96

Consistency checking the experimental data available from the USEPA NCCT CompTox database

Stuart J. Chalk[1], *schalk@unf.edu, Antony J. Williams*[2], *Christopher Grulke*[2]. *(1) Department of Chemistry, University of North Florida, Jacksonville, Florida, United States (2) National Center for Computational Toxicology, USEPA, Research Triangle Park, North Carolina, United States*

The US EPA’s National Center for Computational Toxicology (NCCT) is focused on developing computational estimates of the toxicology of chemicals found in commerce

and in the environment. In order to deliver predictive models NCCT has measured, assembled and delivered an enormous quantity and diversity of data. This includes high-throughput *in vitro* screening data, *in vivo* and functional use data, as data delivered via the CompTox Chemicals Dashboard, a web application providing access to data associated with ~770,000 chemical substances. A subset of these data have been extracted and curated from sources including public and agency databases and scientific publications.

This presentation will present an evaluation of the consistency of the experimental data by conversion of the raw data into the JavaScript Object Notation for Linked Data
(JSON-LD) SciData format and ingestion of the JSON-LD files into a graph database as Resource Description Framework (RDF) triples. The graph database is then searched using SPARQL queries to identify inconsistencies that can then be reviewed and curated.

Conversion of the data was done semi-automatically using PHP and Python scripts to crosswalk the data into a MySQL database and subsequently exported as JSON-LD. As part of the process annotation of compounds in the dataset was augmented with classifications from the ChemOnt ontology. The results of this analysis, pain points encountered, and progress toward automating the workflow using KNIME will be presented. *This abstract does not necessarily represent the views or policies of the U.S. Environmental Protection Agency.*

CINF 97

Literature-based cheminformatics for research in chemical toxicity

Nancy C. Baker[1]*, baker.nancy@epa.gov, Antony J. Williams*[2]*, Thomas Knudsen*[2]*. (1) Leidos, Hillsborough, North Carolina, United States (2) EPA, RTP, North Carolina, United States*

Contained within the 28 million citations in PubMed is an abundance of information about the activity of chemicals in biological systems. At the EPA's National Center for Computational Toxicology we have implemented literature informatics approaches that make effective use of the literature in a variety of applications important to environmental science. For instance, we use text-mining methods to identify chemicals that may cause adverse effects like neurotoxicity or embryonic vascular disruption. Then we can take the resulting list of chemicals and, using the PubMed Abstract Sifter tool, we can explore the linkages and connections to key events in defined adverse outcome pathways (AOPs). Quantitative alterations in embryonic vascular disruption and dysmorphogenesis can be expanded into an AOP for developmental toxicity utilizing information extracted by the Abstract Sifter. The Abstract Sifter technology has been implemented in two forms. One version integrates the richness of PubMed with the flexible data-manipulation capabilities of Microsoft Excel. A web-based implementation

of the technology is available through the EPA CompTox Chemicals Dashboard [https://www.epa.gov/chemical-research/comptox-chemicals-dashboard]. These literature informatics applications help chemists and toxicologists take advantage of the immense knowledge recorded in the chemical literature for predictive toxicology. *This abstract does not necessarily represent U.S. EPA policy.*

CINF 98

Green chemistry and open data

***Jian Zhang**, jiazhang@ncbi.nlm.nih.gov, Evan Bolton. National Institutes of Health, Bethesda, Maryland, United States*

In modern society, chemical substances are used in many areas from laboratory to industry, from consumers to scientists. Safer chemicals or low toxicity chemicals are preferred when there is a choice. EPA and a few other organizations publish a limited amount of information related to safer chemical and green chemistry. PubChem, a public chemical data repository, provides a free, open platform to allow scientific organizations and companies to help share safer chemical information with the community. This talk will present the integrated safer chemical information in PubChem, and discuss the opportunity and challenge how a depositor can publish their green chemistry data in PubChem.

CINF 99

Development of the alternatives assessment dashboard webtool

*Leora Vegosen[2], **Todd Martin**[1], martin.todd@epa.gov. (1) NRMRL, US EPA, Cincinnati, Ohio, United States (2) ORISE, Cincinnati, Ohio, United States*

The goal of alternatives assessment is to identify safer alternatives for chemicals of concern. Comparative chemical hazard assessment is an important component of alternatives assessment. This project aimed to develop a webtool, the Alternatives Assessment (AA) Dashboard, which will display compiled chemical hazard data and enable users to readily compare alternatives on the Environmental Protection Agency
(EPA) Chemistry Dashboard (https://comptox.epa.gov/dashboard).

We obtained chemical hazard data from public online sources including chemical hazard lists, Globally Harmonized System (GHS) scores from several different countries, and a database of quantitative toxicity values. We also obtained predicted values based on quantitative structure activity relationship (QSAR) models using EPA's Toxicity Estimation Software Tool (TEST). We used Java programming to parse the data into a database and to generate hazard scores.

For each data source, we determined a chemical hazard score (Low, Medium, High, or Very High) for each of several hazard endpoints (such as acute toxicity, carcinogenicity, etc.). Our scoring criteria were based on the EPA's Design for the Environment Program (DfE) Alternatives Assessment Criteria for Hazard Evaluation. Other alternatives assessment methodologies such as GreenScreen have built on DfE. We compared advantages and disadvantages of a modified version of the trumping scheme used by GreenScreen and a weighted average method for combining scores from multiple sources into one chemical hazard score for each endpoint.

The beta version of the AA dashboard currently contains over 290,000 score records for more than 85,000 chemicals. This webtool and its underlying data can aid in the pre-prioritization of chemicals under the Toxic Substances Control Act (TSCA) as amended by the Frank R. Lautenberg Chemical Safety for the 21st Century Act (LCSA).

