Letter from the Editor

Thanks for reading the spring 2018 ACS Chemical Information Bulletin (CIB). And a special thank you to all of our authors, editors, and sponsors for contributing to the CIB.

In this issue we have several feature articles including a book review by Bob Buntrock of The Publish or Perish Book series by Anne-Wil Harzing. Bob also wrote a nice tribute to Eugene Garfield’s accomplishments and his personal experiences with Dr. Garfield. Wendy Warr continues her Twenty-five Years Ago piece with an account of the spring 1993 ACS Meeting in Denver, CO. In preparation for the upcoming 2018 spring ACS Meeting in New Orleans, Wendy also included a brief introduction of CINF activity 15 years ago in the 2003 spring ACS Meeting in New Orleans. As I write this, Wendy just sent an e-mail to the CHMINF-L discussion list noting that this year is New Orleans’ 300th birthday. What a great year indeed to meet in New Orleans for the spring 2018 ACS meeting. Our last feature article is a member profile of CINF member Rajarshi Guha.

We decided to bring back the CINF technical program and abstracts in the CIB for New Orleans (at least for now anyway). ACS has moved to a new ACS Online Planner interface for the technical program, and as a result, we felt that printing the program in the CIB could ease the transition for our members. Please note, however, that the program printed in the CIB was generated in late January. For the latest updates, we strongly encourage you to use the official ACS Online Planner.

This Issue of the CIB will be my eighth edited issue over the past six years, and as such, it is likely time to pass editorship of the spring CIB to a new leader. The CIB is an important chemical information publication and it is a huge honor to have served as an editor. Incidentally, the CIB was first published in 1949 as Chemical Literature and all Issues of Chemical Literature and the Chemical Information Bulletin are archived and openly available at The University of North Texas Digital Library. Collections of these publications can be readily found by a title search of “Chemical Literature” or “Chemical Information Bulletin.”

If you are interested in becoming the new spring CIB editor, let me know! I would be happy to work with you to ensure a smooth transition. For my own future, I look forward to taking on new roles in the Division of Chemical Information.

I hope to see many of you in New Orleans!

Vincent F. Scalfani, Editor
The University of Alabama
vfscalfani@ua.edu
CINF Social Networking Events at the Spring 2018 ACS Meeting

The ACS Division of Chemical Information is pleased to host the following social networking events at the Spring 2018 ACS National Meeting in New Orleans, LA.

**Sunday Welcoming Reception & Scholarships for Scientific Excellence Awards**
6:30-8:30 pm, Sunday, March 18th – Lucy's Retired Surf Bar, 701 Tchoupitoulas Street.
Sponsored exclusively by: Clarivate Analytics

Scholarships for Scientific Excellence
Sponsored exclusively by: ACS Publications

**Tuesday Luncheon** (Ticketed Event – Contact Division Chair, Erin Davis)
12:00-1:30 pm Tuesday, March 20th – Camillia/Gardenia Room, Hilton Garden Inn Convention Center.
Sponsored exclusively by: Royal Society of Chemistry

**Speaker:** John H Pardue
Louisiana State University

**Presentation:** TBD
CINF Business Meetings

Saturday, March 17: 12:30-2:30 PM

- Awards Committee: New Orleans Marriott Convention Center, Fleur De Lis
- Education Committee: New Orleans Marriott Convention Center, Mississippi Queen
- Program Committee: New Orleans Marriott Convention Center, Blaine Kern E/F

Saturday, March 17: 3:00-6:00 PM

- Executive Committee: New Orleans Marriott Convention Center, Blaine Kern E/F

Sunday, March 18: 12:00-2:00 PM

- Chemical Structure Association Trust Meeting: Ernest N. Morial Convention Center, Room 336
Chemical Information Program and Meeting Announcements

There are several symposia and meetings of interest to the chemical information community that may or may not be highlighted in the ACS Online Planner:

Sunday, March 18, 2018

CINF 19: Sharing chemical structures: workflows, demonstrations, and discussion.
A variety of chemical information professionals will take place in a lightning style chemical structure sharing demonstration session in CINF 19: Sharing chemical structures: workflows, demonstrations, and discussion.

Session: CINF: Enhance Discovery: Share Chemical Structures
(Organizers: Leah McEwen, Ye Li, and Vincent Scalfani)
Location: River Bend 1, New Orleans Marriott Convention Center
Date & Time: Sunday, March 18 3:20 PM
Duration: 1 hour 45 minutes

Abstract: This session will highlight and demonstrate various workflows for sharing chemical structures as machine readable files and/or depositing chemical structures into databases. We will explore a variety of software, web services, databases, and strategies that help researchers create, organize, store, and share chemical structures. Presenters from the Enhance Discovery: Share Chemical Structures session are invited to demonstrate their own chemical structure creation, sharing, standardization, and/or validation workflows. Attendees are encouraged to bring a laptop or tablet if seeking to participate interactively (when possible). We will conclude with a discussion and identify chemical structure sharing opportunities for the future.

Demonstration speakers will include:

Antony J. Williams and Christopher Grulke
National Center for Computational Toxicology, Environmental Protection Agency, Wake Forest, NC, United States.

Emma Schymanski
Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Belvaux, Luxembourg.

Aurora Costache, Eufrozina A. Hoffmann, András Volford

Vincent F. Scalfani
University Libraries, University of Alabama, Tuscaloosa, AL, United States.
Jian Zhang  
NCBI /NLM/NIH, Bethesda, MD, United States.

Guy Jones  
Royal Society of Chemistry, Cambridge, United Kingdom.

Suzanna Ward  
CCDC, Cambridge, United Kingdom.

Gregory M. Banik  
Bio-Rad Informatics, Philadelphia, PA, United States.

Discussion: DIGChem: FAIR Exchange of Chemical Data

Following the demonstration session, in the same room will be an open meeting with DIGChem: the Data Interest Group for Chemistry to highlight some opportunities to participate in FAIR chemical data exchange initiatives.

Location: River Bend 1, New Orleans Marriott Convention Center  
Date & Time: Sunday, March 18 5:05 PM  
Duration: 55 minutes  
Organizers: Stuart Chalk, Leah McEwen

Description: DIG Chemistry is a global conversation working across organizations to enable accurate exchange of chemical data within a FAIR data environment (FAIR: Findable, Accessible, Interoperable, Reusable). The DIGChem project site was recently launched through a webinar on February 16 (https://sites.google.com/view/digchem). Several working groups will be discussed in more depth, including surveying data publishing guidelines in chemistry journals, formalizing standard formats for open exchange of chemical structures, formulating semantic representation of Gold Book terminology, and the GO FAIR Chemistry Implementation Network. We welcome participation across a wide range of interests within the chemistry community, to hear about your data needs and perspectives, as well as your involvement in our effort to help us meet those needs. We are also interested in participation from other disciplines, in order to make access to chemical research data more available and reusable outside chemistry.

Monday, March 19, 2018

Two CINF symposia sessions are sponsored and recommended by the ACS President, Peter K. Dorhout (Chemical & Engineering News, 2018, 96(4), 39).

1. CINF: Community Sharing of Chemical Safety Data: Yes, No, Maybe?

Location: River Bend 1, New Orleans Marriott Convention Center  
Date & Time: Monday, March 19 8:20 AM  
Organizer: Carmen Nitsche  
Presiders: Carmen Nitsche
Description: Various initiatives are underway that depend on the wider chemistry community to be willing to share safety learnings, but some members of the community are hesitant to exchange such information, for a variety of reasons, including personal embarrassment, IP risk, and unclear fears of legal or regulatory liability. Speakers will discuss safety data sharing and accompanying concerns, so as a community we can assess which concerns are valid, and how to address those that are.

Cooperative Cosponsor: PRES
Nominal Cosponsors: CHAS, CORP

2. CINF: Information Legacy of Eugene Garfield: From the Chicken Coop to the World Wide Web

Location: River Bend 1, New Orleans Marriott Convention Center
Monday, March 19 1:15 PM
Organizers: Helen Lawlor, Michael Qiu, Wendy Warr
Presider: Michael Qiu

Description: Eugene Garfield was an information pioneer. At a time when it could take much more than a year for an article to be published, much less discovered, he created information tools such as Current Contents that allowed for the quick dissemination of information so that scientific discoveries could quickly be built upon. He developed the Science Citation Index that would allow the most cited articles to be easily identified, again so that prior work could be built upon for the benefit of science and society as a whole, and he created chemical information tools that alerted researchers to new compounds, structures, and reactions well before they appeared in other indices, and adopted the use of linear notations to depict chemical structures well in advance of the use of chemical drawing technologies. Also he adopted the digital dissemination of content as soon as it was technologically possible, from magnetic tapes, diskettes, and CD ROMs, all the way through to the World Wide Web. Most importantly, his actions motivated others to follow his lead and build upon his foundation.

Today, information tools such as altmetrics, and search engines such as Google owe their very existence to the creativity, perseverance, and foresight of Eugene Garfield. This symposium will take a look at what his genius developed and how his legacy impacts and pervades the scientific information tools that are used today.

Nominal Cosponsors: HIST, PRES
Financial Cosponsor: Clarivate Analytics
Discussion: OpenStructures: IUPAC Standards project proposal

After the CINF sessions on Monday will be an informal discussion to hear interest for an IUPAC project proposal being developed to formalize standard formats for open exchange of chemical structures.

Location: TBD, will be informal  
Date & Time: Monday, March 19, 5:15 PM  
Duration: 1 hour  
Organizers: Evan Bolton, Leah McEwen

Description: Chemical notations such as InChI and SMILES are complementary, in practice they handle different use cases and needs in the community. The IUPAC InChI project presents an exemplar community process in developing an open standard. Wouldn’t it be great if SMILES could be canonicalized in a standard way, if SMILES from different vendors could obey the same rules? Wouldn’t it be great if SMILES and other chemical structure formats could be developed together as a community? Well they can! We can provide a forum and process to extend the SMILES and connection table families and enable innovation without disruption. Bring your edge cases and join us for an informal discussion on how we can make this a community effort. Contact Evan Bolton (bolton@ncbi.nlm.nih.gov) or Leah McEwen (lrm1@cornell.edu) if you are interested to contribute to this effort. Community relies on ‘U’, let’s advance chemical information together!

Tuesday, March 20, 2018

Please join us for Chemical Information Literacy: Innovation, Collaboration & Assessment, an all-day symposium of talks from instructors, librarians, and vendors who will be sharing how they are successfully supporting the development of chemical information literacy for the next generation of chemists, from one-shot assignments to deeper integration into entire courses and curricula.

Location: River Bend 1, New Orleans Marriott Convention Center  
Date & Time: Tuesday, March 20, 2018, 8:30 AM & 1:30 PM  
Organizers & Presiders: Ye Li, Charity Lovitt, Ginger Szymczak, Teri Vogel  
Nominal Cosponsor: CHED
Applications Invited for CSA Trust Grant for 2018 and 2019.

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 2018. The deadline for receipt of proposals for the 2019 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 2018 Grant is March 30, 2018. Successful applicants will be notified no later than May 9, 2018. The deadline for the 2019 grant is March 29, 2019 and successful applicants will be notified by May 8, 2019.

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 Association Trust: Recent Grant Awardees

2017 – Jesus Calvo-Castro

University of Hertfordshire, England, was awarded a grant to cover travel to present his work at the Fifth International Conference on Novel Psychoactive Substances to be held in Vienna, Austria from August 23-23, 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.

2017 – 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.

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.

2016 – 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.

2016 – 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.

2016 – 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.

2016 – 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. She 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.

2015 – Jack Evans
School of Physical Science, University of Adelaide, Australia. He was awarded a grant to spend two weeks collaborating with the research group of Dr. Francois-Xavier Coudert (CNRS, Chimie Paris Tech).

2015 – Dr. Oxelandr Isayev
Division of Chemical Biology and Medicinal Chemistry, UNC Eshelman School of Pharmacy, University of North Carolina at Chapel Hill. He 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.

2015 – Aleix Gimeno Vives

Cheminformatics and Nutrition Research Group, Biochemistry and Biotechnology Dept., Universitat Rovira i Virgili. He 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. He 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.

2014 – M. José Ojeda Montes

Department of Biochemistry and Biotechnology, University Rovira i Virgili, Spain. She 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 of functional food design.

2014 – Dr. David Palmer

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

2014 – Sona B. Warrier

Departments of Pharmaceutical Chemistry, Pharmaceutical Biotechnology, and Pharmaceutical Analysis, NMIMS University, Mumbai. She 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. He was awarded the 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 took place October 22 – 25, 2013 in Lausanne, Switzerland.
2013 – Dr. Robert S. Paton
University of Oxford, UK. He was awarded the 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 provided Dr. Paton with an opportunity to further his career as a Principal Investigator.

2013 – Dr. Aaron Thornton
Material Science and Engineering at CSIRO in Victoria, Australia. He was awarded the 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 also visited Dr. Maciej Haranczyk, a prior CSA Trust Grant recipient, who is one of the world leaders in virtual screening.

2012 – Tu Le
CSIRO Division of Materials Science & Engineering, Clayton, VIV, Australia. Tu was awarded the 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 the 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. Noel was awarded the 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. Laura was awarded the grant to do three months of research at the University of Innsbruck, Austria.

2008 – Maciej Haranczyk
University of Gdansk, Poland. Maciej was awarded the grant to travel to Sheffield University, Sheffield, UK, for a 6-week visit for research purposes.

2007 – Rajarshi Guha
Indiana University, Bloomington, IN, USA. Rajarshi was awarded the grant to attend the Gordon Research Conference on Computer-Aided Design in August 2007.
2006 – Krisztina Boda
University of Erlangen, Erlangen, Germany. Krisztina was awarded the grant to attend the 2006 spring National Meeting of the American Chemical Society in Atlanta, GA, USA.

2005 – Dr. Val Gillet and Professor Peter Willett
University of Sheffield, Sheffield, UK. They were awarded the 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. Sandra was awarded the grant to purchase equipment needed for her research.

2003 – Prashant S. Kharkar
Institute of Chemical Technology, University of Mumbai, Matunga, Mumbai. Prashant was awarded the 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, UK. Georgios was awarded the 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.
Book Reviews

The Publish or Perish Book(s), Harzing, Anne-Wil; Tarma Software Research Pty Ltd, Melbourne, Australia, 2011.


OK, how did I come to review books that are apparently instructional manuals for a software program, Publish or Perish (PoP), that I have not, and probably will not use? Well, book 3 was cited in some research I was doing in conjunction with reviewing other books on citation analysis and bibliometrics. My article in *J. Chem. Educ.* described comparing citation searching in both Google Scholar and Web of Science (WoS)1 so the subtitle of book 3 was intriguing. I could not find a copy available in libraries in Maine, so I purchased a copy from Amazon. Searching the title also brought books 1 and 2, and so I ordered them as well. Citation searching and analysis is discussed in all three books and the examples use the PoP software.

Per the introduction, the software can be used to organize references and citations for professional reasons, to decide where to publish, and to compile literature via citation searching and analysis. For further information on the software, consult the Harzing website ([harzing.com](http://harzing.com)). Those who use the software may wish to consider writing a review for the *CIB*.

Chapters 1 and 2 of Part 1 are duplicated in Part 2. Chapter 1 is a primer on citation analysis. Sections within the chapter answer “Why citation analysis”, including ranking journals and articles and cautionary notes when using citation analysis: typographical errors, Google Scholar (GS) inaccurate parsing, errors due to non-standard reference formats, and inconsistencies in covering books and conference proceedings both in the literature and in GS. The lack of good correlation between metrics and the true impact of those publishing is described as well as the conundrum of self-citations. The sources for citation research, GS, WoS, and Scopus, are described including history and statistics on methodology and coverage. Another section describes in detail the Journal Impact Factor (JIF) including problems, the h-index, m-quotient, contemporary h-index, individual h-index, and the g-index. The last four indexes are attempts to correct deficiencies in use of the h-index. The subsequent five chapters and six appendices cover doing searches with PoP software. Chapters 1 and 2 conclude with references.

Book 2 begins with the same first 2 chapters as Book 1 and subsequent chapters cover “making your case for tenure or promotion”, evaluation of academics, tips for academic administrators, and decisions on where to submit your publications. The last chapter gives tips on how to conduct searches.
Book 3 was of the most interest to me since it covers more nitty-gritty details on citation searching and analysis within GS and WoS. Chapter 1 is a repeat of the chapter from the other two books and Chapter 2 covers bibliometric research on authors and journals. The advantages and disadvantages of GS and WoS are discussed in detail in Chapter 3 and 4 respectively, including references on previous evaluations. Problems with GS include incorrect author identification (including “phantom authors”) due to problems with diacritics, apostrophes, and print ligatures (ff, fi, and fl).

Chapter 4 presents similar treatment of WoS. For that resource, manual processing of the original material (as opposed to computer processing) yields increased quality of the data. Disadvantages include lack of comprehension in the literature of some disciplines. It is fee-based, expensive, and exhibits slower processing than GS. The results often underestimate the impact of the citations. There are differences in coverage and treatment among journals covered by ISI and those not covered. Deficiencies in the latter include citations only for the first named author and underestimation of citations for academics with “foreign” names.

As a sidebar, when the parent Science Citation Index (SCI) started out, publications were indexed by the first named author only. Subsequent analysis was favored for prominent authors since the custom of the times featured the primary author being listed first, but this custom morphed into the current practice of listing the primary author last. Fortunately, all authors became searchable during the evolution of the SCI and WoS. I wish I had had this resource available when I was writing the *J. Chem. Educ.* Article. ¹

Chapter 5 describes an h-index for journals and Chapter 6 describes citation analysis across disciplines. References conclude both chapters.

The books, especially number 3, are good resources for citation searching and analysis, bibliometrics, and altimetrics, as well as the resources upon which they are based. They may be available in your library, and if not, are fairly inexpensive (prices listed from Amazon).

References

1. Buntrock, R. E. Using Citation indexes, Citation Searching, and Bibliometrics to Improve Chemistry Scholarship, Research, and Administration. *J. Chem. Educ.* 2016, 93 (3), 560-566. DOI: 10.1021/acs.jchemed.5b00451

Robert E. (Bob) Buntrock
Buntrock Associates
Orono, ME
Eugene Garfield

A Tribute to a Great Man and Great Legacy

Like many others, I was saddened to hear of Eugene Garfield's death last year. I hoped that a symposium would be held in his honor and had thought of submitting an abstract but I knew I would not be able to attend the meeting. Instead I asked Vin, editor of this Issue, if a tribute in the CIB would be welcome, and he assured me it was.

Unlike several in our industry, especially those speaking at the symposium, I only met Gene once and I was not a primary customer of his products and services. For example, my organizations were not subscribers to the Science Citation Index (SCI) but we did subscribe to several of the derived products and services. Nevertheless, my usage was significant at all stages of my career.

To my recollection, the only time I met Gene was at the last Chemical Information Gordon Conference in the 1970s (the demise of that conference was very disconcerting but that's another story). Already Gene's prominence and influence in the exploding information industry was well known. My wife Gloria accompanied me to that conference and at the annual Wednesday night lobster dinner we found ourselves seated across from Gene and his companion. After brief introductions and small talk, we set about attacking our half lobsters. Unfortunately this Midwestern rookie managed to spray Gene's companion with juice from his cracking of the lobster claw. We made profuse apologies, but the mood was definitely soured. To Gene's credit, if he remembered the incident, he never acknowledged that in our subsequent conversations about issues of mutual interest, usually via e-mail or the Chemical Information Discussion List (CHMINF-L).

My first introduction to Gene's and ISI services was in graduate school at Princeton via my mentor, E. C. (Ted) Taylor. Like my other mentors, Ted was an advocate of information awareness and usage. He came back from a consulting visit to SK&F in Philadelphia with a gift of a subscription to ACSCA, the current awareness product based on the Science Citation Index (SCI). It featured the ability to track the citation progress of 10 references key to our research. Ted kept five for himself but parceled out the other five to those of us in the group including one key to my thesis research. Occasionally we got hits that we might not have seen or saw earlier than we would have otherwise. This experience helped spark a career long interest in current awareness which became one of my professional specialties.

In addition, Ted strongly encouraged not only doing a thorough literature search for our thesis research, but also to write up the results, publishable in Chemical Reviews (to my knowledge only two of us did so, me and my lab mate Gavin Spence). Ted was also acquainted with SCI via his work with SK&F and said I should supplement my literature search by searching the articles I had found in the SCI. The Princeton subscription to this relatively new product was housed in the Biology Library a half block down Washington Street. I spent a couple of days re-searching my references in the SCI. I never kept notes on how many of the resulting almost 200 references in the review1 were found in the SCI, but I am sure that the number was significant since the topic was somewhat obscure to be...
easily covered by extensive reading of journals or use of *Chemical Abstracts*. The review was my first ever publication, a year after my doctorate and even preceding the publication of my lab research.

Using the SCI in print taught me several lessons. First, someone else citing your key reference might be citing it for reasons other than your interests, but that was usually the minority case and inspection of the retrieved references determined which were relevant. Just to be sure, I “recycled” the new references I retrieved with the SCI and found that this second tier search was not rewarding (especially when done manually) and not worth the effort, probably due to multiple reasons for citing.

After passing my PhD orals, a week later we moved 70 miles west and I began working at Air Products and Chemicals in Allentown PA, in a pesticide synthesis group. The research library was well stocked with resources including several essential journals of organic chemistry, the weekly issues of *Chemical Abstracts*, and several new current awareness products that were the advent of the computerized production of chemical information services. These included *Chemical Titles* from CAS and *Current Contents* (CC) from ISI. Both featured KWIC indexes of permuted titles of articles from the collection of journals covered by each service. However, CC also reproduced the table of contents from journals used for the SCI (3000) which covered several fields of science. When I was doing lab research in industry, synthesis of agricultural chemicals, I found that a broad reading program scanning current journals was the basis for good current awareness (I never knew where my next idea was coming from), but a good reading program must be backed up with good current awareness services, even though often redundant. ISI was a valuable supplier of such services.

In addition, the weekly issues also began with an editorial or essay by Gene on far ranging topics not just in the field of information. These were titled “Current Comments” and they have been archived in the series *Essays of an Information Scientist: Volumes 1-15.*

Some of us readers waggishly referred to Gene’s Current Comments as the “Thoughts of Chairman Gene.” This treasure trove also includes much of the history of both Gene and ISI. More than half of the 56 references in Bonnie Lawlor’s chapter in a recent ACS Symposium Series book are to that archive.

As an example of “off-beat” topics, one Current Comments described the production of ice cream, production of natural vanilla, and the marriage of the two in some brands of ice cream, specifically Breyers Natural Vanilla which still uses natural vanilla pods. After reading that, I have bought that brand ever since when available. Gene was obviously an omnivorous reader with a wide range of interests.

Information topics in Current Comments included journal impact factors (IF) which Gene and ISI developed, most cited papers, refereeing and peer review, and the value of writing and using reviews. Refereeing policy is discussed and controversial to this day and one of Gene’s recommendations was that reviewers should not only review papers, but also sign their reviews. The option to transmit one’s signature on ACS reviewer report forms was present and I always used it with only one repercussion from an author whose manuscript I rejected. Since I had written a review of my thesis work I also valued reviews and wish I could have written more. Later, when my papers were reviewed, I would have appreciated signed reports but I do not recall ever receiving one.
Additional current awareness products from ISI included *Current Abstracts of Chemistry/ Index Chemicus* (CAC/IC) and *Current Chemical Reactions* (CCR) details of which have been described including by Wendy Warr at the upcoming symposium (paper CINF 49 at the ACS meeting in New Orleans). The mission of each current awareness product was to present “new” chemistry, compounds and reactions, with bibliographic citations and structure and reaction schemes copied from the article. Users discovered that “not so new” chemistry was also included with the new chemistry so the coverage was more comprehensive than indicated at face value. Even though I was no longer in the lab when CCR appeared, I continued to keep track of my previous chemistry, both in graduate school and on the job, and since I am a visual learner, the display of structures and reaction diagrams made the reports very usable.

A further aspect of Gene’s interest in publishing, especially the biomedical areas, was his founding of *The Scientist*, a magazine for biomedical and life scientists. Publication has passed to several organizations but still exists both in print and online. Not only research areas are covered, but various career aspects as well.

Since Gene was active professionally, including being President of the American Society for Information Science (ASIS), I would see him at meetings, but as I recall, I never met him again. He did keep track of my contributions, especially on CHMINF-L, and we had several informative exchanges of notes. I only regret that I did not get to talk with him in person and get to know him better. Although organizations I worked for never subscribed to the SCI, my limited experience with citation searching led me to use it wherever I could. I did use citation searching when it became available in the CA file on STN, and later with Google Scholar. I found a new subject to follow and write about (as well as reviewing books) including not only citation searching but also bibliometrics and altmetrics. Those topics along with my career long interest with current awareness make Gene one of my mentors.

The contributions and impact of Gene Garfield and his products and services to scientific information are enormous. It is significant that current awareness was facilitated by the appearance of *Current Contents* even before the development as a publication of the SCI. Gene’s achievements includes early support of current awareness (a key aspect of both my information career and my personal use), invention and development of citation searching in scientific information (later also including the humanities), invention of journal and article Impact Factors (including guidelines on their use and misuse), and the origins of the burgeoning fields of bibliometrics and altmetrics. He was interested in and influential in all aspects of scientific publishing including authorship, publication, value of various types of publications, reviewing and refereeing, searching, and administration. As the upcoming symposium, *Information Legacy of Eugene Garfield: From the Chicken Coop to the World Wide Web*, and various publications³ and interviews⁴ have shown, he was a remarkable man and genius. He has even been given credit for being the “Grandfather of Google.” As I expressed in my paper describing the use of Gene’s products in chemistry pedagogy,⁵ “… the author salutes Eugene Garfield, upon whose genius and shoulders we stand”.

