

Fall 2019 | Vol. 71, No. 3

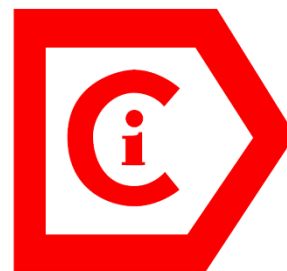
# CHEMICAL INFORMATION BULLETIN

A Publication of the Division of Chemical Information of the ACS

ISSN: 0364-1910



ACS National Meeting & Expo  
Chemistry & Water  
San Diego, California  
August 25-29, 2019



# Chemical Information Bulletin

Fall 2019 — Vol. 71, No. 3

Teri M. Vogel, Editor

Letter from the Editor	3
Message from the CINF Program Chair	4
CINF Symposia List – ACS Fall 2019 Meeting	6
CINF Social Networking Events	8
CINF Meetings and Events	9
Applications Invited for CSA Trust Grants for 2020	10
Report on the Council Agenda	16
New IUPAC Project to Develop SMILES+	18
CINF Member Profile: Ye Li	20
Twenty-five Years Ago in Washington	22
Book Review	24
Sponsor Announcements	26
CINF Officers and Functionaries	35
Contributors to This Issue	38
CINF Technical Program – ACS Fall 2019 Meeting	40
CINF Technical Program with Abstracts – ACS Fall 2019	62

ISSN: 0364–1910

[Cover image](#) is courtesy of Flickr user olegshpyrko (CC BY 2.0 license).

Chemical Information Bulletin

© Copyright 2019 by the Division of Chemical Information of the American Chemical Society

## Letter from the Editor

Welcome to another fall issue. Inside you will find information about the CINF social networking events scheduled to take place at the upcoming fall National Meeting of the American Chemical Society, along with the Technical Program with abstracts (please excuse the formatting). We also have Wendy Warr's report about the 1994 fall meeting in Washington, Bob Buntrock's review of *Applied Chemoinformatics: Achievements and Future Opportunities*, and Donna Wrublewski's interview with Ye Li. For the first time, Donna, Ye, and Steve Wathen will be teaching a pre-ACS workshop titled "Reproducible Data Analysis and Publishing in Chemistry with R" on Saturday, August 24, 2019, 9am-2pm in San Diego, CA.

Please contact [yel@mit.edu](mailto:yel@mit.edu) if you are interested attending. More details about the workshop can be found here:

[https://drive.google.com/file/d/1-KAa5XWngKH1pIRqhMVJ2lb7\\_edFQJqT/view?usp=sharing](https://drive.google.com/file/d/1-KAa5XWngKH1pIRqhMVJ2lb7_edFQJqT/view?usp=sharing).

I would like to thank our generous sponsors for supporting our symposia and receptions, and for providing updates on their products and services. I also wish to thank everyone who contributed articles to this issue, as well as those who gave their time to assist with the copyediting. If you are interesting in writing something for an upcoming issue, please contact me or another *CIB* editor. The winter and summer issues that follow the national meetings are usually packed with symposium reports, so articles for the spring and fall issues are particularly welcome.

I hope to see many of you here in San Diego!

Teri M. Vogel

[tmvogel@ucsd.edu](mailto:tmvogel@ucsd.edu)

## Notice

It was reported on CHMINF-L that longtime CINF member William G. (Bill) Town passed away in June. Bill was awarded the CINF Meritorious Service Award in 2008 for outstanding contributions to the division. He was a division chair, served on a number of CINF committees (including Awards, Program, and Publications), and programmed symposia for multiple national meetings.

A tribute to Bill and his contributions to CINF will be included in an upcoming issue.



(Picture of Bill Town, from Baykoucheva. S.; A Career in Chemistry and Chemical Information? Interview with Bill Town, Chair of the CINF Publications Committee. *Chem. Info. Bull.* **2009**, 61(1), 16-17. <https://acscinf.org/content/career-chemistry-and-chemical-information>.)

## Message the Program Chair

Please register and attend our awesome CINF Division program at the San Diego meeting, August 25 – 28, at the Omni San Diego Hotel at 675 L St, San Diego, CA. You will have your choice of over 170 speakers. There will be three symposia every day, Sunday through Wednesday, along with a new workshop on Saturday, “Reproducible Data Analysis and Publishing in Chemistry with R.” We will recognize the 2019 Herman Skolnik Award recipient Dr. Kimito Funatsu for his contributions to structure elucidation, *de novo* structure generation, and applications of cheminformatics methods to materials design and chemical process control. Topics for our other symposia include text-mining, natural language processing, patent novelty searching, student success, chemical nomenclature, drug discovery, data visualization, open-source tools, materials informatics, extended reality, machine learning, artificial intelligence, crystal structures, and web-based databases. Details are in this *Chemical Information Bulletin (CIB)* and on our website with updates in the online planner - see [https://plan.core-apps.com/acs\\_sd2019/customScreen/aboutShowPhoenix](https://plan.core-apps.com/acs_sd2019/customScreen/aboutShowPhoenix).

Thank you to past program planners, organizers, and speakers. I want especially to acknowledge Erin Davis who was a dynamic chair and has been a long-term program organizer in the area of drug discovery informatics. Tony Williams has also put together amazing programs in the areas of open chemistry databases and sharing content. Michael Qiu has done a wonderful job taking care of the logistics of our social events, ordering the best food, and working with conference staff to ensure it is what we want. Sadly, Erin, Tony and Michael will be taking a break from organizing, and we wish them the best in their future endeavors! ACS CINF is not a lifetime commitment. Join us when you can because it takes many hands to develop an excellent program.

If you will be at the upcoming San Diego national meeting, attend the Program Committee meeting on Saturday August 24, from 12:30 – 2:00 pm in Gallery 1 in the Omni San Diego Hotel to discuss the future program.

Do you want to speak? The Call for Papers for the Spring ACS national meeting in Philadelphia will be open for abstract submission on August 12 - October 14. Please do not wait until the last minute. The Philadelphia meeting will be held March 22 - 26, 2020.

Do you have an idea for a great program for future meetings? The symposium ideas can be sent any time via email to committee members or via this Google form at <https://forms.gle/BuGwyrwx3RdBWXMxGA>. The next opportunity for organizing a program will be for the San Francisco meeting in August 2020.

If your program is ready to be submitted to the Call for Papers for the San Francisco meeting, please use this Google Form at <https://forms.gle/smA6T4Bt1XZhRzGK8>. The symposia titles and descriptions will be due in late November 2019.

Would you like to join the Program Committee? Current committee members are listed on our website at <https://acscinf.org/content/program>. We are always looking for people to sort through the ideas, to organize or assist organizers, and to manage the logistics of putting the program together. It is a great way to network and to keep up with the trends in our field.

See you at the conference!

Sue Cardinal

Chair, Program Committee

[scardinal@library.rochester.edu](mailto:scardinal@library.rochester.edu)

# CINF Symposia – August 25-28, 2019 – Omni San Diego Hotel

Please review the Technical Program at the end of this issue (starting on page 40 for schedule, and starting on page 62 for abstracts) or the ACS online planner for details about each symposium.

## SUNDAY MORNING

Text-Mining & Natural Language Processing for Chemical Information: From Documents to Knowledge

Nothing New Under the Sun: The Practical Challenges of Patent Novelty Searching

Importance of Collaboration to Create Student Success in the Laboratory & Beyond

## SUNDAY AFTERNOON

Text-Mining & Natural Language Processing for Chemical Information: from Documents to Knowledge

Chemical Nomenclature & Representation: Past, Present & Future

Importance of Collaboration to Create Student Success in the Laboratory & Beyond

## SUNDAY EVENING – San Diego Convention Center

CINF Scholarships for Scientific Excellence: Student Poster Competition

## MONDAY MORNING

Driving Drug Discovery via Innovative Data Visualization

Chemical Nomenclature & Representation: Past, Present & Future

Successful Projects Fueled by Open-Source Tools

## MONDAY AFTERNOON

Driving Drug Discovery via Innovative Data Visualization

Chemical Nomenclature & Representation: Past, Present & Future

Materials Informatics

## MONDAY EVENING – San Diego Convention Center

Sci-Mix

## TUESDAY MORNING

Herman Skolnik Award Symposium Honoring Dr. Kimito Funatsu

Extended Reality (XR) in Libraries & Beyond

Drug Discovery: Informatics Approaches

## TUESDAY AFTERNOON

Machine Learning & Artificial Intelligence in Computational Chemistry

One Million Crystal Structures: a Wealth of Structural Chemistry Knowledge

Biologic Informatics

## WEDNESDAY MORNING

Machine Learning & Artificial Intelligence in Computational Chemistry

One Million Crystal Structures: a Wealth of Structural Chemistry Knowledge

Web-Based Chemistry Databases

## WEDNESDAY AFTERNOON

Machine Learning & Artificial Intelligence in Computational Chemistry

One Million Crystal Structures: a Wealth of Structural Chemistry Knowledge

Web-Based Chemistry Databases



## CINF Social Networking Events at the Fall 2019 ACS Meeting



Please Join Us At These  
Division of Chemical Information Events!

The ACS Division of Chemical Information is pleased to host the following social networking events at the Fall 2019 ACS National Meeting in San Diego, CA.

### **Welcoming Reception & Scholarship for Scientific Excellence Posters**

6:30 - 8:30 pm, Sunday, August 25<sup>th</sup> – Exhibit Hall A, San Diego Convention Center

Reception co-sponsored by: **Journal of Chemical Information & Modeling (ACS Publications)**,  
**InfoChem**, and **Thieme Chemistry**.

Scholarships for Scientific Excellence

Sponsored exclusively by: **ACS Publications**



### **Herman Skolnik Award Symposium & Award Presentation Honoring Prof. Kimito Funatsu, University of Tokyo**

Symposium: 8:30am-12:30pm Tuesday, August 27<sup>th</sup> – Grand Ballroom A, Omni San Diego Hotel

Sponsored exclusively by: **Schrödinger**.

### **Herman Skolnik Award Reception**

**Honoring Prof. Kimito Funatsu, University of Tokyo**

Reception: 6:30pm-8:30pm Tuesday August 27<sup>th</sup> – Grand Ballroom C, Omni San Diego Hotel

Sponsored by: **Elsevier Reaxys**, **Google**, and **Wiley**.



## CINF Meetings and Events

### Saturday, August 24: 9:00 am-3:00 pm

- Reproducible Data Analysis and Publishing in Chemistry with R – Omni San Diego Hotel, Gallery 2

### Saturday, August 24: 12:30-2:30 pm

- CINF Awards Committee – Omni San Diego Hotel, Gallery 3A
- CINF Program Committee – Omni San Diego Hotel, Gallery 1

### Saturday, August 24: 3:00-6:00 pm

- CINF Executive Committee – Omni San Diego Hotel, Gallery 1

### Sunday, August 25: 12:00-2:00 pm

- CSA Trust Meeting – Omni San Diego Hotel, Gaslamp 5

### Sunday, August 25: 6:30-8:30 pm

- CINF Poster Session & Welcome Reception – San Diego Convention Center, Exhibit Hall A

### Tuesday, August 27: 6:30-8:30 pm

- Herman Skolnik Award Reception Honoring Kimito Funatsu – Omni San Diego Hotel, Grand Ballroom C



## Chemical Structure Association Trust

### Applications Invited for CSA Trust Grant for 2020

The Chemical Structure Association (CSA) Trust is an internationally recognized organization established to promote the critical importance of chemical information to advances in chemical research. In support of its charter, the Trust has created a unique grant program and is now inviting the submission of grant applications for 2020.

### 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 British pounds 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; e.g. 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;
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

The application deadline for the 2020 grant is March 28, 2020. Successful applicants will be notified no later than May 7, 2020.

### Address for Submission of Applications

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

## Chemical Structure Association Trust: Previous Grant Awardees

2018

**Stephen Capuzzi**, *Division of Chemical Biology and Medicinal Chemistry at the University of North Carolina Chapel Hill Eshelman School of Pharmacy, U.S.A.*, was awarded a grant to attend the 31<sup>s</sup> ICAR in Porto, Portugal from 06/11/2018 to 06/15/2018, where he presented his research entitled “Computer-Aided Discovery and Characterization of Novel Ebola Virus Inhibitors.”

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

**Mark Driver**, *Chemistry Department, University of Cambridge, U.K.*, was awarded a grant to offset costs to attend the 7<sup>th</sup> EUChEMS conference where he will present a poster on his research that focuses on the development and applications of a theoretical approach to model hydrogen bonding.

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

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

2017

**Jesus Calvo-Castro**, *University of Hertfordshire, England*, was awarded a grant to cover travel to present his work at the Fifth International Conference on Novel Psychoactive Substances, to be held in Vienna, Austria from August 23-26, 2017. He works on the development of novel methodologies for the in-the-field detection of novel psychoactive substances (NPS), where chemical structure and information play a crucial role.

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

## 2016

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

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

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

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

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

## 2015

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

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

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

**Aleix Gimeno Vives**, *Cheminformatics and Nutrition Research Group, Biochemistry and Biotechnology Dept., Universitat Rovira i Virgili, Spain*, 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, Hungary*, was awarded a grant for travel to study at the University of Oxford with Dr. Robert S. Paton, a 2013 CSA Trust Grant winner, in order to increase

his experience in the development of computational methodology which is able to accurately model realistic and flexible transition states in chemical and biochemical reactions.

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

**Dr. David Palmer**, *Department of Chemistry, University of Strathclyde, Scotland, U.K.*, was awarded a grant to present a paper at the fall 2014 meeting of the American Chemical Society on a new approach to representing molecular structures in computers based upon ideas from the integral equation theory of molecular liquids.

**Sona B. Warriar**, *Departments of Pharmaceutical Chemistry, Pharmaceutical Biotechnology, and Pharmaceutical Analysis, NMIMS University, Mumbai, India*, 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, U.S.A.*, was awarded a grant for travel to speak on “Structure-property relationships of molecular precursors to organic electronics” at a workshop sponsored by the Centre Européen de Calcul Atomique et Moléculaire (CECAM) that took place October 22 – 25, 2013 in Lausanne, Switzerland.

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

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

#### 2012

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

#### 2011

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

2010

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

2009

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

2008

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

2007

**Rajarshi Guha**, *Indiana University, Bloomington, IN, U.S.A.*, was awarded a grant to attend the Gordon Research Conference on Computer-Aided Design in August 2007.

2006

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

2005

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

2004

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

2003

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

2001

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



# Report on the Council Agenda for August 28, 2019

Council meets on Wednesday, August 28, 2019 beginning, at 8:00 am in the Hilton San Diego Bayfront Hotel, Sapphire Ballroom, San Diego CA. All ACS members are welcome to observe Council meeting, and a special seating area is available.