Comparison of alteratives

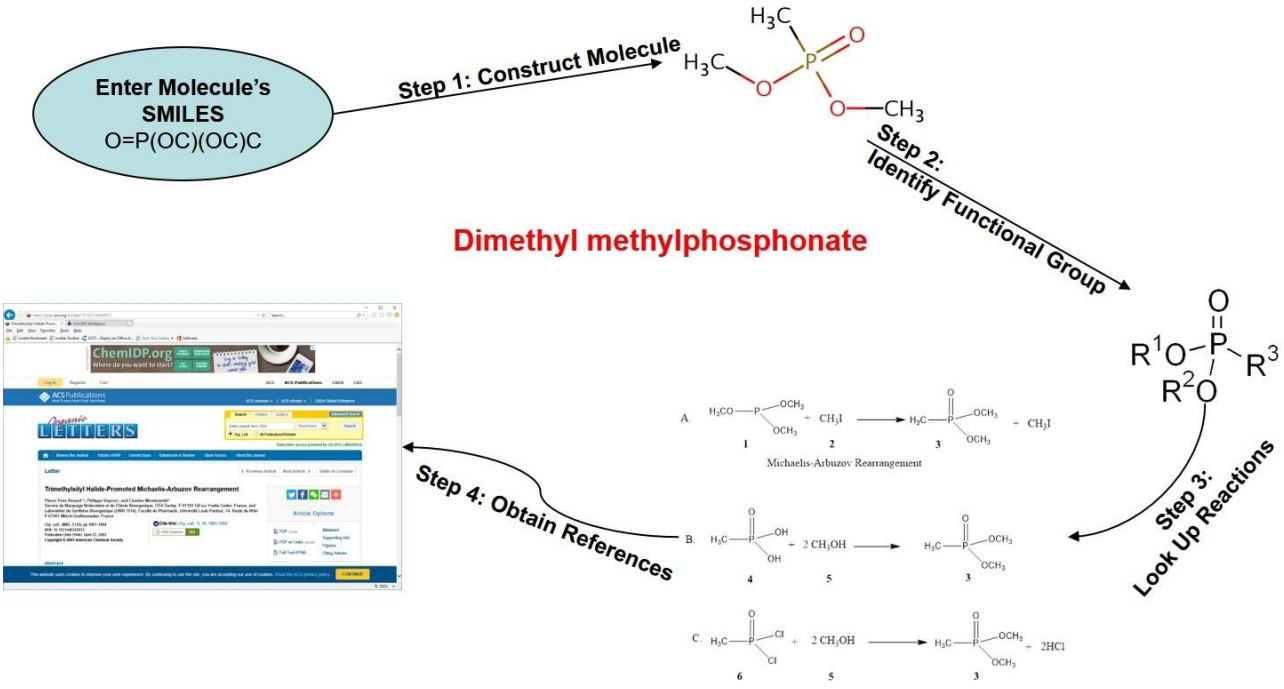
Structure CAS name	Human Health Effects															Ecotoxicity		Fate	
	Acute Mammalian Toxicity			Carcinogenicity	Genotoxicity Mutagenicity	Endocrine Disruption	Reproductive	Developmental	Neurotoxicity		Systemic Toxicity		Skin Sensitization	Skin Irritation	Eye Irritation	Acute Aquatic Toxicity	Chronic Aquatic Toxicity	Persistence	Bioaccumulation
	Oral	Inhalation	Dermal						Repeat Exposure	Single Exposure	Repeat Exposure	Single Exposure							
79-06-1 2-Propenamide	H	M	M	VH	VH	L	M	H	H	H	H	H	H	H	H	H	N/A	L	L
79-01-6 Ethene, trichloro-	L	M	L	VH	VH	N/A	H	H	H	H	H	M	H	H	H	M	M	H	L
108-95-2 Carbolic acid	H	H	H	H	H	H	H	H	H	H	M	H	H	VH	VH	H	L	L	L
50-00-0 Formaldehyde [USP]	H	H	H	VH	H	H	N/A	L	N/A	N/A	H	M	H	VH	VH	H	L	L	L
111-30-8 Glutaraldehyde, >1 - 2% in a non hazardous diluent	H	VH	H	M	H	H	L	L	N/A	H	H	M	H	VH	VH	VH	H	L	L
302-01-2 Hydrazine (see Hydrazine and Hydrazine Sulfate)	H	H	H	VH	VH	N/A	M	M	H	H	H	H	H	VH	VH	VH	VH	L	L
75-21-8 Ethylene Oxide	VH	H	N/A	VH	VH	H	H	H	H	H	H	M	H	H	H	M	L	H	L
7803-57-8 Hydrazine hydrate, or >37 - 64% aqueous solution	VH	VH	H	VH	VH	N/A	M	N/A	H	H	H	H	H	VH	VH	VH	VH	N/A	N/A
101-77-9 4,4'-Diaminodiphenylmethane	H	N/A	VH	VH	H	L	N/A	H	N/A	H	M	H	H	L	H	L	H	L	L
10588-01-9 Chromic acid (H2Cr2O7), disodium salt	H	VH	M	VH	VH	N/A	H	H	N/A	N/A	H	H	H	VH	VH	VH	VH	H	N/A
107-13-1 2-Propenenitrile	H	H	H	VH	H	L	H	H	H	H	H	M	H	H	VH	H	H	H	L
110-91-8 Morpholine, >10 - 25% in a non hazardous diluent	M	M	M	N/A	L	L	N/A	L	N/A	N/A	H	H	N/A	VH	VH	L	M	L	L

CINF 100

Application of chemical informatics to alternatives assessment

William Barrett[1], *barrett.williamm@epa.gov, Sudhakar R. Takkellapati*[1], *Kidus Tadele*[1,3], *Leora Vegosen*[3], *Michael A. Gonzalez*[2]*. (1) Land and Materials Management Division, U.S. Environmental Protection Agency, Cincinnati, Ohio, United States (2) NRMRL/STD/SAB, US Env Protection Agency, Cincinnati, Ohio, United States (3) Oak Ridge Institute for Science and Education, Oak Rdige, Tennessee, United States*

The evaluation of potential alternatives for chemicals of concern (CoC) requires an understanding of their potential human health and environmental impacts during the manufacture, use, recycle and disposal life stages. During the manufacturing phase, the processes used to produce a desired chemical are defined based on the sequence of chemical reactions and unit operations required to produce the molecule and separate it from other materials used or produced during its manufacture. This presentation introduces and demonstrates a tool that links a chemical's structure to information about its synthesis route and the manufacturing process for that chemical. The structure of the chemical is entered using either a SMILES string or the molecule MOL file, and the molecule is searched to identify functional groups present. Based on those functional groups present, the respective named reactions that can be used in its synthesis routes are identified. This information can be used to identify input and output materials for each named reaction, along with reaction conditions, solvents, and catalysts that participate in the reaction. Additionally, the reaction database contains links to internet references and appropriate reaction-specific keywords, further increasing its comprehensiveness. The tool can be developed to facilitate the use of data from the chemical literature and other related software applications can be developed that evaluate the reaction. Potential evaluations that could be performed include identification of alternate reaction routes, solvents, catalysts, reaction conditions and prediction of other reaction products. The chemical manufacturing processing steps can be linked to a chemical process ontology to estimate releases and exposures occurring during the manufacturing phase of a chemical. These types of evaluations would enable a more thorough assessment of chemical alternatives.

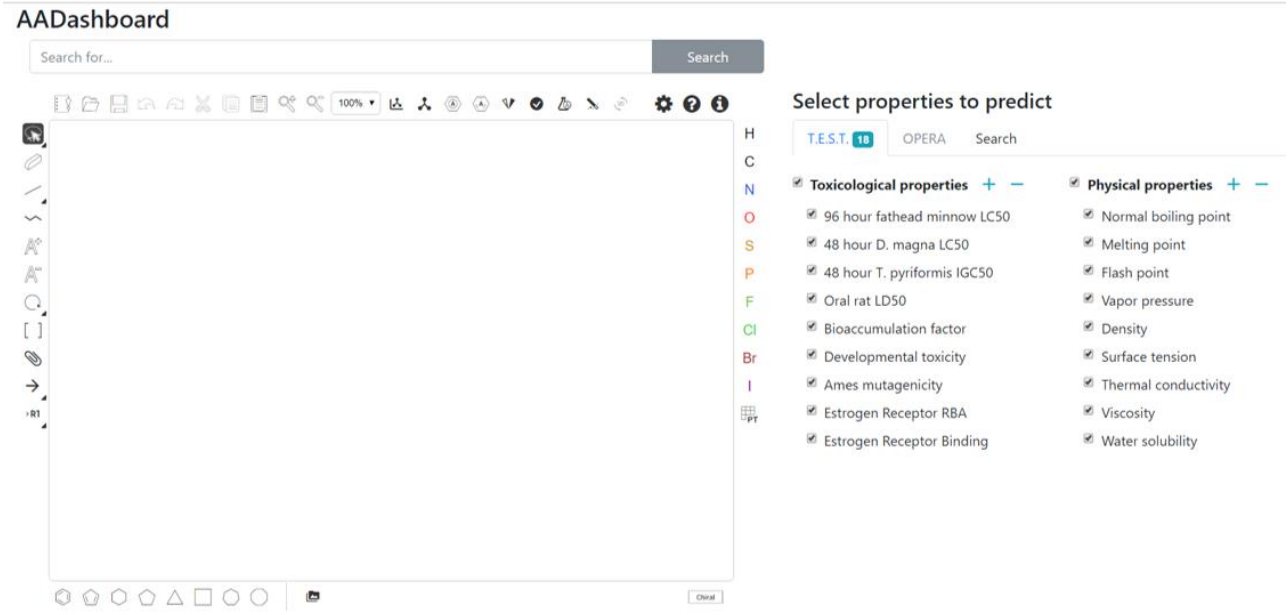


CINF 101

Prediction of toxicity using WebTEST (Web-services Toxicity Estimation Software Tool)