---

³ Various publications have shown Gene’s contributions to bibliometrics and altmetrics.
⁴ Interviews with Gene have highlighted his influence in scientific publishing.
⁵ Paper describing the use of Gene’s products in chemistry pedagogy.
References


5. Buntrock, R. E., Using Citation Indexes, Citation Searching, and Bibliometrics to Improve Chemistry Scholarship, Research, and Administration. *J. Chem. Educ.* **2016**, *93*(3), 560-566.

Addendum

I just learned that E. C. Taylor, Professor of Chemistry Emeritus, Princeton University, passed away in November. Ted was the PhD advisor I cited in this reflection, my mentor both in chemistry and chemical information. He will be sorely missed by many.

Robert E. (Bob) Buntrock
Buntrock Associates
Orono, ME

Editor Note

Readers may also be interested in the following reference:

Twenty-five Years Ago

I cannot believe that I am planning my seventh trip to New Orleans. My records (going back to 1977) show that ACS met there in fall 1987, spring 1996, fall 1999, spring 2003, spring 2008, and spring 2013, and will now meet there again in March 2018. Your esteemed editor suggested that perhaps I should reminisce about the 1999 New Orleans meeting, but he did comment that 19 years was not a typical anniversary. Well, I could write about New Orleans five, 10 or 15 years ago, since all three sound more auspicious anniversaries, but I wrote in the Chemical Information Bulletin July 2017 that I hoped to continue my 25-year history theme when the spring 2018 national meeting approached. So I said that I expected to find some gems in my report on the spring 1993 meeting in Denver, 25 years on. Nevertheless, let us try to appease the editor. What was happening in CINF 15 years ago in New Orleans?

I pulled out a hard copy of my 21st report (shameless plug: details of the 50th report are at http://www.warr.com), and I was not over-excited (quite apart from my jolly post-it note on it saying “dog-eared, mine”). Amazingly, I transcribed 36 talks but most of them are rather COMP-like; I get the feeling that I was proud of my expertise with Microsoft Equation! One passing thought was the number of companies that have been acquired, or have passed on to the great hunting grounds in the sky. Some of you must remember Accelrys, Argenta Discovery, Argonaut Technologies, Barnard Chemical Information, BioFocus, BioWisdom, CENSA (oh, those lengthy ELN symposia that over-ran), ChemNavigator, ChemWeb.com, FIZ CHEMIE Berlin, CAChe Software, InforSense, Inpharmatica, Libraria, LION Bioscience, MDL Information Systems, Metaphorics, NuGenesis Technologies Corporation, Personal Chemistry, Query+, Scimagix, SciQuest, SciTegic, SimBioSys, Symyx Technologies, Synthematix, The Scientific World, Signature BioScience, and Tripos. At least there was one CINF gem: the Journal of the American Chemical Society published its 125th volume in 2003.

Memories from the Denver meeting 25 years ago are so much more fun. I resurrected my meeting report (initially in an old format of Word, in the days when file names were limited to eight characters). A few quotations speak for themselves.

“Bob Massie…gave a motherhood and apple pie statement of CAS’ mission, and then concentrated on the issue of pricing, which he saw as the biggest challenge facing the industry today. CAS will moderate all price increases as far as possible in future. An alternative to search term pricing will appear this summer and over the next 6-18 months other new pricing options will become available.”

“Susan Brown, the new Director of Marketing (replacing Jim Seals) was introduced. She had only been at CAS for about a week, so it was not worth exchanging more than pleasantries with her at this stage. She was previously at Mead. Stu Kaback had been visiting CAS about three weeks before this meeting. He was told how much CAS wanted to improve relations with customers, and so on. He was very worried when he learned that CAS had taken on someone from Mead: Mead apparently has a lousy reputation for customer appreciation. CAS was quick to assure Stu that Sue Brown would be suitable.”
“…the roundtable discussion on producing reports from results, led by Tom Wolff of Amoco. Users should no longer be given an unedited print-out of hits from an online search. A substantive value-added contribution can be made by evaluating citations, emphasizing certain results…and supplying a cover letter describing the search strategy….The next part of the discussion covered whether to do your formatting online or by post-processing…”

“CAS gave a very large and elegant customer appreciation reception on the Tuesday evening…The event partly overlapped with a much smaller MDL reception where we had to suffer a presentation about ISIS on the IBM RISC/6000 AIX…Relations between CAS and MDL have obviously deteriorated since the days of big receptions to mark ISIS/STN Express links. Of course, you don’t give big receptions to announce that you are no longer friends…”

“The PC front end to online DARC, CHEMLINK, written by Fraser Williams to Questel’s precise specification, has never been a success. Mike O’Hara refused to sell it in the United States. He calls it CHEMKLINK.”

“Networking is definitely the rage: Richard Hong knows someone who has three computers networked in his apartment! I recently had at least four telephone calls in one week alone, asking me how you get access to the Internet in England. I myself have access to the Internet through CAS…I use Internet and electronic mail much more than the telephone. I can communicate with academics the world over, with most of the software and hardware companies, and with US industry. The only people I am really isolated from are in the European pharmaceutical industry.”

“VCH had Warr and Suhr’s \textit{Chemical Information Management} on display…The patents section especially has received good reviews. I wish we had rewritten the IT chapter though; it has dated rather, since my first efforts of 1989.”

“COMP Award Symposium Honoring W. Clark Still…The room allocated for this symposium was far too small and I was unable to squeeze in until the end of Bill Jorgensen’s paper.” Plus ça change, plus c’est la même chose.

That seems a fitting note on which to finish. I do hope that I will be able to offer a few comments on the fall 1993 meeting (in Chicago) before the fall 2018 meeting in Boston.

\textit{Wendy Warr  
Wendy Warr & Associates}
CINF Member Profile

Who are you?
Rajarshi Guha

What do you do? (Institution, position, job description/duties)
I’m currently an informatics research scientist within the Division of Preclinical Innovation at the National Center for Advancing Translational Science (NCATS), at the National Institutes of Health (NIH). In this position, my remit is to develop an informatics research program that addresses chemical and biological problems being studied across the center, and develop quantitative approaches to characterizing various aspects of translational science. As an informatics scientist with a broad range of interests, I have had the opportunity to work in a number of areas. A brief summary of some of these areas includes:

- Developing public-facing resources such as Pharos (https://pharos.nih.gov/idg/index), which involves developing databases, web frameworks, API’s, visualizations and so on
- Developing screening infrastructure such as the Trans NIH RNAi screening facility, and more recently implementing the complete informatics infrastructure to support high throughput combination screening (https://tripod.nih.gov/matrix-client/)
- Exploring and developing algorithms and methodologies for predictive modeling and handling of chemical information (as well as other data types such as genomic and bibliographic data). Recent work includes characterizing trends in scaffold properties (https://tripod.nih.gov/ste) and developing methods to predict combination responses
- Providing cheminformatics and modeling support for small molecule development projects in therapeutic areas such as infectious disease and neurodegenerative disorders.

The environment at NCATS is very open, and as a result, the bulk of my work can be freely published. All our tools and data are made available under very permissive licenses.

Why are you in the chemical information field?
My initial interest in chemistry was piqued by the shape of chemical structures in my high school textbooks. From there I pursued undergraduate and graduate degrees in chemistry. But during the same time, I was also growing increasingly interested in computing, especially from the point of view of programming and visualizing things. As I studied chemistry formally, I came to realize that my experimental skills were not very impressive, but when I took my first quantum mechanics course, I realized that this was a way I could study chemistry and computing, and also stay out of the lab! I also credit the Computational Chemistry List (CCL) (http://www.ccl.net/) with providing advice (http://www.ccl.net/cgi-bin/ccl/message-new?1999+06+27+002) on who to study with if I wanted to go down this route.
As a result, after completing my Master’s, I started a PhD in quantum Monte Carlo methods at Penn State University, but soon after starting I realized that the systems I was working on were relatively small and isolated. After learning about other types of computational chemistry, I shifted to Peter Jurs’ group, working on QSAR (quantitative structure–activity relationship) methods. This was attractive because I was able to work with larger datasets as well as be more closely connected to the uses of small molecules in treating diseases. During graduate school I had a lot of freedom to explore machine learning methods, software for cheminformatics, and the associated infrastructure around chemical information. I was greatly influenced by open source and open science efforts, in particular the ones led by Peter Murray-Rust and Christoph Steinbeck. This led to me getting involved with the Chemistry Development Kit (CDK) (https://cdk.github.io/), a Java library for cheminformatics. Since then, open source has been very important to me and I have made efforts to ensure that my work is as freely available as possible, in terms of both data and software.

After finishing my PhD, I joined Indiana University as a post-doc and then as a visiting Assistant Professor of Informatics. During this time, I worked on large scale cheminformatics projects while developing my own research program. While it was a great environment in which to work, the nature of academic cheminformatics is such that access to “fresh” data is limited. This led me to look for other opportunities that not only would let me work in an open manner, but also have access to large sets of real chemical and biological data. As a result I joined the the NIH Chemical Genomics Center (which later became NCATS) at NIH and have been there ever since.

What makes CINF valuable to you? (Include anything relevant, but especially any committees or projects in which you are involved.)

It was during grad school that I joined ACS and learned of CINF. At that time, it felt like the perfect community from which to learn more about ongoing research and open problems. In particular, the conference programming developed by CINF was very useful for me as I looked for problems to work on.

Wanting to get more involved in the community, and seeing some topics not being addressed in the CINF conference programming, I offered to join the CINF Program Committee, then led by Leah McEwen. This was a great opportunity and helped broaden my view of the cheminformatics and chemical information fields and communities. From there, I moved to CINF Assistant Program Chair and then Program Chair, with much help from Leah and the other members of the Committee.

By that time, I had developed close ties to various CINF subcommittees and had a feel for the challenges and opportunities faced by CINF. Thus I moved on to the role of Division Chair. Since serving as chair, I have stayed involved, primarily in the area of programming, and still regularly organize CINF sessions at national meetings.

After 15 years, CINF is very relevant to my interests, and the progress I’ve been able to make in the field is due in large part to the network of colleagues I have met at CINF.
meetings as well as the scientific discussions I have been involved in with CINF colleagues and at meetings arranged by CINF.

CINF Membership Committee
ACS Division of Chemical Information
Notes From Our Sponsors

Division of Chemical Information Sponsors Spring 2018

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 2018 National Meeting in New Orleans, LA, and to support other divisional activities during the year, including scholarships to graduate students in chemical information.

The Division gratefully acknowledges contributions from the following sponsors:

**Gold**
- ACS Publications
- Clarivate Analytics

**Silver**
- Royal Society of Chemistry

**Contributor**
- Bio-Rad Laboratories

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 a sponsor 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 2018
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.*
Announcing ACS Reviewer Lab™, a free online course from ACS Publications, designed to educate researchers on the fundamentals of peer review featuring six modules and interactive exercises.

SIGN UP AT
acsreviewerlab.org
Before there was Clarivate Analytics, there was the Institute for Scientific Information (ISI). Founded nearly 60 years ago by pioneering information scientist Dr. Eugene Garfield, ISI launched a revolution in the indexing and retrieval of scientific and scholarly literature. The antecedents of today’s universally respected Web of Science, and many other Clarivate Analytics tools, took shape under the ISI banner. And although the Institute for Scientific Information became less visible, as the business underwent various corporate iterations and rebranding, its spirit remained, like a thread of DNA.

Now, the foundational, innovative pedigree of the Institute for Scientific Information returns to the fore, as Clarivate Analytics, in its second year as a fully independent company, will re-energize the Institute for Scientific Information within its Scientific and Academic Research Group.

The re-establishment of the Institute for Scientific Information marks a renewed commitment to the original spirit of the company and the development of cutting-edge information resources – now more crucial than ever, as the world deals with an ever-increasing torrent of data. The same originality and vision that created the Science Citation Index and Journal Citation Reports will animate ISI. This energy will enlarge the array of newer resources that have followed those original, flagship products, including the performance-assessment tools InCites and Essential Science Indicators, along with the Emerging Sources Citation Index and its coverage of a rapidly expanding landscape of global research.

As a think-tank and incubator, a platform on which to foster new partnerships and collaborations, and a conduit for the creation of new metrics and analytical tools to meet the ever-expanding volume of information and the shifting dynamics of the digital age, the Institute for Scientific Information will continue the six-decade legacy that underpins Clarivate Analytics.

Needless to say, a revivified Institute for Scientific Information (ISI) demands leadership commensurate with that legacy. Leading the re-establishment and future of ISI will be Samantha Burridge, Director of Strategy and Transformation, bringing with her over 20 years of broad publishing experience. Most recently, Burridge established Nature’s Open Research Group in 2013 as Managing Director, and by 2016 had led the group to become the most highly-regarded and fastest growing open access (OA) business in the world. Burridge will be assisted by a strong internal team and a plethora of new hires, including Dr. Jonathan Adams, who is joining the company, or, more accurately, returning to the company. Adams’s nearly four decades of experience in the application of bibliometrics in research evaluation include the founding of Evidence Ltd in 2000. Prior to that, as policy advisor to the United Kingdom’s Advisory Board for the Research Councils, he worked...
closely with ISI personnel in introducing citation analysis and other evaluative measures to the management of the U.K. science budget, and from 2009 to 2013, Adams was an official part of Thomson Reuters, the immediate forerunner of Clarivate Analytics, as Director of Research Evaluation. He will join the Institute for Scientific Information in April, 2018, from Holtzbrinck Publishing Group, where he currently holds the post of Chief Scientist, Digital Science.

Another key leadership role will be filled by Dr. Nandita Quaderi, who has joined Clarivate Analytics as Editor in Chief of the Scientific and Academic Research Group. An accomplished research scientist, Quaderi earned a PhD in Molecular Genetics from Imperial College London. Her publishing experience includes a stint at Springer Nature, where, under her guidance, *Scientific Reports* became the largest journal in the world. As part of ISI, she will determine the metrics that Clarivate Analytics will use to set and measure quality standards across the Web of Science, taking into account additional content types such as peer-reviewed articles, preprints, datasets, software, etc.

The Institute for Scientific Information (ISI) will benefit from still more experience in publishing and analytics through its engagement with the consulting company SchoolDash. Founded by Dr. Timo Hannay, who was previously the founding managing director of Digital Science, and a director at nature.com, SchoolDash will support ISI in an advisory capacity, as part of a broader mission to promote data science in academia and education.

This first-class leadership team, with Annette Thomas as CEO, is well-known for taking an innovative and customer-centric approach to re-inventing scholarly communication, workflow and assessment. They have come together again to re-establish ISI, motivated by the opportunity to work with the research community and build upon the original, ground-breaking inspiration of founder Dr. Eugene Garfield in new and innovative ways.

Read the full press release
Royal Society of Chemistry News

The Royal Society of Chemistry is a not for profit organization with a heritage that spans 175 years and is an international publishing and knowledge business.

We’re working to shape the future of the chemical sciences, for the benefit of science and humanity.

Introducing Molecular Omics

In 2018 Molecular BioSystems refocused its scope and was relaunched as Molecular Omics.

Molecular Omics publishes molecular level experimental and bioinformatics research in the -omics sciences, including genomics, proteomics, transcriptomics and metabolomics. We also welcome multidisciplinary papers presenting studies combining different types of omics, or the interface of omics and other fields such as systems biology or chemical biology.

Our current free-to-access journals:

Did you know all of our journals have a two-year, free online access period? This gives you a chance to try them out, and see how valuable they could be for the research you do.

- Materials Chemistry Frontiers
  An international, high quality home for studies of a significant nature that further the development of organic, inorganic, composite and nano-materials.

  This journal is part of the Frontiers project, a not-for-profit society partnership between the Chinese Chemical Society and the Royal Society of Chemistry, in collaboration with Institute of Chemistry, Chinese Academy of Sciences.

- Sustainable Energy & Fuels
  Complementing our leading titles Energy & Environmental Science and Journal of Materials Chemistry A, Sustainable Energy & Fuels publishes interdisciplinary research that contributes to the development of sustainable energy technologies, with particular emphasis on new and next-generation technologies.

Royal Society of Chemistry to sell Biochemical Society journals

In 2017, the Royal Society of Chemistry and Portland Press, the wholly-owned publishing arm of the Biochemical Society, agreed a partnership enabling the Royal Society of Chemistry to sell Portland Press journals content.

Speaking from the Frankfurt Book Fair, our commercial director, Dan Dyer said: “I am excited by the opportunities this announcement presents to both the Royal Society of
Chemistry and the Biochemical Society. “With the strategic alliance of the two societies, our customers will benefit by having a single sales contact to discuss all their chemistry and biochemistry content requirements. Working collaboratively in this way will benefit the whole community, by extending our content offering into each other’s respective markets.”

Professor Richard Reece, Chair of Portland Press Board commented: “Portland Press is dedicated to promoting and sharing scientific research, as such I am delighted that our journals will be exposed to a wider market. We can continue to further our aim to offer the best service to the librarian community, whilst increasing the dissemination of our published content to researchers.”

Dr Niamh O’Connor Director of Publishing, Portland Press added: “This aids achievement of the Biochemical Society’s strategic objective to further develop international links and networks, supporting scientists to share their work on a global scale”.

Read and Publish: supporting open access publishing

In 2017 we launched Read & Publish, our popular scheme to support the industry’s transition to OA publishing. With Read & Publish, corresponding authors can publish gold OA in all of our hybrid journals, and their institution gets access to our entire hybrid journal portfolio.

How does it work?

1. **Pay a tailored publishing fee to publish 100% gold open access**
   We calculate an institution’s publishing fee by analyzing the last full years’ publishing output. Every accepted article from corresponding authors can be published gold open access.

2. **Pay a set reading fee to unlock access to every article**
   An institution then pays a set reading fee which gives the library perpetual access rights to the content published in our hybrid portfolio during the term of the contract.

What do our subscribers think?

The Max Planck Digital Library was one of our first customers to sign up to this new agreement:

“We regard this new agreement with the Royal Society of Chemistry as another practical step in the transition from subscription to open access as envisioned in the OA2020 initiative. With this new approach we shift our payments and workflows in a way to make open access the default of publishing for our researchers.”

Ralf Schimmer
Head of Information at the Max Planck Digital Library
So far, around 90% of authors at institutions with Read & Publish have chosen to publish their work OA in our hybrid journal portfolio.

To find out more about any of the above, please get in touch.

New deputy chief executive

Paul Lewis is to join the Royal Society of Chemistry as the new Deputy Chief Executive (Head of House) in March.

An experienced senior leader in the private and not-for-profit sectors, Paul was previously at City & Guilds where he was Executive Director, International and Strategy. He has substantial leadership experience in the education, skills and knowledge sectors with specialist government-to-government and international operating expertise.

Paul has a BSc (Hons) Geographical Sciences, PGCE (qualified teacher status) and MBA (Distinction). He is a graduate of the Royal Military Academy, Sandhurst. Paul is a Freeman of the City of London, a Fellow of the Institute of Leadership and Management, and, a Fellow of the Royal Society for the Arts.

Dr. Robert Parker, Chief Executive of the Royal Society of Chemistry, said: “I am delighted to welcome Paul to the Royal Society of Chemistry. He brings with him valuable knowledge, skills and experience that, in partnership with Helen Pain, our DCEO (Head of Profession) and our wider leadership team, will contribute towards our strategic delivery and achieving our mission to advance excellence in the chemical sciences.”

Paul said: “I am delighted to be joining the highly respected Royal Society of Chemistry and honored to be appointed Deputy Chief Executive. I very much look forward to meeting and working with colleagues, members and stakeholders around the world and feel privileged to be part of the chemical sciences community.”

Paul will be responsible for leading the organization’s publishing, commercial and technology delivery with a strong focus on developing and maintaining our position as a leading, high-quality publisher and knowledge provider, which is core to our Royal Charter and strategic ambitions.

Funding international collaborations for the advancement of science

16 January 2018

We are once again partnering with the Royal Society to offer grants of up to £12,000 as part of the International Exchanges Award, which funds collaborations between researchers in the United Kingdom and Africa.

“International collaborations in research are becoming more and more important for solving global challenges,” says Helen Driver, who leads the Royal Society of Chemistry’s Pan Africa Chemistry Network. “We are proud to fund this exchange, which enables scientists from Africa and the United Kingdom to bring together their knowledge and skills, to mutual benefit.”
The award is given for collaborations between researchers in the United Kingdom and those in Sub-Saharan Africa (excluding South Africa), working on projects in the field of chemistry. The current round of applications closes on 13 March 2018. Apply now.

Picture: Previous award recipients Steven Howdle (second left) and Yonas Chebude (right), with their PhD students Mariana Gamerio (second right) and Yaregal Awoke Genet (left).

Exchanging expertise

In 2016 Professor Steven Howdle, from the University of Nottingham and Dr Yonas Chebude, from Addis Ababa University received £12,000 from the program to collaborate on a research project with an environmental focus. Lucy Lisanti from the Royal Society caught up with them on how their research is progressing, and their plans for the future.

Professor Howdle and Dr. Chebude met and exchanged ideas at an international conference, and this meeting formed the basis of the collaboration between their research groups. Their project concentrates on extracting, purifying and converting oils present in African ironweed into environmentally sustainable biopolymers with potential uses in the pharmaceutical industry and the production of plastics.

‘We have abundant natural resources but we do not have the facilities to implement the ideas. That was the challenge,’ says Dr. Chebude. Ironweed is a prolifically growing natural plant in Ethiopia and the broader Sub-Saharan regions. In Addis, recent work pioneered by Dr. Chebude had demonstrated that the weed contains a significant level of vernonia oil, which can be turned into vernolic acid, a potentially very valuable natural resource and source of bioderived plastics. However, there were significant hurdles to overcome: the yield of vernolic acid from vernonia oil using a conventional solvent extraction method is very low. Meanwhile, in Nottingham, Steve Howdle was working on a similar molecule that is extracted from birch bark using supercritical carbon dioxide, a very gentle and natural extraction methodology.
Learning opportunities

Through the International Exchanges Award, the two groups have brought together their complementary expertise to extract the vernonia oil in order to create a completely new range of bioderived plastics. Using the funding, members of each group have been able to visit their counterpart institution, as well as conduct joint work at unique facilities at Bangor University, U.K. Without such an exchange of skills and expertise the project would not be viable.

In addition to Professor Howdle’s and Dr. Chebude’s input, PhD students from both groups have taken advantage of the learning opportunities. Bilateral visits between the U.K. and Ethiopia have equipped the PhD students with the knowledge, networks and skills necessary to sustain the partnership long after the completion of the award. Beyond the obvious academic benefits, these exchanges also provide the recipients and their teams with an international perspective on challenges they had not previously considered.

“My most important advice to academics is to get rid of the fear of rejection, don’t be afraid of approaching U.K. scientists or other international collaborators,” says Dr. Chebude. He cites this international exchange with increasing his confidence in an academic and research environment, due to both the prestige of the award and the potential impact his research will have on a wide selection of industries. Dr. Chebude suggests that prospective applicants to international exchange program should investigate whether pre-existing collaborations exist between their universities and other organizations. These networks give applicants guidance on who to approach in order to develop a research collaboration.

Forming further collaborations

Such is the success of the working relationship between the two research groups, further collaborations between groups at the two universities have begun to form, inspired by the achievements of Professor Howdle and Professor Chebude. Prof Howdle also commented enthusiastically “…. we have developed new and exciting science that simply would not have happened if we had not been able to follow up our initial meeting by applying for the International Exchanges Award, and this has benefited both of our groups and in particular Yaregal and Mariana.”

In the current world of research, an international outlook is essential for scientists to remain competitive in their field by tackling problems in novel and exciting ways.
Bio-Rad Laboratories, Inc.
Spectroscopy Database and Software Solutions

Bio-Rad is a leader in spectral databases & software for the scientific and chemical information community.

**KnowItAll Spectral Databases and Spectroscopy Software**

Bio-Rad is the leading producer and publisher of spectral databases, with a collection that contains over 2 million spectra (IR, Raman, NMR, NIR, MS, UV-Vis), the world’s largest, covering pure compounds and a broad range of commercial products. These spectral collections are critical when trying to identify or classify unknown spectra.

Bio-Rad’s KnowItAll software ([http://www.knowitall.com](http://www.knowitall.com)) offers comprehensive solutions to identify, search, analyze, and manage spectral data in multiple instrument vendor file formats and techniques (IR, Raman, NIR, NMR, MS, UV-Vis, and chromatography). KnowItAll’s integrated toolsets eliminate the need for multiple software packages and increase overall lab efficiency. Combined with the world’s largest spectral reference database, KnowItAll provides the most advanced technology available for spectral analysis!

Contact us to find out more about the latest release KnowItAll 2018!

**For a trial, please visit** [http://www.knowitall.com/trial2018](http://www.knowitall.com/trial2018)

**Introducing SpectraBase!**

Bio-Rad is pleased to introduce to the chemistry community the new SpectraBase cloud-based spectral repository ([http://spectrabase.com](http://spectrabase.com)) which offers fast text access online to hundreds of thousands of spectra.

**Calling All Campus Librarians and Chemistry Professors:**

*KnowItAll U - Campus-Wide Access to Over 2 Million Spectra*

Bio-Rad’s KnowItAll U offers every student and faculty member, at any computer campus-wide, unlimited access to over 2 million spectra. Users can search the world’s largest spectral collection online or with Bio-Rad’s KnowItAll desktop software for the fastest, most accurate results. Bio-Rad also offers campus-wide access to its KnowItAll IR and Raman Spectral Libraries.

**Please contact us if you are interested in setting up an evaluation for your school** [http://www.knowitall.com/contactus](http://www.knowitall.com/contactus)
SUNDAY MORNING
Section A

New Orleans Marriott Convention Center
River Bend 1

Enhance Discovery: Share Chemical Structures
Financially supported by Chemical Structure Association Trust

Y. Li, V. F. Scalfani, Organizers
L. McEwen, Organizer, Presiding

8:30 Introductory Remarks.