## Elections

- Council will elect five individuals to the Council Policy Committee; the four candidates receiving the most votes will be elected for 2020-2022 term, and the candidate receiving the fifth highest vote will be elected for a one-year term for 2020.
  - Candidates are: George M. Bodner, Joseph A. Heppert, James C. Carver, Lydia E. M. Hines, Dee Ann Casteel, Will E. Lynch. Kenneth P. Fivizzani, Sally B. Peters, Anne M. Gaffney, Margaret J. Schooler
- Council will elect five individuals to the Committee on Committees, with the five candidates receiving the highest numbers of votes elected for the 2020-2022 term.
  - Candidates are: Satinder Ahuja, Sarah M. Mullins, Lisa M. Balbes, Jason E. Ritchie, D. Richard Cobb, Susan M. Schelble, Harry J. Elston, Andrea B. Twiss-Brooks. Emilio X. Esposito, Stephanie J. Watson
- Council will elect five individuals to the Committee on Nominations and Elections, with the five candidates receiving the highest numbers of votes elected for 2020-2022 term.
  - Candidates are: V. Dean Adams, Alan M. Ehrlich, Mark A. Benvenuto, Alan A. Hazari, Michelle V. Buchannan, Amber S. Hinkle, Charles E. Cannon, Thomas H. Lane, Alan B. Cooper, Joseph P. Stoner

## Other Actions

- Council will vote on the Committee on Committees (ConC) recommendations for continuations of selected committees. This review and recommendation is part of a scheduled regular review of committees.
- Council will vote on the recommendation by the Committee on Nominations and Elections (N&E) to realign the distribution of member population within the six electoral districts by transferring the Pittsburgh Local Section from District II to District III to bring both districts within the permissible range for equitable representation.
- Council will vote on whether to approve an update to “The Chemical Professional’s Code of Conduct”.
- Council will vote on a petition to charter a new International Chemical Sciences Chapter for the republic of Georgia.



## For Consideration Only

- Councilors have been asked to consider a petition on Membership and Dues, that will amend Bylaw I, Sec. 3,b; Bylaw II, Sec. 3, a-c; Bylaw XII, Sec. 3, a, d-l – According to the explanation accompanying the petition: “The petitioners seek to amend the ACS Bylaws by moving intact sections relating to the associated benefits of membership, how dues are set, and the dues discounts available into a separate process and procedures document. The Membership Affairs Committee (MAC), with input from the Budget and Finance Committee (B&F), would use this new document to develop and manage these areas, with revisions continuing to require council approval for implementation. The procedures would allow use of the dues escalator but give council the flexibility to make changes (adaptations) not permitted currently.” CINF members who would like additional explanation of this petition are encouraged to contact [meminfo@acs.org](mailto:meminfo@acs.org) and to attend the open meeting of MAC in San Diego. The online planner for the San Diego meeting has MAC meeting information as follows: Location: Sapphire Ballroom KL, Hilton San Diego Bayfront, Sunday, August 25, 3:00 pm.

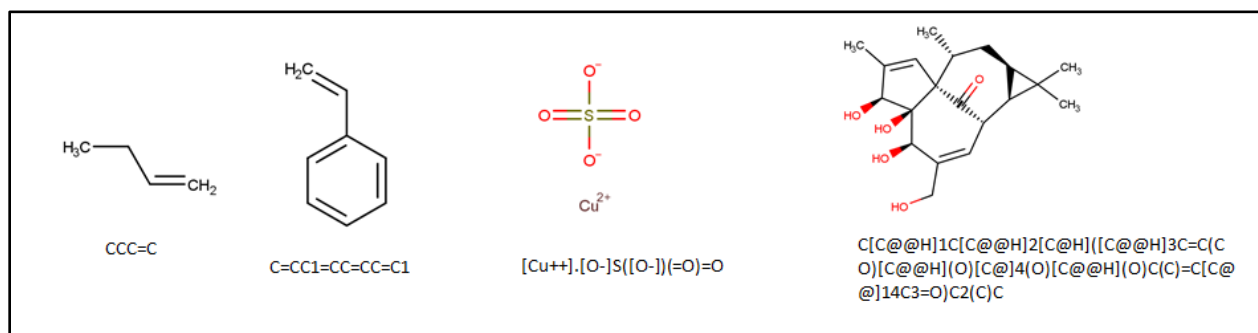
### CINF Councilors

Bonnie Lawlor

Svetlana N. Korolev

Andrea Twiss-Brooks

## New IUPAC Project to Develop SMILES+



Most cheminformaticians will have used, taught or at least be familiar with the SMILES notation, developed by Daylight Chemical Information Systems in the late 1980s. Although SMILES is still widely utilised, unfortunately it is no longer centrally supported or developed to address new requirements and there is no up-to-date documentation. The consequence is that non-standard SMILES dialects and extensions have been developed, usually for very good reasons, by individual organisations to address new use cases. But this evolution of SMILES is hampering interoperability and the sharing of information and data between systems and organisations. The '+' in SMILES+ is to indicate that it's intended that future approved extensions to a core IUPAC SMILES specification will be able to be accommodated.

This new four-year project, being undertaken by a global team led by Vincent Scalfani from the University of Alabama, was approved by IUPAC in April 2019 and reports in to its Committee on Publications and Chemical Standards.

It is sometimes asked why this project is necessary when InChI already exists. InChI and SMILES serve complementary chemical information use cases and having an up-to-date, maintained standard version of SMILES will be an important stepping stone to elaborate on these use cases. To ensure common objectives can be determined and achieved, the SMILES team is in collaboration with the InChI Trust.

Open SMILES, a community sponsored open-standards version of the SMILES language, will form the starting point for a working draft of the IUPAC SMILES+ specification which is being developed openly on GitHub, see

[https://github.com/vfscalfani/IUPAC\\_SMILES\\_plus/blob/master/IUPAC\\_SMILES%2B.asciidoc](https://github.com/vfscalfani/IUPAC_SMILES_plus/blob/master/IUPAC_SMILES%2B.asciidoc).

Early goals for the SMILES+ project team include:

- Reviewing the existing Open SMILES specification, deciding which sections could be removed and identifying obvious gaps which should be addressed.
- Looking at existing documentation and manuals from software providers and building an annotated bibliography of current toolkit SMILES documentation.

- Reaching out to known providers of toolkits which make use of SMILES (CDK, ChemDoodle and OpenBabel are self-declared users of Open SMILES). This will assist development of a test suite of structures, including edge cases.
- Building a list of stakeholder contacts. We are seeking community input from the outset to help determine requirements. This stakeholder group will be engaged throughout the project - please contact Helen Cooke ([helen.cooke100@gmail.com](mailto:helen.cooke100@gmail.com)) if you are interested in joining this group.

Effective communication will be key to the success of this initiative. The project team has this in mind and will present regularly at appropriate events and meetings and communicate through a variety of channels. Look out for us at the InChI Symposium 23-24 August 2019 and the ACS Fall Meeting 25-29 August. Ultimately, the plan is for the specification to be published as an IUPAC recommendation.

Further information about the project is on the IUPAC web site [https://iupac.org/projects/project-details/?project\\_nr=2019-002-2-024](https://iupac.org/projects/project-details/?project_nr=2019-002-2-024).

Dr. Helen Cooke  
CICAG Committee and IUPAC SMILES+ Team

Reprinted with permission from Cooke, H. New IUPAC Project to Develop SMILES+. *CICAG Newsletter*, Summer 2019, p. 6. [http://www.rscicag.org/index.htm/files/CICAG%20Newsletter%20Summer%202019\\_tcm18-251527.pdf](http://www.rscicag.org/index.htm/files/CICAG%20Newsletter%20Summer%202019_tcm18-251527.pdf). Copyright 2019 RSC CICAG.

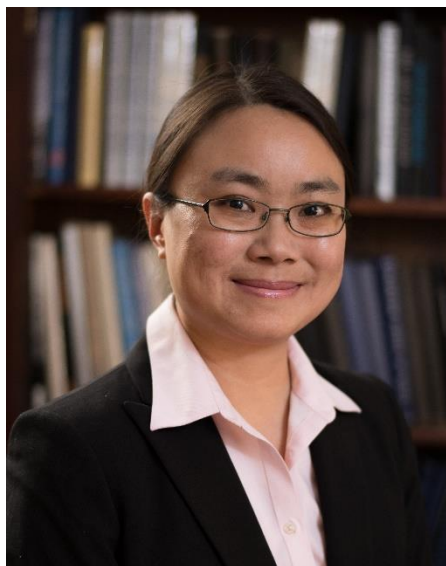
## CINF Member Profile: Dr. Ye Li

### Who are you?

I am a science librarian with intercultural experience. I grew up in southwest China and then studied chemistry at Beijing Normal University. I moved to the United States in 2004 and received both my doctorate in chemistry and master's in library and information Science in 2009. I started my career as a chemistry librarian at the University of Michigan in 2009 and then led scholarly communications initiatives at the Colorado School of Mines from 2016 to 2018.

### What do you do?

I am currently the Librarian for Chemistry and Chemical Engineering at the Massachusetts Institute of Technology. I help students and researchers navigate the chemical information landscape and also partner with them to manage and share their research outputs. My goal is to enable open and reproducible research through cultivating good research practices among chemists.



### Why are you in the chemical information field?

When I was in graduate school as a chemist, I found myself proactively helping my labmates find and organize research information. My fascination with scientific information later prompted me to pursue a master's in library and information science. My background in both fields led to my first job as a chemistry librarian, where I recognized how academic librarians' roles have changed from gatekeepers to facilitators of academic success and interdisciplinary collaborations. My exploration of the scholarly communication domain broadened my horizon to think about all types of research outputs, including research data, throughout the research life cycle. Now, I see my mission in the chemical information field as contributing to the infrastructure, tools, standards, and good practices that enable open and reproducible research in chemistry.

### What makes CINF valuable to you?

CINF has been my professional home because it's where I can connect with respectful and friendly colleagues including librarians and other information professionals. We share the same passion, learn from each other, and collaborate on many fronts. CINF has given me the space not only to share my research and explorations, but also to receive constructive feedback and work with close collaborators. I have been a part of the CINF Education Committee and am now part of the Program Committee. Through coplanning symposiums on chemical information literacy, open access, data sharing, Wikipedia, libraries as innovation centers, and other topics, I learned how to bring different perspectives together to generate dynamic conversations on hot topics during the ACS national meetings. For this upcoming ACS national meeting, I am cohosting

a hands-on workshop titled “Reproducible Data Analysis and Publishing in Chemistry with R”, together with Donna Wrublewski and Steve Wathen. Our hope is to cultivate good data practices that will lead to open and reproducible research among chemists.

*Please visit Ye's ORCID Profile (<https://orcid.org/0000-0001-8361-6916>) for more information about her projects and publications.*

Donna Wrublewski  
Chair, Membership Committee

## Twenty-five Years Ago in Washington

This article is dedicated to the memory of Dr. William Geoffrey Town, March 31, 1943, to June 24, 2019. Bill was a member of the CINF Publications Committee in 1993-1996 and chaired the Awards Committee in 2003-2005, the Nominating Committee in 2001, and the Publications Committee in 2009-2012. He served the division as Chair-Elect in 1999, Chair in 2000, and Past-Chair in 2001, and was Alternate Councilor in 2006-2008. He won the 2008 CINF Meritorious Service Award. The CIB hopes to print an obituary in an upcoming issue.

By coincidence, as I begin my trip down memory lane, and stop at the fall 1994 ACS meeting in Washington, DC, I find that this was the meeting where Bill and I ran joint CINF symposia on competitive intelligence. My own symposium was entitled “Inventing the Future: Competitive Intelligence in Strategic Planning and Decision Making”, and Bill (who then worked for Derwent Publications) organized “Patent Citation Analysis: a Tool for Competitive Intelligence?”

In his Chairman's (*sic*) Message, in the CIB, Gerry Vander Stouw said that the technical program was an intriguing mix of the theoretical and the practical. Two of the symposia were certainly unusual for the division: the graph theory symposium in honor of Alexandru Balaban, and a symposium on the ordering, classification, and display of concepts in chemistry, organized by Sandor Barcza (remember the “Hungarian count” who used to have his jacket slung around one shoulder?). In the latter symposium, Seymour B. Elk spoke on orismology, and what constitutes a ring (not the same as a “cycle” in graph theory). Some of you will remember “all those papers with lots of rings in them” that used to appear in the *Journal of Chemical Information and Computer Systems (JCICS)*. *JCICS* was strong on graph theory in those days.

There was much talk of the Internet, the information superhighway, and the impact on publishing. One CINF symposium was devoted to copyright, and there was discussion on “electrocopying”, that is, electronic copying as opposed to photocopying. The writing was on the wall for the myriads of CD-ROM products available in fall 1994. The possibilities for online full text were being investigated. Michael Lesk of Bellcore spoke about the CORE “electronic chemical journals” experiment, in which Cornell University, ACS, CAS, Bellcore, and OCLC were collaborators.

CJACS Plus was a new STN service offering page images from 1992-1994 from 23 ACS journals. It allowed display and print of full articles, as scanned page images, using STN Express, as CAS offline prints, or by fax. The documentation said, “Communication packages other than STN Express can also be used so long as Kermit is available and you have a 14,400 bps modem and CompuServe (for example). You can use any computer and software that can capture TIFF images compressed in Group 4 fax format. It takes about two minutes a page on Internet.”

Jeffrey Spring of ACS and co-authors, presented a poster on processing author manuscripts received on diskette. In January 1993, less than 5% of manuscripts were received in softcopy; by July 1994 that had risen to about 45%.

In computational chemistry, neural networks were the “in thing”. The Gasteiger-Zupan book on neural networks in chemistry had been published in 1993. The COMP program in Washington included a symposium on artificial intelligence (AI) and innovative computing in chemistry and drug design. I myself gave a talk on computer-aided structure elucidation. I was very interested in CSEARCH and SpecInfo at the time. Other papers covered molecular similarity, QSAR, pharmacophore mapping, *de novo* design, and other topics. Fast forward 25 years, and AI and big data are being hyped in a big way.

In the exhibition, Greg Banik, then at ISI, was showing the planned Reaction Citation Index, for use with Current Chemical Reactions and REACCS. Two “electronic notebooks” were on show: Helix Research Station and the Forefront Group’s Virtual Notebook System. The era of ELNs was beginning.

MDL had no booth but had a hospitality suite, shared with BIOSYM Technologies, in a nearby hotel. My report notes that “MDL and BIOSYM seem to be attached like a horse and carriage nowadays.” MSI had just bought BIOCAD. Eventually, all these companies ended up in the Accelrys camp. In 1994, Tripos Associates became independent of Evans and Sutherland, and became a publicly listed company, Tripos, Inc. That company, sadly, folded some years later.