Todd Martin[1], *martin.todd@epa.gov, Antony J. Williams*[3], ***Valery tkachenko***[2], *tkachenko.valery@gmail.com. (1) NRMRL, US EPA, Cincinnati, Ohio, United States (2) Science Data Software, Rockville, Maryland, United States (3) NCCT, US EPA, RTP, North Carolina, United States*

A Java-based web service was developed within the US EPA's Chemistry Dashboard to provide real time estimates of toxicity values and physical properties. WebTEST can generate toxicity predictions directly from a simple URL which includes the endpoint, QSAR method, and SMILES string. An API is now available to allow prediction of toxicity from GET and POST commands. A web interface was created to allow users to make predictions for single chemicals (by drawing chemicals in a Ketcher chemical structure editor or by searching for chemicals in the US EPA's Chemistry Dashboard). A batch mode has been developed to enable prediction of multiple chemicals. Previously calculated results for over 700 thousand chemicals were stored in a database to improve response time. WebTEST now accesses the Chemical Transformation Simulator (CTS) web-service to predict environmental transformation products which will yield more comprehensive comparisons of chemical alternatives.



CINF 102

Case study in quantitative GenRA predictions using repeated dose toxicity studies from ToxRefDB v2.0

George Helman[1]*, helman.george@epa.gov, Grace Patlewicz*[2]*, Imran Shah*[2]*, Katie Paul Friedman*[2]*, Ly Pham*[1]*, Sean Watford*[1]*. (1) Oak Ridge Institute for Science and Education, Oak Ridge, Tennessee, United States (2) U.S Environmental Protection Agency, Durham, North Carolina, United States*

Computational approaches have recently gained popularity in the field of read-across to automatically fill data-gaps for untested chemicals. Previously, we developed the generalized read-across (GenRA) tool, which utilizes *in vitro* bioactivity data in conjunction with chemical descriptor information to derive local validity domains for selection of existing legacy data to qualitatively predict hazard based on study data summarized in US's EPA ToxRefDB v1.0. Following the success in qualitatively predicting hazard by study type and target organ or system, a clear need surfaced: quantitative predictions of the potency of possible hazard based on analogous chemicals. To address this need, we modified the GenRA workflow to make quantitative predictions for point of departure (POD) values obtained from ToxRefDB v2.0. To evaluate GenRA predictions, we aggregated oral Lowest Observed Adverse Effect Levels (LOAEL) for 1,049 chemicals by endpoint category, namely effects grouped as: systemic, developmental, reproductive, and/or cholinesterase-related. The mean LOAEL values for each chemical were converted to log10 molar equivalents. For defining chemicals with minimum similarity to the target chemical, Morgan chemical fingerprints

with a minimum Jaccard similarity threshold of 0.5 and a maximum of 10 nearest neighbors were used. The predicted systemic, developmental, reproductive, and cholinesterase LOAEL had R^2 (and RMSE) values of 0.38 (1.34), 0.22 (0.78), 0.14 (0.93), and 0.25 (0.7), respectively. These findings suggest that the GenRA workflow can be updated to include quantitative predictions, and that due to the relatively limited chemical set and hazard data available, it may be helpful to provide the user with the range of possible POD values for a given endpoint category for use in screening-level assessments.

CINF 103

Enhancement of acute toxicity prediction by multi-task learning

Sergey Sosnin[1]*, sergey.sosnin@skoltech.ru, Dmitry Karlov*[1]*, Igor V. Tetko*[2]*, Maxim V. Fedorov*[1]*. (1) Skolkovo Institute of Science and Technology, Moscow, Russian Federation (2) Research Center For Environmental Health (Gmbh), Institute Of Structural Biology, Neuherberg, Germany*

One of the biggest challenges in drug discovery is the prediction of acute toxicity of drug candidates. Approximately 30% of compounds do not pass the first stage of clinical trials, mostly because of the underestimation of the adverse events during pre-clinical stages. That fact motivates us to search the ways of enhancing the performance of existing methods. It should be noted that toxicity can be represented by different endpoints: it can be measured for different species using different types of administration, etc., and it is questionable if the knowledge transfer between endpoints is possible. In this study, we proposed a new multi-task deep neural network architecture and proved that our method overperforms both single-task DNN and other machine learning methods. To create a dataset we extracted acute toxicity data for more than 87 000 compounds belonging to 29 toxicity endpoints from the Registry of Toxic Effects of Chemical Substances. Each endpoint was characterized by several parameters: animal species, administration, and description of exposure. A number of sets of chemical descriptors have been calculated in the OCHEM platform and used as inputs for machine learning methods. We compared both multi-task deep neural networks (to predict a profile of toxicity for all endpoints simultaneously) and single-task machine learning methods: single-task neural networks, XGboost, Random Forest, k-nearest neighbors. Our experiments showed that multi-task deep neural networks had achieved remarkably better averaged quality (RMSE = 0.72) than both single-target neural networks (RMSE = 0.83) and other methods.

CINF 104

Prediction of toxicity: Deep learning with small and imbalanced datasets

Gerhard F. Ecker, *gerhard.f.ecker@univie.ac.at, Jennifer Hemmerich, Ece Asilar. Dept. Pharmaceutical Chemistry, University of Vienna, Wien, Austria*

Neural networks, used in deep learning, have become a popular tool for various prediction tasks. Whilst up to now traditional machine learning approaches are commonly used for activity predictions, the Merck Kaggle competition and the Tox21 Challenge were won by deep neural network architectures. The results of the two challenges indicate that deep learning could improve the predictivity, especially for complex endpoints such as hepatotoxicity.