8:40 1. Sharing chemical structures from university theses and dissertations on Institutional Repositories and PubChem. V.F. Scalfani, P. Rupar, M.S. Alexander, D.G. Williams


9:55 Intermission.

10:10 4. Deposit chemical structures into PubChem - a public data repository. J. Zhang, A. Gindulyte, B. Shoemaker, P. Thiessen, E. Bolton

10:35 5. Supporting the community to share chemical structures. G. Jones, R. Kidd

11:00 6. Chemical structures: more than just meta-data? M.G. Hicks

11:25 Discussion.

Section B

New Orleans Marriott Convention Center
Natzhez

Drug Discovery: Cheminformatic Approaches
Cosponsored by COMP
E. Davis, Organizer, Presiding

9:00 7. Synthetically Accessible Virtual Inventory (SAVI) – Next version: Toward more needles and less haystack. H. Patel, Y. Pevzner, W. Ihlenfeldt, M.C. Nicklaus

discovery. **Y. Djoumbou Feunang**, D. Wishart

**9:50 9.** CCCTK (Compute Cure for Cancer ToolKit): An open source drug discovery platform for design of novel anti-cancer agents. **M. Karthikeyan**, r. vyas

**10:15** Intermission.

**10:30 10.** Practical machine learning methods for QSAR and QSAR predictions. **V. Tkachenko**, A. Korotcov, R. Zakharov, B. Sattarov, A. Mitrofanov

**10:55 11.** Making virtual REAL: synthetically feasible compounds and their exploration in docking screens. **Y. Moroz**

**11:20 12.** Deep learning approaches for detecting high-throughput screening false positives. **M. Matlock**, T. Hughes, S. Swamidass


**Drug Design**

**Structure & Ligand-Based Design**
Sponsored by COMP, Cosponsored by CINF

**Machine Learning for Catalysis Research**
Sponsored by CATL, Cosponsored by CINF and COMP

**Marriage of Machine Learning, Knowledge Representation & Chemical Sciences**

**Data Mining & Frameworks for Chemical Discovery**
Sponsored by COMP, Cosponsored by CINF and PHYS

**Large-Scale Applied Molecular Modeling**
Sponsored by COMP, Cosponsored by CINF and MEDI

**SUNDAY AFTERNOON**

**Section A**

New Orleans Marriott Convention Center
River Bend 1

**Enhance Discovery: Share Chemical Structures**
Financially supported by Chemical Structure Association Trust

L. McEwen, Organizer
Y. Li, V. F. Scalfani, Organizers, Presiding

**1:15** Introductory Remarks.

**1:20 14.** Sharing chemical structures with peer-reviewed publications. Are we there yet? **A.J. Williams**

**1:35 15.** Integration of Markush structures into EPA's DSSTox database to represent and enumerate UVCB substances. **C. Grulke**, A.J. Williams, A. Richard

**1:55 16.** Curating and sharing structures and spectra for the environmental community. **E. Schymanski**, A.J. Williams

**2:15 17.** ChemAxon's chemical structure file formats for better data storage and search. **A.D. Costache**, E.A. Hoffmann, A. Volford

**2:40 18.** NCI/CADD CACTUS web server: Tool for connecting chemical structures. **M.C. Nicklaus**

**3:05** Intermission.

**3:20 19.** Sharing chemical structures:

Section B
New Orleans Marriott Convention Center Natzhez
Fragrances, Food & Cheminformatics

R. J. Bienstock, J. A. Bikker, Organizers, Presiding

1:15 Introductory Remarks.

1:20 20. Global structure diversity and chemical space of food chemicals. N.N. Trujillo-Minero, J.L. Medina-Franco


2:00 22. Using structural data in the food and fragrance industries. S. Ward, A. Sarjeant, S. Vyas, P.I. Andrews

2:20 Intermission.

2:30 23. Predicting human olfactory perception from molecular structure. J. Mainland

2:50 24. Using molecular fields to understand molecular determinants of the olfaction process. M. Slater


3:30 26. QSAR/QSPR models to support fragrance ingredient molecular design. D. Chekmarev, J. Kattas, J.A. Bikker

3:50 Intermission.

4:00 27. Translating methods from pharma to fragrances and flavors. T. Mansley, E. Champness, P. Hunt, N. Foster, M.D. Segall


4:40 29. Computational analysis towards understanding food dye and coloring agent genotoxicity. R.J. Bienstock, L. Perera, M.A. Pasquinelli

Machine Learning for Catalysis Research
Sponsored by CATL, Cosponsored by CINF and COMP

Marriage of Machine Learning, Knowledge Representation & Chemical Sciences
Artificial Intelligent Searching of Chemical Space
Sponsored by COMP, Cosponsored by CINF and PHYS

Large-Scale Applied Molecular Modeling
Sponsored by COMP, Cosponsored by CINF and MEDI

MONTDAY MORNING

Section A

New Orleans Marriott Convention Center River Bend 1

Community Sharing of Chemical Safety Data: Yes, No, Maybe?
Cosponsored by CHAS, CA and PRES‡

C. I. Nitsche, Organizer, Presiding
What's all this fuss about data sharing? In search of improved laboratory safety. C.I. Nitsche

Promoting safety culture through sharing - a Dow perspective. M.E. Jones, L. Seiler, C. Mapes

Chemical and laboratory safety – the role of scholarly publishers. C. Toro, S.B. Tegen

Chemical safety data in the Handbook of Chemistry and Physics. J. Rumble, D.R. Lide, F. Macdonald

Safety sharing culture: Learning from the aviation industry. T. Zoeller

Intermission.

Information/practice sharing forums among American Chemistry Council member companies. I. McGee, D. Sandidge

Experiences with learning experience reports. W.B. Tolman

Balancing act: Protecting all interests. S. Addlestone, S. Christman

Parsing the “lessons learned” space: Layers of opportunities and challenges. R. Stuart

Panel Discussion.

Section B
New Orleans Marriott Convention Center Natzhez

Workflows & Cheminformatics

R. J. Bienstock, G. Landrum, Organizers, Presiding

Open-source web tools for modeling and design tracking: Workflows facilitating collaborative drug discovery. K.W. Lexa, J.A. Feng

Building upon chemical similarity - methods to extend instance-based learning in cheminformatics. T.H. Luechtefeld

Chemical workflows supporting automated research data collection. V. Tkachenko, R. Zakharov, F. Prior, A.V. Kabanov, A. Tropsha

Intermission.

Diamond, XChem and CCP-CompMedChem: Creating user-focused tools and workflows for structure-based drug design. A. Bradley, R. Skyner, F. von Delft


Interactive and reproducible data analysis with the open-source KNIME Analytics Platform. G. Landrum

Intermission.

Using Python to streamline access to the Cambridge Structural Database through new workflows. P. Sanschagrin, S. Vyas


LiveDesign: Integrative molecular modeling and cheminformatics for collaborative drug design. E. Davis, T. Garland

Computational Catalyst Design for Energy Conversion & Storage
Advances in Theory, Computational Models & Approaches
Sponsored by COMP, Cosponsored by CATL and CINF

Insights into Structure, Function, Dynamics & Evolution of Enzymatic Mechanisms from Computational Simulation
Sponsored by COMP, Cosponsored by CINF, MEDI and PHYS

Drug Design
QSAR & Docking
Sponsored by COMP, Cosponsored by CINF

Machine Learning for Catalysis Research
Sponsored by CATL, Cosponsored by CINF and COMP

Marriage of Machine Learning, Knowledge Representation & Chemical Sciences

Applied Machine Learning: Molecular Dynamics, Materials & Virtual Screening
Sponsored by COMP, Cosponsored by CINF and PHYS

Large-Scale Applied Molecular Modeling
Sponsored by COMP, Cosponsored by CINF and MEDI

MONDAY AFTERNOON
Section A
New Orleans Marriott Convention Center River Bend 1

Information Legacy of Eugene Garfield: From the Chicken Coop to the World Wide Web
Cosponsored by HIST and PRES

Financially supported by Clarivate Analytics

H. A. Lawlor, W. A. Warr, Organizers
M. Qiu, Organizer, Presiding

1:15 Introductory Remarks.


1:50 49. From the Index Chemicus Registry System to SciFinder and beyond. W.A. Warr

2:20 50. Eugene Garfield: The father of chemical text mining and artificial intelligence (AI) in cheminformatics. R.A. Sayle

2:50 Intermission.

3:00 51. Eugene Garfield’s legacy and its impact on the culture of research. S. Baykoucheva

3:30 52. Beyond citations: What are new ways to assess content that will extend the assessment toolbox? T.A. Carpenter

4:00 53. Novel research and its scientific and technological impact. J. Wang

4:30 54. Clarivate Analytics: Building on the Garfield Legacy with web of science. J. Testa

5:00 Concluding Remarks.

Section B
New Orleans Marriott Convention Center Natzhez

Workflows & Cheminformatics

R. J. Bienstock, G. Landrum, Organizers, Presiding
1:30  55. Quality data to quality models.  

   T. Hesketh, P. Hunt, M.D. Segall, E. Champness, T. Mansley

1:50  56. Development and implementation of Amgen Small Molecule Projects Spotfire Report (ASMPSR) to streamline Hit2Lead and lead optimization process.  

   L. Jia, S. Geuns-Meyer, H. Gao, T.G. Hopper, M. Southern, B. Lanman, Y. Sun

2:10  Intermission.

2:20  57. Automating matched molecular pair analysis of bioactivity and solubility data.  

   F. van den Broek, M. Clark

2:40  58. Reaction and chemistry data blending.  

   M. Fischer, J. Buckley, F. van den Broek

3:00  59. Integrated life science data and the power of workflows.  

   J. Gurinova, D. Digles, G.F. Ecker

3:20  Intermission.

3:30  60. Automated workflows for data curation and standardization of chemical structures for QSAR modeling.  

   K. Mansouri, A. McEachran, C. Grulke, A. Richard, R. Judson, A.J. Williams

3:50  61. Integrated visualization of the research landscape of proteins.  

   J. Gurinova, G.F. Ecker

Computational Catalyst Design for Energy Conversion & Storage

Development of Electro- & Photocatalysts  

Sponsored by COMP, Cosponsored by CATL and CINF

Insights into Structure, Function, Dynamics & Evolution of Enzymatic Mechanisms from Computational

Simulation  

Sponsored by COMP, Cosponsored by CINF, MEDI and PHYS

Drug Design

Molecular Property  

Sponsored by COMP, Cosponsored by CINF

Machine Learning for Catalysis Research  

Sponsored by CATL, Cosponsored by CINF and COMP

Marriage of Machine Learning, Knowledge Representation & Chemical Sciences  

Deep Learning for Deep Chemical Understanding  

Sponsored by COMP, Cosponsored by CINF and PHYS

Large-Scale Applied Molecular Modeling  

Sponsored by COMP, Cosponsored by CINF and MEDI

MONDAY EVENING

Section A

Ernest N. Morial Convention Center  

Halls D/E

Sci-Mix  

E. Alvaro, Organizer

8:00 - 10:00

62. Analysis of food pairings based on food metabolites.  

   R. Reed, T. Neumann


   F. Li
64. Teaching chemical information literacy: A 21st-century skill using 21st-century tools.  
C. Hoffner

S. Wussow, C. Jaeger, M. Buchholz

66. Triangulation of repurposing candidates for orphan diseases.  
J. Gurinova, D. Digles, G.F. Ecker

Y. Djoumbou Feunang, D. Wishart

68. Organizing the pathways and products of nerolidyl diphosphate into a searchable database.  
R. Patnayakuni, C. Hamann

69. Designing CDK2 inhibitors using the molecular chimera approach.  
B. Cook, D. Fourches

70. D-Peptide Builder: A web-based application to enumerate the chemical space of peptides.  
B. Diaz Eufracio, J.L. Medina-Franco, O. Palomino-Hernández

71. De-novo drug design with deep reinforcement learning.  
M. Popova, O. Isayev, A. Tropsha

72. One-click QSAR: A universal approach for developing accurate models.  

TUESDAY MORNING

Section A

New Orleans Marriott Convention Center
River Bend 1

Chemical Information Bulletin, 2018, 70 (1)
11:20 80. Building competency: Scaffolding information literacy skills throughout the chemistry curriculum. **C. Cowden**

11:40 Concluding Remarks.

Section B

New Orleans Marriott Convention Center Natzhez

**Cheminformatics Resources & Software Tools Supporting Environmental Chemistry**
Cosponsored by COMP and ENVR

G. Patlewicz, A. J. Williams, Organizers, Presiding

8:30 81. Predictive computational techniques for chemical risk assessment. **D. Fourches**


9:10 83. OPERA: A free and open source QSAR tool for predicting physicochemical properties and environmental fate endpoints. **K. Mansouri**, C. Grulke, R. Judson, A.J. Williams

9:30 84. WebTEST (Web-services Toxicity Estimation Software Tool). **T. Martin**, V. Tkachenko, A.J. Williams


10:10 Intermission.


11:30 89. Use of 2D chemical structure and bioactivity profiles to generate chemical categories within an Adverse Outcome Pathway network. **M. Nelms**, S.W. Edwards


**Computational Catalyst Design for Energy Conversion & Storage**

Development of Homogeneous & Heterogeneous Catalysts
Sponsored by COMP, Cosponsored by CATL and CINF

**Insights into Structure, Function, Dynamics & Evolution of Enzymatic Mechanisms from Computational Simulation**
Sponsored by COMP, Cosponsored by CINF, MEDI and PHYS

**ACS Award for Computers in Chemical & Pharmaceutical Research:**
Symposium in honor of Jürgen Bajorath
Sponsored by COMP, Cosponsored by CINF and MEDI

TUESDAY AFTERNOON

Section A

New Orleans Marriott Convention Center
River Bend 1

Chemical Information Literacy:
Innovation, Collaboration & Assessment
Cosponsored by CHED

Y. Li, C. Lovitt, G. V. Szymczak, T. M. Vogel, Organizers, Presiding

1:30 Introductory Remarks.

1:35 91. Importance of synthetic speech mark-up language in the teaching of chemistry concepts in a multi-sensory way. C.A. Supalo, A.E. Neybert

1:55 92. Introducing cheminformatics early - prepares your students for success. M. Pozenel


2:55 Intermission.

3:15 95. Innovative instruction program for materials science and engineering undergraduate students. S.J. Redalje

3:35 96. Evolution of natural product total synthesis: Mapping pathways through literature searching. L. McEwen

3:55 97. Instructional scaffolding of information literacy skills in a problem-based learning context. Y. Li, G.V. Szymczak

4:15 98. Teaching information literacy through the chemistry laboratory. C.E. Flener-Lovitt, B. Finley, A. Berger

4:35 Discussion.

4:55 Concluding Remarks.

Section B

New Orleans Marriott Convention Center
Natzhez

Cheminformatics Resources & Software Tools Supporting Environmental Chemistry
Cosponsored by COMP and ENVR

G. Patlewicz, A. J. Williams, Organizers, Presiding


2:10 101. Enhancing exchange of environmental data between EPA and FDA. Y. Borodina, S. Winfield, B. Pruitt

2:30 102. Prediction of pKa from chemical structure using free and open-source tools. V. Tkachenko, N. Cariello, A. Korotcov, K. Mansouri, A.J. Williams

2:50 103. Cheminformatics tools for supporting environmental chemistry. Y. Djoumbou Feunang, D. Wishart

3:10 Intermission.
3:30 104. Investigating ligand and structure-based modeling followed by mixture toxicity prediction of per-and polyfluoroalkyl substances: A virtual screening approach. S. Kar, S. Ghosh, J.R. Leszczynski


4:10 106. Adding complex expert knowledge into chemical databases: Transforming surfactants in wastewater. E. Schymanski, C. Grulke, J. Hollender, A.J. Williams

4:30 107. Prioritizing anthropogenic chemicals in drinking water sources through combined use of mass spectrometry based exposure data and ToxCast toxicity data. A.M. Brunner, M.M. Dingemans, K.A. Baken, A.P. van Wezel

Computational Catalyst Design for Energy Conversion & Storage

Development of Homogeneous & Heterogeneous Catalysts
Sponsored by COMP, Cosponsored by CATL and CINF

Insights into Structure, Function, Dynamics & Evolution of Enzymatic Mechanisms from Computational Simulation
Sponsored by COMP, Cosponsored by CINF, MEDI and PHYS

ACS Award for Computers in Chemical & Pharmaceutical Research: Symposium in honor of Jürgen Bajorath
Sponsored by COMP, Cosponsored by CINF and MEDI

WEDNESDAY MORNING
Section A
New Orleans Marriott Convention Center
River Bend 1

General Papers
E. Alvaro, Organizer, Presiding

8:30 108. Units of Measure Interoperability Service (UMIS): FAIR units for FAIR data. S.J. Chalk, R.J. Hanisch

8:55 109. RA21: Improving access to scholarly resources, from anywhere, on any device. R. Youngen


9:45 Intermission.

10:00 111. Prediction of emission and absorption spectra for Eu²⁺-doped inorganic phosphors based on stoichiometric information. H. Nakano, K. Tanaka, T. Miyao, K. Funatsu, R. Shirasawa, S. Tomiya


**Drug Design**

**Algorithm, Tool & Web Service**
Sponsored by COMP, Cosponsored by CINF

**Structure-Based Drug Design for GPCRs**
Sponsored by COMP, Cosponsored by CINF and MEDI

**WEDNESDAY AFTERNOON**

**Drug Design**

**Molecular Dynamics**
Sponsored by COMP, Cosponsored by CINF

**Structure-Based Drug Design for GPCRs**
Sponsored by COMP, Cosponsored by CINF and MEDI

**THURSDAY MORNING**

**Drug Design**

**Agonist, Antagonist & Macrocycle**
Sponsored by COMP, Cosponsored by CINF

**OTHER SYMPOSIA OF INTEREST:**

**ACS Award for Computers in Chemical & Pharmaceutical Research:** Symposium in honor of Jürgen Bajorath
(see COMP, Tue)

**Marriage of Machine Learning, Knowledge Representation & Chemical Sciences**
(see COMP, Sun, Mon)

**Open Resources for Automated Structure Verification & Elucidation**
(see ANYL, Sun)
CINF 1

Sharing chemical structures from university theses and dissertations on Institutional Repositories and PubChem

Vincent F. Scalfani, vincent.scalfani@gmail.com, Paul Rupar, Mary S. Alexander, Donald G. Williams. (1) Chemistry, University of Alabama, Tuscaloosa, Alabama, United States (2) University Libraries, University of Alabama, Tuscaloosa, Alabama, United States

This presentation will discuss our initial efforts with sharing chemical structure data from The University of Alabama theses and dissertations. We are working with chemistry graduate students and faculty to create machine-readable chemical structure data files (SDFs) alongside submitted theses and dissertations. These SDFs contain the key synthesized substances from the thesis or dissertation and include the associated chemical connection tables, local identifiers, names, SMILES, and InChI line notations. The generated machine-readable SDFs are then shared on our Institutional Repository alongside the original thesis/dissertation, and more broadly on PubChem, allowing for greater management, discoverability, and reuse of the chemical structure data. We will discuss our initial workflows for generating and sharing the SDFs, as well as considerations, challenges, and opportunities for the future.

CINF 2

Impact of stereochemistry on sharing chemical structures

Gregory M. Banik, gregory_banik@bio-rad.com, Muthukumarasamy Karthikeyan, Karl Nedwed, Keith Kunitsky, Michelle D’Souza, Ty Abshear. (1) Bio-Rad Laboratories, Inc., Informatics Division, Philadelphia, Pennsylvania, United States (2) Digital Information Resource Centre, CSIR National Chemical Laboratory, Pune, India

From the chair and boat representations of six-membered rings to Fischer and Haworth projections to pseudo-3D perspective drawings, chemists have been drawing 2D versions of 3D molecules for over a century and a quarter. Humans are able to accurately interpret the implicit 3D information of these 2D chemical structure representations, but the interpretation by
computer software has been less accurate. A related issue is the depiction of chemical structures with implicit stereochemical information using generally accepted drawing conventions that humans understand but software does not (for example, stereochemistry of steroid rings). The explosion of chemical structures in electronic format has heightened these problems as it relates to sharing and interlinking electronic chemical structure resources.

To address the issue, chemical information scientists have used software-based “standardization” of chemical structures. This includes the conversion of Haworth projections and chair and boat representations to regular hexagons with wedge and hashed bonds, sometimes removing or misinterpreting the implicit stereochemistry of the original representation. The alternative to software-based standardization is software-based interpretation, that is, software that accurately interprets the original 3D intent of the 2D drawing in exactly the same way that a chemist would. Case studies of misinterpretation resulting from software-based standardization will be discussed and will be contrasted with software-based interpretation of the same cases using a new technology. Finally, the results of a new external validation study of the technology will also be presented.

CINF 3

Sharing chemical data through a structural database

Suzanna Ward1, ward@ccdc.cam.ac.uk, Amy Sarjeant2, Ian Bruno1, Matthew P. Lightfoot1. (1) The Cambridge Crystallographic Data Centre, Cambridge, United Kingdom (2) Cambridge Crystallographic Data Centre, Piscataway, New Jersey, United States

Depositing and sharing chemical structures within publicly accessible curated databases can present challenges but doing so greatly increases the discoverability of original research and can enable scientists worldwide to derive knowledge from the data.

This presentation will reflect on our experiences in creating the Cambridge Structural Database (CSD), a domain specific database of over 915,000 small molecule crystal structures. We will detail the methods that we use to facilitate data deposition, storage and accessibility and describe the workflows that we have in place to manage data at the CCDC. During the presentation we will highlight important aspects of creating the CSD that help ensure that the database is comprehensive and up to date and that each individual dataset is readily discoverable.

One key aspect of CCDC workflows is integration with publisher workflows that aim to associate data and article from submission through peer review to publication. These integrations allow data to be reviewed alongside the manuscript prior to publication and ensures that datasets are accessible at the point an article is published.

A second important aspect is the assignment of standardised chemical connectivity and 2D structural information from the atomic coordinates in the deposited files using the knowledge in the CSD. This is an essential step for creating a standardised, consistent database of validated structures that ensures researchers’ data are discoverable and can be re-used across disciplines.

We will conclude the presentation by reflecting on what we have learnt from the process of sharing chemical data over the
years, highlight some recent developments at CCDC in this area, and discuss how services could evolve in the future to further support the community in the sharing of their chemical data.

**CINF 4**

**Deposit chemical structures into PubChem - a public data repository**

*Jian Zhang, jiazhang@ncbi.nlm.nih.gov,*  
*Asta Gindulyte, Ben Shoemaker, Paul Thiessen, Evan Bolton. NCBI/NLM/NIH, Bethesda, Maryland, United States*

PubChem is a public data repository specialized in the field of chemical information, bioactivity data for tested chemical substances and more. PubChem’s upload system accepts several chemical structure annotations format including SDF (mol), smiles, and InChI string. The system archives all original chemical structure drawing, known as substance, and at the same time to standardize all chemical entities into compounds which are unique in PubChem’s compound database. PubChem allows users to deposit chemical records, and provides structure search functionality to search the whole database and/or subsets. This presentation will discuss the various chemical structure upload methods and searching hints.

**CINF 5**

**Supporting the community to share chemical structures**

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

We will discuss our experiences in the different aspects of chemical structure sharing, covering ChemSpider, our involvement in the Open PHACTS and PharmaSea projects, and especially our publications. Most recently, we have introduced a new data publishing route via the EPSRC National Chemical Database Service to allow researchers to deposit and publish machine readable chemical structure files and their associated spectra under open licences, in association with a research publication or as an independent dataset. Each submitted structure is validated using the Chemistry Validation and Standardization Platform (CVSP) and assigned a shareable DOI, alongside associated spectra and metadata.

**CINF 6**

**Chemical structures: More than just meta-data?**

*Martin G. Hicks, mhicks@beilstein-institut.de. Beilstein Institut, Frankfurt, Germany*

In terms of communicating chemical research information, chemical structures can be something of a chimera. Bearing in mind the fundamental role of molecules in chemical research this is not only surprising, but looking forward to a more highly interconnected research infrastructure, something that needs to be urgently addressed and improved. In publications we require chemical structures to be human readable, but not necessarily complete, and having normalized and idealized 2D projected coordinates. All trained chemists understand what is meant by these diagrams, but machines have great difficulty reading them correctly. Conversely, taking machine readable structure representations, with correct 3D coordinates, electron densities etc., and turning them into the expected 2D projection for chemists to read is also non-trivial. Yet the workflow from bench or computer, to repository, to publication to database either requires such transformations to routinely work correctly,
or that multiple structure representations to be held in parallel. How can we correct the situation? Do we need a new format, or would this only bring incremental advantages until such a time as our fundamental understanding of chemistry has improved?

**CINF 7**

**Synthetically Accessible Virtual Inventory (SAVI) – Next version: Toward more needles and less haystack**

*Hitesh Patel*¹, hiteshetal379@gmail.com, *Yuri Pevzner*², *Wolf-Dietrich Ihlenfeldt*², *Marc C. Nicklaus*³. (1) Computer-Aided Drug Design Group, Chemical Biology Laboratory, Center for Cancer Research, National Cancer Institute, Frederick, Maryland, United States (2) Xemistry GmbH, Konigstein, Germany

The Synthetically Accessible Virtual Inventory (SAVI) project is an international collaboration to computationally generate a very large number of reliably and inexpensively synthesizable novel screening sample structures. SAVI is premised on the idea that the expansion into synthetically accessible chemical space is more cost-effective than attempting an often nearly impossible synthesis of computer-proposed structures on a per-project basis. SAVI's strength lies in the three of its major components: A set of richly annotated computer-encoded chemical transforms going beyond reactant substructure matching and incorporating the chemical context of each virtual reaction; a highly annotated and set of readily available starting materials (building blocks); and a powerful chemoinformatics toolkit putting the first two together.