I had written a review of the Windows version of the Current Facts in Chemistry CD-ROM for the February 1995 issue of *DATABASE* magazine and was impressed with the new interface, identical with that for in-house CROSSFIRE, which was on show at the ACS meeting. Beilstein Information Systems promised CROSSFIRE Online for 1995 with the same interface. I will be able to tell you about that if, in six months’ time, I am able to produce an article about the meeting in Anaheim, CA, in spring 1995, but, for an even earlier CIB, I must work on an obituary. Please join me in remembering Bill Town, my dear friend and colleague of over 40 years.

Rest in peace, Bill.

Wendy Warr

## Book Review

*Applied Chemoinformatics: Achievements and Future Opportunities*, Engel, T., Gasteiger, J., Eds.: Wiley-VCH, Weinheim, Germany, 2018, 617 pp + xxvi, ISBN 978-3-527-34201-3 (paperback), \$175.

This tome/textbook is a companion to another book on chemoinformatics (1), previously reviewed (2). Like the former, it has multiple authors, 52 in this case. This book has 14 chapters, but several have subsections, some of those qualify as subchapters (Chapter 6 has 13), and all chapters and sub-chapters conclude with an “Essentials” sidebar or a conclusion, sections on available software and web services and/or a selected reading list. Every chapter also has a list of chapter-specific references (Wendy Warr is cited several times), and there is an overall index at the end of the book.

The focus is obviously on computer aspects of chemoinformatics, but more “classic” information topics and methods are included, making this book of additional interest to CINF members, librarians, information specialists, and those who may not necessarily be computer or IT experts. The foreword and introductory section of chapter 1 are identical to those of the previous book (1).

Chapter 2 covers QSAR and QSPR. As an editorial note, I am pleased that QSPR (Quantitative Structure-Property Relationships) are treated as a broader class of QSAR, that is, activity is properly considered a property. This is something this outsider was proposing from the early days of QSAR 40 years ago). Flow charts for modeling process are shown. Application of individual approaches for the many properties of molecules that are not directly predictable, as well as machine learning and validation of models, are considered.

Chapter 3 covers prediction of physicochemical properties, including melting points and  $pK$  values, and estimation of  $\log P$  and the use of nonlinear methods. Chapter 4 covers chemical reactions and consists of three subchapters, including reaction prediction and biochemical pathways. The former includes two interesting sections titled “What is a good reaction?” and “Can we trust the literature?”. The latter displays the “Biochemical Pathways” wall chart, describes searching methods for relevant databases, and displays the resulting SOMs (self-organizing maps). Chapter 5 describes structure-spectrum correlations and computer-assisted structure elucidation. (We have come a long way in 50 years when we chemistry students struggled with Silverstein and Bassler). Spectra covered are IR, NMR, and mass spectra (but not UV/visible).

Chapter 6, on drug discovery, is the longest and most complex with 13 subchapters. These subsections are Overview; Bridging Information on Drugs, Targets and Diseases; Chemoinformatics in Natural Product Research; Chemoinformatics of Chinese Herbal Medicines; PubChem; Pharmacophore Perception and Applications; Prediction, Analysis, and Comparison of Active Sites; Structure-Based Virtual Screening; Prediction of ADME Properties; Prediction of Xenobiotic Metabolism; Chemoinformatics at the CADD Group of the National Cancer Institute; Uncommon Data Sources for QSAR Modeling; and Future Perspectives of Computational Drug Design.



Chapter 7 deals with Computational Approaches in Agricultural Research including estimation of adverse effects, risk assessment, and registration. (My first two jobs were in agricultural chemical synthesis, more than 40 years ago. Once again, we have come a long way). Chapter 8 covers chemoinformatics in modern regulatory science. Chapter 9 deals with chemoinformatics in analytical chemistry, including translation of data into information. Chapter 10 covers chemoinformatics in food science. Topics considered include food information databases, structure-flavor relationships, and “flavor cliffs” (when two, otherwise-similar compounds include one flavored and one not).

Chapter 11 describes computational approaches to cosmetic products, including skin models. Chapter 12 covers applications in materials science. Such applications are more difficult than those in molecules. Mathematical descriptions of materials are covered, as are pitfalls in the processes. Chapter 13 covers process control and soft sensors. The latter are not analyzers; they monitor process variables in real time. In Chapter 14, Gasteiger describes future advances in well-established fields, emerging fields, and the renaissance of some fields.

As with the first book, this book should be available in every chemistry department and library.

#### References

(1) *Chemoinformatics: Basic Concepts and Methods*, Engel, T., Gasteiger, J., Eds.; Wiley-VCH, Weinheim, Germany, 2018

(2) *CIB*, 71 (2), Summer 2019, p. 19-20, [https://bulletin.acscinf.org/PDFs/CIB\\_71\\_2.pdf](https://bulletin.acscinf.org/PDFs/CIB_71_2.pdf).

R. E. Buntrock  
Buntrock Associates



## Division of Chemical Information Sponsors Fall 2019



The American Chemical Society Division of Chemical Information is very fortunate to receive generous financial support from our sponsors to maintain the high quality of the Division's programming, to promote communication between members at social functions at the ACS Fall 2019 National Meeting in San Diego, CA, 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:**

<b>Gold</b>	<b>ACS Publications</b> <b>Schrödinger</b>
<b>Silver</b>	<b>Google</b> <b>Wiley</b>
<b>Bronze</b>	<b>Chemical Abstracts Service</b> <b>Elsevier Reaxys</b> <b>Thieme Chemistry</b>
<b>Contributor</b>	<b><i>Journal of Chemical Information &amp; Modeling</i> (ACS Publications)</b> <b>Bio-Rad Laboratories</b> <b>InfoChem</b>

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

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

Graham Douglas  
Chair Pro Tem, Fundraising Committee 2019  
Email: [Sponsorship@acscinf.org](mailto:Sponsorship@acscinf.org)  
Tel: 510-407-0769

[The ACS CINF Division is a non-profit tax-exempt organization with taxpayer ID no. 52-6054220.](#)



Congratulates Professor Kimito Funatsu,  
University of Tokyo,

on receiving the 2019 Herman Skolnik Award  
of the ACS Division of Chemical Information  
for his contributions to structure elucidation, *de novo* structure generation  
and applications of cheminformatics methods to materials design and  
chemical process control.

Thank you for your contributions.

## Google Patents and Google BigQuery

### Making chemistry information universally accessible and useful

Google Patents is a resource for providing public access to worldwide intellectual property. It continues to expand its reach, now offering content from over 100 countries, translations of most patent content into English, and patent search integrated with Google Scholar. Advanced search options include similarity searching and machine classification of non-patent literature.



Developments to make Google Patents chemistry- and life-science-aware are of particular interest to the scientific and legal communities. Curated data is being made publicly available in Google BigQuery, a massive, cloud-based relational database with the potential to significantly enhance the way the scientific community uses information.

BigQuery processes billions of rows and terabytes of data in tens of seconds, and scales to petabytes of data at low cost. Data providers contribute database tables for public access, or protect their tables with access control lists (ACLs) for private or subscriber access.

Data from tables uploaded and maintained by different providers can be jointly analyzed with SQL so users can focus on analyzing data instead of maintaining database infrastructure. Efforts are accelerating to increase the scientific content in BigQuery, with broad applications across multiple domains.

Google Patents and Google BigQuery are providing a publicly-available platform that leverages content from disparate sources to cross-map scientific content with capabilities beyond what is easily done today.

Learn more:

<https://support.google.com/faqs/answer/6390996>

<https://cloud.google.com/bigquery/>

<https://cloud.google.com/bigquery/docs/>

Explore great inventions at  
<http://patents.google.com>



Congratulations to Prof. Kimito Funatsu!

*Molecular Informatics* and Wiley-VCH would like to congratulate Kimito Funatsu on receiving the 2019 Herman Skolnik Award. Prof. Funatsu has frequently published his excellent research in *Molecular Informatics* and is a highly valued author. He has also served as a guest editor and currently serves on the Editorial Advisory Board.

Enjoy the following recent, free articles by Prof. Funatsu and his co-workers:

- [Novel Electrotopological Atomic Descriptors for the Prediction of Xenobiotic Cytochrome P450 Reactions](#) Kazuma Kaitoh, Masaaki Kotera, Kimito Funatsu; *Mol. Inf.* **2019**, DOI: 10.1002/minf.201900010
- [Random Forest Model with Combined Features: A Practical Approach to Predict Liquid-crystalline Property](#) Chia-Hsiu Chen, Kenichi Tanaka, Kimito Funatsu; *Mol. Inf.* **2019**, 38, 1800095. DOI: 10.1002/minf.201800095
- [Identification of Bioactive Scaffolds Based on QSAR Models](#) Tomoki Nakagawa, Tomoyuki Miyao, Kimito Funatsu, *Mol. Inf.* **2018**, 37, 1700103. DOI: 10.1002/minf.201700103

[Molecular Informatics](#) is an international publication of high-quality, interdisciplinary research on all molecular aspects of bio/cheminformatics and computer-assisted molecular design. *Molecular Informatics* presents methodological innovations that lead to a deeper understanding of ligand-receptor interactions, macromolecular complexes, molecular networks, design concepts, and processes that demonstrate how ideas and design concepts lead to molecules with a desired structure or function, preferably including experimental validation. The journal's scope includes, but is not limited to the fields of drug discovery and chemical biology, protein and nucleic acid engineering and design, the design of nanomolecular structures, strategies for modeling of macromolecular assemblies, molecular networks and systems, pharmaco and chemogenomics, computer-assisted screening strategies, and novel technologies for the *de novo* design of biologically active molecules. As a unique feature, *Molecular Informatics* publishes "Methods Corner" review-type articles, which feature important technological concepts and advances within the scope of the journal.

Wiley and *Molecular Informatics* are proud to be sponsoring the Division of Chemical Information.

### About Wiley

Wiley drives the world forward with research and education. Through publishing, platforms, and services, we help students, researchers, universities, and corporations to achieve their goals in an ever-changing world. For more than 200 years, we have delivered consistent performance to all of our stakeholders. The company's website can be accessed at [www.wiley.com](http://www.wiley.com).

WILEY-VCH

molecular  
informatics  
models – molecules – systems

WILEY

## Lundbeck selects Reaxys and Reaxys Medicinal Chemistry to improve internal and external data integration and access



*Big Data project will create comprehensive database of chemistry and bioactivity related content to help increase the visibility, reusability and actionability of existing information.*

**New York, February 26, 2019**

Elsevier, the information analytics business specializing in science and health, announced today that Lundbeck, the global pharmaceutical company specializing in brain diseases, has selected Elsevier's Reaxys and Reaxys Medicinal Chemistry (RMC) to support the ambitions of its research organization and accelerate data sharing and mining.

The current implementation will provide Lundbeck scientists with integrated access to vital chemical and biological information from the Reaxys and the Reaxys Medicinal Chemistry databases, as well as internal Lundbeck data, from a single, seamless interface. Both organizations have committed for multiple years, and envision projects involving procurement and inventory solution integration, predictive modelling and analytical solutions helping researchers to accelerate their work.

“Our goal is to develop therapies for complex brain diseases,” said Ludovic Tranholm Otterbein, Director Research Informatics & Operations at Lundbeck. “To achieve this, we need to further refine our analytical capabilities around drug discovery and get even more value from our existing data by reusing information we’ve generated over the years. The challenge has been to break down our internal data siloes and increase data interoperability so that this could become a reality.”

Life science companies today are struggling with the question of how to manage the vast amounts of data they are generating. The fragmented nature of life sciences R&D, which often involves having dozens of labs with hundreds of scientists scattered across different geographies, can make it very difficult for firms to know what experiments have already been conducted. As a result, researchers frequently end up unnecessarily duplicating experiments already done by their colleagues. Even when experimental data are captured by ELNs or LIMS, this information is not often available to other research teams, leading to wasted time and resources.

“All pharma companies know that data are the lifeblood of research, the challenge they have is ensuring that they can navigate those data to meaningfully apply it,” said Cameron Ross, Managing Director, Life Science Solutions, Elsevier. “They need intuitive tools which enable them to understand the essence of large datasets, so that they can apply the new knowledge to find better cures. We’ve drawn on our decades of experience with data management and taxonomies to develop solutions to make the lives of researchers as easy as possible and help them spend less time searching and more time innovating.

“Our project with Lundbeck also involves working together to integrate Reaxys solutions into its existing research ecosystem. This means not just providing out-of-the-box technical support but the right customizations and hands-on experience, as well, so that researchers can get optimal value out of the data.”

---

Reaxys retrieves literature, compound properties, and chemical reaction data faster than any other solution and, together with RMC, offers pharma companies a 'one-stop shop' chemistry ecosystem, thanks to its integration capabilities, innovative APIs, and bioactivity data. The Reaxys platform contains over 240 years of unparalleled chemistry content, including: 119 million organic, inorganic and organometallic compounds; 46 million chemical reactions; 500 million published experimental facts; 16,000 chemistry related periodicals; Asian, European, and U.S. patents; and six indexing sources for a cross-disciplinary view of chemistry. RMC offers access to data from a vast repository of peer-reviewed journal articles and patents and is interoperable with Reaxys.

---

### **About H. Lundbeck A/S**

H. Lundbeck A/S (LUN.CO, LUN DC, HLUY) is a global pharmaceutical company specializing in brain diseases. For more than 70 years, we have been at the forefront of neuroscience research. We are tirelessly dedicated to restoring brain health, so every person can be their best.

An estimated 700 million people worldwide are living with brain diseases, and far too many suffer, due to inadequate treatment, discrimination, a reduced number of working days, early retirement, and other unnecessary consequences. Every day, we strive for improved treatment and a better life for people living with brain diseases. We call this “Progress in Mind”. Read more at [www.lundbeck.com/global/about-us/progress-in-mind](http://www.lundbeck.com/global/about-us/progress-in-mind).

### **About Elsevier**

Elsevier is a global information analytics business that helps scientists and clinicians to find new answers, reshape human knowledge, and tackle the most urgent human crises. For 140 years, we have partnered with the research world to curate and verify scientific knowledge. Today, we're committed to bringing that rigor to a new generation of platforms. Elsevier provides digital solutions and tools in the areas of strategic research management, R&D performance, clinical decision support, and professional education; these include ScienceDirect, Scopus, SciVal, ClinicalKey, and Sherpath. Elsevier publishes over 2,500 digitized journals, including *The Lancet* and *Cell*, 39,000 e-book titles, and many iconic reference works, including *Gray's Anatomy*. Elsevier is part of RELX, a global provider of information-based analytics and decision tools for professional and business customers.

[www.elsevier.com](http://www.elsevier.com)

## Thieme's latest Science of Synthesis updates shed light on photocatalysis in organic synthesis and other topics

Science of Synthesis continues as the most up-to-date and comprehensive resource for synthetic chemists worldwide with the latest two updates.