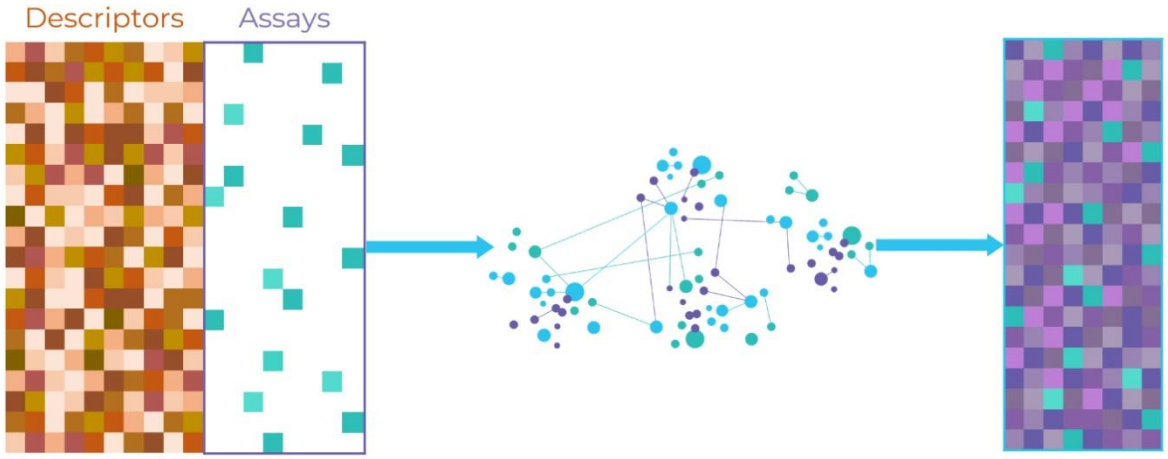
However, a major problem in this area is the small size and imbalancy of available toxicity datasets. In order to overcome overtraining and a bias towards the majority class, traditional machine learning approaches oversample the minority class (e.g. SMOTE) or apply cost sensitive learning. Other techniques used comprise bagging or boosting. For deep neural networks these techniques are rarely used as they increase the training time or lead to even smaller datasets. With the aim to solve both, imbalancy and insufficient size of the training set, we use conformational sampling along with 3D based description of the molecules. This allows to increase the size of the dataset without creating artificial samples. First results indicate that this is a versatile method for applying deep learning technologies on specific toxicity endpoints such as drug induced liver injury.

CINF 105

Imputing compound activities based on sparse and noisy data

*Thomas Whitehead[2], Ben Irwin[1], Peter A. Hunt[1], **Matthew D. Segall**[1], matthew.d.segall@gmail.com, Gareth Conduit[2,3]. (1) R&D, Optibrium Limited, Cambridgeshire, United Kingdom (2) Intellegens Limited, Cambridge, United Kingdom (3) Cavendish Laboratory, University of Cambridge, Cambridge, United Kingdom*

We describe a novel deep learning neural network method and its application to impute assay activity values. Unlike conventional machine learning approaches, this method can accept both sparse, noisy bioactivity data and molecular descriptors as input, enabling it to learn directly from correlations between activities measured in different assays as well as structure-activity relationships. Furthermore, the model provides a robust estimate of the confidence in each prediction, enabling attention to be focused on only the most accurate results. Using case studies on public domain data sets, we show that the neural network method outperforms traditional quantitative structure-activity relationship models, results which are improved further by focusing only on the most confident predictions.



CINF 106

Machine learning in the context of bioactivity

*Jochen Sieg, **Matthias Rarey**, rarey@zbh.uni-hamburg.de. University of Hamburg, Hamburg, Germany*

The search for bioactive compounds is the key step in early drug discovery. Among other techniques, the similarity principle (in the form of matched molecular pairs or free energy prediction), structure-based virtual screening, and of course experimental high-throughput screening are applied. In this talk, our results related to the use of deep learning and other machine learning techniques in these three design scenarios are summarized. How well does machine learning on matched molecular pairs affinity data perform? What signals do deep learning-based scoring functions for protein-ligand docking capture and how do they compare to classical ML methods? How can we make use of ML in the evaluation of experimental screening data? In summary, we will demonstrate that standard validation schemes on frequently used data sets like DUD-E are non-conclusive in many cases. It remains a challenge for the future to develop novel validation procedures and datasets to truly compare ML-based methods for bioactivity prediction.

CINF 107

ML and AI in the design of new drug lead compounds

***Shahar Keinan**, skeinan@gmail.com, William J. Shipman, Elizabeth H. Frush, Ed Addison. Cloud Pharmaceuticals, Durham, North Carolina, United States*

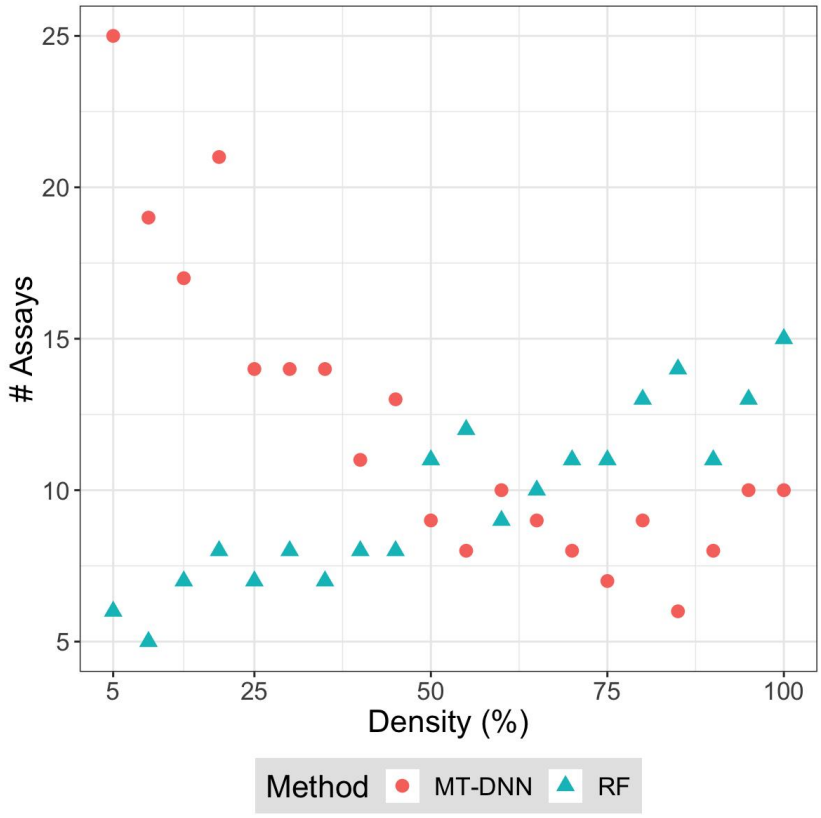
The hype surrounding the use of Machine Learning (ML) and Artificial Intelligence (AI) can be found in almost every field today. It is perhaps most prevalent in healthcare, and

in drug discovery specifically. But in the rush to deploy AI for drug discovery, we should not overlook the obvious. Unlike text and image analysis, where data sets are abundant, the available data sets in chemistry tend to be sparse and small. In order to successfully use ML in Drug Design, one must choose the right dataset to begin with. An example of such set, that has been tested and deployed with multiple ML algorithms, is the EPO TOX21 set. However, such datasets are rare. We suggest here the use of Augmented Intelligence, the application of AI methods, such as big data and ML, to enhance computational chemistry and other non-AI algorithms and information. We will demonstrate the benefits of using Augmented Intelligence for drug design on several examples. One such example is the prediction of solvation energy of different conformers. Conformers are vectorized by Coulomb Matrix and applying Deep Tensor Neural Network algorithms, resulting in a reduction of the computational cost of deploying quantum chemistry to decide which conformers are relevant.CINF 108

Influence of compound profiling matrix density on the performance of multi-task deep neural networks and random forest models

Raquel Rodríguez Pérez[2,1], *raquel.rp13@gmail.com, Jurgen Bajorath*[2]*. (1) Medicinal Chemistry, Boehringer Ingelheim, Biberach an der Riss, Germany (2) Life Science Informatics, University of Bonn, B-IT, Bonn, Germany*