For the SAVI version released as of the time of this writing, nearly 300 million novel structures, each annotated with a one-step synthetic route and a set of properties important in current drug design, were created *in silico*, using about 377,000 building blocks and 14 highly annotated transforms. We will present updated transforms, an expanded building block set, and other advances in the SAVI methodology and strategy to improve the reliability of the proposed synthetics routes, i.e. to reduce the “haystack” of synthetically non-tractable molecules in SAVI. We will also briefly touch on a planned SAVI GUI with structure and property search capabilities. We will present the current status as well as scientific and technical challenges of the SAVI project.

**CINF 8**

**BioTransformer: An accurate, freely available tool for predicting small molecule metabolism—applications in drug discovery**

*Yannick Djoumbou Feunang*¹, djoumbou@ualberta.ca, *David S. Wishart*¹,². (1) Biological Sciences, University of Alberta, Edmonton, Alberta, Canada (2) Computing Science, University Of Alberta, Edmonton, Alberta, Canada

Metabolism has long been recognized as a significant factor in the success of generally time- and cost-expensive drug development projects, as it impacts the pharmacology, bioavailability, and toxicology of drugs. In fact, the metabolism of a drug can produce metabolites that have significantly different pharmacological and/or toxicological profiles, compared to their parent. Understanding how chemicals are metabolized within humans is therefore crucial to filter out bad candidates in early drug development phases, and reduce attrition rates. Over the last two decades, *in silico* prediction has raised increasing interest in the drug industry, and thus, sparked the development of several
software tools and web servers for rapid, accurate, and free or low-cost prediction of small molecule metabolism. Unfortunately, even with the growing abundance of in silico metabolism prediction tools, there continue to be a number of significant limitations, especially with regard to their performance, their utility and their accessibility. In particular: 1) very few tools predict more than the sites of metabolism; 2) only a small number of tools provide predicted structures, and those that do generally place restrictions on their distribution; 3) very few of the tools make their databases or training sets available; 4) none of the comprehensive prediction tools are freely available; 6) many tools seriously over-predict metabolites and have remarkably high false positive rates; and 7) none of the tools combine phase I, phase II, gut microbial metabolism, and promiscuous metabolism together. To address these issues, we have developed BioTransformer, a freely available software tool for the prediction of small molecule metabolism.

BioTransformer combines a knowledge-based approach and machine learning to predict the metabolism of small endogenous and exogenous molecules in humans (phase I and phase II) as well as their gut microbiota. Moreover, it integrates a module for predicting the microbial degradation of small molecules in soil and aquatic environments. In this presentation, I will provide a detailed description of BioTransformer, and some of its most recent applications, including the expansion of the DrugBank and HMDB databases, as well as a collaborative study of polyphenol and terpenoid metabolism by humans and their gut microbiota. BioTransformer is freely available at https://bitbucket.org/djoumbou/.

CINF 9

CCCTK (Compute Cure for Cancer ToolKit) An open source drug discovery platform for design of novel anti-cancer agents

Muthukumarasamy Karthikeyan¹, karthinc@gmail.com, renu vyas². (1) Digital Information Resource Centre, CSIR National Chemical Laboratory, Pune, India (2) Bioengineering, ADT University, Pune, Maharashtra, India

We developed an opensource tool for Drug Discovery platform to handle BigData (Cancer Specific Molecular, Literature, Clinical, Genomic, Proteomic, Metabolomic, Transcriptomics data from publicly available FDA, NIH, NCI etc resources) automatically in a High Performance computing environment for the design of novel anti-cancer drugs utilizing machine learning and artificial intelligent systems. In the past two decades we have established chemoinformatics infrastructure and opensource based specialized molecular informatics applications for the design of novel molecules with potential bioactivities, derived from experimental data using computational simulation and modeling approach. Even though the computational and informatics tools were developed in isolation for specific purposes, and commercial tools are not able to meet current requirement of integrated approach with multi-dimensional data (numeric, textual, image, compressed, encrypted, both in structured and un-structured formats, RDBMS etc) . The opensource compute cure for cancer platform is compatible with high performance computing system to handle such large scale multi-dimensional data and efficient algorithms using deep-learning and machine learning architecture. The above system would also use most of the
open-source tools and publicly available databases directly from respective resources dynamically. The system will be made open-source (via Sourceforge / git) for the benefit of researchers who can add value to the software with additional modules to handle complex tasks in the background. The user friendly GUI would be provided with workflow pipeline to handle both sequential and parallel tasks. The current prototype is capable of handling all chemoinformatics tasks (chemical name to structure, Generation of molecular scaffolds and functional groups from drugs, leads related to treatment of cancer, generation of large collection of virtual library of novel molecules from the collection of known drugs and lead molecules for further studies.

CINF 10

Practical machine learning methods for QSPR and QSAR predictions

Valery Tkachenko, tkachenko.valery@gmail.com, Alexander Korotcov, Rick Zakharov, Boris Sattarov, Artem Mitrofanov. Science Data Software, LLC, Rockville, Maryland, United States

In the past several years machine learning techniques have played an important role and become absolute necessity in the modern drug discovery process. Multiple methods for predicting physicochemical and chemo-biological endpoints have proven their robustness and significantly improve our current state of understanding of molecular features/properties associated with some specific pharmacological features. Despite a good number of drug discovery supporting toolkits and methods available to public, academy and pharma there is a demand to have a tool which can combine mining/curation of the heterogeneous chemical data and multiple sophisticated molecular machine learning algorithms. This kind of toolkit have to be able to train models using a variety of machine learning algorithms with minimum user intervention or/and have access to a ready to use pre-trained models. In this study we have evaluated our toolbox (Open Science Data Repository, currently under development) for data curation and machine learning modelling for drug discovery. Different heterogeneous publicly available datasets related to Tuberculosis, Malaria, Bubonic plaque, Chagas disease, and others have been used to tune and train multiple machine methods including traditional methods such as Naïve Bayes, k-Nearest Neighbors, Random Forest, Boosted Decision Trees, Regularized Logistic Regression, and Support Vector Machines, as well as novel deep learning methods with Neural Networks models of different complexity. A wide range of model evaluation metrics such as Receiver Operating Characteristic, Area Under Curve, F1-score, Cohen’s kappa, Matthews correlation coefficient have been used to evaluate and compare machine learning models performance. A variety of commonly used in cheminformatics molecular descriptors for compounds representation was built in our methods, thus an additional layer of tuning by searching of the best molecular descriptor for a particular model can be used. Most of the models performed pretty well and the developed workflows are
ready to be used for QSPR and QSAR. Moreover, all already tuned and trained models from this study are ready to use for public and can be found on https://figshare.com/s/0286924045d50441bf98. We strongly believe that the modern in silico approaches combined with advances in data mining, curation, and machine learning methods will only accelerate the drug discovery processes.

CINF 11

Making virtual REAL: Synthetically feasible compounds and their exploration in docking screens

Yurii Moroz, ysmoroz@gmail.com, ChemBioCenter, National Taras Shevchenko University of Kyiv, Kyiv, Kyiv, Ukraine

Molecular docking is limitless in selecting compound sets to screen because both in-stock and virtual compounds are dockable. While virtual molecules cover larger chemical space than the stock ones and represent a pool of novel structures and chemotypes, the major issue with the virtual entities, however, is whether the molecules can be accessed synthetically within the acceptable time period at the reasonable price.

A solution to the virtual problem is to utilize in-stock qualified building blocks into in-house validated reactions. We have called this synthesizable space – REAL database (REAL = readily accessible). REAL database comprises derivatives of 120 thousand stock building blocks combined via 119 reaction schemes that have been tested in more than 3 million experiments affording 2.3 million stock compounds. REAL database covers most of the medicinal chemistry relevant chemical space including REAL building blocks, REAL fragments, REAL drug-like, REAL covalent modifiers, REAL PPI modulators. Common features of the compounds within the database enable the transition from the fragments to the lead-like or the drug-like molecules simplifying optimization of a hit and generation of hit follow-up libraries. We have demonstrated the applicability of REAL compounds to initial hit finding in several projects.

CINF 12

Deep learning approaches for detecting high-throughput screening false positives

Matthew Matlock², matt.matlock@gmail.com, Tyler Hughes², S Joshua Swamidass¹,². (1) Washington University, Saint Louis, Missouri, United States (2) Pathology and Immunology, Washington University in St. Louis, Saint Louis, Missouri, United States

Drug discovery programs often begin with high throughput assays (HTS) that screen vast chemical libraries to uncover potential scaffolds, from which medicinal chemists formulate drugs or biochemical probes. Unfortunately, a large proportion of compounds recovered in these screens are later proved to be false positives. These compounds often exhibit broad, non-specific activity in many assays due to undesirable chemical interactions with assay components. This has led to publications identifying “selective” probes.
that are later shown to be promiscuously active against many targets. These bad behaviors come in a variety of forms, including covalent reactivity, redox cycling, fluorescence, and aggregation. Identifying and removing these problematic compounds from screening libraries is critical to improve the success of these experiments, the quality of chemical libraries, and the availability of selective chemical probes.

While properly controlled experiments are critical for characterizing these false positives, *in silico* techniques provide chemists with additional tools to use in selecting candidates from HTS assays. Pan-Assay Interference Compound (PAINs) filters, a set of substructure filters, is the most commonly used *in silico* technique. While useful, PAINs filters cannot provide mechanistic explanations of assay interference. We show that a deep learning model of small-molecule covalent reactivity with biological macromolecules can be repurposed to detect false positives in screening assays. This model compliments PAINs filters in two key ways. First, the model can discriminate between false positives and non-promiscuous compounds among compounds matching the same PAINs filter. Second, the model can identify cases where observed assay interference is due to chemical structures not associated with a PAINs filter. Overall, our analysis suggests that deep-learning approaches to modeling promiscuous behavior may complement current methods for identifying PAINs.

CINF 13

Virtual screening and fast resynthesis of hits: Anchor.Query

Alexander Doemling¹, a.s.s.domling@rug.nl, David Koes³, Carlos J. Camacho². (1) Department of Drug Design, University of Groningen, Groningen, Netherlands (2) University of Pittsburgh, Pittsburgh, Pennsylvania, United States (3) Computational and Systems Biology, University of Pittsburgh, Pittsburgh, Pennsylvania, United States

AnchorQuery (http://anchorquery.csb.pitt.edu) is a web application for rational structure-based design of protein-protein interaction (PPI) inhibitors. A specialized variant of pharmacophore search is used to rapidly screen libraries consisting of more than 31 million synthesizable compounds biased by design to preferentially target PPIs. Every library compound is accessible through one-step multi-component reaction (MCR) chemistry and contains an anchor motif that is bioisosteric to an amino acid residue. The inclusion of this anchor not only biases the compounds to interact with proteins, it also enables a rapid, sub-linear time pharmacophore search algorithm. AnchorQuery provides all the tools necessary for users to perform online interactive virtual screens of millions of compounds, including pharmacophore elucidation and search, and enrichment analysis. The platform will be exemplified with the discovery of allosteric inhibitors of PDK1.
CINF 14

Sharing chemical structures with peer-reviewed publications. Are we there yet?

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

In the domain of chemistry one of the greatest benefits to publishing research is that data can be shared. Unfortunately, the vast majority of chemical structure data associated with scientific publications remain locked up in document form, primarily in PDF files or trapped on webpages. Despite the explosive growth of online chemical databases and the overall maturity of cheminformatics platforms, many barriers stifle the exchange of chemical structures via publications. These challenges include incomplete support by accepted standards (especially InChI) for advanced stereochemistry, organometallic compounds and generic “Markush” representations, the difference between human-readable and computer-readable forms of data, and challenges with the computer representation of chemical structures. To address these obstacles to chemical structure sharing, US EPA National Center for Computational Toxicology scientists are using a combination of cheminformatics applications and online repositories to distribute chemical structure data associated with their publications. This presentation will describe how EPA-NCCT chemical structure data that is amenable to indexing and distribution are shared and highlight the benefit of open data sharing for modeling, data integration, and increasing research impact.

CINF 15

Integration of Markush structures into EPA’s DSSTox database to represent and enumerate UVCB substances

Christopher Grulke, grulke.chris@epa.gov, Antony J. Williams, Ann Richard. National Center for Computational Toxicology, Environmental Protection Agency, Wake Forest, North Carolina, United States

Many chemicals of interest in the domain of environmental chemistry fall in the class of substances referred to as Unknown, Variable composition, Complex reaction product or Biological origin (UVCB) substances. These chemicals generally cannot be represented with a single structure and, therefore, pose a significant challenge when trying to accurately characterize and share research results via chemical databases that rely primarily on distinct chemical structures to index and merge substance information. For this reason, it is not uncommon for public chemical databases to include questionably mapped structures associated with names or registry numbers that indicate a UVCB. To provide a more accurate documentation of the substance and clarify these data linkages, EPA’s DSSTox project has started employing Markush representations to represent UVCBs. Whereas the use of a Markush representation is more accurate than a single “representative” structure, it yields a new set of problems, including difficulty with determining uniqueness, storage format limitations, and inconsistencies in different software’s handling of the representations. It has, however, enabled the automated linking of a UVCB with the associated, well-defined chemical components which can be auto-enumerated within the software package employed. It also provides an interpretable depiction of the substance,
and clarifies the linkage of research results to the appropriate chemistry. Additionally, it greatly increases the efficiency of collecting information associated with all components of a UVCB from within EPA databases, providing the ability to quickly identify and gather the compendium of data associated with either the UVCB or any of its structural components.

CINF 16

Curating and sharing structures and spectra for the environmental community

Emma Schymanski\textsuperscript{2}, emma.schymanski@eawag.ch, Antony J. Williams\textsuperscript{1}. (1) National Center for Computational Toxicology, Environmental Protection Agency, Wake Forest, North Carolina, United States (2) Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Belvaux, Luxembourg

The increasing popularity of high mass accuracy non-target mass spectrometry methods has yielded extensive identification efforts based on spectral and chemical compound databases in the environmental community and beyond. Increasingly, new methods are relying on open data resources. Candidate structures are often retrieved with either exact mass or molecular formula from large resources such as PubChem, ChemSpider or the EPA CompTox Chemistry Dashboard. Smaller, selective lists of chemicals (also called “suspect lists”) can be used to perform more efficient annotation. Mass spectral libraries can then be used to increase the confidence in tentative identification. Additional metadata (e.g. exposure and hazard information, reference and data source information) can be extremely useful to help identify substances of high environmental interest. Exchanging information and “sharing structural linkages” between these resources requires extensive curation to ensure that the information is shared correctly, yet many valuable datasets arise from scientists and regulators with little official cheminformatics training. This talk will cover curation efforts undertaken to map spectral libraries (e.g. MassBank, EU, mzCloud) and suspect lists from the NORMAN Suspect Exchange (http://www.norman-network.com/?q=node/236) to unique chemical identifiers associated with the US EPA CompTox Chemistry Dashboard. The curation workflow takes advantage of years of experience, as well as contact with the original data providers, to enable open access to valuable, curated datasets to support the environmental community and scientists beyond (e.g. https://comptox.epa.gov/dashboard/chemical_lists). This work enables sharing high quality open data with the community for reuse and repurposing.

CINF 17

ChemAxon’s chemical structure file formats for better data storage and search

Aurora D. Costache\textsuperscript{1}, aurora.costache@gmail.com, Eufrozina A. Hoffmann\textsuperscript{2}, András Vollford\textsuperscript{2}. (1) ChemAxon, Lexington, Massachusetts, United States (2) ChemAxon, Budapest, Hungary

There are a variety of file formats for storing information about molecular structures, but one of the major challenge is to store all the necessary parameters which describes effectively a chemical molecule, or even a reaction, without losing important information. That can further facilitates accurate search between data stored in a database for example. In this presentation, we introduce ChemAxon’s solution for better storage
and exchange of information on molecular structures and properties. We will focus on two file formats: ChemAxon Extended SMILES and Marvin Document (MRV). ChemAxon Extended SMILES is a line notation which extends the well-known SMILES notation with features which are not represented in the regular SMILES strings. One of the advantages of extended SMILES file format from ChemAxon is the ability to store enhanced stereochemical representation information which offers accurate searches. It is developed keeping in mind that the generated string should be readable by old SMILES capable systems which are not familiar with the new features. Of course in this case the features stored in the extended part are not usable for such systems, but still the features supported by basic SMILES format are there and valid.

On the other hand the MRV (Marvin Document) file format based on the open source CML (Chemical Markup Language) format, provides the ability to store all the molecule structure information, all the relevant query information, rendering parameters, text and graphical elements. Both file formats from ChemAxon are human readable text formats that allow users to check the stored information without any additional converters.

CINF 18

NCI/CADD CACTUS web server: Tool for connecting chemical structures

Marc C. Nicklaus, mn1@helix.nih.gov. Nci Frederick Bldg 376 RM 207, Nati Inst Health Ft Detrick, Frederick, Maryland, United States

We present select services of the NCI CADD Group’s web server CACTUS (CADD Group Chemoinformatics Tools and User Services). The Chemical Identifier Resolver (CIR) functions as a resolver for a large number of chemical structure identifiers. It provides conversion of an input structure identifier into another structure identifier or a variety of representations, ranging from identifiers such as InChI[Key] through structure drawings to molecular properties such as molecular weight. CIR can be used via a web form or a simple URL API. The Chemical Structure Lookup Service (CSLS) provides the user with information about, and links to, occurrence of an input structure in more than 100 databases, both public and commercial. These and other CACTUS services can help in working with structure files and notations, thereby aiding chemistry students, educators, researchers, and librarians in their sharing and discovery of chemical structures.

CINF 19

Sharing chemical structures: Workflows, demonstrations, and discussion

Vincent F. Scalfani, vincent.scalfani@gmail.com, Leah McEwen, lrm1@cornell.edu, Ye Li, yeli@mines.edu, Emma Schymanski, emma.schymanski@eawag.ch, Aurora Costache, acostache@chemaxon.com, Eufrozina A. Hoffmann, ehoffmann@chemaxon.com, András Volford, avolford@chemaxon.com, Antony J. Williams, tony27587@gmail.com, Christopher Gruulke, grulke.chris@epa.gov, Jian Zhang, jiazhang@ncbi.nlm.nih.gov, Suzanna Ward, ward@ccdc.cam.ac.uk, Guy Jones, jonesg@rsc.org, Gregory M. Banik, gregory_banik@bio-rad.com. (1) Arthur Lakes Library, Colorado School of Mines, Golden, Colorado, United States (2) Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Belvaux, Luxembourg (3) National Center for Computational Toxicology, Environmental...
This session will highlight and demonstrate various workflows for sharing chemical structures as machine-readable files and/or depositing chemical structures into databases. We will explore a variety of software, web services, databases, and strategies that help researchers create, organize, store, and share chemical structures. Presenters from the Enhance Discovery: Share Chemical Structures session are invited to demonstrate their own chemical structure creation, sharing, standardization, and/or validation workflows. Attendees are encouraged to bring a laptop or tablet if seeking to participate interactively (when possible). We will conclude with a discussion and identify chemical structure sharing opportunities for the future.

**CINF 20**

**Global structure diversity and chemical space of food chemicals**

Nicole N. Trujillo-Minero, Jose L. Medina-Franco, jose.medina.franco@gmail.com. Universidad Nacional Autonoma de Mexico, Mexico City, Mexico

It has been recognized that the chemoinformatic analysis of food chemicals is limited even though public data bases are available. Computational methods, resources, and concepts already developed could be used for this analysis. Herein we discuss the insights of a comprehensive chemoinformatic analysis of a major public food chemical database with more than 2000 compounds. The global chemical space of the database was compared with an FDA approved compounds database and a FDA approved compounds database for diabetes treatment using a PCA analysis of physicochemical properties. To analyze chemical structures and diversity, they were considered three major criteria: physicochemical properties, molecular scaffolds and fingerprints of different design. The evaluation was conducted with two open servers recently developed in our group: Platform for Unified Molecular Analysis (PUMA) and Consensus Diversity Plots (CDPlots). These applications are part of the DIFACQUIM Tools for Chemoinformatics (D-Tools) available at www.difacquim.com.

The conclusions of this study that will be presented at talk represent a further advance in the field of Foodinformatics.
CINF 21

**Food additives data integration in PubChem**

**Jian Zhang**, jiazhang@ncbi.nlm.nih.gov, **Paul Thiessen**, **Asta Gindulyte**, **Evan Bolton**. **NCBI/NLM/NIH, Bethesda, Maryland, United States**

Food additives are substances that added to food to perform various functions such as to improve and preserve the taste, texture, nutrition and appearance of food. For centuries, people used salt to preserve meats, added herbs and spices to improve the flavor of foods. There are thousands of substances that are used as food additives. Today, agencies maintain food additives databases in several countries. Pubchem, as a chemical information hub, has integrated food related chemical information from several resources including FDA CFSAN, EU Food Improvement Agents, FAO/WHO JECFA. This presentation will discuss how we integrate the data into PubChem from various data formats, and the data accessing methods.

CINF 22

**Using structural data in the food and fragrance industries**

**Suzanna Ward**, ward@ccdc.cam.ac.uk, **Amy Sarjeant**, Shyam Vyas, **Philip I. Andrews**. (1) **The Cambridge Crystallographic Data Centre, Cambridge, Cambs, United Kingdom** (2) **Cambridge Crystallographic Data Centre, Piscataway, New Jersey, United States**

Thanks to the exemplary approach to data sharing by the crystallographic community the Cambridge Structural Database (CSD) now contains over 915,000 entries and the rate of growth continues to rise. This rich resource of organic and metal-organic crystal structures helps us to expand our knowledge across many research fields and is already used by some in the food industry. This short talk will explore what food and fragrance molecules are currently in the CSD and attempt to identify the overlap between compounds in the CSD and public databases of food and fragrance molecules. Along the way we will highlight some of the fascinating structures we find, from the different polymorphs of chocolate to the structures derived from fine wines.

We will go on to look at how our structural informatics tools could be used on some of these molecules and explore how the food and fragrance industries could use these tools to gain new insights to help create the next food sensation or number one selling scent. We will conclude by looking to the future and asking you what we can do to help these industries benefit more from cheminformatics methods.

CINF 23

**Predicting human olfactory perception from molecular structure**

**Joel Mainland**, jmainland@monell.org. (1) **Monell Chemical Senses Center, Philadelphia, Pennsylvania, United States** (2) **Neuroscience, University of Pennsylvania, Philadelphia, Pennsylvania, United States**

In vision and hearing, the wavelength of light and frequency of sound are highly predictive of color and tone. In contrast, predicting the smell of a novel molecule from its chemical structure has not been solved. In other words, there is not a scientist or perfumer in the world who can view a novel molecular structure and predict how it will smell. Using models based on physicochemical properties, we can predict the intensity of a novel...
molecule at any concentration \( r = 0.89, p < 0.001 \), and the pleasantness \( r = 0.71, p < 0.001 \) and quality \( r = 0.55, p < 0.001 \) of single molecules in external validation. In summary, current models are now capable of looking at a novel molecular structure and predicting how it will smell.

**CINF 24**

**Using molecular fields to understand molecular determinants of the olfaction process**

*Martin Slater*, martin@cresset-group.com. Cresset Discovery Services, Litlington, Cambridgeshire, United Kingdom

The use of molecular fields in pharmaceutical discovery is well established for virtual screening, scaffold hopping, and QSAR. Uniquely, they offer the ability to break molecular recognition down into electrostatic, shape and hydrophobic components that can inform all aspects of ligand-based design. Unlike fingerprint or other 2D descriptors, molecular fields offer the opportunity to generate generic, target-based models of activity that encompass multiple ligand chemistries. As such the additional cost of working in 3D (the need to consider conformation) is compensated by the wider applicability and robustness of the models obtained.

The olfaction process is governed by a similar set of recognition events that operate in the biological targets for pharmaceutical discovery. However, fragrance molecules generally are restricted by a unique set of properties that separate them in chemical space from traditional pharmaceutical ligands. They are smaller and more compact, often with a higher fraction of \( \text{Sp}3 \) atoms, and tend to be feature poor. Modeling their activity becomes challenging as traditional pharmacophores are not detailed enough to capture the subtleties and 2D fingerprints, are too sensitive to changes in chemistry, and immune to the changes in chirality that are fundamental to fragrance perception. In this respect, modeling fragrance molecules brings many of the same challenges that are faced in fragment-based drug discovery.

I will show that using molecular fields can closely relate molecules with similar biological response regardless of underlying chemistry. I will look at the classification of molecules based on their electrostatic and shape properties and show that we can discriminate between them. Advantages and problems of this approach will be discussed, and parallels drawn with the modeling of fragments in drug discovery, showing where the Cresset approach can be beneficial.

**CINF 25**

**Designing new fragrance ingredients meeting specific technical requirements**

*David T. Stanton*, stantondt@cinci.rr.com, Johan Smets, An Pintens, Hugo Denutte. (1) Fabric & Home Care, Procter & Gamble, Cincinnati, Ohio, United States (2) Fabric & Home Care, Procter & Gamble, Brussels, Belgium (3) Flavor & Fragrance, F&HC, Procter & Gamble, Brussels, Belgium

Consumer product formulations are complex and designed to meet demanding performance requirements related to characteristics such as cleaning, substrate modification, and shelf life. The performance requirements for fragrance ingredients used in these products are similarly demanding. They must delivered to the required consumer touch point and provide the intended benefit such malodor control or imparting freshness. However it is not always possible to find
existing fragrance ingredients that meet all these requirements for a given new consumer product. Thus, potential new fragrance ingredients were designed using computational chemistry models to meet specific technical requirements. The models used and the molecules designed will be described, and their technical performance will be outlined.