Version 4.13 (March 2019):



### [Photocatalysis in Organic Synthesis, edited by B. König](#)

The Science of Synthesis Reference Library covers expert-evaluated content focusing on subjects of particular current interest. The latest volume in this series introduces the key basic concepts of photophysics and describes typical laboratory set-ups for photoredox catalysis, an area that has developed rapidly over the past 15 years. Written by pioneers and leaders in the field, the volume addresses newcomers and experts alike, presenting a collection of the most useful, practical, and reliable methods of photocatalysis. Thus, synthetic chemists are provided with the necessary insights to apply the new, visible-light-based tools immediately and reliably. In addition to detailed descriptions of key photocatalytic transformations, including representative experimental procedures, the volume includes a range of industrial case studies.

### [Knowledge Updates 2018/4](#)

The SOS Knowledge Updates complement the SOS Reference Library to make Science of Synthesis the complete information source for the modern synthetic chemist. Among the highlights is the major update by A. Gagnon et al. on the synthesis and application of bismuth compounds, an area that has seen much growth in recent years. Also of interest are the updates on carbamic acids and esters by J. Podlech, on tetraheterosubstituted methanes with a carbon-halogen bond by R. Zimmer et al., on various alkenylsulfur compounds by R. Kawecky, and on oxetanes and oxetan-3-ones by R. A. Croft and J. A. Bull.

### [Links to SynOne](#)

Related Thieme content can now be reached via links to and from Science of Synthesis to Thieme Chemistry's discovery tool **SynOne** (<https://synone.thieme.com>).



Version 4.14 (July 2019):

[Knowledge Updates 2019/1](#)

This release includes:

- A new chapter on the combination of gold catalysis with enzyme catalysis, organocatalysis, or transition-metal catalysis (I. Celik, S. Hummel, and S. F. Kirsch). Such dual catalytic approaches involving gold have received much attention in recent years as they can lead to unprecedented reactivity; this review presents recent highlights in the area.
- A completely revised chapter on lithium amides (C. T. Nieto, J. Eames, and N. M. Garrido). The synthesis and applications of both achiral and chiral lithium amides are reviewed, with a particular focus on the use of chiral lithium amides in asymmetric processes.
- A major update on the synthesis of isoquinolines, isoquinoline N-oxides, and isoquinolinium salts (B. S. Pilgrim and M. J. Tucker), with an emphasis on transition-metal-catalyzed routes to these important heterocyclic systems.
- Updates on the synthesis of imidic acids, isoureas, guanidines, and their derivatives (J. Podlech).

To get access to Science of Synthesis or a free trial, please visit: <http://sos.thieme.com>. For more information about Science of Synthesis, please visit the website at <http://www.thieme-chemistry.com/sos/>.

## *Journal of Chemical Information & Modeling* (ACS Publications)

*Journal of Chemical Information and Modeling* is excited to partner once again with the Division of Chemical Information. The editors and staff work hard to serve the needs



of CINF members and the global group of scientists engaged in developing new methodologies in the fields of chemical informatics and molecular modeling. This, we hope, is reflected in the many initiatives that we have launched: impactful special issues, Application Notes, and establishment of our Early Career Board. We invite your submissions for our upcoming special issues focusing on molecular simulation in Latin America, Cryo-EM, and new trends in virtual screening. As always, we are open to hearing your ideas. Please contact me at [eic@jciim.acs.org](mailto:eic@jciim.acs.org) anytime! Thanks so much for all your support, and all the best in 2020!

Kennie Merz  
Editor-in-Chief

# CINF Officers and Functionaries

## Chair

Elsa Alvaro  
Northwestern University  
[elsa.alvaro@northwestern.edu](mailto:elsa.alvaro@northwestern.edu)

## Chair-Elect

Jeremy Garritano  
University of Virginia  
[jq9jh@virginia.edu](mailto:jq9jh@virginia.edu)

## Past-Chair

Erin Davis  
Schrödinger, Inc.  
[erindavis@gmail.com](mailto:erindavis@gmail.com)

## Secretary

Tina Qin  
Harvard University  
[qinnamsu@gmail.com](mailto:qinnamsu@gmail.com)

## Treasurer

Stuart Chalk  
University of North Florida  
[schalk@unf.edu](mailto:schalk@unf.edu)

## CINF Councilors

Bonnie Lawlor  
[chescot@aol.com](mailto:chescot@aol.com)

Andrea Twiss-Brooks  
University of Chicago  
[atbrooks@uchicago.edu](mailto:atbrooks@uchicago.edu)

Svetlana N. Korolev  
University of Wisconsin, Milwaukee  
[skorolev@uwm.edu](mailto:skorolev@uwm.edu)

## CINF Alternate Councilors

Rachelle Bienstock  
RJB Computational Modeling LLC  
[rachelleb1@gmail.com](mailto:rachelleb1@gmail.com)

Charles Huber  
University of California, Santa Barbara  
[huber@library.ucsb.edu](mailto:huber@library.ucsb.edu)

Jeremy Garritano  
University of Virginia  
[jg9jh@virginia.edu](mailto:jg9jh@virginia.edu)

#### Archivist/Historian

Bonnie Lawlor  
[chescot@aol.com](mailto:chescot@aol.com)

#### Awards Committee Chair

Rajarshi Guha  
Vertex Pharmaceuticals  
[rajarshi.guha@gmail.com](mailto:rajarshi.guha@gmail.com)

#### Careers Committee Chair

Neelam Bharti  
Carnegie Mellon University  
[nbharti@andrew.cmu.edu](mailto:nbharti@andrew.cmu.edu)

#### Communications and Publications Committee Chair

Graham Douglas  
[communications@acscinf.org](mailto:communications@acscinf.org)

#### Education Committee Chair

Grace Baysinger  
Stanford University  
[graceb@stanford.edu](mailto:graceb@stanford.edu)

#### Finance Committee Chair

Stuart Chalk  
University of North Florida  
[schalk@unf.edu](mailto:schalk@unf.edu)

#### Fundraising Interim Committee Chair

Graham Douglas  
[communications@acscinf.org](mailto:communications@acscinf.org)

#### Membership Committee Chair

Donna Wrublewski  
Caltech Library  
[dtwrub@caltech.edu](mailto:dtwrub@caltech.edu)

## Nominating Committee Chair

Erin Davis  
Schrödinger, Inc.  
[erinsdavis@gmail.com](mailto:erinsdavis@gmail.com)

## Procedures Chair

Bonnie Lawlor  
[chescot@aol.com](mailto:chescot@aol.com)

## Program Committee Chair

Susan Cardinal  
University of Rochester  
[scardinal@library.rochester.edu](mailto:scardinal@library.rochester.edu)

## Webmasters

Rachelle Bienstock  
RJB Computational Modeling LLC  
[rachelleb1@gmail.com](mailto:rachelleb1@gmail.com)

Stuart Chalk  
University of North Florida  
[schalk@unf.edu](mailto:schalk@unf.edu)

## *Chemical Information Bulletin* Editor – Spring Issue

Kortney Rupp  
Lawrence Livermore National Laboratory  
[kortneyrupp@gmail.com](mailto:kortneyrupp@gmail.com)

## *Chemical Information Bulletin* Editor – Summer Issue

David Shobe  
Patent Information Agent  
[avidshobe@yahoo.com](mailto:avidshobe@yahoo.com)

## *Chemical Information Bulletin* Editor – Fall Issue

Teri Vogel  
University of California San Diego  
[tmvogel@ucsd.edu](mailto:tmvogel@ucsd.edu)

## *Chemical Information Bulletin* Editor – Winter Issue

Judith Currano  
University of Pennsylvania  
[currano@pobox.upenn.edu](mailto:currano@pobox.upenn.edu)

# Fall 2019 CINF Bulletin Contributors

## Articles and Features

Robert E. Buntrock  
Wendy Warr  
Donna Wrublewski

## Columns and Reports

Andrea Twiss-Brooks  
Sue Cardinal  
Helen Cooke  
Rajarshi Guha

## Sponsor Information

Graham Douglas

## Production

Judith Currano  
Svetlana Korolev  
David Shobe  
Teri Vogel  
Wendy Warr

This page left intentionally blank

# CINF Technical Program - ACS Fall 2019 Meeting

S. Cardinal, *Program Chair*

## SUNDAY MORNING

Section A

Omni San Diego Hotel  
Grand Ballroom A

### Text-Mining & Natural Language Processing for Chemical Information: From Documents to Knowledge

R. J. Bienstock, *Organizer*  
J. L. Nauss, *Organizer, Presiding*

**9:00** Introductory Remarks.

**9:05 CINF 1.** Enhancing data-driven summarization of relations between chemicals, genes, proteins, and diseases based on text mining of biomedical literature. **L. Zaslavsky**, A. Gindulyte, P. Thiessen, E. Bolton

**9:25 CINF 2.** Introducing automated polymer data extractor tool. **C. Dai**, K. Schmidt, D.Y. Zubarev, V.A. Piunova, K. Suruguchi, D.P. Sanders

**9:45 CINF 3.** Extraction of polymer-related information in the cheminformatics tool CIRCA. **K. Schmidt**, C. Dai, T.D. Griffin, K. Suruguchi, D.Y. Zubarev, V.A. Piunova, N. Park, J. Hedrick, L.C. Anderson, D.P. Sanders

**10:05** Intermission.

**10:20 CINF 4.** Abstract recommendation system: beyond word-level representations. V. Korolev, **A. Mitrofanov**, B. Sattarov, V. Tkachenko

**10:40 CINF 5.** Current challenges in text-mining for chemical information. **R.A. Sayle**, J.W. Mayfield, N. O'Boyle

**11:20 CINF 6.** MOLVEC: Open source library for chemical structure recognition. T. Peryea, D. Katzel, T. Zhao, N. Southall, **D. Nguyen**

**11:40** Concluding Remarks.

## SUNDAY MORNING

Section B

Omni San Diego Hotel



Grand Ballroom D

## **Nothing New Under the Sun: The Practical Challenges of Patent Novelty Searching**

Cosponsored by CHAL<sup>‡</sup> and CPRM<sup>‡</sup>

Financially supported by Patent Information Users Group (PIUG)

S. R. Adams, E. S. Simmons, *Organizers, Presiding*

**8:10** Introductory Remarks.

**8:20 CINF 7.** Patent novelty searching: Scope of prior art. **E.S. Simmons**

**8:50 CINF 8.** Quality of indexing for non-patent literature and its implications for retrieval of relevant prior art. **S.R. Adams**

**9:20 CINF 9.** Finding what others miss: Comprehensive prior-art resources for chemical substance searching. **J. Zabilski**

**9:50** Intermission.

**10:00 CINF 10.** Practical pointers for conducting prior art searches for small molecules, a patent practitioner's point of view. F.J. Koszyk, **X. Pillai**

**10:30 CINF 11.** What's new with scientific content in Google patents. **I. Wetherbee, S. Boyer, J. Frommer, V. Mehta, B. Arneson**

**11:00 CINF 12.** Non-patent literature citations in EPO opposition and limitation proceedings. **S.R. Adams**

**11:30 CINF 13.** Inventions for sale? Navigating the on-sale bar under the America Invents Act. **J.L. Krieger**

## **SUNDAY MORNING**

Section C

Omni San Diego Hotel  
Grand Ballroom E

## **Importance of Collaboration to Create Student Success in the Laboratory & Beyond**

Financially supported by CAS

S. P. Kuhn, *Organizer*

M. Pozenel, *Presiding*

**8:30** Introductory Remarks.

**8:35 CINF 14.** Science librarian, a chemical educator, and an EHS professional walk into a lab: Laboratory safety as a collaborative teaching tool. **L.R. McEwen, S.B. Sigmann, R. Stuart**

**9:00 CINF 15.** Innovative teaching collaboration provides students with practical drug discovery

experience. **T.E. Mansley**, R.L. Broadrup, B. Perry, **H. Ahamed**, **T. Aramburu**

**9:25 CINF 16.** Driving student success through undergraduate internships in biopharma. **N. Hawryluk**

**9:50 CINF 17.** Evolving data needs in chemistry and bio-sciences: What role can a librarian play in creating student success? **S. Ramachandran**, K. Howell

**10:15 CINF 18.** Training the biomedical workforce for long-term career success. **A. Bankston**

**10:40 CINF 19.** Business of student success. **J. Rosenberg**

**11:05** Panel Discussion.

**11:55** Concluding Remarks.

## **Bibliography of Chemistry**

### **Chemical Bibliography**

Sponsored by HIST, Cosponsored by CINF

## **Immersive Virtual Reality for Molecular Design**

Sponsored by COMP, Cosponsored by CHED, CINF and COMSCI

## **Nanoinformatics: Information & Data Sciences Applied to Nanomaterials Synthesis, Properties & Biological Effects**

### **Nanoinformatics for Nanomedicines**

Sponsored by COLL, Cosponsored by CINF

## **SUNDAY AFTERNOON**

Section A

Omni San Diego Hotel  
Grand Ballroom A

## **Text-Mining & Natural Language Processing for Chemical Information: From Documents to Knowledge**

R. J. Bienstock, *Organizer*

J. L. Nauss, *Organizer, Presiding*

**1:30** Introductory Remarks.

**1:35 CINF 20.** Automatic identification of relevant chemical compounds from patents. S.A. Akhondi, H. Rey, M. Schwörer, **M. Maier**, J. Toomey, H. Nau, G. Ilchmann, M. Sheehan, M. Irmer, C. Bobach, M. Doornenbal, M. Gregory, J. Kors

**1:55 CINF 21.** Augmenting manual curation of chemical patent information in the Derwent World Patents Index. **A. Klein**, S. McGhee, J. Hookes

**2:15 CINF 22.** Journey continues: Addition of French, Russian, Chinese, Korean, and Japanese patents to PATENTSCOPE ChemSearch. **J. Eiblmaier**, **C. Mazenc**, **D. Geppert**, **L. Isenko**, **H. Saller**

**2:35** Intermission.

**2:50 CINF 23.** Automating chemical structure and inhibition data extraction from patents: Text-mining approach. F. Costa, I. Haldoupis, **J.L. Nauss**, A. Hinton

**3:10 CINF 24.** BioAssay express: Creating and exploiting assay metadata. **P. Cheung**, A. Clark, J. Darlington

**3:30 CINF 25.** Building fast, robust, and reliable prediction models using very large biological data sets. **L. Weber**, H. Boehm

**3:50** Concluding Remarks.

## **SUNDAY AFTERNOON**

Section B

Omni San Diego Hotel  
Grand Ballroom D

### **Chemical Nomenclature & Representation: Past, Present & Future**

Cosponsored by HIST and NTS<sup>‡</sup>

Financially supported by International Union of Pure and Applied Chemistry (IUPAC); Chemical Structure Association Trust (CSA Trust); InChI Trust

G. Grethe, H. A. Lawlor, L. R. McEwen, *Organizers*

M. M. Rogers, *Organizer, Presiding*

**1:15** Introductory Remarks.

**1:20 CINF 26.** IUPAC and its role in the development of chemical nomenclature and structure representation. **R. Hartshorn**