Deep learning (DL) methods are experiencing increasing interest in many fields including drug discovery. For example, multi-task DL (MT-DL) is used for quantitative structure-activity relationships (QSARs) and virtual screening (VS) applications, among others. Although some promising results have been reported for DL and MT-DL, it remains to be determined in which application scenarios and under which conditions MT-DL becomes superior to standard ML algorithms such as support vector machine
(SVM) or random forest (RF).
We have investigated the performance of MT deep neural networks (MT-DNN) and RF for activity predictions using a large compound profiling matrix extracted from screening data. In particular, the influence of training data sparseness on model performance was systematically explored. Therefore, sub-matrices of varying density (5-100%) were generated and used to build MT-DNN and RF models. With such models, an independent test set consisting of complete compound activity profiles was predicted. When performances of RF and MT-DNN were averaged over all targets, MT-DNN yielded slightly superior results. However, when exploring the relative performance of the methods, there was no consistent advantage of MT-DNN over per-target RF models under varying density conditions. Only when training data was very sparse, MT-DNN models outperformed RF for individual assays (Figure 1). However, RF met or partly exceeded the prediction performance of MT-DNN at increasing density levels. For prediction tasks with sparse training data, MT-DNN should be preferentially explored.



Comparison of AUC performance for individual targets using MT-DNN and RF. Reported are the mean number of targets in which one method outperformed the other on the basis of AUC at varying matrix density.

CINF 109

Many possible roles of deep learning in drug discovery: Separating truth from hype

Robert Abel*, robert82a@gmail.com, Karl Leswing, Kyle Marshall, Joshua Staker, Carolyn McQuaw, Steven Jerome, Sayan Mondal, Sathesh Bhat. Schrödinger Inc., New York, New York, United States*

Deep learning methods have been successfully applied to multiple challenging problems including language translation, image processing, and voice recognition, among others. These successes have motivated extensive investigations to determine where application of deep learning could most benefit preclinical drug discovery. We will here report our own successful use of deep learning to identify hit molecules, build more predictive ADMET models, improve the efficiency of classical scoring techniques, identify suspicious experimental data, and expedite data extraction and curation from public sources. We will also comment more generally on the types of drug discovery problems likely to be a good fit for deep learning, and those problems likely outside the reach of the technique.

CINF 110

Industry perspective: Deep learning for QSAR models

Jie Shen, jshen@lilly.com. Eli Lilly and Company, Indianapolis, Indiana, United States

Deep learning has drawn a lot of attentions in different areas including drug discovery. In pharmaceutical industry, one of the impactful applications of deep learning is to predict the molecular activities, including physical chemical properties, potency, pharmacokinetics, and toxicity, using molecules' chemical structure. This is known as quantitative structure activity relationship (QSAR) modeling. Generally, a machine learning algorithm is used to model such relationship. Deep learning is expected to outperform other machine learning algorithms, especially with big data sets. This presentation summarized some of our recent works focusing on answering following questions:

1. What advantages deep learning can bring to us?
2. What is the best way for molecular structure representation?
3. How to conduct hyperparameter tuning for deep learning QSAR models?
4. How to improve the generalization ability for deep learning QSAR models?

In order to answer those questions, we have conducted systematic evaluations on 24 industry ADME data sets with over 50000 deep learning QSAR models. We have also explored several ways of building deep learning based QSAR models with large industry ADME datasets, including adding different types of noises into the data and model, using multi-task learning and transfer learning. Based on the data and analysis, we have proposed a best practice of building deep learning QSAR models.

CINF 111

OPERA models for physicochemical properties, environmental fate and toxicological endpoints to support regulatory purposes

Kamel Mansouri[1], 30752563@acs.org, Richard Judson[2], Antony J. Williams[2], Nicole Kleinstreuer[3]. (1) Integrated Laboratory Systems Inc, Morrisville, North Carolina, United States (2) USEPA, Rtp, North Carolina, United States (3) NICEATM/NIEHS, Research Triangle Park, North Carolina, United States

The increasing number of publicly available databases with large amounts of data is facilitating computational modeling efforts in different fields. However, for QSAR/QSPR modeling, the quality of experimental data and chemical structure information remains a big challenge. In order to build a suite of robust predictive QSAR/QSPR models to support regulatory purposes, all collected datasets for the open structure-activity relationship application (OPERA) project have undergone extensive curation procedures.

Automated workflows were designed and applied to select only high-quality data, and associated chemical structures were standardized to generate QSAR-ready forms prior to modeling. The five OECD principles for QSAR modeling were followed to provide scientifically valid, high accuracy models with minimum complexity that yield mechanistic interpretation, when possible. Current OPERA models cover a wide variety of physicochemical properties, environmental fate and toxicity endpoints including logP, water solubility, pKa, plasma-protein binding, endocrine disrupting activity, and acute oral toxicity. Technical and performance details are described in OECD-compliant QSAR model reporting format reports (QMRF reports). OPERA is constantly updated with new data and models and can be used via the open source standalone application on GitHub (https://github.com/kmansouri/OPERA) and its predictions can be accessed through the EPA's CompTox Chemicals Dashboard (https://comptox.epa.gov/dashboard) and NICEATM's Integrated Chemical Environment (https://ice.ntp.niehs.nih.gov/). *This abstract does not necessarily represent the views or policies of any federal agency.*

CINF 112

Applications of a chemotype-enrichment approach to the ToxCast data landscape and beyond: Inverting the SAR paradigm

Ann Richard[1], richard.ann@epa.gov, Ryan Lougee[2], Christopher Grulke[1], Nancy C. Baker[3], Jun Wang[4], Antony J. Williams[1]. (1) National Center for Computational Toxicology, US EPA, Research Triangle Park, North Carolina, United States (2) National Center for Computational Toxicology, US EPA ORISE Student Trainee, Research Triangle Park, North Carolina, United States (3) Leidos, Hillsborough, North Carolina, United States (4) National Health & Environmental Effects Research Laboratory, US EPA ORISE Post Doctoral Fellow, Research Triangle Park, North Carolina, United States

Traditional structure-activity relationship (SAR) approaches to modeling in toxicology are bounded and determined by the chemicals in the tested space, and are limited by such factors as size and structure diversity of that tested space, mechanistic diversity of the activity endpoint, and percentage of actives. A chemotype (CT)-enrichment approach inverts this paradigm and creates a static chemical abstraction layer consisting of structural features that define local chemistry domains upon which diverse test sets and biological outcomes (or binary categorical assignments of any sort) can be projected. A recently developed Chemotype-Enrichment Workflow (CTEW), based on a publicly available set of ToxPrint chemical features, has been applied to a broad range of datasets spanning EPA's Computational Toxicology research programs. The results

have facilitated a global view of ToxCast high-throughput screening results for over 1000 assays as projected by enrichments within ToxPrint chemistry domains. From the vantage point of defined chemistry, various patterns are revealed, including the degree to which CT-biological activity signals are present in datasets not amenable to global SAR analysis, areas of CT-assay promiscuity, and areas of CT inactivity and possibly related analytical QC fails. Interesting patterns in CT enrichments can also be observed upon "tuning" of the biological signal within time-series for a given assay, and upon tuning of the CTs from more to less specific representations. CTEW analyses of a sample of individual bioassay datasets will be shown to illustrate the power of the approach to probe chemical-bioactivity patterns. *This abstract does not necessarily represent the views or policies of the U.S. Environmental Protection Agency.*