CINF 26

QSAR/QSPR models to support fragrance ingredient molecular design

Dmitriy Chekmarev, Jeffrey Kattas, Jack A. Bikker, jack.a.bikker@iff.com. Research & Development, International Flavors and Fragrances, Union Beach, New Jersey, United States

Fragrance ingredients are used in creations that range from fine perfumes to common household products such as cleaners and cat litter. The performance of each ingredient is determined by the physical and psychophysical properties of the ingredient. Intense ingredients with high vapor pressure provide high impact that rapidly dissipates. In contrast, ingredients with lower intensity and vapor pressure may have a more sustained contribution to scent performance. In choosing to design or select a new ingredient, the synthetic chemist will consider both the desired odor character and the intended performance profile. To support this effort, models of such properties as vapor pressure, solubility, Henry’s Law, biodegradation, and ingredient intensity have been developed. Various algorithms, descriptor spaces, and software platforms (e.g. Biovia Pipeline Pilot, Optibrium Auto-Modeller, R) have been applied to internal and external data sets to generate and validate models. A comparison of select outcomes will be discussed.

CINF 27

Translating methods from pharma to fragrances and flavors

Tamsin Mansley, tamsin@optibrium.com, Edmund Champness, Peter Hunt, Nick Foster, Matthew D. Segall, matthew.d.segall@gmail.com. Optibrium Ltd, Cambridge, United Kingdom

The pharma sector has generated a wealth of experience in cheminformatics methods that are used in the optimisation of small, ‘drug like’ molecules. While there are differences in the chemistries used to develop flavors and fragrances and the optimisation objectives of these projects, many computational methods can be translated from pharma to guide the design and selection of compounds in this context and improve efficiency and productivity. The properties that describe molecules in these fields are typically different, but both disciplines have the goal of quickly targeting compounds with a balance of properties for the project’s objectives.

We will discuss approaches to compound selection and design, including chemical space analysis, property prediction and multi-parameter optimisation, comparing and contrasting datasets and models from pharma with those in flavors and fragrances. This will be illustrated by case studies to build and apply robust QSAR models predicting relevant properties, design and prioritisation of new compound ideas and analysis of chemical spaces for selection of compounds, using fragrances and flavors datasets.

CINF 28

Predicting color stability of fragrance ingredients in consumer products based on molecular structure
Fragrances often comprise relatively diverse sets of compounds. The components of a consumer product chassis can be similarly diverse. The color stability of fragranced consumer products is a persistent problem, and efforts to identify the compounds involved are complicated by all this diversity. Thus, methods were investigated to predict which components of a fragrance were related to the color instability from the molecular structure of the fragrance ingredient. A model will be described that classifies fragrance ingredients according to the likelihood that they will cause color instability in common consumer product chassis.

CINF 30
What's all this fuss about data sharing? In search of improved laboratory safety

Carmen I. Nitsche, cnitsche@swbell.net. Pistoia Alliance/CINforma Consulting, Long Branch, New Jersey, United States

Data sharing remains a difficult proposition in many settings. While not all data is proprietary or in need of confidentiality, we tend to work in environments where it is much easier to just say no, rather than figure out how to get to a yes on data sharing. Is safety data different? Should we feel an obligations to share in the name of safety? In this paper we will consider some examples of community sharing projects and set the stage for the panel discussion.

CINF 31
Promoting safety culture through sharing - a Dow perspective

Mark E. Jones, acs_mj@mjphd.net, Lori Seiler, Cathy Mapes. (1) R&D, The Dow Chemical Company, Midland, Michigan, United States (2) The Dow Chemical Company, Midland, Michigan, United States (3) TECS/Fiber and Polymer Science, North Carolina State University, Raleigh, North Carolina, United States

Sudan I, 1-phenylazo-2-naphthol (C.I. Solvent Yellow 14 and Solvent Range R) is a synthetic azo dye which is used as a coloring agent. Its use in food products is illegal, as it is classified as a mutagen, however it has frequently been found in food products such as chili powder, paprika, curry powder, saffron, and palm oil (M. Coulet, Food and Chemical Toxicology 48 (2010) S106–S111 ). Between 2003 and 2005, over 600 food products containing Sudan dyes were recalled in the UK. The international agency for research on cancer (IARC) has found limited experimental evidence for carcinogenicity and there is no adequate data in humans. NTP experimental study tests have yields mixed results. We will report computational modeling and analysis of Sudan dye-DNA interactions and interactions with DNA repair enzymes to shed some light on mechanisms of Sudan dye genotoxicity.
Dow has had a long history of innovation around safety culture and policy. It promotes collaboration and sharing as a way for continuous improvement in the safety arena. We will discuss our efforts to support academia in the safety realm, and how we navigated the legal and regulatory landscape to offer this community support.

CINF 32

Chemical and laboratory safety: The role of scholarly publishers

Carlos Toro, c_toro@acs.org, Sarah B. Tegen. American Chemical Society, Washington, District of Columbia, United States

Promoting a culture of safety within a research environment should be a central priority for everyone involved in the design and execution of scientific experiments. The devastating aftermath that follows a laboratory accident underscores the importance of setting clear safety guidelines that all stakeholders support, implement, and audit regularly. As a scholarly publisher, the American Chemical Society is keenly aware of its role in promoting safety. We will briefly discuss what role editors, authors, and reviewers of scholarly publications serve in guiding safety behaviors, and some the initiatives ACS Publications has undertaken to support the safe practice of chemical sciences around the globe.

CINF 33

Chemical safety data in the Handbook of Chemistry and Physics

John Rumble1, jumbleusa@earthlink.net, David R. Lide2, Fiona Macdonald3. (1) R&R Data Services, Gaithersburg, Maryland, United States (2) Taylor and Francis, Boca Raton, Florida, United States (3) CRC Handbook of Chemistry and Physics, Gaithersburg, Maryland, United States

The rich variety of data important to the safe use of chemicals makes it difficult to summarize chemical safety concisely. For a number of years, the CRC Handbook of Chemistry and Physics has drawn on expert advice to provide a set of high quality of tables that provide easy access to many different data types. Fields covered include typical chemical safety concerns – flammability, water reactivity, explosion (shock) hazards, chemical carcinogens, and chemical incompatibility. Because chemical safety also occurs in laboratory settings with advanced instruments, physical data on laser hazards and ionizing radiation are given. As good chemical safety practices start with using correct procedures and appropriate equipment, detailed information is given on selection of laboratory gloves, garments and respirators as well as equipment as such gas cylinders, fume hoods, and biosafety cabinets. Recently a section of nanomaterial safety has been added. The variety of information makes the CRC Handbook both an important introduction to chemical safety for entering students and new researchers and a trusted source for practitioners and persons starting new research directions. Most of the data tables contain extensive references to additional data resources with more detailed information. The availability of the CRC Handbook in both print and online formats allows it to reach the world-wide chemistry community. Sharing these data, which have been reviewed by acknowledged experts and taken from authoritative resources, enables diverse communities to share factual information in an effective manner.
CINF 34

Safety sharing culture: Learning from the aviation industry

Thomas Zoeller, tom.zoeller@csb.gov. U.S. Chemical Safety and Hazard Investigation Board, Washington, District of Columbia, United States

As the chemical industry wrestles with issues of safety data sharing, it is worth considering that other industries have faced similar challenges and have found ways to collaborate for the benefit of the wider community. We will consider how the aviation industry uses data collection and information sharing, such as the Aviation Safety Information Analysis and Sharing (ASIAS) program, to be an effective tool for proactively addressing safety issues.

CINF 35

Information/practice sharing forums among American Chemistry Council Member Companies

Irene McGee¹, irene.mcgee@covestro.com, David Sandidge². (1) Covestro LLC, Pittsburgh, Pennsylvania, United States (2) American Chemistry Council, Washington, District of Columbia, United States

Participation in Responsible Care is a condition of membership for American Chemistry Council (ACC) members and Responsible Care Partner companies. ACC member and Responsible Care Partner companies pledge to improve environmental, health, safety and security (EHS&S) performance for facilities, processes and products throughout the entire operating system. Responsible Care companies have an employee safety record that is five times better than the average of the U.S. manufacturing sector as a whole, and almost three times better than the business of chemistry overall. Member companies also are committed to open and transparent reporting and submit annual data on their progress toward meeting performance measure goals. This information is made publicly available on the ACC website.

In 2015 the ACC formed a taskforce and expert panels to develop recommendations to address continual improvement in industry performance in the areas of process safety and occupational safety. A key recommendation was for the formation of Information/Practice Sharing Forums, where ACC members could share not only incident learnings, but good and effective practices that individual companies have implemented, all shared in a legally safe environment. In 2017, two forums were completed, one for Process Safety and one for Occupational Safety. These forums will continue in an effort to enhance safety and recognize that an effective means to improve safety performance throughout the industry is to learn from each other.

This paper will describe the framework the ACC put together so that these sharing sessions could occur, as well as the benefits of the sharing sessions for the member companies.

CINF 36

Experiences with learning experience reports

William B. Tolman, wtolman@umn.edu. Chemistry Dept, University of Minnesota, Minneapolis, Minnesota, United States

At the University of Minnesota, efforts to enhance the culture of safety in research laboratories have been led by the Joint Safety Team (JST), an organization developed and run by students and postdoctoral associates
in the Departments of Chemistry and Chemical Engineering & Material Science. Recognizing the need to better share experiences of incidents/situations in the laboratories (including but not limited to “near misses” of accidents), the JST implemented an anonymous reporting system whereby community members provide “learning experience reports.” I will discuss how the program was initiated, how the reports are managed, examples of reports made, how these reports are disseminated, difficulties encountered in convincing students to submit them, and feedback received about them.

CINF 37

Balancing act: Protecting all interests

Steve Addlestone, saddleston@eastman.com, Stacia Christman. Eastman Chemical Company, Kingsport, Tennessee, United States

Having a strong safety culture requires learning from incidents and near misses. Although it is important to share this information at least internally, legal concerns sometimes require an approach that balances competing interests, such as enforcement concerns. There still can be room to learn from past events while still acknowledging these legal issues.

CINF 38

Parsing the “lessons learned” space: Layers of opportunities and challenges

Ralph Stuart, rstuartcih@me.com. Dept of Env Hlth Safety, Keene State College, Keene, New Hampshire, United States

There are many different types of laboratory incidents we can learn from, ranging from “near misses” to full blown Chemical Safety Board root cause investigations. This presentation will propose a taxonomy for these events and identify the information needs to support development of different types of Lessons Learned based on the audiences the Lessons will be shared with.

CINF 39

Open-source web tools for modeling and design tracking: Workflows facilitating collaborative drug discovery

Katrina W. Lexa, lexa@dnli.com, Jianwen A. Feng. Denali Therapeutics, San Francisco, California, United States

Placing modeling tools in the hands of synthetic chemists provides a mechanism for immediate feedback on proposed compounds, resulting in rapid refinement of nascent designs, and thus increasing the pace of potentially significant impacts on drug discovery.

Existing tools for design screening are tied to license-based software, severely restricting their availability and utility. We will present our new open-source web tool for such screening, which uses smina, RDKit, and chemalot for docking, shape, and strain energy calculations. Our web tool provides users with a 3D overlay of their proposed compounds relative to a reference target-ligand complex and also generates a table with key information relating these designs to the other
proposed and synthesized compounds within the same project.

In addition, we will present our new open-source tool for tracking novel compound proposals within a project. New designs are uploaded and immediately accessible to all members of the project team. By utilizing RDKit/chemalot for physiochemical property calculations, our web-based design tracker enables prioritization of designs as well as tracking of the rationale for each design through the entire timeline of the project. Common or recurring design themes that emerge from this tracking may then be tagged for follow-up.

Implementing design tools and tracking as part of the discovery workflow will encourage and empower chemists to rapidly test hypotheses with minimal resource use, facilitate exploration of novel chemical space, and foster collaborative, team-driven drug discovery. These open-source web-based tools for ligand modeling and design tracking are available on GitHub.

CINF 40

Building upon chemical similarity - methods to extend instance based learning in cheminformatics

Thomas H. Luechtefeld¹,², tom@toxtrack.com. (1) Environmental Health Engineering, Johns Hopkins School of Public Health, Baltimore, Maryland, United States (2) Underwriters Laboratories, Northbrook, Illinois, United States

The UL Cheminformatics Toolbox (ULCT) is a new suite of tools for toxicology and cheminformatics resulting from collaboration between UL and researchers from Johns Hopkins University. Initially created to address regulatory needs in Europe (REACH) the tool was built using chemical similarity based approaches and publically available toxicology data from nearly 10,000 chemicals registered under REACH.

Since its creation the toolbox continues development along two main avenues: data source integration and algorithm improvement. Data magnitude increase is responsible for most recent gains in ULCT performance. As such, integration of more data in a fast, safe manner is the priority task. Algorithm improvements aim at increasing the scale of training data available to ULCT algorithms.

The question of data reliability is paramount in toxicology. Algorithms trained on biased data are doomed to replicate those biases. Toxicology struggles with large imbalances in training data and failures to achieve consensus in chemical labels.

In toxicity there is typically greater comfort in labelling a compound as ‘positive’ than ‘negative’ for a given hazard. This leads to an abundance of chemicals with positive hazard labels and a relative lack of negatives. We discuss the results of resampling methods on training datasets to help handle label imbalances.

Many cheminformatics models rely on a single data repository. The ECHA C&L data is a large repository for hazard labeling, but even within this repository there are chemicals with discordant labels (different submissions record different results for the same compound). We review the impacts of different approaches to consensus labelling.

Successful management of data integration greatly increases the amount of data available for algorithm training. Traditional models target single endpoints and only use training data that have been
labeled for their respective endpoint. Single target models result in a failure to leverage large datasets and generate a zoo of algorithms each with their own specific implementations. We discuss how large datasets can be used to (1) build new chemical representation through deep learning and auto-encoders (2) build multi-target algorithms that can apply lessons learned in modeling one endpoint to other endpoints (transfer learning).

CINF 41

Chemical workflows supporting automated research data collection

Valery Tkachenko¹, tkachenko.valery@gmail.com, Rick Zakharov¹, Fred Prior³, Alexander V. Kabanov³, Alexander Tropsha². (1) SCIENCE DATA SOFTWARE, LLC, Rockville, Maryland, United States (2) Univ of North Carolina, Chapel Hill, North Carolina, United States (3) Eshelman School of Pharmacy, University of North Carolina Chapel Hill, Chapel Hill, North Carolina, United States (4) University of Arkansas for Medical Sciences, Little Rock, Arkansas, United States

Acquisition of data from public sources is inefficient, time consuming and limited in scope. The NIH has recently posted its intention to financially support data deposition by investigators through the ‘data sharing plan’ for each funded proposal. However, this plan also points to a current weakness of the centralized data sharing and acquisition as all laboratories use different data collection and formatting approaches. These inconsistencies in data formatting by individual labs leads to the need to invest significant resources in data curation and interpretation by the technical staff involved in the maintenance of the centralized data collection resource such as CaNanoLab or Nanomaterial Registry.

It would be far more efficient and useful if there were a standardized data collection and deposition template with standard key terms (such as Minimal Information about Nanomaterials, MIAN) that could be modified to add new or important additional data or parameters for each investigator. These new features could be ultimately adopted in the classification scheme and guide the scope of the expanding database. This approach would be a win-win as it would enable structure for the investigators laboratory, consistency in data reporting and a means of transmitting data to the database in parallel to publication to eliminate the acquisition step from the process.

In this talk we will outline our experience building Open Science Data Repository, a federated database system for direct acquisition, curation and management of research data, including nanomaterial data capture, transformation, and streamlined submission to nanomaterial knowledgebases. The key part of the system is microservices based architecture which exposes RESTful API suitable for direct integration into Workflow Management Systems as well as built-in modules facilitating and enforcing various lab-specific standard operating procedures.

CINF 42

Diamond, XChem and CCP-CompMedChem: Creating user-focused tools and workflows for structure-based drug design

Anthony Bradley¹², anthony.richard.bradley@gmail.com, Rachael Skyner¹, Frank von Delft¹². (1) Diamond Light Source, Oxford, United Kingdom (2) University of Oxford, Oxford, United Kingdom

Since 2015 the XChem beamline at
Diamond (http://www.diamond.ac.uk/Beamlines/Mx/Fragment-Screening.html) has been performing high-throughput fragment screening using X-ray crystallography; 1000 compounds individually screened in a week. This world-wide user programme is used by ~30 academic groups per year. Such groups, having found 10s of experimentally determined, are then tasked with finding best-practice for designing novel compounds in a structure-based drug design programme. Many computational tools exist to optimise such hits, however finding cost-effective and usable solutions is challenging. For example, many open-source tools are available to be used but require guidance and knowledge to use and interoperate between.

These problems have led to the development of the CCP CompMedChem (http://www.ccp-cmc.org/) - a CCP for in silico small molecule compound design. Built up of academic method developers, industrial computer-aided drug design groups and software providers. CCP CMC aims are threefold. 1) Facilitate naive user access to computational chemistry best-practice and tools. 2) Build an academic route-to-market for tools and technologies. 3) Build a forum that allows novel method development that speaks to the needs of pharma. In this talk I will discuss XChem’s scientific challenge, the technical solutions (Python Luigi and Docker) we are building, the partnerships (such as CCP CMC) we have established and the future plans we envisage.

CINF 44

Interactive and reproducible data analysis with the open-source KNIME Analytics Platform

Gregory Landrum, greg.landrum@gmail.com. KNIME AG, Basel, Switzerland

One of the many factors driving the usage of workflow tools in the fields of computational chemistry and cheminformatics (as well as in informatics/data-science groups in general) is the desire to be able to reproduce the steps...
carried out while doing an analysis. In a research environment this desire for reproducibility can conflict with the fact that data exploration and interactive analyses are often required. The open-source KNIME Analytics Platform allows data exploration and interactive analysis to be made reproducible without placing an undue burden on the user. In order to make this as easy and comprehensive as possible, and recognizing that we will never be able to provide every tool our users may want, KNIME makes it straightforward to integrate with other software tools. Capabilities can be added to KNIME in a number of ways including creating new nodes, using Python and/or R snippets, invoking web services, and calling out to external command-line applications.

Because KNIME itself is already quite well known in the computational chemistry and cheminformatics communities, I will present a very short overview of KNIME followed by a couple of examples highlighting some of the new tools for interactive data analysis and working with Python that have recently been added to the open source platform. Throughout the presentation a strong emphasis will be placed on reproducibility.

**CINF 45**

**Using Python to streamline access to the Cambridge Structural Database through new workflows**

*Paul Sanschagrin, paul@sanschagrin.com, Shyam Vyas. Cambridge Crystallographic Data Centre, Piscataway, New Jersey, United States*

The Cambridge Structural Database (CSD) contains crystal structures for over 915,000 organic and metal-organic compounds providing a vast amount of information about structural trends, geometric tendencies and intermolecular interactions. Such data can be used across many fields of chemical inquiry. Traditionally, the only way to access the tools used to mine and analyze this data was through the CSD-System suite of desktop software produced by the Cambridge Crystallographic Data Centre (CCDC). While powerful, it is not always easy to do larger scale analysis or to link CSD-System functionalities together with external software packages. To this end, CCDC developed and released the CSD Python API version 1.0 in November 2015. This API (Application Programming Interface) is under continued development and is a key way that users will access the CSD going forward. This presentation will talk briefly about the API and will then present examples in which it has been used to tie together and streamline specific analyses. Included will be examples within the CCDC application environment, with 3rd party Python libraries and modules, and with larger workflow tools such as Pipeline Pilot. An example application of using the CSD Python API to build a basic web-based tool will also be shown.

**CINF 46**

**Floe: Parallel, heterogeneous workflows on the cloud**

*Paul C. Hawkins, phawkins@eyesopen.com, Andrew G. Shewmaker, Jharrod W. LaFon, Allan G. Skillman, Robert W. Tolbert. OpenEye Scientific, Santa Fe, New Mexico, United States*

Two major problems continue to beset modern cheminformatics: scale and collaboration. Scale has become increasingly important in many aspects of cheminformatics yet generating, storing, searching and analyzing massive datasets still present major problems for existing workflow tools. Parallelization of computation within a workflow engine...
would ease some of the bottlenecks that arise with "big data" but this has hitherto proven difficult to implement. Access to optimal hardware for a particular calculation within a workflow (for example the use of CPUs and GPUs within the same workflow) is not yet available in existing workflow engines. Most importantly, current workflow engines do not easily allow access to the rapidly growing, and increasingly inexpensive, computational resources of the cloud. Collaboration is an increasingly central part of the modern research environment, with multiple groups and organizations contributing to a project; for example drug discovery increasingly relies upon interactions between academia, CRO’s, biotechs and large pharma companies. Efficiently and effectively capturing and sharing data from such disparate sources is difficult enough; capturing and broadcasting expertise developed at diverse sites is much more challenging and presents a substantial unmet need in pharma and in many other types of organization.

Here we introduce Floe, the OpenEye cloud-native workflow engine, designed to enable cloud-based cheminformatics and drug discovery. Floe enables straightforward generation of powerful workflows that deploy cutting edge science and technology from diverse sources, backed by the functionally unlimited computing power of the cloud and via a graphical development environment. These workflows can easily be shared across and between organizations in a simple yet secure manner, allowing the rapid and effective capture and broadcasting of expertise. Floe enables trivially simple parallelization of workflow components and easily permits the integration of CPU and GPU-based algorithms in a single workflow to provide maximal processing power without any need for expertise in parallel programming. Examples of the use of Floe in several high compute demand problems will be discussed.

**CINF 47**

**LiveDesign: Integrative molecular modeling and cheminformatics for collaborative drug design**

**Erin Davis**, erinsdavis@gmail.com, Trip Garland. Schrödinger Inc, New York, New York, 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 workflow use cases.
Eugene Garfield: The man and his legacy

Helen A. Lawlor, chescot@aol.com. Retired, Radnor, Pennsylvania, United States

Eugene Garfield was a complex and remarkable man. He played the role of employer, mentor, friend, and role model to many people around the world and throughout his lifetime he was unwaveringly loyal to those who came to know him well. He also left a legacy far beyond the concept of citation indexing and bibliometrics, and the ideas that he developed during the latter half of the last century continue to fuel advances in cheminformetrics and information science. This presentation will provide a glimpse of the man who was Eugene Garfield and a brief overview of his many accomplishments through the eyes of an early and long-term employee of the Institute for Scientific Information (ISI® – now Clarivate Analytics) who ultimately had the blessing of counting Eugene Garfield among her friends.

From the Index Chemicus Registry System to SciFinder and beyond

Wendy A. Warr, wendywarr.com. Warr & Associates, Crewe, Cheshire, United Kingdom

Gene Garfield is probably best known as a pioneer of citation analytics, but some of us remember him as an early leader in the indexing of chemical information. Index Chemicus was launched in 1960, four years before the official launch of the Science Citation Index. It began as a current awareness service with fragment codes, molecular formulas, and structural diagrams. Wiswesser Line Notations (WLNs) were added later. The Index Chemicus Registry System (ICRS) was created in 1968. My first job in chemical information was in the Experimental Information Unit at the University of Oxford where I learned to code structures into WLN as part of a project looking into the feasibility of making Index Chemicus available to U.K. universities. In the 1970s, the foundations of substructure searching were laid. Structure search systems based on WLN were used by many companies in-house at that time, and ICRS offered the possibility of structure search of the literature using the same software. Nowadays, graphics-based systems are widely used, but line notations and fragment codes have not completely disappeared. Index Chemicus itself is now part of the Web of Science online. Chemical literature searching is dominated by SciFinder and Reaxys. This talk will draw lessons from the history of chemical structure handling, and will touch on some of the issues that we still have to tackle.

Eugene Garfield: The father of chemical text mining and Artificial Intelligence (AI) in cheminformatics

Roger A. Sayle, roger@nextmovesoftware.com. NextMove Software, Cambridge, United Kingdom

Although perhaps more widely known for his later work on bibliometrics, Eugene Garfield (1925-2017) is also acknowledged within the chemical text mining community as the first person to describe the entity recognition of chemical names (and chemical formulae) in free text, and their extraction as chemical formulae for indexing and search. What is now considered a branch of Natural Language Processing (NLP) and Artificial Intelligence
(AI), used for mining the current Big Data of journal articles and patents, actually traces its origins back to Eugene’s 1962 PhD thesis from the University of Pennsylvania’s Department of Linguistics. Modern “name-to-structure” software and chemical text mining tools not only owe their ancestry to algorithms first described over 50 years ago, but the ability of those original approaches to semantically resolve chemicals, and to handle ambiguous or generic structures places them at what is considered the state-of-the-art even today. In this talk we describe recent advances in text mining of chemical formulae and explain why Eugene Garfield’s pragmatic approach to indexing chemical documents is in many ways superior to the recent vogue (fad?) of applying deep learning using LSTM (long short-term memory) recurrent neural networks or conditional random fields (CRFs) to chemical text mining.

**CINF 51**

**Eugene Garfield’s legacy and its impact on the culture of research**

*Svetla Baykoucheva*, sbaykouc@umd.edu. STEM Libraries, University of Maryland, College Park, Maryland, United States

Eugene Garfield’s ideas have had a tremendous impact on science, culture, politics, and individual scientists. The Science Citation Index (SCI), which Garfield created, and the subsequent products based on it (such as Web of Science and Journal Citation Reports) have changed how researchers gather scientific information, perform research, and communicate their findings. As they are evaluated, hired, promoted, and funded on the basis of the impact of their work, researchers have become more aware of the importance of having their papers cited and published in high-impact journals. The “impact” thinking has spread to funding agencies, university administrators, editors, librarians, and even politicians. This paper will discuss the many ramifications of Eugene Garfield’s ideas and legacy and how they have changed the culture of research.