**1:45 CINF 27.** IUPAC brief guides on nomenclature: Summary of the key nomenclature principles addressed in IUPAC colored books. **M.M. Rogers**

**2:10 CINF 28.** Evolution of CAS nomenclature: Past, present, and future. **M.A. Strausbaugh**

**2:35 CINF 29.** Updating the Braille Code of Chemical Notation 1997. **P. Verhalen**

**3:00** Intermission.

**3:15 CINF 30.** Carbon nanotube nomenclature: Challenges of naming emerging materials. **E. Mansfield**

**3:40 CINF 31.** Chemical nomenclature from books to computers: ACD/Name and IUPAC Division VIII. **A. Yerin**

**4:05 CINF 32.** IUPAC, nomenclature, and chemical representation: From the perspective of a worldwide structural database. **M.P. Lightfoot**, I. Bruno, C. Tovee, S. Ward, S. Wiggin

**4:30 CINF 33.** Chemical representation: Toolbox for human and machine collaboration. **L.R. McEwen**, E. Hepler-Smith

## **SUNDAY AFTERNOON**

Section C

Omni San Diego Hotel  
Grand Ballroom E

### **Importance of Collaboration to Create Student Success in the Laboratory & Beyond**

Financially supported by CAS  
S. P. Kuhn, *Organizer*  
M. Pozenel, *Presiding*

**1:30** Introductory Remarks.

**1:35 CINF 34.** Helping students stand out in the academic job market. **R.J. Gilliard**

**2:00 CINF 35.** Connecting the dots across academia and industry to ensure skill alignment. **M. Grandbois**

**2:25 CINF 36.** Creating digital learning objects for chemistry. **Y. Sevryugina**

**2:50 CINF 37.** RA21: Secure, seamless access for research. **R. Youngen**

**3:15 CINF 38.** Experiences in scientific information literacy education. **J. Ji**

**3:40 CINF 39.** Identifying the different definitions of student success between young scientists, faculty and administration in academia and hiring managers in industry. **M. Pozenel**

**4:05** Concluding Remarks.

## 150 Years of the Periodic Table

Sponsored by HIST, Cosponsored by CINF, INOR<sup>‡</sup> and PRES

## Nanoinformatics: Information & Data Sciences Applied to Nanomaterials Synthesis, Properties & Biological Effects

### Nanoinformatics for Nanomaterials

Sponsored by COLL, Cosponsored by CINF

## SUNDAY EVENING

Section A

San Diego Convention Center  
TBD

### CINF Scholarships for Scientific Excellence: Student Poster Competition

Financially supported by ACS Publications  
E. Alvaro, M. Qiu, *Organizers*

**6:30 - 8:30**

**CINF 40.** Withdrawn.

**CINF 41.** Crystal-structure prediction via basin-hopping global optimisation employing tiny periodic simulation cells and multipole expansion. C. Burnham, P. Samanta, **M. Ghaani**, N. English

**CINF 42.** CDD vault: Complexity simplified. **J. Darlington**, W.W. Smith, B.A. Bunin

**CINF 43.** Withdrawn.

**CINF 44.** Medicinal chemistry based measure of R group similarity. **N. O'Boyle**, R.A. Sayle

**CINF 45.** Systematic pipeline for automated structure-based molecular design: Beyond the static picture of hepatic organic anion transporting polypeptides. **A. Tuerkova**, B. Zdrzil

**CINF 46.** Public database supporting evidence-based exposomics. **R.R. Sayre**, J. Wambaugh, K. Phillips, A.J. Williams, C. Grulke

## MONDAY MORNING

Section A

Omni San Diego Hotel

Grand Ballroom A

## Driving Drug Discovery via Innovative Data Visualization

D. F. Ortwine, *Organizer*  
P. Beroza, *Organizer, Presiding*

**8:30** Introductory Remarks.

**8:35 CINF 47.** Coupling the 1D, 2D, and 3D data worlds to facilitate drug discovery. **D.F. Ortwine**

**9:05 CINF 48.** Using knowledge graphs for prediction and visual hypothesis generation in drug discovery. **D.J. Wild**

**9:35 CINF 49.** Visualizing relationships between protein targets, GO annotations and diseases via dynamic network representations. **B. Zdrazil, L. Richter, N. Brown**

**10:05** Intermission.

**10:20 CINF 50.** How a visual vocabulary defines what you see in your data. **R. Guha**

**10:50 CINF 51.** Molecular viz: I feel the need...the need for speed...and usability. **J. Boström**

**11:20 CINF 52.** Is virtual reality useful for visualizing and analyzing molecular structures? **T.E. Ferrin**

## MONDAY MORNING

Section B

Omni San Diego Hotel  
Grand Ballroom D

## Chemical Nomenclature & Representation: Past, Present & Future

### Challenges & Opportunities in Chemical Representation

Cosponsored by HIST and NTS<sup>†</sup>  
Financially supported by International Union of Pure and Applied Chemistry (IUPAC); Chemical Structure Association Trust (CSA Trust); InChi Trust  
L. R. McEwen, M. M. Rogers, *Organizers*  
G. Grethe, H. A. Lawlor, *Organizers, Presiding*

**8:15** Introductory Remarks.

**8:20 CINF 53.** Chemical structure standardization and synonym filtering in PubChem. **S. Kim, P. Thiessen, Q. Li, B. Yu, E. Bolton**

**8:45 CINF 54.** Challenges in chemical registration system migrations, and how to deal with them. **G. Blanke**

**9:10 CINF 55.** Making a hash of it: Advantage of selectively leaving out structural information. **N. O'Boyle**, R.A. Sayle

**9:35 CINF 56.** Crafting persistent identifiers and structure-based representations in DSSTox as surrogates for chemical names to better support interoperability in computational environments. **C. Grulke**, A. Richard, A.J. Williams

**10:00** Intermission.

**10:15 CINF 57.** Classification of reactions by type or name. **G. Grethe**, J. Eiblmaier, H. Kraut, D. Kunzman, P. Loew

**10:40 CINF 58.** UDM: Enabling exchange of comprehensive reaction information. **F. van den Broek**, G. Blanke

**11:05 CINF 59.** Reimagining IUPAC recommendations as a chemical ontology for semantic chemistry. **S.J. Chalk**

**11:30 CINF 60.** Publishing FAIR spectral data and chemical structures: Report from the NSF workshop in Orlando. **L.R. McEwen**, **V.F. Scalfani**

## **MONDAY MORNING**

Section C

Omni San Diego Hotel  
Grand Ballroom E

### **Successful Projects Fueled by Open-Source Tools**

R. J. Bienstock, *Organizer, Presiding*

**8:30** Introductory Remarks.

**8:35 CINF 61.** SciWalker: Comprehensive ontology-based chemical search. **L. Weber**, C. Bobach, F. Berthelmann, T. Boehme, S. Boyer, M. Irmer, K. Kruse, U. Laube, J. Ludwig, A. Pueschel, C. Ruttkies, I. Wetherbee

**9:00 CINF 62.** RDKit: Open-source cheminformatics from machine learning to chemical registration. **G. Landrum**

**9:25 CINF 63.** How the RCDK enables open source cheminformatics in R: From fingerprints to mass spectra. **R. Guha**, E.L. Schymanski, T. Schulze, M.A. Stravs

**9:50** Intermission.

**10:00 CINF 64.** Applying commonly overlooked corrections to DFT frequency calculations with GoodVibes. **G. Luchini**, R.S. Paton

**10:25 CINF 65.** Analysis of the acid/base profile of natural products as starting points of epigenetic drug discovery. **M.G. Santibanez-Moran**, J. Naveja, B. Pilón-Jiménez, M. Rico-Hidalgo, D. Manallack, J.L. Medina-Franco

**10:50 CINF 66.** Pharos: Open-source target illumination platform. **T. Sheils**, D. Nguyen, V. Siramshetty, N. Southall, T.I. Oprea

**11:15** Intermission.

**11:25 CINF 67.** Using open data, services, and source software to deliver the EPA CompTox Chemicals Dashboard. **A.J. Williams**, C. Grulke, K. Mansouri, J. Dunne, J. Edwards

**11:50 CINF 68.** Facilitating community-based chemical curation by providing an open source version of the DSSTox chemical and list registration software that supports the EPA CompTox Chemicals Dashboard. **C. Grulke**, A.J. Williams, A. Singh, J. Dunne, J. Edwards, A. Richard

## **Connecting Professionalism, Safety & Ethics: Opportunities & Challenges**

Sponsored by CHAS, Cosponsored by CINF

## **150 Years of the Periodic Table**

Sponsored by HIST, Cosponsored by CINF, INOR<sup>‡</sup> and PRES

## **MONDAY AFTERNOON**

Section A

Omni San Diego Hotel  
Grand Ballroom A

## **Driving Drug Discovery via Innovative Data Visualization**

D. F. Ortwine, *Organizer*  
P. Beroza, *Organizer, Presiding*

**1:30** Introductory Remarks.

**1:35 CINF 69.** Matched molecular pair (MMP) and matched molecular series (MMS) visualizations for drug discovery. **C. Keefer**

**2:05 CINF 70.** Using DOCK and ZINC to visualize ultra-large chemical libraries. **J.J. Irwin**, B. Shoichet, L. Jiankun, T.E. Balius, R.A. Sayle, I. Singh, A. Levit, Y. Moroz, M. O'Meara, C. Dandarchuluun, B. Wong, J. Young, K. Tang

**2:35 CINF 71.** Emerging AI and machine learning approaches for designing novel chemicals and materials with the desired properties. M. Popova, O. Isayev, **A. Tropsha**



**3:05** Intermission.

**3:20 CINF 72.** Visualizing structure-based deep learning scoring functions for protein-ligand interactions. **D. Koes**

**3:50 CINF 73.** Cheminformatics-powered visualization methods of complex multidimensional SAR data. **D. Fourches**

**4:20 CINF 74.** Data visualization for compound library enhancement: Application of artificial intelligence algorithms from computer chess. **R.A. Sayle**, N. O'Boyle, N. Zorn, R. Affentranger

## **MONDAY AFTERNOON**

Section B

Omni San Diego Hotel  
Grand Ballroom D

### **Chemical Nomenclature & Representation: Past, Present & Future**

#### **InChI'ng Forward**

Cosponsored by HIST and NTS<sup>‡</sup>

Financially supported by International Union of Pure and Applied Chemistry (IUPAC); Chemical Structure Association Trust (CSA Trust); InChi Trust

G. Grethe, H. A. Lawlor, M. M. Rogers, *Organizers*

L. R. McEwen, *Organizer, Presiding*

M. G. Hicks, *Presiding*

**1:15** Introductory Remarks.

**1:20 CINF 75.** Reaction InChI (RInChI): Present and future. **G. Blanke**, J.M. Goodman, G. Grethe, H. Kraut

**1:45 CINF 76.** Chemical mixtures: File format, open source tools, example data, and mixtures InChI derivative. **A. Clark**, P. Cheung, J. Darlington, L.R. McEwen

**2:10 CINF 77.** Organometallics: InChI'ng forwards to better representations and happier chemists. **I. Bruno**, C. Batchelor, J.M. Goodman, G. Blanke

**2:35 CINF 78.** Names for structural variability: Alkanes from maximum efficiency to the limits of existence. **J.M. Goodman**

**3:00** Intermission.

**3:15 CINF 79.** IUPAC SMILES+ specification: Proposed community effort to advance interoperability of the SMILES chemical structure representation. **V.F. Scalfani**, L.R. McEwen, C. Grulke, E. Bolton, G. Landrum, H. Cooke, I. Yamada, J.J. Irwin, J.L. Medina-Franco, M.Q. Olozabal, O. Koepler, S. Richardson

**3:40 CINF 80.** InChI open education resource (OER). **R.E. Belford**, E.C. Bucholtz, S.P. Wathen, M.A. Walker, J. Cuadros, T. Gupta, N. Brown, V.F. Scalfani

**4:05 CINF 81.** Keeping up the momentum: Brief report from the InChI San Diego workshop. **R.J. Boucher**, **R. Kidd**, I. Bruno, S.R. Heller, L.R. McEwen

**4:20** Discussion.

## **MONDAY AFTERNOON**

Section C

Omni San Diego Hotel  
Grand Ballroom E

### **Materials Informatics**

H. Senderowitz, A. Tropsha, *Organizers, Presiding*

**1:00** Introductory Remarks.

**1:05 CINF 82.** Fast and accurate interatomic potential models by genetic programming. A. Hernandez, A. Balasubramanian, F. Yuan, S. Mason, **T. Mueller**

**1:30 CINF 83.** Accelerating design of inorganic materials with machine learning and AI. **O. Isayev**

**1:55 CINF 84.** Deep learning from crystallographic representations of periodic systems. **P.M. Maffettone**, A.I. Cooper

**2:20 CINF 85.** Application of machine learning tools for the analysis of combinatorial libraries of all metal-oxides photovoltaic cells. **H. Senderowitz**, A. Yosipof, O. Kaspi

**2:45** Intermission.

**3:00 CINF 86.** Database of low-energy cluster structures for atomically precise nanoclusters across the periodic table calculated using density functional theory. **P. Lile**, T. Mueller

**3:25 CINF 87.** Self-assembly of metal-organic frameworks. **Y.J. Colon**, A. Guo, L.W. Antony, K. Hoffmann, J.J. De Pablo

**3:50 CINF 88.** Accelerated discovery of high-refractive-index polyimides via *First-Principles* materials modeling and informatics. **J. Hachmann**

**4:15 CINF 89.** Experiment specification, capture and laboratory automation technology (ESCALATE): Software pipeline for automated chemical experimentation and data management, with application to metal halide perovskite discovery. **J. Schrier**

**4:40 CINF 90.** Standardization of structural representation of polymers used in medicinal products. **Y. Borodina**, I. Filippov, T. Peryea, Y. Pevzner

## Connecting Professionalism, Safety & Ethics: Opportunities & Challenges

Sponsored by CHAS, Cosponsored by CINF

## 150 Years of the Periodic Table

Sponsored by HIST, Cosponsored by CINF, INOR<sup>‡</sup> and PRES

### MONDAY EVENING

Section A

San Diego Convention Center  
TBD

#### Sci-Mix

S. K. Cardinal, *Organizer*

**8:00 - 10:00**

**18, 40-42, 44-46, 80, 84, 86.** See Previous Listings.

**100, 114, 133, 135, 155, 158, 172.** See Subsequent Listings.

### TUESDAY MORNING

Section A

Omni San Diego Hotel  
Grand Ballroom A

#### Herman Skolnik Award Symposium Honoring Dr. Kimito Funatsu

Cosponsored by PROF  
Financially supported by Schrödinger  
S. K. Cardinal, K. Funatsu, *Organizers*  
M. Sugimoto, *Presiding*