CINF 113

Framing chemical safety and risk management: Ontological perspectives from laboratory procedures and incident reports

*Cogan M. Shimizu[2], **Leah R. McEwen**[1], lrm1@cornell.edu. (1) Clark Library, Cornell University, Ithaca, New York, United States (2) Wright State University, Dayton, Ohio, United States*

Chemical safety is the practice of assessing risk from chemical hazards and applying management strategies. In laboratory practice this involves recognizing hazards and assessing risk associated with experimental procedures. Exploring the components of laboratory procedures can help frame when hazards might arise, what factors could be involved and potential impact of any adjustments. Several analog methods exist for evaluating chemical hazards that can be mapped into models for chemical procedures to help facilitate the process of risk assessment. Building on previous ontology pattern (ODP) design work on semantic trajectories, a pattern module (tailored to a domain and use-case) is proposed for chemical laboratory procedures. Current work is exploring the relationship of the laboratory level at which the researcher operates with the molecular level of chemical reactivity that can lead to hazardous situations. This pattern work can be incorporated into collections of chemical data and procedural information, including electronic laboratory notebooks, to support risk assessment and safety documentation.

CINF 114

Evaluation of the chemotype-enrichment workflow as a tool for independent evaluation biological activity thresholds and a comparison with traditional QSAR methods

***Ryan Lougee**[1], rrlougee@ncsu.edu, Ann Richard[2], Christopher Grulke[3]. (1) NCCT, US Environmental Protection Agency, Durham, North Carolina, United States (2) MD 205-*

01, US EPA, Research Triangle Park, North Carolina, United States (3) Zachary Piper Solutions, New Hill, North Carolina, United States

The Chemotype-Enrichment Workflow (CTEW), an automated univariate analysis tool, was developed to explore activity enrichments within local chemistry domains associated with EPA's ToxCast bioassay results. The current implementation of the CTEW employs the publicly available ToxPrint chemotype (CT) set to define a fixed chemical abstraction layer for use in exploring individual assay datasets, as well as for comparing enrichment results globally, across assay endpoints and datasets. For each assay, the CTEW processes an input file consisting of a list of structures and binary hit-calls [1,0], generates CT fingerprints, and applies a set of defined statistical thresholds to identify CT enrichments in positive [1] and negative [0] activity space. ToxPrints offer a set of visualizable chemical features that can facilitate interpretation of results and support hypothesis generation. CTEW enrichment results can be considered an "attribute" of each assay dataset(s) and, as such, can be used to probe biological activity threshold assumptions (usually expert-based) that define the associated hit-calls. In addition, CTEW results can identify local activity enrichments in small, noisy, or imbalanced datasets that are less amenable to global Quantitative Structure-Activity Relationship (QSAR) analysis. In our first case study, we employed the CTEW to examine the overall impact on CTEW enrichment results when analyzing ToxCast assay results with and without application of a biological cytotoxicity filter. Cross validation is used to evaluate the cytoxicity filter and non-filtered results across the ToxCast Assays, and examine their activity against a random sample. Secondly, to bridge our understanding of the benefits and limitations of the CTEW approach in relation to global QSAR modeling, we compared balanced accuracy statistics of the full set of enriched CTs for each ToxCast assay dataset with the results from gradient-boosted random forest QSAR models applied to the same ToxCast assay datasets. *This abstract does not reflect U.S. EPA policy.*

CINF 115

Case study in quantitative GenRA predictions using acute oral toxicity

George Helman[2]*, helman.george@epa.gov, Imran Shah*[1]*, Grace Patlewicz*[1]*. (1) U.S. Environmental Protection Agency, Durham, North Carolina, United States (2) Oak Ridge Institute for Science and Education, Oak Ridge, Tennessee, United States*

Computational approaches have recently gained a lot of traction in the field of read-across. Previously we developed our generalized read-across (GenRA) approach, which utilizes in vitro bioactivity data in conjunction with chemical descriptor information to derive local validity domains to facilitate read-across prediction. Predictions are made using the similarity weighted average of nearest neighbors. GenRA has been used to predict up to 574 different apical effects as a binary call from repeat-dose toxicity studies available in ToxRefDB. Here we evaluate the application of GenRA for

quantitative predictions, specifically using a large dataset of rat oral acute LD50 toxicity data. These data were constructed by mining and merging a number of publicly available resources and comprise 21,210 unique LD50 values for 15,698 discrete chemicals. We used 9293 chemicals for which structures were available in the US EPA CompTox Chemicals Dashboard to make GenRA predictions of LD50 values. GenRA predictions used Morgan chemical fingerprints with a minimum Jaccard similarity threshold of 0.5 and a maximum of 10 nearest neighbors over the entire dataset. We evaluated confidence in GenRA predictions using four-fold cross-validation testing producing r^2 values 0.56±0.3. These findings show the utility of GenRA for making quantitative predictions of toxicity.

CINF 116

Comprehensive computational approach for predicting human skin sensitization as suggested alternative to animal testing

Eugene Muratov[3], murik@email.unc.edu, Vinicius M. Alves[3], Joyce Borba[2], Rodolpho Braga[1], Stephen Capuzzi[3], Arhur C. Silva[1], Nicole Kleinstreuer[4], Carolina H. Andrade[1], Alexander Tropsha[3]. (1) Faculty of Pharmacy, Federal University of Goias, Goiania, Goias, Brazil (2) Federal University of Goias, Chapel Hill, North Carolina, United States (3) Chemical Biology & Medicinal Chemistry, University of North Carolina - Chapel Hill, Chapel Hill, North Carolina, United States (4) NIEHS, RTP, North Carolina, United States

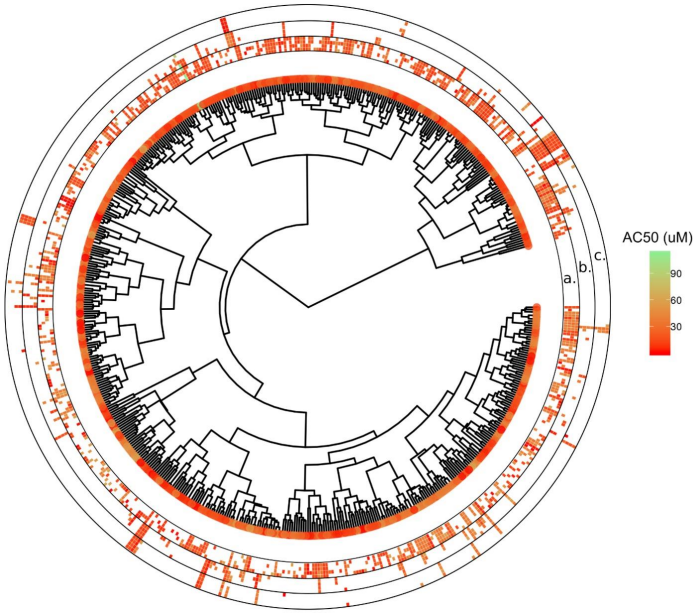
Traditionally, the skin sensitization potential of chemicals has been assessed using animal models. However, the preferred animal testing for chemical safety assessment has shown major drawbacks, such as poor correlation with human response, high cost and slow outcome rate. Moreover, due to growing ethical, political, and financial concerns, sustainable alternatives to animal testing need to be developed. As publicly available skin sensitization data continues to grow, computational approaches are expected to reduce or replace animal testing for the prediction of human skin sensitization potential. Here, we report on new comprehensive computational approach for predicting human skin sensitization that integrates multiple Quantitative Structure-Activity Relationship (QSAR) models developed on *in vitro*, *in chemico*, animal, and human data, into a final Naïve Bayes model. QSAR models were generated using human repeat insult patch tests (HRIPTs), human maximization tests (HMTs), as well as mouse Local Lymph Node Assay (LLNA) data for skin sensitization. In addition, we also include data for three validated alternative methods, named Direct Peptide Reactivity Assay (DPRA), KeratinoSens, and the human Cell Line Activation Test (h-CLAT). Models were developed and rigorously validated according to the best practices of QSAR modeling using open-source tools. Predictions obtained from these models were then used to build a Naïve Bayes model for predicting human skin sensitization with high externally estimated predictivity: balanced accuracy (89%), sensitivity (94%), positive predicted value (91%), specificity (84%), and negative predicted value (89%). All the developed models could be accessed at Pred-Skin 2.0 Web-portal at: http://5.189.141.142:5002/.