**CINF 52**

**Beyond citations: What are new ways to assess content that will extend the assessment toolbox?**

*Todd A. Carpenter*, tcarpenter@niso.org. National Information Standards Organization (NISO), Baltimore, Maryland, United States

Citations provide a useful tool for assessing the quality and impact of one particular type of research output. But there are many ways in which quality and impact can be assessed, as well as a range of scholarly outputs that are being produced. And rather than simplifying assessment down to a single number, what are ways that we can extend the toolbox by which we assess impact. Over the past five years, new forms of impact assessment have been developed and deployed, which are adding color to the practice of scholarly assessment. This talk will explore the foundation of these metrics in Dr. Garfield’s work, what is the future of altmetrics, and the area of assessment more broadly.

**CINF 53**

**Novel research and its scientific and technological impact**

*Jian Wang*, jian.wang@kuleuven.be. Leiden University, Leiden, Netherlands

Following the combinatorial novelty perspective, novelty in science can
be measured by examining whether a publication makes new combinations of referenced journals and how distant are the newly-paired journals to each other. Using WoS articles in 2001, we study the relationship between novelty and citation impact. We find that novel research demonstrates a high risk/high gain profile: Novel papers are more likely to be a top 1% highly cited paper in the long run, to inspire follow on highly cited research, and to be cited in a broader set of disciplines, but at the same display a higher variance in their citations. We also observe delayed recognition of novel papers which are less likely to be top cited in the short run. In addition, novel research is more highly cited in “foreign” fields but not in their “home” field. Finally, novel papers are published in journals with a lower Impact Factor, compared with non-novel papers, ceteris paribus.

In addition, patent references to the scientific literature provides a paper trail of knowledge flow from science to innovation. We investigate the relationship between novelty and technological impact. We find that novel publications are more likely to be directly cited by patents and also indirectly by other scientific publications which are cited by patents. Within the set of scientific papers cited at least once by patents, there are no additional significant differences in the speed or the intensity of the technological impact between novel and non-novel scientific prior art, but the technological impact from novel science is significantly broader, covering more diverse technological fields and reaching technology fields previously non-impacted.

CINF 54

Clarivate Analytics: Building on the Garfield legacy with Web of Science

James Testa, james.testa@Clarivate.com

Clarivate Analytics, Philadelphia, Pennsylvania, United States

As a pioneer in automated indexing and retrieval of information, Eugene Garfield developed the numerous citation databases that have changed how scientists search and assess scholarly literature. Today, Clarivate Analytics, as the steward for Web of Science, is advancing scientific indexing and analytics beyond the journal with expanded indexing of regional content, patents, data sets, and funding data. In this session, Clarivate will set the stage for the future evolution of citation indexing based on the legacy of Dr. Garfield.

CINF 55

Quality data to quality models

Travis Hesketh1,2, travis@optibrium.com, Peter Hunt1, Matthew D. Segall1, Ed Champness1, Tamsin Mansley1. (1) R&D, Optibrium Limited, Cambridgeshire, United Kingdom (2) Chemistry, University of Strathclyde, Glasgow, United Kingdom

Public domain databases, such as ChEMBL and PubChem, are excellent sources from which to obtain compound data for modelling of target activities or other properties. However, despite the best attempts of the maintainers of these databases, significant pre-processing of the raw data downloaded from these databases is necessary before high quality models can be built. Issues include the mixing of qualified and unqualified values; high variability between different measurements of the same activity; and duplicate, identical values of an activity for the same compound, resulting from multiple publications of the same measurement.
In this presentation, we will describe an automated workflow to pre-process raw downloads from ChEMBL to generate a clean data set, for example filling in missing standardised activity values, checking comparability between different assays and discarding qualified data, where appropriate, while retaining as much of the initial information as possible. The resulting sets are ready for building of quantitative structure-activity relationship (QSAR) models. The Python code implementing this workflow is available for download.

We will also describe how the clean data set can be used as input to an automated process for generation and validation of QSAR models. This builds multiple models using different statistical and machine learning methods, compares the resulting models to identify the best model and then validates the best model using an external independent test set. The combined workflow makes it easy to quickly develop, validate and deploy high quality models from the rich data available in the public domain.

This workflow will be illustrated with example models for target activities and absorption, distribution, metabolism and elimination (ADME) properties.

**CINF 56**

**Development and implementation of Amgen Small Molecule Projects Spotfire Report (ASMPSR) to streamline Hit2Lead and lead optimization process**

**Lei Jia**¹, ljia@amgen.com, Stephanie Geuns-Meyer², Hua Gao², Timothy G. Hopper², Mark Southern¹, Brian Lanman¹, Yaxiong Sun². (¹) Amgen, Thousand Oaks, California, United States (2) Amgen Inc, Cambridge, Massachusetts, United States

Effective analysis and interpretation of data is the key to move medicinal chemistry programs forward. A typical drug discovery effort includes hundreds or thousands of compounds and may employ hundreds of assays. In combination, millions of data points are often involved in the “hit to lead” and “lead optimization” process. Cheminformatics tools implemented in workflows are essential for leveraging those data to facilitate compound design, understand Structure-Activity Relationships (SAR), formulate and test hypotheses, and drive optimal decisions in medicinal chemistry programs.

A set of inhouse Pipeline Pilot protocols were developed for a full workflow including: retrieving compound and assay data from internal data warehouses, physiochemical property calculation, predictive ADME modeling, managing compound design ideas, and data processing. The protocols are modular for maximum flexibility under the scope of individual medicinal chemistry programs. Data are refreshed daily. The output data are imported through active links within Spotfire for visualization and analysis. Add-on functions were developed inside Spotfire to allow medicinal chemist to use advanced computational chemistry tools outside Spotfire for further design and analysis. This application has become a key component in Amgen small molecule drug discovery process.

**CINF 57**

**Automating matched molecular pair analysis of bioactivity and solubility data**

**Frederik van den Broek**², broek@gmx.ch, Matthew Clark¹. (¹) Elsevier, Philadelphia, Pennsylvania, United States (2) Elsevier, Amsterdam, Netherlands
Most matched molecular pair analysis (MMPA) is focused on analysis of bioactivity data. The availability of workflow tools which include MMPA tools, has made it much easier to extend this powerful cheminformatics concept to other parameters relevant in (medicinal) chemistry. We will present MMPA analysis not limited to bioactivity data, but also including other parameters (e.g. solubility) using KNIME workflows and the rich data sets extracted from a large body of scientific literature.

CINF 58

Reaction and chemistry data blending

Markus Fischer^1, markus.fischer.01@gmail.com, Jannise Buckley^2, Frederik van den Broek^2, broek@gmx.ch. (1) Elsevier, Frankfurt am Main, Germany (2) Elsevier, Amsterdam, Netherlands (3) Dassault Systemes Biovia Corp., San Diego, California, United States

Combining heterogeneous data from various disparate data sources into one coherent and normalized dataset is the essential, but often overlooked prerequisite for building an analytics workflow. Furthermore, these workflows are of growing importance in the era of data-driven discovery by supporting a researcher to become a knowledge scientist. We will present use cases which show the power of scientific workflows to distil insights from the rich chemistry data sources available to academia and industry, such as Reaxys, OpenPhacts, and ChEMBL. Built by making use of the latest features in Pipeline Pilot and using techniques such as statistics and machine learning, these workflows provide efficient tools to inform the scientific decision-making process.

CINF 59

Integrated life science data and the power of workflows

Jana Gurinova, Daniela Digles, Gerhard F. Ecker, gerhard.f.ecker@univie.ac.at. Dept of Pharmaceutical Chemistry, University of Vienna, Wien, Austria

The public availability of large data sources such as ChEMBL and the Open PHACTS Discovery Platform allows immediate retrieval of large data sets for different sorts of research questions. However, submitting complex queries and analyzing the data retrieved still is time consuming and prone to errors. In this context, the use of workflow engines such as KNIME or Pipeline Pilot allows to submit complex queries and enables to simultaneously query different domains, such as compounds, targets, pathways, and diseases. Within this contribution case studies for the exploitation of linked open data for the development of ligand-transporter interaction models and their use for predicting complex in vivo toxicity endpoints will be presented. Furthermore, we will outline KNIME workflows for transporter profiling and drug repurposing. Finally, we will present ToxPHACTS, which is based on a KNIME workflow for toxicological read across.

CINF 60

Automated workflows for data curation and standardization of chemical structures for QSAR modeling

Kamel Mansouri^2,^1, mansourikamel@gmail.com, Andrew McEachran^2, Christopher Grulke^2, Ann Richard^3, Richard Judson^3, Antony J. Williams^3. (1) ScitoVation, Research Triangle Park, North Carolina, United States (2) NCCT, US EPA (ORISE), Research Triangle Park,
Large collections of chemical structures and associated experimental data are publicly available and can be used to build robust QSAR models for applications in different fields. One common concern is the quality of both the chemical structure information and associated experimental data. Here we describe the development of automated KNIME workflows to both, assist in the curation of data and to standardize the chemical structures according to a set of standard rules. The publicly available PHYSPROP physicochemical properties and environmental fate datasets were used as case studies to reveal commonly encountered errors and develop a set of rules to correct them. The workflow first assembles structure–identity pairs using up to four provided chemical identifiers, including chemical names, CASRNs, SMILES, and MolBlocks. Problems detected included errors and mismatches in chemical structure formats, identifiers, and various structure validation issues, including hypervalency and stereochemistry descriptions. Subsequently, a structure standardization KNIME workflow was used to generate “QSAR-ready” forms prior to calculating molecular descriptors. This workflow performs a series of operations on the 2D structures including desalting, stripping stereochemistry, standardizing tautomers and nitro groups, correcting valence, neutralizing when possible and removing duplicates. A machine learning procedure was applied to evaluate the impact of this curation process. The models based on the curated data and standardized structures showed statistically improved predictive performance. These workflows were used to curate and standardize the full list of PHYSPROP datasets that were used to develop OPERA models available on the EPA’s CompTox Chemistry Dashboard (https://comptox.epa.gov). They were also applied on thousands of other datasets that were used in international consortiums such as CERAPP and CoMPARA. The QSAR-ready workflow was modified to generate “MS-ready structures” to support mass spectrometry non-targeted analysis. All workflows, data, and models are open-source and freely available on GitHub (https://github.com/kmansouri) for further usage and integration by the scientific community.

CINF 61

Integrated visualization of the research landscape of proteins

Jana Gurinova, jana.gurinova@univie.ac.at, Gerhard F. Ecker. Department of Pharmaceutical Chemistry, University of Vienna, Vienna, Austria

There are many databases that provide information about proteins such as their topology, as well as information related to these proteins, such as their connections to diseases, drugs or patents. This wealth of information is however seldom visualized and even less so across multiple databases. For an integrated view of the research landscape of proteins we developed a KNIME workflow capable of visualizing a family of proteins as a phylogenetic tree, annotated with the above mentioned datasets. When looking at the annotation with diseases, drugs and patents, we observe three cases for proteins that are associated with a number of diseases. A protein that - besides diseases - is associated with a number of drugs and patents, or no drugs but a number of patents (meaning that presumably research towards drugs in under way), or, most interestingly, a
To facilitate research into these proteins which often lack a crystal structure, we implemented a visualization of the protein’s topology, so as to provide ideas for a homology template by comparing the protein of interest to others within the phylogenetic tree.

The workflow starts with a multiple sequence alignment web service (Clustal Omega) which creates the phylogenetic tree, which is then passed to iTOL, a web service capable of visualizing phylogenetic trees and diverse annotations. The annotations are derived from DisGeNET (diseases), DrugBank (drugs) and SureChEMBL (patents). The topology information is retrieved from UniProt. By combining these web services and databases we arrive at an integrated view of the research landscape of proteins as well as a starting point for a selection of a homology model template.

CINF 73

Strategic outreach into chemistry and chemical engineering research groups at UC Berkeley

Kortney K. Rupp1,2, kortneyrupp@berkeley.edu, (1) UC Library, University of California, Berkeley, Berkeley, California, United States (2) Research IT, Lawrence Berkeley National Laboratory, Berkeley, California, United States

In an increasing digital information environment, subject liaisons for chemistry and related fields face an uphill battle cultivating outreach opportunities with faculty, graduate students and postdoctoral researchers. Research groups within departments often exist as individual communities and offer a unique opportunity to learn about research taking place and how members of those communities seek and transfer information.

As a brand new subject liaison, I have launched a series of outreach efforts to integrate my presence into the Departments of Chemistry and Chemical and Biomolecular Engineering, learn about research, and reassess library services. Of these efforts attending departmental graduate student seminars have been the most valuable experiences up to this point. These seminars provide insights into group work that allow me to connect with faculty on a topic they care very much about—their research and the opportunity to provide mentoring and training for their students. From these interactions I then contact an administrative staff member, graduate student or the faculty member directly to ask to attend an upcoming research group meeting.

In this presentation I will discuss the feedback I received from my initial inquiries, the networking and contact
methods that have been most effective, and the challenges I have faced throughout the process. The main purpose of the initial group meeting attendance is to listen and learn. Based on the initial group meeting, tailored suggestions are made about which library services to promote and what instruction would be most valuable for that specific group. For example, one positive outcome has led to embedded library instruction within a synthesis group. New first-year graduate students and postdoctoral researchers will receive a library resource overview and in-depth training using the Reaxys interface for organic synthesis from both the subject liaison and vendor representative. Based on an assessment of these initial sessions, plans for future outreach will be proposed.

CINF 74

Catalyzing chemical information literacy through ChemOnCampus

Daniel Christe, christe.daniel@elsevier.com. Elsevier, Philadelphia, Pennsylvania, United States

Learning in the digital age is increasingly contextual, embodied, and on-demand. This macrotrend of technologically-driven empowerment, coupled with the strong evidence basis for active learning pedagogies over didactic teaching modalities has produced a shift in recent years toward active learning pedagogical approaches in STEM courses. To help educators navigate this shift, Elsevier launched the ChemOnCampus program to provide resources and activities across the undergraduate and graduate spectrum – this includes a game-based learning approach to literature search, a flipped learning model “research sprints”, and authorship workshops at the graduate/undergraduate levels. The talk will emphasize collaboration between students, faculty, librarians, and information analytics providers to enhance chemical information literacy. The presenter will share actionable insights from a new flipped learning model called Research Sprints, which are integrative, fast-paced, active learning experiences emphasizing creativity, collaboration, and communication in which teams “sprint” to harvest information needed to solve a given challenge, harnessing chemical information search tools such as Reaxys and Knovel. The participants collaborate to harvest information and create graphical abstracts to communicate findings through oral presentations and via social media channels (e.g. Mendeley, Twitter). To-date, research sprints have been tested at the primary, secondary, and post-secondary, and graduate levels.

CINF 75

Creating bonds across campus: A general chemistry information literacy initiative

Derek Behmke, dbehmke@ggc.edu, Adrienne Harmer, aharmer@ggc.edu. Georgia Gwinnett College, Lawrenceville, Georgia, United States

The Chemistry and Research Services Library faculty at Georgia Gwinnett College have worked together for the last eight years to embed information literacy in the first-year introductory general chemistry course sequence that is required for all students with STEM majors at the college. The library’s work to adopt and adapt the ACRL Framework has revitalized our cooperative efforts and has led to some major revisions and expansions of our collaborative information literacy initiative. The initiative has recently expanded to include a scaffolded approach to information literacy centered on a series
of course embedded research projects in two courses that require students to research and present about their findings related to nitrates in local water supplies. The instructional model uses a flipped approach combined with exciting new in-class activities and assessments that guide students on an exploration of how to assess and evaluate information through the lens of understanding “Information Creation as a Process” and “Searching as a Strategic Exploration”. This presentation will introduce the collaborative pedagogical work, cover the instructional design process and methods, and discuss recent efforts to assess how much and how well chemistry students are learning the information literacy knowledge practices and dispositions they need to succeed in their STEM fields.

CINF 76

Helping users RAMP up by learning about chemical safety information resources

Grace Baysinger, graceb@stanford.edu. Robin Li and Melissa Ma Science Library, Stanford Libraries, Stanford, California, United States

Before working at the bench, it is important for researchers to Recognize hazards, Assess and Minimize hazardous risks, and Prepare for emergencies (RAMP). Chemical safety information resources provide tools to help with this evaluation. While users are familiar with the chemistry literature, they often struggle to find chemical safety information. This presentation will cover key resources and useful search strategies. To date, outreach efforts to help familiarize students, faculty, and staff with chemical safety information resources include holding workshops, meeting with lab groups, and participating in a health and safety fair by asking users to “stump the librarian.”

CINF 77

Using SurveyMonkey and Qualtrics to assess student learning in information literacy programs for undergraduate and graduate chemistry courses

Svetla Baykoucheva1, sbaykouc@umd.edu, Amanda J. Schech2, Elizabeth C. Griffith3. (1) STEM Libraries, University of Maryland, College Park, Maryland, United States (2) Department of Chemistry and Biochemistry, University of Maryland, College Park, Maryland, United States

Assessment of student learning is often difficult to perform during information literacy instruction, especially in large undergraduate chemistry courses. This paper describes how online assignments, prepared with SurveyMonkey and Qualtrics, were used in undergraduate and graduate chemistry courses at the University of Maryland College Park to evaluate student ability to efficiently find literature and properties of chemical compounds and write lab reports. The assignments were graded, and the grades were included in the total grade of the students for the course. Students performed very well on the assignments and provided valuable feedback. This information literacy instruction program was designed and implemented in close collaboration between the librarian and the course instructors.

CINF 78

Building new information literacy collaborations at MTSU between library faculty and chemistry faculty

Judith M. Iriarte-Gross1, Mary Ellen Sloane2, maryellen.sloane@mtsu.edu, Denise Fitzgerald-Quintel1, Heather Listhartke1, Christy Groves1. (1) Middle
Library faculty at MTSU, with the support of an instructional development grant, developed 16 lesson plans for science courses. The lesson plans utilize the Association of College and Research Libraries (ACRL) Information Literacy Framework to develop knowledge practices and dispositions across the spectrum of the Framework. The Library faculty worked with Chemistry faculty to refine the lesson plan template and learning outcomes. The lesson plans also identify high impact pedagogies that are part of MTSU’s quality enhancement plan program.

The presenters will discuss ways to integrate the ACRL Information Literacy Framework into the curriculum in meaningful, comprehensive, and practical ways. We will also discuss assessment of student learning and developing partnerships between Library faculty and chemistry faculty.

CINF 79

Continuous evolution of delivering chemical information literacy through the stepping stone approach

Andrew A. Yeagley¹, chemist.4.life@gmail.com, Melissa C. Rhoten¹, Sarah E. Porter³, Benjamin Topham³. (1) Longwood University, Farmville, Virginia, United States (2) Chemistry, Longwood University, Farmville, Virginia, United States (3) Chemistry and Physics, Longwood University, Farmville, Virginia, United States

It has been four years since our attempt to map scientific information literacy into our four year undergraduate curriculum. The stepping stone approach was a conscious effort to align and involve all faculty at Longwood University in a unified approach to delivering informational literacy skills to our graduates. We built upon existing scaffolding and streamlined content delivery to scaffold the information the students were introduced to. Components of informational literacy in our courses have been constantly undergoing improvements and we have had our first cohort of students move through the program. We will focus on discussing our most recent additions to the program and student feedback on their preparedness with regard to informational literacy.

CINF 80

Building competency: Scaffolding information literacy skills throughout the chemistry curriculum

Chapel Cowden, chapel-cowden@utc.edu. UTC Library, University of Tennessee at Chattanooga, Chattanooga, Tennessee, United States

Many library instruction programs are moving to embed information literacy skills throughout the curriculum of individual departments in a systematic way—often referred to as “scaffolding”, as skills are introduced and then built upon throughout a student’s time in the program. Just such a task was undertaken at the University of Tennessee at Chattanooga to scaffold information literacy skills for Chemistry and Biochemistry majors.

Rather than just randomly select classes for library instruction, or wait until a faculty member contacts the library for instruction, scaffolding allows the librarian to look more deeply at departmental curricula and pinpoint crucial areas for information literacy engagement. This engagement
does not always have to be instruction, but could take many forms: providing or consulting on assignments, interacting with students in online course management systems, providing alternative content (videos, podcasts, worksheets, etc.).

Following a description of the scaffolding program for Chemistry and Biochemistry majors at UTC, we will explore benefits and drawbacks to scaffolding and how you can work to scaffold information literacy skills with the Chemistry departments engaged with at your own institutions.

**CINF 81**

**Predictive computational techniques for chemical risk assessment**

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

Environmental chemists increasingly rely on cheminformatics methods and tools for chemical risk assessment, especially when it comes to identifying those chemicals that need to be experimentally tested in priority. As both the quantity and diversity of experimental chemical biological data are growing, cheminformatics tools are expected to fully assist in giving unprecedented biological insight in the mechanisms leading to toxicity and achieving better prediction performances for untested chemicals. Such additional knowledge becomes extremely relevant for chemical risk assessment and any executive decision regarding the potential toxicants. In this presentation, we will discuss several cheminformatics methods and associated software to (i) conduct hierarchical QSAR modeling that fully integrates both classification and regression models to afford higher prediction accuracy, (ii) screen a library of environmental chemicals using structure-based docking for prioritizing the compounds to be tested experimentally, and (iii) study the dynamic receptor-ligand interactions for toxicants using 4D fingerprints computed from molecular dynamics trajectories.

**CINF 82**

**Identifying and prioritizing chemicals for evaluation of potential endocrine bioactivity and exposure**

Alicia Frame, frame.alicia@epa.gov, Kristan Markey, Alaa S. Kamel, Sean Watford, Katie Paul-Friedman, Richard Judson, Antony J. Williams. (1) National Center for Computational Toxicology, Environmental Protection Agency, Wake Forest, North Carolina, United States (2) US EPA, Alexandria, Virginia, United States (3) Office of Chemical Safety and Pollution Prevention, US EPA, Washington, District of Columbia, United States (4) USEPA, Rtp, North Carolina, United States

The US Environmental Protection Agency (EPA) Endocrine Disruptor Screening Program (EDSP) uses a two-tiered approach to determine a chemical’s potential to interact with estrogen, androgen, and thyroid hormone pathways in humans and/or wildlife. Currently, EPA is tasked with evaluating potential endocrine activity of all pesticide chemicals and substances found in drinking water to which a substantial population may be exposed. EDSP has identified a “Universe of Chemicals” (UoC) – based on existing pesticide active and inert ingredients, regulated Safe Drinking Water Act (SDWA) contaminants, and the Contaminant Candidate List (CCL) Universe. The current UoC is overly broad, and contains numerous compounds which are of limited interest to EPA, difficult to evaluate for endocrine disrupting activity, or for
which relevant exposures are outside EPA’s regulatory authority. Because the requirements for testing, data review, and weight-of-evidence determination require substantial temporal, financial, animal, and staff resources, a workflow was developed to eliminate and/or defer many irrelevant substances as part of prioritization of chemicals with increased potential for exposure and endocrine activity. The EDSP Dashboard, developed as part of EPA’s CompTox dashboards initiative, is intended to provide greater transparency in the selection and prioritization process and will provide agency scientists and external stakeholders with a tool to rapidly explore data that guides agency decisions, including the revised UoC.

The refinement of the EDSP UoC is concurrent with the implementation of the EDSP21 work plan, aimed at incorporation of alternative scientific approaches to more rapidly screen chemicals for their ability to interact with the endocrine system. The EDSP21 approach utilizes high throughput bioassays, computational models, and exposure predictions to screen and prioritize the UoC for potential endocrine bioactivity. Ultimately, the prioritization of the UoC will enable more complete implementation of EDSP21 by focusing EDSP screening and prioritization resources on only the most relevant chemicals.

CINF 83

 OPERA: A free and open source QSAR tool for predicting physicochemical properties and environmental fate endpoints

Kamel Mansouri1, mansourikamel@gmail.com, Christopher Grulke2, Richard Judson2, Antony J. Williams2. (1) ScitoVation, Research Triangle Park, North Carolina, United States (2) National Center for Computational Toxicology, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, United States

Collecting the chemical structures and data for necessary QSAR modeling is facilitated by available public databases and open data. However, QSAR model performance is dependent on the quality of data and modeling methodology used. This study developed robust QSAR models for fourteen physicochemical properties and environmental fate endpoints that can be used for regulatory purposes. Publicly available data were collected from the PHYSPROP database among other sources. These data sets have undergone extensive curation using an in-house automated workflow to enhance the quality of the data. The chemical structures were standardized to “QSAR-ready form” prior to calculation of the molecular descriptors. The modeling procedure was based on the five OECD principles for QSAR models to produce reliable yet simple models. Genetic algorithms were used to select the most pertinent and mechanistically interpretable descriptors (from 2 to 15 with an average of 11 descriptors). The sizes of the modeled datasets varied from 150 chemicals for biodegradability half-life to 14,050 chemicals for logP, with an average of 3222 chemicals across all endpoints. The optimal models were built on randomly selected training sets (75%) and validated using 5-fold cross-validation (CV) and test sets (25%). The CV Q² of the models varied from 0.72 to 0.95 with an average of 0.86 and an R² test from 0.71 to 0.96 with an average of 0.82. Modeling and performance details were described in QSAR model reporting format (QMRFs) and validated by the European Commission’s Joint Research Center (JRC) for OECD compliance. All models are delivered as a free, open source/open
data application called OPERA (OPEn structure-activity Relationship App) used to predict properties for ~750,000 chemicals. The predicted data are freely available on the EPA's CompTox Chemistry Dashboard (https://comptox.epa.gov).