**8:30 CINF 91.** Monitoring progress in lead optimization. **J. Bajorath**

**8:55 CINF 92.** Electronic-structure informatics using 3D descriptors of molecules. **M. Sugimoto**

**9:20 CINF 93.** Fast evaluation of potential synthesis routes using DFT calculations on the basis of

Transition State Data base (TSDB). **K. Hori**

**9:45 CINF 94.** Development using materials informatics in Japanese companies. **Y. Uchi**

**10:10 CINF 95.** Prediction and control of vacuum deposition process by data-driven method. **Y. Takeda, Y. Zushi, T. Ogushi, E. Kuribe**

**10:35** Intermission.

**10:40 CINF 96.** Designing synthesizable bioactive compounds with chemistry-savvy machine intelligence. **G. Schneider, D. Merk, F. Grisoni, A. Button, L. Friedrich, J.A. Hiss, P. Schneider**

**11:05 CINF 97.** Activity landscape and its application to molecular design. **K. Hasegawa**

**11:30 CINF 98.** Data-driven drug discovery and medical treatment by machine learning. **Y. Yamanishi**

**11:55 CINF 99.** Development of data driven chemistry in chemistry and chemical engineering. **K. Funatsu**

**12:20** Award Presentation.

## **TUESDAY MORNING**

Section B

Omni San Diego Hotel  
Grand Ballroom D

### **Extended Reality (XR) in Libraries & Beyond**

S. K. Cardinal, *Organizer*  
M. Qiu, N. Ruhs, *Organizers, Presiding*

**8:15** Introductory Remarks.

**8:25 CINF 100.** Application of extended reality (XR) technologies in the academic library to support innovative research and instruction in the physical sciences and engineering disciplines. **E. Cabada, M.C. Schlembach**

**8:55 CINF 101.** Deploying a VR workstation and molecular visualization at Caltech library. **T.E. Morrell, D. Wrublewski**

**9:25 CINF 102.** Librarians and extended reality: Enhancing teaching and learning in the chemical sciences. **S. Putnam, M.M. Nolan, E. Williams**

**9:55** Intermission.

**10:10 CINF 103.** Using XR to teach about chemical lab safety. **S. Ramachandran, R. Broyer, S.**

**Cutchin, S. Fu**

**10:40 CINF 104.** Digital collections at Cal Poly Pomona and the California State University campuses. **J. Selco**

**11:10** Panel Discussion.

**11:40** Concluding Remarks.

## **TUESDAY MORNING**

Section C

Omni San Diego Hotel  
Grand Ballroom E

### **Drug Discovery: Informatics Approaches**

E. Davis, *Organizer, Presiding*

**8:15 CINF 105.** Discovery of novel inhibitors of human galactokinase by virtual screening. **M. Shen**, X. Hu

**8:40 CINF 106.** Measuring R group similarity using medicinal chemistry data. **N. O'Boyle**, R.A. Sayle

**9:05 CINF 107.** Mechanism and prediction of UGT metabolism. **M. Öeren**, P. Hunt, D.J. Ponting, M. Segall

**9:30** Intermission.

**9:45 CINF 108.** Signals lead discovery as the Corteva Cheminformatics Workbench. **D. Tomandl**, S. Smith, J. Wilmot

**10:10 CINF 109.** Probing allosteric modulators of AMP-activated protein kinase. **X. Hu**, J.J. Marugan, W. Zheng

**10:35 CINF 110.** Underlying scientific evidence discovery for FDA orphan drug designations from the GARD integrative knowledge graph: Towards drug discovery for rare diseases. **Q. Zhu**, D. Nguyen, N. Southall

### **Connecting Safety, Education, Training & Productivity in Analytical Laboratories**

Sponsored by ANYL, Cosponsored by CCS, CHAS<sup>‡</sup>, CINF and PRES

## **TUESDAY AFTERNOON**

## Section A

Omni San Diego Hotel  
Grand Ballroom A

### Machine Learning & Artificial Intelligence in Computational Chemistry

#### Drug Discovery

T. Robertson, *Organizer, Presiding*

**1:30** Introductory Remarks.

**1:35 CINF 111.** Explore, exploit, and extrapolate: How AI-driven SAR navigation facilitates lead optimisation in drug discovery. **D. Marcus**, C. Luscombe, S. Pickett, S. Senger, D. Green

**2:05 CINF 112.** AI-driven drug design across the discovery spectrum: Case studies. **J.H. Griffin**

**2:35 CINF 113.** What compound to synthesize next? How machine learning and artificial intelligence impact compound optimization. **D. Kuhn**, K. Preuer, M. Krug, G. Klambauer, S. Hochreiter, F. Rippmann

**3:05** Intermission.

**3:25 CINF 114.** Pretraining deep learning molecular representations for property prediction. **B. Liu**, W. Hu, J. Leskovec, P. Liang, V.S. Pande

**3:55 CINF 115.** Modeling protein flexibility with conformational sampling improves ligand pose and bioactivity prediction. **K.A. Stafford**, J. Sorenson, I. Wallach

**4:25 CINF 116.** Machine learning for the discovery of  $\alpha_v\beta_6$  integrin antagonists. **J.D. Hirst**, S. Oatley, E. Guest, T. Gaertner, S.J. MacDonald

**4:55** Concluding Remarks.

## TUESDAY AFTERNOON

### Section B

Omni San Diego Hotel  
Grand Ballroom D

### One Million Crystal Structures: A Wealth of Structural Chemistry Knowledge

M. Stahl, *Organizer*  
H. Abourahma, I. Bruno, *Organizers, Presiding*

**1:30** Introductory Remarks.

**1:40 CINF 117.** One million crystal structures in the CSD: Cause for celebration, cause for consideration. **R. Taylor**, I. Bruno

**2:10 CINF 118.** Leveraging the CSD's one million structures in course-based undergraduate research experience. **H. Abourahma**, A. Sarjeant

**2:35 CINF 119.** Use of the Cambridge Structural Database in the undergraduate chemistry curriculum. **A.T. Royappa**

**3:00 CINF 120.** Examining research data through a crystal lens: Teaching students about primary data, data representation, and data management using crystal structure databases. **J.N. Currano**

**3:25** Intermission.

**3:40 CINF 121.** Materials genome approach to functional materials discover using the CSD. **K.R. Cousins**, **S.B. Rodriguez**

**4:05 CINF 122.** Building a collection of metal–organic frameworks in the Cambridge Structural Database for materials discovery. **D. Fairen-Jimenez**

**4:30 CINF 123.** Improved crystal structure determination from powder diffraction data using the Cambridge Structural Database system. **K. Shankland**, E. Kabova, J. Cole

**4:55 CINF 124.** Million opportunities: Using the CSD to design color changing molecular switches. **P.R. Raithby**

## **TUESDAY AFTERNOON**

Section C

Omni San Diego Hotel  
Grand Ballroom E

### **Biologic Informatics**

R. J. Bienstock, *Organizer, Presiding*

**1:30** Introductory Remarks.

**1:35 CINF 125.** Biologics information in PubChem. **J. Zhang**, P. Thiessen, T. Cheng, B. Shoemaker, E. Bolton, N. O'Boyle, R.A. Sayle

**2:00 CINF 126.** Current progress in HELM representation, integration, and data migration. **D. Deng**, T. Yuan, J. Lee, R. Hotchandani

**2:25 CINF 127.** Representational and algorithmic challenges in biologic informatics 2019. **R.A. Sayle**, N. O'Boyle

**2:50** Intermission.

**3:00 CINF 128.** Notation for identification of glycans contained in glycoproteins, glycolipids, and other biomolecular structure data. **I. Yamada**, N. Miura, S. Tsuchiya, K.F. Aoki-Kinoshita

**3:25 CINF 129.** Trends in biologics research and development: Analytic studies based on CAS-curated data. **C.Y. Liu**, Y. Li, Y. Deng

**3:50 CINF 130.** Biosequence searching: How CAS is expanding workflow solutions for IP searchers and beyond. **R.J. Walczak**

## **WEDNESDAY MORNING**

Section A

Omni San Diego Hotel  
Grand Ballroom A

### **Machine Learning & Artificial Intelligence in Computational Chemistry**

#### **Materials Science**

T. Robertson, *Organizer*  
Y. An, *Presiding*

**8:30** Introductory Remarks.

**8:35 CINF 131.** Pesticide quantitative structure-greenhouse-activity relationship models. **D. Tomandl**, C. Klittich, J. Herbert, N.M. Satchivi, D. Demeter

**9:05 CINF 132.** Pore volumes and surface areas of metal-organic frameworks as descriptors for materials discovery. **A. Mroz**, C.H. Hendon

**9:35 CINF 133.** Artificial neural network-based approach to thermodynamic property estimation. **R. Van de Vijver**, P. Plehiers, P. Verberckmoes, G.B. Marin, C.V. Stevens, K. Van Geem

**10:05** Intermission.

**10:25 CINF 134.** Persistent homology for chemical applications: Story of birth and death. **K.D. Vogiatzis**, A. Cherne, J.K. Kirkland, V. Maroulas, C. Putman Micucci, J. Townsend

**10:55 CINF 135.** ML models that both learn and teach chemistry via partitioning reactive trajectories by reaction product in phase space using support vector machines. **G. Grazioli**, S. Roy, C.T. Butts

**11:25 CINF 136.** Deep neural network model for MD-level packing density predictions and its application in the study of 1.5 million organic molecules. **J. Hachmann**

**11:55** Concluding Remarks.



## WEDNESDAY MORNING

Section B

Omni San Diego Hotel  
Grand Ballroom D

### One Million Crystal Structures: A Wealth of Structural Chemistry Knowledge

I. Bruno, M. Stahl, *Organizers*  
H. Abourahma, *Organizer, Presiding*

**9:00** Introductory Remarks.

**9:05 CINF 137.** Pervasive approximate symmetry in *P1* and high-*Z'* organic crystals: Implications for crystal nucleation. **C.P. Brock**

**9:30 CINF 138.** One million crystal structures: One million disappearing polymorphs waiting to happen? **J. Helfferich**, J. van de Streek, M. Neumann

**9:55 CINF 139.** What the Cambridge Structural Database tells us about hydrates. **J. Werner**, J.A. Swift

**10:20 CINF 140.** Energetics of co-crystal formation: Informing prediction through combining the database with large scale simulations and machine learning. **G.M. Day**, C.R. Taylor, D. McDonagh, W. Fyffe, C. Skylaris

**10:45** Intermission.

**11:00 CINF 141.** Using knowledge-based tools to evaluate solid-form design and risk assessment. **B. Sandhu**, C.B. Aakeroy, S.M. Reutzel Edens, A. Sarjeant, S. Vyas

**11:25 CINF 142.** What did the CSD ever do for drug discovery? **J. Liebeschuetz**

**11:50 CINF 143.** Improved structure-based drug design with one million small molecule crystal structures. **B. Kuhn**

## WEDNESDAY MORNING

Section C

Omni San Diego Hotel  
Grand Ballroom E

### Web-Based Chemistry Databases

A. J. Williams, *Organizer*  
C. Grulke, A. Williams, *Presiding*

**8:00** Introductory Remarks.

**8:05 CINF 144.** Computational database for first-row transition metals. **K. Basemann**, A. Leffel, A.D. Sadow, T.L. Windus

**8:25 CINF 145.** Open chemistry: Democratizing web-based chemistry databases. **M.D. Hanwell**, C. Harris, A. Genova, M. El Khatib, M. Haghighatlari, J. Hachmann, W. Dejong

**8:45 CINF 146.** KinaMetrix: Web resource to investigate kinase conformations and inhibitor space. **R. Rahman**, P.M. Ung, A. Schlessinger

**9:05 CINF 147.** Structure-based search of chemical libraries with Pharmit. **D. Koes**

**9:25 CINF 148.** Molecular malthusianism? Next three logs of the growth of purchasable chemical space. **J.J. Irwin**

**9:45** Intermission.

**10:00 CINF 149.** Searching for similar reactions and molecules using the power of graph databases and the graph edit distance metric. **V. Delannee**, M.C. Nicklaus

**10:20 CINF 150.** Helping chemists identify new opportunities during chemistry research: Building and turning high-quality data into actionable insights. **J. Swienty Busch**

**10:40 CINF 151.** Challenges and opportunities of delivering structural data on the web. **M.P. Lightfoot**, I. Bruno, S. Ward

**11:00 CINF 152.** Withdrawn.

**11:20 CINF 153.** PubChem: Improving access to chemical information on the web. **A. Gindulyte**

## **WEDNESDAY AFTERNOON**

Section A

Omni San Diego Hotel  
Grand Ballroom A

### **Machine Learning & Artificial Intelligence in Computational Chemistry**

#### **New Methods**

T. Robertson, *Organizer*  
K. Marshall, *Presiding*

**1:30** Introductory Remarks.

**1:35 CINF 154.** Neural network potential for modeling radical reactions. **R. Messerly**, P. St. John, A.E. Roitberg, S. Kim

**2:05 CINF 155.** Predicting NMR in real-time through message-passing neural network. **Y. Guan**, R.S. Paton

**2:35 CINF 156.** Practical applications of deep learning to imputation of drug discovery data. T. Whitehead, B. Irwin, P. Hunt, **M.D. Segall**, G. Conduit

**3:05** Intermission.

**3:25 CINF 157.** Protein binding site fingerprinting for activity screening in machine learning. **B. Bergman**, K.A. Stafford, D. Bernard, S. Schroedl

**3:55 CINF 158.** Multiagent consensus equilibrium (MACE) for addressing the scaling challenges of computational chemistry. **J.R. Ulcickas**, G.J. Simpson

**4:25 CINF 159.** New approach to regression uncertainty analysis and applications to drug design. **M. Waldman**, R. Clark

**4:55** Concluding Remarks.

## WEDNESDAY AFTERNOON

Section B

Omni San Diego Hotel  
Grand Ballroom D

### One Million Crystal Structures: A Wealth of Structural Chemistry Knowledge

H. Abourahma, M. Stahl, *Organizers*

I. Bruno, *Organizer, Presiding*

**1:30** Introductory Remarks.

**1:35 CINF 160.** Insights from CSD crystallographic data applied to drug discovery. **N. Nevins**

**2:00 CINF 161.** What fragment hit to follow and how? Using hotspots to prioritise chemistry resources. M. Smilova, P. Curran, C. Radoux, W. Pitt, J. Cole, A. Bradley, **B. Marsden**

**2:25 CINF 162.** Traversing interoperability: Drug development harnessing the CSD and PDB. **A. Brink**

**2:50** Intermission.

**3:05 CINF 163.** Semantic representation of CIF files: Mining crystal structures in the CSD. **S.J. Chalk**

**3:30 CINF 164.** Learning from a database of a million crystalline materials. **R.I. Cooper**

**3:55 CINF 165.** From structure to crystallisation and manufacturing: Journey in applications of the CSD. **C.C. Wilson**