CINF 117

Predicting chemical-assay interference using Tox21 qHTS data

Alexandre Borrel[1], alexandre.borrel@univ-paris-diderot.fr, Ruili Huang[4], Menghang Xia[4], Keith Houck[2], Richard Judson[3], Nicole Kleinstreuer[1]. (1) Biostatistics and Computational Biology Branch, NIEHS, Morrisville, North Carolina, United States (2) MC D343-03, US Environmental Protection Agency, Research Triangle Park, North Carolina, United States (3) USEPA, Rtp, North Carolina, United States (4) NCATS, NIH, Bethesda, Maryland, United States

The federal Tox21 consortium produces quantitative high-throughput screening (qHTS) data on thousands of chemicals across a wide range of assays covering critical biological targets. Many of these assays, and those used in other in vitro screening programs, rely on luciferase and fluorescence-based readouts which can be susceptible to signal interference by certain chemical structures resulting in false positive outcomes. Included in the Tox21 portfolio are assays specifically designed to measure interference in the form of luciferase inhibition and autofluorescence via multiple wavelengths (red, blue, and green) and under various conditions (cell-free and cell-based, two cell types). Out of 8,305 unique chemicals tested in the Tox21 interference assays, percent actives ranged from 0.5% (red autofluorescence) to 9.9% (luciferase inhibition) after filtering for curve class, efficacy, and cytotoxicity cutoffs. Bimodal potency distributions were observed among active chemicals, potentially corresponding to specific and non-specific activity. Self-organizing maps and hierarchical clustering were used to relate chemical structural clusters to interference activity profiles. Multiple machine learning algorithms were applied to predict assay interference based on molecular descriptors and chemical properties. The best performing predictive models (accuracies of ~80%) are being included in a web-based tool that will allow users to predict the likelihood of assay interference for any new chemical structure. *This abstract does not reflect official EPA or NIH policy.*



Hierarchical clustering of active chemicals on autofluorescence assays for a. blue, b. green and c. red wavelengths. Hierarchical clustering is realized using a Euclidean distance computed from a set of non-redundant molecular descriptor (165) and using the Ward method. AC50 for each chemical is represented in mM using the color scale shown.

CINF 118

Methods for *in silico* screening of use and exposure data in authority databases

***Stellan Fischer**, fischer.stellan@gmail.com. Swedish Chemicals Agency, Stockholm, Sweden*

High throughput screening of chemical exposure data need access to formalised information of use. Important sources of information for chemical use are the national authorities. The accessibility of the data in public is however limited. The Swedish Chemicals Agency has therefore developed methods to facilitate exposure of confidential data on chemical use from regulatory databases into formalised exposure estimations. Companies putting chemicals on the Swedish market has to notify the Swedish Chemicals Agency. Algorithms where developed to automatically transform the registered use information, which is classified information, into an exposure index, representing non-confidential information on release and exposure estimations of chemicals. The main purpose with developing this exposure index was as a supporting tool for prioritization of authority work. The index concept was later on adopted by other European countries keeping similar national product registers.

These declassified exposure estimates are now annually published on internet. The exposure index tool has also recently been found useful in collaboration with analytical laboratories in non-target or suspect screening. A national market list (incl. exposure estimates) of chemicals was developed for scientists to adapt in their development of suspect lists for matching against their generated MS data. These findings indicate that close collaboration between scientists and regulatory authorities is a promising way forward for enhancing identification of emerging pollutants in non-target screening. Further, an extended version of the Swedish *market list* including the entire EU market is now under development within the scientific network NORMAN. The purpose is to support the identification work of tentative candidates in LC-HRMS in suspect screening. It is now currently being tested in several collaboration projects within EU. In addition, another less explored data source for estimating chemical exposure are official patent databases. All patents have to be categorized according to a harmonized technical categorisation system. These categories can be used for exposure estimations. Pilot studies have recently been conducted to explore this field.

CINF 119

Novel nanodescriptors applied in QNAR: Combination of virtual nanomaterial library and geometrical structure of nanomaterial

Xiliang Yan[1,2]*, yanxiliang1991@gmail.com, Alexander Sedykh*[3]*, Wenyi Wang*[2]*, Bing Yan*[4]*, Hao Zhu*[2,5]*. (1) School of Chemistry and Chemical Engineering, Shandong University, Jinan, Shandong, China (2) Center of Computational and Integrative Biology, Rutgers University, Camden, New Jersey, United States (3) Sciome, Durham, North Carolina, United States (4) School of Environmental Science and Engineering, Shandong University, Jinan, Shandong, China (5) Chemistry Department, Rutgers Univesity, Camden, New Jersey, United States*

Computational modeling has been widely used for decades to predict bioactivities of small molecules in the drug discovery procedure. However, the applicability of computational modeling to deal with nanomaterials (NMs) is limited due to the complexity of NM structures. The lack of suitable descriptors, which can correctly quantify nanostructures and represent the nanostructural diversity, prevents the development of predictive computational models for NMs. To address this challenge, a computational workflow was established in this study to 1) virtually construct diverse gold nanoparticle (GNP) libraries; 2) develop predictive Quantitative Nanostructure Activity Relationship (QNAR) models based on newly designed nanodescriptors; and 3) validate the universal applicability of these new nanodescriptors.

The core of this modeling procedure is a novel algorithm to employ geometrical chemical descriptors to quantify nanostructures based on constructed virtual gold nanoparticles (vGNPs). The applicability, validity and feasibility of this method were proofed by modeling six in-house GNP datasets and one external dataset, which consist of different nanostructures and were experimentally tested for various nano-bioactivities (i.e., GNP-enzyme interactions, nano cellular uptake and nano-induced reactive oxygen species (ROS)) and physicochemical properties (i.e., logP and zeta potential), respectively. Associated vGNP libraries were first developed based on these GNP datasets. After generating nanodescriptors from all vGNPs, QNAR models were constructed using various machine learning approaches for all the modeling sets. The high predictabilities of all developed QNAR models suggest this workflow can be used as a universal tool for nanomaterial modeling and virtual screening purpose.