CINF 84

WebTEST (Web-services toxicity estimation software tool)

Todd Martin¹, martin.todd@epa.gov, Valery Tkachenko², Antony J. Williams³. (1) National Center for Computational Toxicology, Environmental Protection Agency, Wake Forest, North Carolina, United States (2) Science Data Software, LLC, Rockville, Maryland, United States (3) National Risk Management Research Laboratory, Environmental Protection Agency, Cincinnati, Ohio, United States

A Java-based web service is being developed within the US EPA’s Chemistry Dashboard to provide real time estimates of toxicity values and physical properties, to generate alternative assessment (AA) hazard profiles to compare chemical alternatives, and to suggest sustainable synthesis routes.

Currently WebTEST can generate toxicity predictions directly from a simple URL. For example, to estimate water solubility one can enter https://comptox.epa.gov/dashboard/web-test/WS?smiles=CCO into a web browser. Web interfaces are being created to allow users to make predictions for single chemicals (e.g. by drawing chemicals in a Ketcher chemical structure editor or by searching for a chemical in the Chemistry Dashboard) and for batches of chemicals (e.g. by loading a structure data file). Previously calculated chemicals will be stored in a database to minimize the display of results.

Alternatives assessment frameworks evaluate chemical alternatives in terms of human health effects, ecotoxicity, and fate. Example categories include acute mammalian toxicity, acute aquatic toxicity, and bioaccumulation, respectively. Online data sources such as Japan’s National Institute of Technology (NITE) can be utilized to obtain GHS (Global Harmonization System) scores for comparing alternatives. Data gaps can be filled using the toxicity models within WebTEST (e.g. the oral rat LD₅₀ model). Finally, WebTEST will provide suggested “green” synthesis routes for chemicals identified as alternatives to a chemical of concern. Synthesis routes can be generated by analyzing the key functional groups present in a molecule and then querying a reaction ontology library.

CINF 85

Literature-based cheminformatics for research in chemical toxicity

Nancy Baker², baker.nancy@epa.gov, Thomas Knudsen³, Antony J. Williams¹. (1) National Center for Computational Toxicology, Environmental Protection Agency, Wake Forest, North Carolina, United States (2) Leidos, Hillsborough, North Carolina, United States (3) National Center for Computational Toxicology, US EPA, Research Triangle Park, North Carolina, United States

PubMed is the largest freely available
Chemical Information Bulletin, 2018, 70 (1)

source of published literature available online with access to 27 million citations (as of October 2017). Contained within the literature is an abundance of information about the activity of chemicals in biological systems. Literature informatics approaches applied to chemical toxicity research can help researchers make use of the literature in highly effective ways. At the EPA's National Center for Computational Toxicology, in addition to our text-mining capabilities, we have developed a novel approach to article retrieval in our PubMed Abstract Sifter. The Abstract Sifter is a document retrieval tool that integrates the richness of PubMed with the powerful data-manipulation capabilities of Microsoft Excel. Results from a PubMed search are imported directly into an Excel sheet where the end-user can then use a novel “sifter” methodology for quick, agile relevance ranking of articles. The tool also enables article triage capabilities through easy tagging and noting functionality. Triaged citations can be exported to external software such as reference management tools. The Abstract Sifter can also provide a high-level view of a corpus of literature for a defined set of entities such as chemicals. This “landscape” view helps researchers assess the volume of literature in any given subject area to help with project scoping and chemical ranking. A version of the Abstract Sifter is also implemented in the web-based CompTox Chemistry Dashboard developed by the EPA's National Center for Computational Toxicology. Both versions of the tool will be demonstrated and discussed.

CINF 86

Chemotype-enrichment workflow:
A univariate analysis workflow for exploring chemical feature enrichments across EPA's ToxCast chemical-assay landscape

Ryan Lougee1, rrlougee@ncsu.edu, Ann Richard1, Christopher Grulke2. (1) MD 205-01, US EPA, Research Triangle Park, North Carolina, United States (2) Zachary Piper Solutions, New Hill, North Carolina, United States (3) NCCT, ORISE US EPA, Durham, North Carolina, United States

EPA's ToxCast library, spanning more than 4000 diverse chemical structures (>8000 including Tox21 chemicals), is designed to cover the environmental toxicity and chemical exposure landscape of interest to EPA. Each ToxCast chemical has been screened in tens to hundreds of in vitro high-throughput screening (HTS) assays, and in vivo toxicity data are available for over 1000 of the chemicals. A principal aim of the ToxCast program is to use data across these different domains to build models for predicting potential toxicity or exposure, and for prioritizing limited testing resources. However, largely because of their chemically and mechanistically diverse contents, ToxCast/Tox21 data sets pose challenges to traditional global structure-activity relationship (SAR) modeling approaches. Codifying local chemistry domains within the inventory, through use of publicly available fingerprinting methods, can serve to amplify SAR signals within these domains. An automated set of command line tools, known as the Chemotype-enrichment workflow (CTEW), has been developed to identify enriched fingerprint features (known as chemotypes) across ToxCast chemicals and corresponding HTS and in vivo assay results. The workflow generates fingerprints (e.g., ToxPrints, MACCS, PubChem) for EPA's DSSTox database content (>1M structures), as well as newly introduced structures. For each of the >800 ToxCast/Tox21 HTS assay datasets, a fingerprint file of the test set structures was first queried, then a discretized assay endpoint (1,0) vector
was used to identify enriched features, handle duplicate assay chemicals, and finally generate an enrichment table, employing statistical thresholds of Odds Ratio >3, Fisher’s Exact p-value <0.05, and a minimum of 3 active chemicals. The approach offers an intuitive, flexible complement to traditional SAR methods, with results that are easily interpreted and anchored to defined chemical features, and that can productively guide more targeted SAR investigations.

CINF 87

New developments in delivering public access to data from the National Center for Computational Toxicology at the EPA

Antony J. Williams1, tony27587@gmail.com, Christopher Grulke1, Andrew McEachran2, Grace Patlewicz1, Imran Shah1, John Wambaugh1, Richard Judson1, Ann Richard1, Jeff Edwards1. (1) National Center for Computational Toxicology, Environmental Protection Agency, Wake Forest, North Carolina, United States (2) Oak Ridge Institute of Science and Education (ORISE), Durham, North Carolina, United States

Researchers at EPA’s National Center for Computational Toxicology integrate advances in biology, chemistry, and computer science to examine the toxicity of chemicals and help prioritize chemicals for further research based on potential human health risks. The goal of this research program is to quickly evaluate thousands of chemicals, but at a much reduced cost and shorter time frame relative to traditional approaches. The data generated by the Center includes characterization of thousands of chemicals across hundreds of high-throughput screening assays, consumer use and production information, pharmacokinetic properties, literature data, physical-chemical properties as well as the predictive computational modeling of toxicity and exposure. We have developed a number of databases and applications to deliver the data to the public, academic community, industry stakeholders, and regulators. This presentation will provide an overview of our work to develop an architecture that integrates diverse large-scale data from the chemical and biological domains, our approaches to disseminate these data, and the delivery of models supporting predictive computational toxicology. In particular, this presentation will review our new CompTox Chemistry Dashboard and the developing architecture to support real-time property and toxicity endpoint prediction.

CINF 88

Actualizing research into practical tools: A case study of GenRA, a new read-across tool

George Helman2,1, helman.george@epa.gov, Grace Patlewicz1, Imran Shah1. (1) National Center for Computational Toxicology, US EPA, Research Triangle Park, North Carolina, United States (2) Oak Ridge Institute for Science and Education, Oak Ridge, Tennessee, United States

Read-across, a popular data gap filling technique traditionally relies on a thorough expert driven assessment. This can lead to inconsistent predictions, has limitations in terms of the numbers of chemicals that can be evaluated and offers little insight into the generalizability of the approach or its performance. We sought to evaluate the baseline performance of read-across for a large set of chemicals in a systematic, objective and reproducible manner and to provide quantitative measures of uncertainty for the predictions derived. The approach developed,
generalized read-across (GenRA), relies on chemical descriptor information and/or \textit{in vitro} bioactivity data (derived from high throughput screening data from ToxCast) to derive read-across predictions of toxicity effects in \textit{in vivo} repeat-dose toxicity studies. In translating the approach into a web-based tool, we anchored its development around the typical category workflow and structured this into a dynamic grid interface. The default starting point is to identify source analogues with associated \textit{in vivo} data on the basis of chemical fingerprints. The next step is to analyze the scope and quantity of available data (both \textit{in vitro} and \textit{in vivo}). The third step generates a data matrix in order to evaluate the analogues—in terms of their consistency and concordance of effects across the different toxicity effects. The final steps involve generating a GenRA prediction and exporting the predictions in a flat file. Recent work has investigated physicochemical similarity and its impact in improving the read-across performance from the baseline GenRA. Using the Lipinski rule of 5 parameters, the performance for target organs aggregated over all study types was evaluated. In general, filtering structural analogues on the basis of physicochemical parameters led to a decrease in performance, whereas expanding the analogue identification to optimize both physicochemical and structural similarity resulted in improved performance. We describe the functionality and features of the GenRA web application which has recently been released in open beta including ongoing refinements such as point of departure prediction and extending the chemical space to cover all of the DSSTox inventory. GenRA offers a novel and practical means of being able to perform objective read-across that can be helpful in screening level hazard assessments.

CINF 89

Use of 2D chemical structure and bioactivity profiles to generate chemical categories within an adverse outcome pathway network

Mark Nelms$^{1,2}$, nelms.mark@epa.gov, Stephen W. Edwards$^2$. (1) ORISE, Oak Ridge, Tennessee, United States (2) U.S. EPA, Durham, North Carolina, United States

The Adverse Outcome Pathway (AOP) framework has emerged to capitalise on the vast quantity of mechanistic data generated by alternative techniques, as well as advances in systems biology, cheminformatics, and bioinformatics. AOPs provide a scaffold onto which mechanistic data can be organised to establish a connection between a molecular initiating event (MIE) and an adverse outcome. The MIE is the initial interaction between chemical and biological systems. Understanding the details of this interaction between chemicals and biological molecules can help discern the structural and/or physico-chemical properties that may be required to perturb the MIE. This information can enable chemicals to be grouped based upon their ability to perturb an MIE using both structural and functional criteria such as a conserved structural fragment, \textit{in vitro} activity connected to the MIE, and/or toxicological data connected to downstream key events in the AOP. Subsequently, the use of chemical groups, alongside information relating to the associated MIE/AOP(s), can be used for various applications. These include: hazard/risk assessment, prioritisation of chemicals for further testing, and/or identification of the AOP network(s) most likely to be of concern if chemicals co-occur within the environment. Upon development of chemical groups, data gap filling methods such as read-across can be
used to provide predictions for chemicals within a chemical group that currently lack relevant toxicological data. This work covers how 2D chemical structure information and bioactivity profiles can be utilised to group chemicals, and how the inclusion of bioactivity profiles may help in the refinement of these chemical groupings for prioritisation and/or hazard/risk assessment. Additionally, this presentation will discuss how chemical categories can be used in conjunction with AOP networks to guide various aspects of mixtures risk assessment.

CINF 90

COSMOS DB and ChemoTyper: Public resources for managing, sharing and exploring toxicity data and chemical space

James F. Rathman, rathman.1@osu.edu, Chihae Yang, Christof Schwab, Oliver Sacher, Aleksey Tarkhov, Joerg Maruszczyk, Thomas Kleinoeder, Aleksandra Mostrag, Bruno Bienfait.

COSMOS Database, maintained through the COSMOS DataShare Point, is a means for managing and sharing toxicity and chemical data to assist in the safety assessment of cosmetics-related substances under the paradigm of non-animal testing strategies. This public resource comprises a total of 102,285 substance records spanning 31 toxicity endpoints searchable within COSMOS DB. Also included in the database is a comprehensive and openly-accessible cosmetics inventory derived from 15 sources, including regulatory inventories such as CosIng, Korean Cosmetics Industry Institute, US Personal Care Products Council, and Cosmetics Ingredient Review, as well as the European Food Safety Authority (EFSA). The database houses the critical point of departure data identified by the COSMOS TTC project in collaboration with the ILSI Europe, as well as other TTC datasets such as those established by Munro et al. (non-cancer) and CPDB (cancer). The COSMOS DB thus functions as the de facto clearing house for TTC datasets, which can be exported from the database. Similarities and differences between databases can be analyzed by exploring their chemical spaces, both in terms of molecular structures and properties. CORINA Symphony Community Edition provides a set of molecular properties suitable for examining similarity in property space, while ChemoTyper provides the ToxPrints, a diverse set of chemotypes for structure representation, ideal for comparing datasets in structure space. In addition to their use in chemical space exploration, these public resources can be used for designing structural rules specific to toxicity endpoints or biological activities of interest.

CINF 91

Importance of synthetic speech mark-up language in the teaching of chemistry concepts in a multi-sensory way

Cary A. Supalo, csupalo@purdue.edu, Ashley E. Neybert.

Synthetic speech mark-up language (SSML) impacts how text-to-speech renderings of chemical and other customized spoken words can be spoken by commercially available speech output.
programs. This capability is valuable for students with visual impairments that rely on speech output as a major tool to access textbooks. Currently synthetic speech output speaks chemical expressions as word and not by their proper chemical names. For example, NaCl is pronounced, “NaCl,” instead of sodium chloride. It uses the sequence of consonants and vowels to interpret them as words. SSML serves as a powerful multi-sensory access tool for students with print disabilities. However, the access technology industry does not widely support the use of SSML in commercially available products. Electronic book publishers and other on-line course content providers are encouraged to support SSML. This will open doors of opportunity for access to technical content in chemistry and other sciences.

CINF 92

Introducing cheminformatics early - prepares your students for success

Mindy Pozenel, mpozenel@cas.org. Marketing, CAS, Columbus, Ohio, United States

Over 50% of STEM undergraduates change their major to something outside of science and technology. There is increasing pressure to make sure your students are engaged and better prepared for their future careers. Chemistry Class Advantage helps prepare organic chemistry students for a successful career by cultivating chemical information literacy and grabbing their attention with real-world examples. By integrating chemical information searching in undergraduate organic chemistry classes, students inherently get exposed to cheminformatics and begin to understand the application of chemistry to their everyday life through the literature. Lessons have been developed with chemistry faculty across the US so they can easily map to the topics in popular textbooks and integrate into curriculum for the entire semester or a couple extra credit assignments. By using real-world scenarios in the literature, students become more engaged and better understand their domain. To accomplish these goals, Chemistry Class Advantage can help:

- Coordinate information literacy with the concepts in the textbook;
- Build new connections between professors, students and the librarians by introducing chemical information search tools early in the academic career.
- Prepare students for future success by connecting classroom concepts to real-world research and provide them the skills to analyze this research.

CINF 93

What can you teach using PubChem?

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

Every year a substantial amount of public money is invested on scientific research projects, resulting in a large amount of open data freely available in the public domain. PubChem is an example of public repositories of chemistry open data, with more than 235 million substance descriptions, 93 million unique chemical structures, and one million biological assay experiments, collected from more than 550 data sources. It is one of the top five most visited chemistry web sites in the world, with more than 2 million unique users per month at peak. About half of them are between ages 18 and 24, suggesting that PubChem is heavily used by undergraduate or graduate students at academic institutions although
it is originally developed to serve the biomedical research community. Unfortunately, many teachers as well as students merely view PubChem as a free online reference work used in traditional chemistry courses, without realizing new opportunities that PubChem provides for the chemical education community in the age of big data. This presentation discusses important aspects of PubChem as cheminformatics education resource, including data quality and accuracy, data provenance and governance, structure standardization, terminologies, and so on. It also presents various PubChem tools and services for search, analysis, and download, that can be useful in cheminformatics classroom. In addition, we will discuss how PubChem was used in the Cheminformatics OLCC (On-Line Chemistry Courses), an intercollegiate hybrid course co-taught by residential teaching faculty from multiple campuses and online guest lecturers with cheminformatics expertise.

CINF 94

Reaxys and you: Working together to prepare future chemists

Norah Xiao, n.xiao@elsevier.com, Juergen Swienty-Busch. Elsevier, Washington D.C., District of Columbia, United States

Preparing students with information literacy skills nowadays is challenging, as our students grow up with a Google mindset and apply this to how they search for answers every day. Do our students search, evaluate and use the information effectively? This is more complicated and more crucial to chemistry information, as knowledge is highly specialized. Are students ready to solve real world problems related to their life now and later in the career? For educators, how to incorporate information tools into teaching? Working with educators and librarians, we have been focusing on addressing these key areas in teaching and learning chemistry, and have been striving to make it easier, more effective and streamlined to support teaching and learning needs to prepare next generation chemists.

CINF 95

Innovative instruction program for materials science and engineering undergraduate students

Susanne J. Redalje, curie@u.washington.edu. Reference and Research Div, University of Washington, Seattle, Washington, United States

Materials Sciences and Engineering 311 is a course required of all University of Washington undergraduate MSE students. Rotating through various modules, the students begin learning to use the equipment and techniques which are important to their field. Along with learning to use the laboratory equipment, the Instructor emphasizes the corresponding information needs of the students as they pursue the field. For several years now, the Instructor has worked with the chemistry subject specialist to develop an information module which includes techniques for finding data and property information and finding, understanding, and ethically using scientific research articles. The subject librarian teaches multiple sessions of the information modules, generally consisting of no more than 5-6 students, allowing for a great deal of interaction between the librarian and the students.

CINF 96

Evolution of natural product total synthesis: Mapping pathways through literature searching
Leah McEwen, lrmt@cornell.edu. Clark Library, Cornell University, Ithaca, New York, United States

Beginning graduate students research discovery stories of total synthesis or mechanistic studies that have an interesting literature trail over time, such as Sharpless dihydroxylation, total synthesis of strychnine, taxol, and others. Students are expected to prepare (in groups) an annotated chronology of key references and present three slides on the pathway evolution. Students also work on a structure based reaction searching exercise with classic databases such Reaxys, Science of Synthesis, SciFinder and the Encyclopedia of Organic Reactions. The two part class is semi-flipped with portions of the assignment due before the lecture for students to explore the databases on their own terms and experience searching challenges to discuss further in class. This presentation will review the development of this literature exploration as a collaboration between the organic chemistry professor and the chemistry librarian.

CINF 97

Instructional scaffolding of information literacy skills in a problem-based learning context

Ye Li, yeli@mines.edu, Ginger V. Szymczak, gshultz@umich.edu. (1) Arthur Lakes Library, Colorado School of Mines, Golden, Colorado, United States (2) Department of Chemistry, University of Michigan, Ann Arbor, Michigan, United States

In the course of problem-based learning, students seek outside information related to the problem they are solving. With the abundant availability of information technology, especially Google, students intuitively go beyond their textbook and assigned reading. However, in a prior study, we found that when information literacy skills are not taught explicitly taught in this context students will conduct unproductive searches and use unreliable information. In so doing, their ability to effectively address the posed problem may be hindered. These prior research results were used to inform the development of instructional modules, which scaffold information literacy skills into problem-based learning curriculum. The modules model for students how to find and evaluate information when working to solve a synthetic organic chemistry problem. The modules were evaluated using a pre-post test of information literacy, focus group interviews, and evaluation of student work. Through the evaluation, we will improve the intervention modules and articulate how problem solving needs interact with the development of information literacy skills.

CINF 98

Teaching information literacy through the chemistry laboratory

Charity E. Flener-Lovitt, chariteach@gmail.com, Brandon Finley, Alyssa Berger, berg31@uw.edu. (1) Physical Sciences, University of Washington Bothell, Redmond, Washington, United States (2) University of Washington Bothell, Bothell, WA, Washington, United States

In an information driven economy, it is necessary to develop information literacy skills, especially when searching for information about chemicals. An online assignment, “Identifying Expert Sources of Information”, was developed for the second quarter lab to help students explore and evaluate the types of online resources that provided information about
chemicals. Students first chose a molecule that had a use as a health supplement. Then they reported the type of information available from a web search engine, Wikipedia, ChemSpider, and a scientific database. For each source, students described the information available according to the 5 W’s (Who, What, Where, When and Why) and wrote an ACS style reference for each source. At the end of the assignment, students were asked to complete a 100-200 word reflection about finding authoritative sources of chemical information. In subsequent lab, students were explicitly asked to search for information from these resources and provide appropriate references to them in their lab reports.

This assignment was developed while revising the general chemistry lab curriculum at the University of Washington Bothell to incorporate guided inquiry and sociocultural context in meaningful ways. In this talk, we will discuss how this assignment fits within the lab curriculum and provide a preliminary assessment of the effectiveness of this assignment. Assessment will be provided through student self-evaluation of their information literacy skills and instructor evaluation of student performance.

CINF 99

Machine learning methods for chemical properties and toxicity-based endpoints prediction using open source libraries

Valery Tkachenko, tkachenko.valery@gmail.com, Alexander Korotcov, Rick Zakharov, Boris Sattarov, Artem Mitrofanov. Science Data Software, LLC, Rockville, Maryland, United States

In the last decade there is an increasing interest in using in silico tools for potential risk assessment of newly released chemicals due to the large number of chemicals enter the market yearly and the big uncertainty on their possible hazardous effects. Different tools and methods based on machine learning techniques already exist and were used in a wide range of applications starting from quantitative structure-property relationships and expanding into predictive toxicology. There is a lot of historical data accumulated across multiple databases which is publicly available and can be used with novel machine learning methods. Unfortunately, due to different datasets, metrics and validation strategies, the significant gaps remain in both the quantity and quality of data available coupled with optimal predictive methods. This work is an attempt to develop a multitask system which can serve as searchable curated collections of multiple chemical datasets and ready to use novel machine learning methods solely built using open source frameworks and libraries. We have implemented a set of self-tuned, using grid search and k-fold validation, traditional machine learning methods (shallow methods) such as Naïve Bayes, k-Nearest Neighbors, Random Forest, Boosted Decision Trees, Regularized Logistic Regression, and Support Vector Machines base on open source Scikit-learn (http://scikit-learn.org/)
The novel Deep Neural Networks models of different complexity have been also implemented using Keras (https://keras.io/), a deep learning open library, and a Tensorflow (www.tensorflow.org) as a backend. The machine learning models were trained and evaluated to predict measures of toxicity from the physical characteristics of the structure of chemicals using the same datasets as in the Toxicity Estimation Software Tool (https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test). The Deep Learning models showed very good performance evaluation characteristics and were found to be useful in predicting of toxicological and physicochemical parameter endpoints. The results of this work support an optimistic view that some of current obstacles in cheminformatics can be overcome by using Deep Learning methods.

CINF 100

Exploring chemical space using ChemMaps.com

Alexandre Borrel1,2, aborrel@ncsu.edu, Nicole Kleinstreuer2,3, Denis Fourches1,2
1Bioinformatics Research Center, NC State University, Raleigh, North Carolina, United States
2DIR/BCBB, NIEHS, RTP, North Carolina, United States
3NICEATM, NIEHS, RTP, North Carolina, United States

The need for navigating chemical space has become more important due to the increasing size and diversity of chemical biological databases (e.g., ChemSpider, DrugBank, ChEMBL, ToxCast). To do so, modelers typically rely on projection techniques applied to series of quantitative molecular descriptors directly computed from two-dimensional chemical structures. However, the multiple cheminformatics steps required to compute and visualize chemical space are technical, necessitate coding 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 was developed to browse and visualize the space of 2,000 FDA-approved drugs and over 6,000 drug candidates. Each compound was initially characterized using a large set of molecular descriptors including 1D-2D RDKIT descriptors and 3D PADEL descriptors (238 1D-2D and 44 3D after removing correlated descriptors). Principal Component Analysis was used to project compounds in three-dimensional space, where compounds’ coordinates in the first two dimensions were calculated using 1D-2D descriptors, and the third dimension (Z axis) was determined using 3D descriptors only. To optimize the representation of the space and the interactive, user-friendly navigation experience, we developed the ChemMaps.com webserver using modern 3D-optimized web technologies such as HTML5, JavaScript, and CGI. The chemical coverage is now being expanded to include environmental chemical space based on the U.S. EPA TSCA inventory, as well as toxicological categorizations based on curated animal study data and predictive high-throughput screening signatures. Users accessing ChemMaps.com can immediately explore the entire compound library using a responsive, mouse-based, easy-to-use navigation tool. Since all information and coordinates are pre-computed, the browsing is instantaneous and does not require computational skills. Similar to searching Google Maps for a specific address, users can search ChemMaps via a dedicated search bar (e.g., name, indications, pharmacological class, toxicity values) and visualize the space with options to zoom in on chemical “neighborhoods”. Additional browsing, searching, and exporting
options are underway, including tools to support read-across and chemical risk assessment.

CINF 101

Enhancing exchange of environmental data between EPA and FDA

Yulia Borodina¹, yulia.borodina@fda.hhs.gov, Sarah Winfield¹, Brittany Pruitt². (1) FDA, Silver Spring, Maryland, United States (2) Office of Chemical Safety and Pollution Prevention, US Environmental Protection Agency, Washington, District of Columbia, United States

The Environmental Protection Agency (EPA)’s mission is to protect human health and the environment in the United States. The Food and Drug Administration (FDA) is responsible for protecting and promoting the public health by ensuring, among other things, the safety of human food and animal feed by administering and enforcing the Federal Food, Drug, and Cosmetic Act (FD&C Act) and several related public health laws. The EPA and the FDA, have certain related objectives in carrying out their respective food safety, public health, and associated regulatory, marketing, trade, and research activities. It is desirable to enhance the exchange of information between EPA and FDA. For this goal, the electronic data standards developed for FDA’s pharmaceutical labels are being adjusted to accommodate EPA’s pesticide active ingredient, label and the pesticide tolerance data. The details and the status of this initiative will be discussed.