**4:20 CINF 166.** New frontiers beyond one million: New horizons for structural chemistry. **J. Harter**, I. Bruno

**WEDNESDAY AFTERNOON**  
Section C

Omni San Diego Hotel  
Grand Ballroom E

**Web-Based Chemistry Databases**

A. J. Williams, *Organizer*  
C. Grulke, A. Williams, *Presiding*

**1:30 CINF 167.** US EPA CompTox Chemicals Dashboard: Integrating chemistry and biology data to serve computational toxicology and environmental science. **A.J. Williams**, C. Grulke, A. Richard, R. Judson, G. Patlewicz, I. Shah, J. Wambaugh, K. Paul-Friedman, J. Dunne, J. Edwards

**1:50 CINF 168.** Lessons learned in building the CompTox Chemicals Dashboard: Engineering a more sustainable web-based chemical database. **C. Grulke**, A.J. Williams, A. Singh, J. Dunne, A. Richard, J. Edwards

**2:10 CINF 169.** In the world of free, is there room for subscription solutions? **J.W. Taylor**, M.A. Pozenel

**2:30 CINF 170.** Google BigQuery for analysis of scientific datasets. **S. Boyer**, **I. Wetherbee**, L. Weber, J. Frommer

**2:50 CINF 171.** Google BigQuery for analysis of scientific datasets: Interactive exploration and analysis of the data using KNIME analytics platform. **G. Landrum**, M. Pawletta, J. Prinz

**3:10** Intermission.

**3:25 CINF 172.** ChEMBL, SureChEMBL and UniChem: Web-based chemistry databases for drug discovery and chemical research. **R. Arcila**

**3:45 CINF 173.** CIRCA: Your cheminformatics assistant. **T.D. Griffin**, E.W. Louie, S. Boyer, L. Anderson, K. Schmidt, D.P. Sanders

**4:05 CINF 174.** Transforming quality chemical data handbooks into web-based chemistry databases. **J. Rumble**, F. Macdonald

**4:25 CINF 175.** Does bigger mean better in the world of chemistry databases? **A.J. Williams**, C. Southan

**4:45 CINF 176.** Evolution of the public chemistry databases: Past and the future. **V. Tkachenko**, R. Zakharov



# CINF Technical Program with Abstracts - ACS Fall 2019 Meeting

## CINF 1

### Enhancing data-driven summarization of relations between chemicals, genes, proteins, and diseases based on text mining of biomedical literature

**Leonid Zaslavsky**, *zaslavsk@ncbi.nlm.nih.gov*, Asta Gindulyte, Paul Thiessen, Evan Bolton. National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, United States

PubChem knowledge panels summarize important relationships between chemicals, genes, proteins and diseases uncovered by analysis of term co-occurrences in biomedical literature [1]. A part of PubChem compound, target and disease pages [2], they are built using automatic annotations of PubMed records by natural language processing (NLP) software LeadMine [3], and term matching to PubChem and other public databases. The most relevant information is identified using statistical analysis, summarized, and visualized using relevance-based sampling, allowing a compact, highly informative and easy-to-comprehend knowledge representation.

Our ongoing project includes quality enhancements in both intelligent data analysis and interpretation of annotations provided by NLP software. In statistical analysis, we introduced an adjusted entity co-occurrence score, with a correction for random co-occurrence of frequently-annotated entities applied to the record count. To improve the annotation quality, we are working on more accurate type identification and matching to database records through more sophisticated postprocessing that includes processing of annotations of different types together, taking context into account, using curated data, and iteratively improving annotation of the whole database by discovering term-use patterns.

Besides improving quality of knowledge summarization in PubChem papers, our efforts allow to better understand the typical use, rank synonyms, correct submitter-provided names, and verify submitter-provided links in our databases, helping to automate biocuration efforts.

## CINF 2

### Introducing automated polymer data extractor tool

**Chunlei Dai**, *chunlei.dai@ibm.com*, Kristin Schmidt, Dmitry Y. Zubarev, Victoria A. Piunova, Krishnakumar Surugucchi, Daniel P. Sanders. IBM Research - Almaden, San Jose, California, United States

Many break-through technologies require new polymers, however the time to develop, optimize and certify the new materials usually takes much longer than to develop the new technology. At the same time, an exorbitant amount of scientific studies get published each year, making it impossible for any given researcher to stay on top of all publications in their field. Knowledge of structure-property relationships of polymers can accelerate the discovery of new materials. However, it is difficult to extract chemistry-knowledge based information, because of the expressions are different for each scientist. Therefore, it is beneficial to use NLP process pipeline to understand them and extract the data.

A first step in the process is to recognize chemical entities and their abbreviations. We then use table extractor to get all the information from tables (structured data) and rule based syntax to get the information from the text (non-structured data). We use our system then to normalize the data then associate them to build a structured searchable dataset. Based on chemical dictionary, table extractor and rule based NLP extract tool (SystemT), we can extract chemical entities and their associated properties, techniques, reactions, roles and relationships from scientific documents, reports, and patents. Our system is an automated, chemistry knowledge based cognitive data extraction and relation build tool.

In this presentation, we will introduce a polymer data extractor based on IBM's SystemT. We will show the process how the data is been extracted and associated together.

### CINF 3

#### Extraction of polymer-related information in the cheminformatics tool CIRCA

**Kristin Schmidt**<sup>1</sup>, *schmidkr@us.ibm.com*, **Chunlei Dai**<sup>1</sup>, **Thomas D. Griffin**<sup>1</sup>, **Krishnakumar Surugucchi**<sup>1</sup>, **Dmitry Y. Zubarev**<sup>2</sup>, **Victoria A. Piunova**<sup>3</sup>, **Nathan Park**<sup>1</sup>, **James Hedrick**<sup>1</sup>, **Laura C. Anderson**<sup>1</sup>, **Daniel P. Sanders**<sup>4</sup>. (1) IBM Research, San Jose, California, United States (2) Dept. of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts, United States (3) IBM Almaden Research Center, Los Gatos, California, United States (4) IBM Research - Almaden, San Jose, California, United States

Molecules are a fundamental building block of life, critical to many aspects of scientific and technical work. Informatics related to molecules, compounds, and materials, their properties, structure, and behavior, can provide a significant acceleration to scientific and technical innovation. A first step in this process is to extract relevant information from unstructured data such as patents and journal articles. Current tools focus on the text mining of small molecule names using either dictionary and rule-based models or statistical machine learning methods. Extracting information about polymers and their synthetic parameters is rare, and so far either inaccurate or incomplete. In this talk, we will introduce a polymer annotator based on IBM's SystemT, which adopts an algebraic approach. We will show examples of extracted data, compare performance to manual data extraction and present how the annotator is incorporated into the cheminformatics tool CIRCA.

### CINF 4

#### Abstract recommendation system: beyond word-level representations

**Vadim Korolev**, **Artem Mitrofanov**, *mitrofjr@gmail.com*, **Boris Sattarov**, **Valery Tkachenko**. *Science Data Software, Rockville, Maryland, United States*

Public repositories containing a diverse chemical and biological data are one of the main sources of knowledge for further biomedical research. Unfortunately, extraction and transforming these data into a well-interpretable form is a complex task. Ongoing community efforts are mainly focused on the analysis of co-occurrences of terms, text annotation based on terms similarity and related tasks.

Here we present an approach based on natural-language processing techniques, which is intended to shift the focus of a search for similar texts on chemical topics from word- to document-level. PubMed records were used to implement word2vec and doc2vec models. Generated text representations can be used to search for similar abstracts; similarity is more dependent on this representation than co-presence of the certain terms (neighbor compounds, similar publication date, etc.).

Document-level clustering was also implemented to provide insight into PubMed text corpus structure. This approach can serve as an alternative to standard topic modeling techniques for discovery of the hidden semantic features in unsupervised manner.

### CINF 5

#### Current challenges in text-mining for chemical information

**Roger A. Sayle**, *roger@nextmovesoftware.com*, **John W. Mayfield**, **Noel O'Boyle**. *NextMove Software, Cambridge, United Kingdom*

Named Entity Normalization (NEN) is the task of determining entities mentioned in text. NEN is different from Named Entity Recognition (NER) in that NER identifies the occurrence or mention of a named entity in text by does not identify which specific entity it is. The distinction is subtle but critical for extracting chemical information from scientific literature, where understanding which chemical an author is referring to is vital. Perhaps disappointingly, this distinction also means that many recent advances in NER, such as word2vec, LSTM neural networks and conditional random fields, have little relevance in practice.

A significant technical challenge with NEN is the need to represent the entity/referent computationally. For many terms there exist suitable ontologies and identifiers, and for chemicals there connection tables and canonical

notations such as InChI and SMILES. Hence, the cutting edge of chemical information text mining is not just about the syntax of entities, but also semantics.

In this presentation, we describe several entity types at the limits of the current state-of-the-art. These range from handling chemical notations, such as HELM, WLN and molecular line formulae, through more complex entities such as scientific journal names and clinical trials (NCT numbers) to more abstract concepts such as the generic terms alkanes, heterocycles, alloys, solvents or zintl phases. The presentation will also cover chemical nomenclature beyond the traditional IUPAC-like nomenclatures handled by existing name-to-structure software, including inorganic nomenclature, peptides, protein variants and mutants, and even mixtures.

In the words of Ludwig Wittgenstein, "The limits of my language means the limits of my world". Increasing what can be handled by chemical information text mining requires not only extending the language it recognizes, but also the world that it can represent.

## CINF 6

### MOLVEC: Open source library for chemical structure recognition

*Tyler Peryea, Daniel Katzel, Tongan Zhao, Noel Southall, **Dac-Trung Nguyen**,  
nguyenda@mail.nih.gov. NCATS, NIH, Rockville, Maryland, United States*

As the volume of chemical structures published in the literature continues to grow at an ever-increasing pace, the need to systematically capture this data in a machine-readable manner (e.g., for the purposes of information retrieval and knowledge extraction) has never been more urgent. While there have been ongoing efforts to encourage publishers to require that chemical structures be made available in a machine-readable format (e.g., InChI, SMILES), much of the relevant chemical structures remain embedded within journal articles as digital images. There are a number of tools that can convert chemical structure images into machine-readable formats with high accuracy. However, the availability of these tools is rather limited in that they are either commercial, lack source code, or require complex package dependencies to build. Herein we report on our recent effort in developing a robust and fast Java library for converting any chemical structure image into a variety of 2D molecular vector (MOLVEC) formats (e.g., MDL, InChI with coordinates, etc.). The library is (i) compact with minimal dependencies (a self-contained executable JAR file can be as small as 10MB), (ii) relatively fast (3-4 images per second), and (iii) accurate over a wide range of image quality and resolution (achieved over 90% accuracy on the USPTO dataset). MOLVEC can be used standalone or embedded within a web service. It is also currently being distributed as part of the G-SRS substance registration software and is heavily used by FDA staff to register APIs, impurities and other chemical entities found in IND applications and other regulatory submissions. The complete source code and test benchmarks for MOLVEC are available at <https://spotlite.nih.gov/ncats/molvec>.

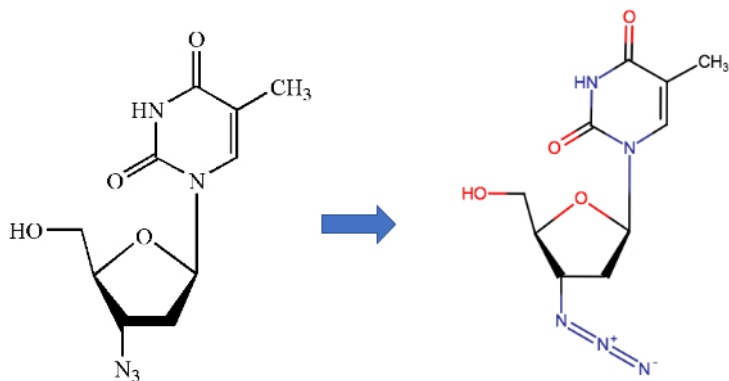


Image to 2D molecular vector format conversion with MOLVEC.



## CINF 7

### Patent novelty searching: Scope of prior art

**Edlyn S. Simmons**, *edlyns@earthlink.net*. Simmons Patent Information Service, LLC, Fort Mill, South Carolina, United States

Researchers are barred from receiving patents if their inventions and discoveries are disclosed in the prior art. Although it is common knowledge that patents and journal articles must be searched to avoid claiming subject matter within the prior art, other categories of disclosure are also bars to patentability. This presentation will discuss the scope of prior art defined by patent laws and the expansion of the scope of prior art under United States patent law that resulted from the America Invents Act of 2011 .

## CINF 8

### Quality of indexing for non-patent literature and its implications for retrieval of relevant prior art

**Stephen R. Adams**, *stephen.adams@magister.co.uk*. Magister Ltd, Victoria, Roche, Cornwall, United Kingdom

Most citations in official patent office search reports consist of prior publications from the patent literature. One reason for this is the convenience of searching in patent collections, across different technical fields in multiple languages, by using either customised classifications which have been developed over many years to optimise retrieval, or electronic full text. It is comparatively rare to find collections from the non-patent literature corpus (journals, conference proceedings, reports, preprints) which have been indexed to the same level of precision. This leads to a relatively poor retrievability of non-patent literature, compared to patents. However, the modern patent information specialist is faced with a legal mandate to consider the universal state of the art, which gives equal weight to all prior publications, irrespective of their source. Some of the implications of the quality of indexing will be considered, both for the patent applicant and the granting offices.

## CINF 9

### Finding what others miss: Comprehensive prior-art resources for chemical substance searching

**John Zabilski**, *jzabilski@cas.org*. CAS, Groveport, Ohio, United States

The challenge of prior art searching for patent novelty is compounded when looking for chemical information. Comprehensive chemical substance searches are particularly challenging as substances have complex names, are frequently covered as structural images, can be defined as derivatives or intermediates without a complete name, or are represented by a generic structure. This talk will compare sources for chemical non-patent literature and discuss resources including indexed content and structure-based approaches that help identify hidden substances unnamed in the original document.

## CINF 10

### Practical pointers for conducting prior art searches for small molecules, a patent practitioner's point of view

**Francis J. Koszyk<sup>1</sup>**, **Xavier Pilla<sup>2</sup>**, *xpillai@leydig.com*. (1) Leydig, Voit & Mayer, Ltd, Chicago, Illinois, United States (2) Leydig Voit Mayer Ltd, Chicago, Illinois, United States

This presentation will provide a brief overview of best practices for conducting prior art searches in the area of small molecules from the perspective of the patent practitioner, with a view towards patentability and patent validity investigations. Topics to be discussed include selection of appropriate databases, structuring of search queries, and interpretation of search results.