CINF 120

Reaction library for predicting direct phototransformation products of aquatic organic contaminants

Chenyi Yuan[1], yuan.211@osu.edu, Caroline T. stevens.caroline@epa.gov[2], Eric J. Weber[2]. (1) Oak Ridge Institute for Science and Education (ORISE), Athens, Georgia, United States (2) US EPA, Athens, Georgia, United States

Cheminformatics-based applications have been actively developed to predict transformation pathways of environmental contaminants, especially biotransformation pathways. Direct photolysis, through which sunlight-absorbing compounds undergo transformation, is an important degradation pathway in aquatic systems, but lacks publicly available predictive tools. We systematically reviewed more than 600 journal publications and 150 regulatory reports and encoded the identified reaction sites and about 140 pathways into a reaction library using a cheminformatics-based software platform. The execution of this reaction library has been internally evaluated to predict plausible primary phototransformation products for diverse contaminants such as pesticides, pharmaceuticals, and energetic compounds. The reaction pathways are tentatively prioritized based on product formation rates wherever information is available. This reaction library will be the first publicly available one of its kind through execution of the Chemical Transformation Simulator. Special caution needs to be taken while using the library since photochemical reactivity can change dramatically with slight modification of the compound structure.

CINF 121

Cheminformatics and non-targeted analysis: A two-way street

Elin M. Ulrich[1], *ulrich.elin@epa.gov, Jon Sobus*[1], *Seth Newton*[1], *Christopher Grulke*[2], *Ann Richard*[2], *Randolph Singh*[4], *Andrew McEachran*[3], *Katherine Phillips*[1], *Kamel Mansouri*[3], *John Wambaugh*[2], *Kristin K. Isaacs*[1], *Antony J. Williams*[2]. *(1) National Exposure Research Laboratory, US Environmental Protection Agency, Research Triangle Park, North Carolina, United States (2) National Center for Computational Toxicology, US Environmental Protection Agency, RTP, North Carolina, United States (3) National Center for Computational toxicology, US EPA / ORISE, Research Triangle Park, North Carolina, United States (4) US EPA/NERL, Oak Ridge Institute for Science and Education, RTP, North Carolina, United States*

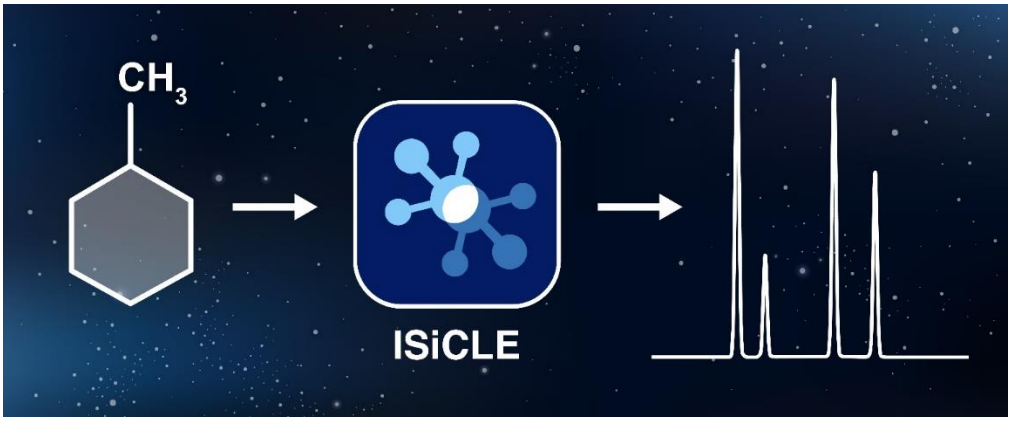
Non-targeted analysis (NTA) methods use high-resolution mass spectrometry to better understand the identity of a wide variety of chemicals present in environmental samples (and other matrices). NTA methods are considered a qualitative analysis type, forgoing chemical standards (often because they are not available), and producing semi-quantitative concentration data using surrogate calibration techniques. Data collection for NTA experiments focuses on being as inclusive as possible without having a preconceived notion as to what chemicals would be of interest and allows retrospective queries of the data. Analysis of the resultant mass spectrometry information: 1) relies on cheminformatics to identify and rank chemicals, and 2) yields data that are used by cheminformaticians and decision makers to address environmental problems. In the first case, the CompTox Chemicals Dashboard is used to propose and select chemical structures supported by mass spectrometry data, either by single substances or batch searching using accurate mass/chemical formula. In the second case, NTA experiments produce lists of chemicals detected by a specific analysis type and suspected of being present in a particular environmental matrix. Those NTA data are then used to generate models to predict chromatography or ionization conditions likely to detect a particular molecule or the probability that similar compounds would be present in the same media. One source of such data is the EPA's Non-Targeted Analysis Collaborative Trial (ENTACT), which involves the blinded and unblinded analysis of nearly 1,300 unique compounds combined into ten synthetic mixtures. The chemicals selected for ENTACT were originally procured for high throughput toxicity testing in the EPA's ToxCast program. The wealth of information obtained from the analysis of the mixtures by at least sixteen laboratories is staggering. This information will be compiled into a publicly available database and will serve to improve our understanding of previously collected toxicity information, assist in optimizing analytical methods, and help predict the behavior of substances within the same chemical space We will report on progress in delivering the database for community access as well as on work to expand cheminformatics support for NTA such as providing tools and data to allow comparison of computationally generated fragmentation spectra for over 800,000 chemicals to experimentally acquired spectra.

CINF 122

Elucidation of chemical dark matter using 'standards-free' small molecule identification

Ryan Renslow[1,2], rrenslow@gmail.com, Sean Colby[1], Dennis Thomas[1], Jamie Nunez[1], Yasemin Yesiltepe[2], Niranjan Govind[1], John R. Cort[1], Justin Teeguarden[1]. (1) Pacific Northwest National Laboratory, Richland, Washington, United States (2) The Gene and Linda Voiland School of Chemical Engineering and Bioengineering, Washington Statue University, Pullman, Washington, United States

Through innovations in computational quantum chemistry, we have developed a platform to overcome a significant obstacle in the field of metabolomics: the absence of methods for accurate, comprehensive identification of small molecules without relying on authentic chemical standards. Our pipeline, the *in silico* chemical library engine (ISiCLE), uses a large-scale computational quantum chemistry platform for calculating chemical properties, which exploits PNNL's high-performance computational chemistry software, NWChem. We investigated positional and geometric metabolite isomers, analyzed using ion mobility spectrometry-mass spectrometry, and found that ISiCLE was significantly more accurate (errors < 1-2%) at calculating CCS values compared to other methods, in part due to Boltzmann weighting of hundreds of candidate conformers by relative energy. This level of accuracy enabled us to even distinguish *cis*/*trans* isomers. For novel molecule structure elucidation, ISiCLE employs density functional theory (DFT) techniques to calculate NMR chemical shifts of molecule sets, with custom options for different solvents, nuclei, and user-selected chemical shift reference compounds. NMR chemical shift predictions were validated with experimental data from 300 molecules available in the literature. ^{1}H and ^{13}C chemical shifts were calculated with eight levels of DFT theory, with RMSD errors reaching 0.8 ppm and 5 ppm, respectively. Furthermore, we tested ISiCLE on conformers obtained using DFT-based *ab initio* molecular dynamics, demonstrating the ability to reduce chemical shift errors to less than 0.1 ppm (^{1}H) and 2 ppm (^{13}C) using Boltzmann weighting of calculations for hundreds of conformers. Here, we discuss several additional applications of our "standards-free" metabolomics approach, including the identification of environmental degradation products, the separation of molecular isomers, the characterization of complex blinded mixtures, and the correction of mislabeled isobaric isosteres, all without relying on authentic standards.


CH3
ISiCLE