CINF 102

Prediction of pKa from chemical structure using free and open-source tools

Valery Tkachenko², tkachenko.valery@gmail.com, Neal Cariello⁴, Alexander Korotcov², Kamel Mansouri³, Antony J. Williams¹. (1) National Center for Computational Toxicology, Environmental Protection Agency, Wake Forest, North Carolina, United States (2) Science Data Software, LLC, Rockville, Maryland, United States (3) ScitoVation, Research Triangle Park, North Carolina, United States (4) Integrated Laboratory Systems, Research Triangle Park, North Carolina, United States

The ionization state of a chemical, reflected in pKa values, affects lipophilicity, solubility, protein binding and the ability of a chemical to cross the plasma membrane. These properties govern the pharmacokinetic parameters such as absorption, distribution, metabolism, excretion and toxicity and thus pKa is a fundamental chemical property and is used in many models of chemical toxicity.

Experimentally determining pKa is not feasible for high-throughput assays. Predicting pKa is challenging and existing models have been developed only using restricted chemical space (e.g., anilines, phenols, benzoic acids, primary amines) and lack of a generalized model impedes ADME modeling.

No free and open source models exist for heterogeneous chemical classes, however, several proprietary programs exist. In this work, pKa open data bundled with DataWarrior (http://www.openmolecules.org/) were used to develop predictive models for pKa. After data cleaning, there were ~3100 and ~3900 monoprotic chemicals with an acidic or basic pKa, respectively. 1D and 2D chemical descriptors (AlogP, Topological polar surface area, etc) in addition to 12 fingerprints (presence or absence of a chemical group) were generated using
PaDEL software. Three datasets were used: acidic, basic and acidic and basic combined.

13 datasets were examined, the 1D/2D descriptors and 12 fingerprints. Using the Extreme Gradient Boosting algorithm showed that the MACCS and Substructure Count fingerprints yielded the best results, with models showing an R-Squared of ~0.78 and a RMSE of 1.42.

Recently, Deep Learning models have showed remarkable progress in image recognition and natural language processing. To determine if the Deep Learning algorithms would increase model performance we examined the datasets and found that the Deep Learning models were somewhat superior than Extreme Gradient Boosting with an R-Squared of ~0.80 and an RMSE of ~1.38.

CINF 103
Cheminformatics tools for supporting environmental chemistry

Yannick Djoumbou Feunang¹, djoumbou@ualberta.ca, David S. Wishart¹,². (1) Biological Sciences, University of Alberta, Edmonton, Alberta, Canada (2) Computing Science, University Of Alberta, Edmonton, Alberta, Canada

The contribution to the human chemical exposome is determined by the exposure to exogenous chemicals (e.g. foods, pharmaceuticals and personal care products, environmental contaminants and pollutants), and endogenous chemicals (e.g. lipids, vitamins, hormones) that are produced or modified during genetically programmed events (e.g. amino acid and lipid metabolism) or in response to various stimuli (e.g. intake of foods, drugs, poisons). Understanding the effects of such exposures requires: 1) the availability of various types of data (chemical, metabolomic, functional, toxicological), preferably in a structured way; 2) a very good understanding of how chemicals are metabolized or degraded in various environments; and 3) tools to identify those chemicals and their products from mass spectrometry measurements. This is particularly important, as a significant portion of the chemicals constituting the dark matter of the human exposome (~3,000,000 chemicals) are believed to be transformation products of known chemicals. Moreover, the biosynthesis or metabolism of a chemical can often explain its function, distribution, and toxicity within an organism. In this presentation, I will describe a number of tools and databases that we are developing to address this needs. In particular, I will present 1) ChemOnt, an (extension of ClassyFire’s) ontology for classifying chemicals based on their structure, functional role, health effects, physical appearance, and route-of-exposure; 2) BioTransformer, a software tool for the prediction of small molecule metabolism in humans, their gut microbiota, as well as in the soil and aquatic microbiomes; 3) CFM-ID 3.0, a software tool designed to accurately predict mass spectra for rapid compound identification; 4) ContaminantDB, a comprehensive electronic database of nearly 100,000 known chemical contaminants; and 5) HMDB 4.0, the newest release of the Human Metabolome Database. Moreover, I will describe some applications to illustrate how these tools can be used to support environmental chemistry.
CINF 104

Investigating ligand and structure-based modeling followed by mixture toxicity prediction of per-and polyfluoroalkyl substances: A virtual screening approach

Supratik Kar1, supratik.kar@icnanotox.org, Shinjita Ghosh2, Jerzy R. Leszczynski1. (1) Department of Chemistry, Physics and Atmospheric Sciences, Jackson State University, Jackson, Mississippi, United States (2) School of Public Health, Jackson State University, Jackson, Mississippi, United States

Exposure to poly- and perfluoroalkyl substances (PFASs), an emerging class of endocrine disrupting halogenated pollutants, has been linked to thyroid toxicity in human populations across the globe. The PFASs can compete with thyroxine (T4) for binding to the human thyroid hormone transport protein transthyretin (TTR) which may lead to reduce thyroid hormone levels leading to endocrine disrupting activity. In this background, twenty-four PFASs, together with structurally similar natural fatty acids binding capacity in radioligand-binding assay values were modeled with classification- and regression-based quantitative structure-toxicity relationship (QSTR) to identify the responsible structural features and fragments of these diverse classes of PFASs. Additionally, docking study performed employing complex of TTR with bound 2,6-difluorobiphenyl-4-carboxylic acid (PDB: 2F7I) to constitute the receptor model for human TTR provided corroborating evidence for these binding interactions and indicated multiple high-affinity modes of binding. Further, developmental toxicity data on zebrafish (Danio rerio) embryos of single PFAS and tertiary mixtures were modeled employing QSTR tool. The computed models, as well as consensus model, are then employed for toxicity prediction of binary and tertiary mixtures which have no experimental data to check mixtures mode of toxicity and their possible threshold of toxicity for future risk assessment (Figure 1). Further, for virtual screening, we have employed twenty-four PFASs from our previous work to build a huge external dataset consists of single (24), binary (276) and tertiary (2024) mixtures (equal ratio mixtures) by permutation of the studied chemicals. Then, all developed models are employed for predicting this external dataset for interpretation of toxicity threats for individuals and mixtures along with identification of diverse range and combination of toxicity thresholds (Figure 2). We found that chemicals within these mixtures displayed concentration addition suggesting a similar mode of toxic action. Importantly, next-generation chemicals were less acutely toxic singly and in mixtures than their first generation counterpart. Not only that, mixtures of PFASs showed less toxicity than single PFAS. The developed in silico models, therefore, provide an understanding of important structural attributes of these chemicals and may serve as an efficient query tool for screening of large database.

CINF 105

Chemoinformatic approach to identification of antiviral components in humic substances

Alexey Orlov1,2, mase@qsar.chem.msu.ru, Alexander Y. Zherebker1, Anastasia A. Eletskaya2,3, Victor S. Chemikov2, Liubov I. Kozlovskaya2, Yuri V. Zhemov4, Vladimir A. Palyulin1, Dmitry I. Osolodkin1,2, Irina V. Perminova1. (1) Department of Chemistry, Lomonosov Moscow State University, Moscow, Russian Federation (2) Institute of Poliomyelitis and Viral Encephalitides,
Humic substance (HS) is a heterogeneous supramolecular ensemble of natural organic compounds formed by oxidative decomposition of biomacromolecules in soils, coals, waters, etc. Inhibitory effect of HS on reproduction of a wide range of DNA and RNA viruses reveals a possibility to apply HS against human pathogens with no approved therapeutics. However, the exact structural composition of HS cannot be determined by modern analytical techniques and structure and the nature of HS bioactivity remains unknown, thus complicating the process of development HS-based drugs. We investigated inhibitory effect of ten HS samples on enteroviruses and flaviviruses in cell-based assays and developed a technique for an active component identification using Fourier-transform ion cyclotron resonance mass spectrometry FTICR MS data analysis and chemical database search. All HS samples inhibited tick-borne encephalitis virus (TBEV) reproduction with EC50 values in range of 0.1-1 μg/ml and exhibited no cytotoxicity up to 10 μg/ml. The same HS samples did not inhibit reproduction of enteroviruses. Observed specificity of the samples was attributed to possible presence of compounds with a specific anti-TBEV activity. Molecular composition of all HS samples was explored by ESI FTICR MS. More than 6.5k unique formulae were identified and spectral data were analysed using van Krevelen diagrams and heatmaps generated based on samples formulae fingerprints. Analysis revealed subtle differences in the formulae composition of samples. Chemoinformatic approach was further employed to identify putative active components structures. We developed a methodology for identifying active components of HS samples using publicly available data from ChEMBL version 23. Formulae presented in the samples were searched in ChEMBL and related structures and bioactivity entries were retrieved and standardised. Compounds corresponding to formulae present in HS samples are mostly represented by typical natural compounds, such as flavonoids, coumarins, etc. Antiviral activity profiles were analysed for flaviviruses and enteroviruses and several compounds were suggested as potential active components of the HS samples. This methodology can be applied for any mixture of small molecules with known formulae composition and spectrum of bioactivity.

CINF 106

Adding complex expert knowledge into chemical databases: Transforming surfactants in wastewater

Emma Schymanski1, emma.schymanski@eawag.ch, Christopher Grulke3, Juliane Hollender1, Antony J. Williams3. (1) Eawag, 8600 Dübendorf, Switzerland (2) Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Belvaux, Luxembourg (3) National Center for Computational Toxicology, Environmental Protection Agency, Wake Forest, North Carolina, United States

The increasing popularity of high mass accuracy non-target mass spectrometry methods has yielded extensive identification efforts based on chemical compound databases. Candidate structures are often retrieved with either
exact mass or molecular formula from large resources such as PubChem, ChemSpider or the EPA CompTox Chemistry Dashboard. Additional data (e.g. fragmentation, physicochemical properties, reference and data source information) is then used to select potential candidates, depending on the experimental context. However, these strategies require the presence of substances of interest in these compound databases, which is often not the case as no database can be fully inclusive. A prominent example with clear data gaps are surfactants, used in many products in our daily lives, yet often absent as discrete structures in compound databases. Linear alkylbenzene sulfonates (LAS) are a common, high use and high priority surfactant class that have highly complex transformation behaviour in wastewater. Despite extensive reports in the environmental literature, few of the LAS and none of the related transformation products were reported in any compound databases during an investigation into Swiss wastewater effluents, despite these forming the most intense signals. The LAS surfactant class will be used to demonstrate how the coupling of environmental observations with high resolution mass spectrometry and detailed literature data (expert knowledge) on the transformation of these species can be used to progressively “fill the gaps” in compound databases. The LAS and their transformation products have been added to the CompTox Chemistry Dashboard (https://comptox.epa.gov/) using a combination of “representative structures” and “related structures” starting from the structural information contained in the literature. By adding this information into a centralized open resource, future environmental investigations can now profit from the expert knowledge previously scattered throughout the literature.

CINF 107

Prioritizing anthropogenic chemicals in drinking water sources through combined use of mass spectrometry based exposure data and ToxCast toxicity data

Andrea Mizzi M. Brunner¹, brunner@hifo.uzh.ch, Milou M. Dingemans¹, Kirsten A. Baken¹, Annemarie P. van Wezel¹.². (1) Chemical Water Quality and Health, KWR Watercycle Research Institute, Nieuwegein, Netherlands, Netherlands (2) Copernicus Institute of Sustainable Development, Utrecht University, Utrecht, Utrecht, Netherlands

Advancements in high-resolution mass spectrometry (HRMS) based screening methods have enabled a shift from target to non-target analyses to detect chemicals in water samples. The multitude of suspect chemicals resulting from such non-target screenings need to be prioritized for further identification and potential inclusion into monitoring programs. Here, we compare a strategy developed for the prioritization of chemicals in Dutch raw and drinking water samples based on semi-quantitative exposure data from HRMS (Sjerps et al. 2016) to a strategy based on high-throughput in vitro toxicity data. Between 2007-2014, 151 Dutch water samples, including waste water treatment plant effluent, surface water, ground water and drinking water, were collected. HRMS non-target screening analyses detected >7000 structures in these samples which could be linked to >1000 suspects from a curated suspect list of >5000 EU and water relevant chemicals. These suspects were subsequently prioritized based on exceedance of the Threshold of Toxicological Concern (TTC). Here, rather than using this discrete scale, we ranked suspects based on their semi-quantitative total concentration, expressed
as internal standard equivalents. We then compared both the TTC prioritization and the continuous ranking of the chemicals to a prioritization based on chemical-specific \textit{in vitro} toxicity data from the publicly available EPA ToxCast database. Using the 5th percentile of the AC50 values from all ToxCast assays in a hypothesis-free approach, and from assays considered most water relevant in parallel, >500 suspects could be ranked based on their toxicity with respect to a total of >1000 different assay endpoints. The comparison showed that different prioritization strategies resulted in a different ranking of suspect chemicals. We therefore propose a novel prioritization scheme that combines both exposure and toxicity data and takes advantage of their complementarity to prioritize suspects in water samples.

CINF 108

\textbf{Units of Measure Interoperability Service (UMIS): FAIR units for FAIR data}

\textit{Stuart J. Chalk}, schalk@unf.edu, Robert J. Hanisch. (1) Department of Chemistry, University of North Florida, Jacksonville, Florida, United States (2) Office of Data and Informatics, National Institute of Standards and Technology, Gaithersburg, Maryland, United States

The growing focus on Findable Accessible Interoperable and Reuseable (FAIR) data is important making research data readily available are reusable by the scientific community. To this point however, little attention has been given to the equally important representation of units of measure that must go along with FAIR data such that we can all interpret and integrate heterogeneous data sets. The Units of Measure Interoperability Service (UMIS) is a NIST project focused on bringing together existing computer representations of units of measure so that data can be normalized when needed without fear of losing meaning or misrepresenting data when formally published or openly shared. This paper will describe the UMIS repository, the content currently loaded into the repository, services available for users, and system features. A discussion of the future development will also be presented, including the progress toward making UMIS aligned with the GO FAIR Initiative and supporting the FAIR data infrastructure.

CINF 109

\textbf{RA21: Improving access to scholarly resources, from anywhere, on any device}

\textit{Ralph Youngen}, ryoungen@gmail.com. ACS, Washington, District of Columbia, United States

One year following the formation of Resource Access for the 21st Century (RA21), the joint STM and NISO initiative, the RA21 task force will report on its efforts to align and simplify pathways to subscribed content across participating scientific platforms. The primary objective of the initiative is to improve the user experience when accessing scholarly content by offering seamless, non-traditional authentication means that are not dependent on workflow or location. This session will explore the benefits offered by anonymous forms of authentication that enhance security, permit customizations, and enable the collection of accurate usage analytics from both the library and publisher perspectives. The issues of accessibility, privacy, and security are addressed through three multi-stakeholder pilot projects which will be showcased during the session. Feedback from user surveys and early
recommendations emerging from the three pilot results will be shared. Results are being collected now and focus on guidance around the user experience as well as best practices for the security and privacy of user data. The final results of this project will be a set of best practice guidelines based on the real-world experienced developed through the pilots.

CINF 110

Rapid collection of experimental physicochemical property data to inform various models and testing methods

Chantel I. Nicolas\textsuperscript{1,5}, chantel.nicolas@gmail.com, Kamel Mansouri\textsuperscript{1,5}, Katherine Phillips\textsuperscript{4}, Christopher Grulke\textsuperscript{3}, Ann Richard\textsuperscript{3}, Antony Williams\textsuperscript{2}, James Rabinowitz\textsuperscript{1}, Kristin Isaacs\textsuperscript{4}, Alice Yau\textsuperscript{2}, John Wambaugh\textsuperscript{3}. (1) ScitoVation, RTP, North Carolina, United States (2) Southwest Research Institute, San Antonio, Texas, United States (3) USEPA NCCT, Research Triangle Park, North Carolina, United States (4) USEPA NERL, Research Triangle Park, North Carolina, United States (5) ORISE, Oak Ridge, Tennessee, United States

In order to determine the potential toxicological effects, toxicokinetics, and route(s) of exposure for chemicals, their structures and corresponding physicochemical properties are required. With this data, the risk for thousands of environmental chemicals can be prioritized. However, as there are limitations on the availability on experimental data, we have attempted to efficiently fill these data gaps by generating new data for 200 structurally diverse compounds. This set of compounds were rigorously selected from the USEPA Distributed Structure-Searchable Toxicity Database (DSTTox). Evaluated in this pilot study are rapid experimental methods to determine five physicochemical properties including the log of the octanol:water partition coefficient ($\log(K_{ow})$), vapor pressure, water solubility, Henry’s law constant, and the acid dissociation constant. For the majority of the compounds in this study, experiments were successful for at least one property, with $\log(K_{ow})$ yielding the largest return (176 values). Using ToxPrint Chemotypes, it was observed that the presence of 21 structural features may have played an overall role in rapid estimation failures. Where available, the new estimates were compared with previous measurements in order to illustrate the consistency of new experimental data with traditional measurement methods. As quantitative structure-property relationship (QSAR) models are relied upon for filling huge gaps in physicochemical property information, we evaluated 5 suites of QSARs for their predictive ability and chemical coverage or applicability domain of the new experimental estimates. Accurate measurements of these properties are crucial for facilitating better exposure predictions in two ways: 1) direct parameterization of exposure models; and 2) construction of physicochemical property QSARs with a wider applicability domain, whereby their resulting predictions can be used to parameterize exposure models in the absence of experimental data.

CINF 111

Prediction of emission and absorption spectra for Eu$^{2+}$-doped inorganic phosphors based on stoichiometric information

Hiroshi Nakano\textsuperscript{1}, hiroshi.b.nakano@sony.com, Kenichi Tanaka\textsuperscript{2}, Tomoyuki Miyao\textsuperscript{1}, Kimito Funatsu\textsuperscript{2}, Raku Shirasawa\textsuperscript{1}, Shigetaka Tomiya\textsuperscript{1}. (1) Advanced Technology Research Division, Sony,
Peak wavelengths of emission and absorption spectra of Eu²⁺-doped inorganic phosphors have been predicted. The prediction models are based only on stoichiometric information of the host materials. We developed and used the original descriptors focusing on the inorganic materials. Gaussian process regression and descriptor selection algorithm are used to build the prediction models for screening of inorganic phosphors. They are simple and practical enough for experimentalists to explore the host materials. The peak wavelengths of emission and absorption spectra are predicted with the error of 139 meV and 166 meV, respectively. The importance analysis of descriptors is understood theoretically with the current of the physical mechanism deciding the photoluminescence properties. Additionally, we propose the possible phosphor candidates for next-generation television systems.

CINF 112

Deep learning approach to computational chemistry of lanthanides

Valery Tkachenko¹, tkachenko.valery@gmail.com, Artem Mitrofanov²,¹, Peter Matveev², Alexander Korotcov¹, Rick Zakharov¹. (1) Science Data Software, LLC, Rockville, Maryland, United States (2) Chemistry Department, Moscow State University, Moscow, Russian Federation

Rare-earth elements are widely used in all kinds of gadgets nowadays. But similarity of their chemistry and natural occurrence force us to develop new effective methods of their separation. Experimental search of new effective and selective ligands requires significant material and human resources, so computational methods are commonly used for estimating potential ligand efficiency. And a question of complexation of rare-earth elements is still the most important in terms of ligands design. Here we would like to introduce new hybrid approach for in silico ligand design based on combination of accurate quantum chemistry (QC) and novel quantitative structure-property relationship (QSPR) methods. QC methods require a lot of computational resources and usage of supercomputer equipment, while QSPR ones can be used on personal computer. On the other hand, lack of experimental data, doesn't allow us to train accurate model for QSPR. Appropriate combination of the approaches allow us to avoid their weak sides. Proposed method is based on experimental data as well as QC calculations and can be easily scaled to other chemical complexation problems.

CINF 113

Mapping of antiviral chemical space with ViralChEMBL: Use cases and new findings

Dmitry I. Osolodkin³, dmitry_o@qsar.chem.msu.ru, Alexey Orlov²,¹, Anastasia Nikitina³,¹, Liubov I. Kozlovskaya³, Vladimir A. Palyulin³, Dragos Horvath², Alexandre Varnek². (1) Department of Chemistry, Lomonosov Moscow State University, Moscow, Russian Federation (2) UMR 7140 CNRS/University of Strasbourg, Strasbourg, France (3) Institute of Poliomyelitis and Viral Encephalitides, Chumakov FSC R&D IBP RAS, Moscow, Russian Federation

Correct annotation of data is a crucial factor defining the quality of analysis and predictions in chemoinformatics. In the field of antiviral activity data, we...
have developed a scheme for taxonomy annotation of ChEMBL assays, which allowed us to create the ViralChEMBL database. In this database all data points are linked to the taxa defined by the International Committee for Taxonomy of Viruses (ICTV), and antiviral activity profiles are available for more than 250,000 compounds, comprising more than 600,000 data points.

In this presentation we describe the new version of ViralChEMBL, incorporating the latest ChEMBL data and the latest version of ICTV taxonomy. We also show several use cases, where chemical space maps and graphs are employed to depict and mine information contained in ViralChEMBL.

Generative Topographic Mapping (GTM) is a well-known approach to dimensionality reduction and chemical space visualization. Universal GTMs (trained in order to maximize their proficiency to predict a vast panel of unrelated biological properties) were previously suggested as a useful tool for intuitive analysis of structure-property relations in chemical space. One of the advantages of the GTM approach is the ability to calculate responsibility patterns (RPs), defining compounds by their probabilities to reside on particular nodes.

We applied the universal maps for the analysis of ViralChEMBL data and for the prediction of new antiviral compounds. Thus, the similarity between structures can be visualized and GTM can be used as a tool for compound repurposing.

DrugBank and an in-house database were placed on the maps, to identify compounds that could be repurposed as antivirals because they reside in areas with high populations of validated antivirals. Structures sharing the same RPs were retrieved and manually analysed. The use of consensus RPs (RPs from more than one map) allowed filtering out pairs of unrelated compounds, accidentally sharing a common location with respect to a single map. Several compounds were selected for further experimental antiviral activity evaluation against the representatives of Flavivirus and Enterovirus genera.
2018 CINF Officers and Functionaries

Chair
Erin Davis
Schrödinger
erinsdavis@gmail.com

Chair-Elect
Elsa Alvaro
Northwestern University
elsa.alvaro@northwestern.edu

Past-Chair
Rachelle Bienstock
National Institute of Environmental Health Sciences
rachelleb1@gmail.com

Secretary
Tina Qin
Vanderbilt University
qinnamsu@gmail.com

Treasurer
Rob McFarland
Washington University
rmcfarland@wustl.edu

CINF Councilors
Bonnie Lawlor
chescot@aol.com

Andrea Twiss-Brooks
University of Chicago
atbrooks@uchicago.edu

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

CINF Alternate Councilors
Carmen Nitsche
CINforma
carmen@cinformaconsulting.com

Charles Huber
University of California, Santa Barbara
huber@library.ucsb.edu

Archivist/Historian
Bonnie Lawlor
See Councilor

Audit Committee Chair
TBD

Awards Committee Chair
TBD

Careers Committee Chair
Neelam Bharti
University of Florida
neelambh@ufl.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
Rob McFarland
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
Schrödinger
erinsdavis@gmail.com

2017–2018 Program Committee Chairs
Elsa Alvaro
Northwestern University
elsa.alvaro@northwestern.edu

Rachelle Bienstock
National Institute of Environmental Health Sciences
rachelleb1@gmail.com

Tellers Committee Chair
Susan Cardinal
see Careers Committee Chair

Chemical Information Bulletin Editor Spring
Vincent F. Scalfani
The University of Alabama
vfscalfani@ua.edu

Chemical Information Bulletin Editor Summer
Judith Currano
University of Pennsylvania
currano@pobox.upenn.edu

Chemical Information Bulletin Editor Fall
Teri Vogel
UC San Diego Library
tmvogel@ucsd.edu

Chemical Information Bulletin Editor Winter
David Shobe
Patent Information Agent
avidshobe@yahoo.com

Webmaster
Erin Davis
Schrödinger
erinsdavis@gmail.com

Assistant Webmaster
TBD
Spring 2018 CIB Contributors

Articles and Features
Bonnie Lawlor
Donna Wrublewski
Robert E. (Bob) Buntrock
Wendy A. Warr

Sponsor Information
Graham Douglas

Production
Bonnie Lawlor
Elsa Alvaro
Teri Vogel
Vincent F. Scalfani
Wendy A. Warr
Schedule of Future ACS National Meetings

<table>
<thead>
<tr>
<th>Meeting</th>
<th>Date</th>
<th>Year</th>
<th>City, State</th>
<th>Theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>256th</td>
<td>Aug. 19–23</td>
<td>2018</td>
<td>Boston, MA</td>
<td>Nanoscience, Nanotechnology &amp; Beyond</td>
</tr>
<tr>
<td>257th</td>
<td>Mar. 31–Apr. 4</td>
<td>2019</td>
<td>Orlando, FL</td>
<td>Chemistry for New Frontiers</td>
</tr>
<tr>
<td>258th</td>
<td>Aug. 25–29</td>
<td>2019</td>
<td>San Diego, CA</td>
<td>Chemistry of Water</td>
</tr>
<tr>
<td>260th</td>
<td>Aug. 23–27</td>
<td>2020</td>
<td>San Francisco, CA</td>
<td>Chemistry from Bench to Market</td>
</tr>
<tr>
<td>261st</td>
<td>Mar. 21–25</td>
<td>2021</td>
<td>San Antonio, TX</td>
<td>TBA</td>
</tr>
<tr>
<td>262nd</td>
<td>Aug. 22–26</td>
<td>2021</td>
<td>Atlanta, GA</td>
<td>TBA</td>
</tr>
<tr>
<td>263rd</td>
<td>Mar. 20–24</td>
<td>2022</td>
<td>San Diego, CA</td>
<td>TBA</td>
</tr>
</tbody>
</table>