## CINF 11

### What's new with scientific content in Google patents

**Ian Wetherbee**<sup>1</sup>, *wetherbee@google.com*, **Stephen Boyer**<sup>2</sup>, *skboyer@gmail.com*, **Jane Frommer**<sup>2</sup>, **Vihang Mehta**<sup>1</sup>, **Broderick Arneson**<sup>1</sup>. (1) Google, Mountain View, California, United States (2) Collabra Inc, San Jose, California, United States

Since its inception in 2006, Google Patents has been a resource for providing public access to worldwide intellectual property. Google Patents continues to expand its reach, now offering content from over 100 countries, translations of most patent content into English, and patent search integrated with Google Scholar. Advanced search options include similarity searching and machine classification of non-patent literature documents. This talk will focus on recent developments to make Google Patents chemistry and life-sciences aware, which will be of particular interest to the chemical, pharmaceutical and legal communities.

## CINF 12

### Non-patent literature citations in EPO opposition and limitation proceedings

**Stephen R. Adams**, *stephen.adams@magister.co.uk*. *Magister Ltd, Victoria, Roche, Cornwall, United Kingdom*

The European Patent Office allows for two administrative processes by which granted claims may be modified in scope. In both cases, prior art found in the non-patent literature can be used to persuade the Office that original claims are unjustified, and need to be amended. A study of the relative proportion of patent to non-patent literature cited during these proceedings illustrates the importance of non-patent literature as part of the state of the art. If adequate searches are done in the non-patent literature, either by the applicant prior to filing or the examining office during prosecution, it can lead to better quality granted patents which are able to stand up to subsequent challenges.

## CINF 13

### Inventions for sale? Navigating the on-sale bar under the America Invents Act

**Justin L. Krieger**, *jkrieger@kilpatricktownsend.com*. *Roberts Mlotkowski Safran Cole, P.C., McLean, Virginia, United States*

Many commentators and the US Patent & Trademark Office initially interpreted America Invents Act (AIA) section 102(a) as fundamentally changing the "on-sale" bar by requiring an offer for sale to make the "claimed invention . . . available to the public" in order to constitute a prior art event. The US Supreme Court, however, recently confirmed in *Helsinn v. Teva* that the law may remain substantially unchanged. Join registered patent attorney Justin Krieger in discussing the nuances of the on-sale bar under the AIA as interpreted by this important Supreme Court decision. This presentation will provide an overview of the on-sale bar before and after enactment of the AIA, secret offers for sale, geographical considerations and the AIA grace period.

## CINF 14

### Science librarian, a chemical educator, and an EHS professional walk into a lab: Laboratory safety as a collaborative teaching tool

**Leah R. McEwen**<sup>1</sup>, *lrm1@cornell.edu*, **Samuella B. Sigmann**<sup>2</sup>, *sigmannsb@appstate.edu*, **Ralph Stuart**<sup>3</sup>, *ralph@rstuartcih.org*. (1) Clark Library, Cornell University, Ithaca, New York, United States (2) Chemistry, Appalachian State University, Boone, North Carolina, United States (3) Dept of Env Hlth Safety, Keene State College, Keene, New Hampshire, United States

As emphasized in the ACS Guidelines for Bachelor's Degree Programs, chemistry students are tasked with learning a number of laboratory skills relevant to the practice of science in addition to chemistry concepts [1]. These skills reflect the ACS Professional Core Values and ultimately converge on developing competency in safe and responsible conduct of research with a broad awareness of social context. Helping students to master and integrate these cognate skills with their evolving chemistry knowledge involves significant interdisciplinary support. The authors - a science librarian, a chemical educator, and an EHS professional - have been exploring opportunities to blend their distinct but complementary expertise in support of student learning. This presentation will describe key concepts and materials arising from this collaboration over the past several years, and the incorporation of these outcomes into safety initiatives throughout the ACS.

## CINF 15

### **Innovative teaching collaboration provides students with practical drug discovery experience**

**Tamsin E. Mansley**<sup>1</sup>, *tamsin@optibrium.com*, **Robert L. Broadrup**<sup>2</sup>, **Benjamin Perry**<sup>3</sup>, **Hassan Ahamed**<sup>2</sup>, *hahamed1@haverford.edu*, **Tomás Aramburu**<sup>2</sup>, *taramburu@haverford.edu*. (1) Optibrium, Ltd, Cambridge, United Kingdom (2) Chemistry, Haverford College, Malvern, Pennsylvania, United States (3) Drugs for Neglected Diseases Initiative (DNDi), Geneva, Switzerland

Practical experiments in undergraduate chemistry labs frequently focus on reinforcing a specific synthetic route or apparatus that has been introduced during a seminar class. Whilst important, these experiments lack context and do not address the wider set of skills a chemist will require to thrive in a scientific career outside academia.

Haverford College has reinvented their 'Superlab' undergraduate chemistry course to incorporate partnerships with Optibrium, provider of drug discovery software StarDrop™, and Drugs for Neglected Diseases Initiative (DNDi). The new course provides third year students with hands-on experience in an active Leishmaniasis drug discovery project run through the Open Synthesis Network at DNDi.

We will discuss how this collaborative approach provides students with a practical understanding of the challenges faced in a drug discovery project. Students are taught by, and collaborate with, professionals who are experts in the drug discovery field and they have access to the state-of-the-art software with which to develop their ideas. Synthetic skills are honed through synthesis of potential drug candidates, for which students develop their own routes. The resulting compounds are tested by DNDi, adding to the available structure-activity data as the project progresses and enabling the students to contribute to the discovery of potential new therapeutics for a critical disease affecting the developing world. Students also experience working within a drug discovery team and develop the teamwork and collaboration skills required to thrive in their future career.

## CINF 16

### **Driving student success through undergraduate internships in biopharma**

**Natalie Hawryluk**, *nhawryluk@celgene.com*. Global Health, Celgene, San Diego, California, United States

Through Celgene's summer intern program, undergraduates experience multiple components of drug discovery and learn how to work with research project teams. Interns within Celgene Global Health conduct hands-on labwork, learn medicinal chemistry principles, search scientific literature, utilize scientific database, and contribute to publications. Our approach to mentoring, and hands-on student experiences along with student outcomes will be presented.

## CINF 17

### **Evolving data needs in chemistry and bio-sciences: What role can a librarian play in creating student success?**

**Shalini Ramachandran**, *shalinir@usc.edu*, **Karen Howell**. Libraries, University of Southern California, Los Angeles, California, United States

In this talk, we present our perspectives as librarians at the University of Southern California (USC) in engaging with science and engineering students as they negotiate a changing landscape of information and data literacy. One of us has liaison responsibilities in chemistry and the biological sciences and the other is the head of USC's busiest library, Leavey Library, which provides a wide array of programming and services to undergraduate and graduate students. Together, we organized a data manipulation workshop this spring, targeted toward faculty and graduate students in chemistry and biology. Many science and engineering fields generate a lot of research data and, increasingly, students and faculty wish to know how to manage and manipulate data. We decided to organize this workshop based on administering a survey to faculty and students, the results of which showed a high level of interest in learning data skills. Our presentation will focus on what we learned from collaboration with faculty, students, and an external organization (Data Carpentry). Using the data we gathered before and after the workshop, we assess how our experience organizing this workshop can be used to help students succeed in a data-intensive environment. Some of the questions we plan to explore are: 1) How has experimentation changed in the past decade that requires evolving skills on the part of students? 2) What are the differences between faculty, graduate, and undergraduate students in terms of data literacy? 3) What role can librarians play in optimizing student success in the lab and beyond?

## **CINF 18**

### **Training the biomedical workforce for long-term career success**

*Adriana Bankston, abankston81@gmail.com. Future of Research, Abington, Massachusetts, United States*

Sustainability of the research enterprise depends on a well-trained biomedical workforce that is both educated in performing rigorous research and can gain valuable skills required for career success. Adequately training the workforce requires partnerships among various groups, including universities, funding agencies, industry and government. At the same time, it is critical that scientists from all backgrounds have access to high-quality academic training as they progress in their careers. Therefore, systemic change in academia is required for ensuring long-term success of an entire generation of scientists. Proposed systemic changes in academia include better pay, adequate mentoring, leadership opportunities and increased value for the intellectual contributions of young scientists to the research enterprise. Increasing transparency around career outcomes for graduate students and postdoctoral researchers is an additional step towards enabling young scientists to make better informed career decisions. Future of Research aims to empower young scientists to advocate for change in many of these areas in an evidence-based manner. Our organization has made significant impact in advocating for increased postdoctoral salaries, resulting in institutional change. We are currently focused on improving the mentoring landscape within institutions, ensuring meaningful leadership experiences for young scientists in scientific societies, and celebrating their contributions to the manuscript peer review process. These are a few examples of areas where systemic change in academia is needed. We have also advocated for career outcomes transparency and highlighted the efforts of others who made an impact in this space. We hope this work results in meaningful actions among key partners towards improving training for the biomedical workforce and ensuring long-term career success for promising young scientists.

## **CINF 19**

### **Business of student success**

*Jennifer Rosenberg, jrosenberg@cas.org. CAS, Columbus, Ohio, United States*

Universities continue to expand their approaches to enabling student success in today's competitive Academic environments. The development of programs, tools, and resources to meet the needs of students along their Academic careers is paramount to those strategies. These strategies have an interesting parallel to the Customer Success methodologies being employed by many corporations as a business differentiator and competitive advantage. This session will explore how academic institutions may benefit from these same methodologies, as they strive to serve their students in the areas of onboarding, support, development and retention.

## **CINF 20**

## Automatic identification of relevant chemical compounds from patents

Saber A. Akhondj<sup>5,2</sup>, Hinnerk Rey<sup>1</sup>, Markus Schwörer<sup>1</sup>, **Michael Maier**<sup>1</sup>, *m.maier@elsevier.com*, John Toomey<sup>3</sup>, Heike Nau<sup>1</sup>, Gabriele Ilchmann<sup>1</sup>, Mark Sheehan<sup>2</sup>, Matthias Imer<sup>4</sup>, Claudia Bobach<sup>4</sup>, Marius Doornenba<sup>2</sup>, Michelle Gregory<sup>2</sup>, Jan Kors<sup>5</sup>. (1) Elsevier Information Systems GmbH, Frankfurt, Germany (2) Elsevier B.V., Amsterdam, Netherlands (3) Elsevier Limited, London, United Kingdom (4) OntoChem GmbH, Halle/Saale, Germany (5) Erasmus MC, Rotterdam, Netherlands

In commercial research and development projects, public disclosure of new chemical compounds often takes place in patents. Only a small proportion of these compounds are published in journals, usually a few years after the patent. Patent authorities make available the patents but do not provide systematic continuous chemical annotations. Content databases such as Elsevier's Reaxys provide such services mostly based on manual excerpts, which are time-consuming and costly. Automatic text-mining approaches help overcome some of the limitations of the manual process. Different text-mining approaches exist to extract chemical entities from patents. The majority of them have been developed using sub-sections of patent documents and focus on mentions of compounds. Less attention has been given to relevancy of a compound in a patent. Relevancy of a compound to a patent is based on the patent's context. A relevant compound plays a major role within a patent. Identification of relevant compounds reduces the size of the extracted data and improves the usefulness of patent resources (e.g. supports identifying the main compounds). Annotators of databases like Reaxys only annotate relevant compounds. In this study, we design an automated system that extracts chemical entities from patents and classifies their relevance. The goldstandard set contained 18 789 chemical entity annotations. Of these, 10% were relevant compounds, 88% were irrelevant and 2% were equivocal. Our compound recognition system was based on proprietary tools. The performance (F-score) of the system on compound recognition was 84% on the development set and 86% on the test set. The relevancy classification system had an F-score of 86% on the development set and 82% on the test set. Our system can extract chemical compounds from patents and classify their relevance with high performance. This enables the extension of the Reaxys database by means of automation.

### CINF 21

#### Augmenting manual curation of chemical patent information in the Derwent World Patents Index

**Andrew Klein**, *andrew.klein@clarivate.com*, Susan McGhee, Justine Hookes. Clarivate Analytics, Saint Paul, Minnesota, United States

The Derwent World Patents Index (DWPI) database is well known for producing manually indexed records of specific and general chemical substances in patent documents. Processing more than a million new patent documents per year requires extraordinary efforts, and would be impossible without a robust information processing workflow. This paper will detail the role that text mining and natural language processing plays to assist with the Derwent editorial process, including the named entity recognition systems themselves, their place in the overall workflow, and the value that this indexing provides for searching patents by chemical and polymeric substances.

### CINF 22

#### Journey continues: Addition of French, Russian, Chinese, Korean, and Japanese patents to PATENTSCOPE ChemSearch

**Josef Eiblmaier**<sup>1</sup>, *je@infochem.de*, **Christophe Mazenc**<sup>2</sup>, *christophe.mazenc@wipo.int*, **Dorothee Geppert**<sup>1</sup>, *dg@infochem.de*, **Larisa Isenko**<sup>1</sup>, *li@infochem.de*, **Heinz Saller**<sup>3</sup>, *hs.consultant@infochem.de*. (1) InfoChem GmbH, Munich, Germany (2) World Intellectual Property Organization, Geneva, Switzerland (3) Dr. Saller Beratungs- und Beteiligungs GmbH, Stockdorf, Germany

PATENTSCOPE is a free patent search system offered by the World Intellectual Property Organization (WIPO), covering over 66 million patent documents.

The last big enhancement of the system was the addition of chemical search capabilities, accomplished using

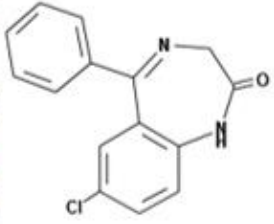


InfoChem's automatic text- and image-mining technologies.

An automatic workflow was developed and put into operation allowing real-time, multi-modal chemical text annotation and image recognition.

Since the number of patent applications in some Asian countries increases rapidly, this process was enhanced to also handle French, Russian, Korean, Japanese and Chinese files in the most recent project phase. This talk will report the results and will address technical challenges encountered such as OCR quality, heterogeneity of sources, scalability, performance and parallelization.

洋)、7-硝基-5-苯基-1H-1,4-苯并二氮杂革-2(3H)-酮(硝基安定)、7-氯-5-苯基-1H-1,4-苯并二氮杂革-2-(3H)-酮(甲西洋)、去甲左啡诺、6-二甲氨基-4,4-二苯基-3-己酮(去甲美沙酮)、去甲吗啡、二苯派己酮、属于罂粟(Papaver somniferum)的汁液(鸦片)、7-氯-3-羟基-5-苯基-1H-1,4-苯并二氮杂革-2(3H)-酮(奥沙西洋)、(顺式-反式)-1-甲基-11b-苯基哌啶并[3,2-d][1,4]苯并二氮杂革-6-(5H)-酮(奥沙3-吗啡烷酮(羟考酮)、氧吗啡酮、属于罂粟物种(包括setigerum亚种)的阿片全碱、2-亚氨基-5-苯基-4-哌啶酮(pernoline)、1,2,3,4,5,6-六氢-6,11-二甲基-3-(3-哌啶在辛)、5-乙基-5-(1-甲基丁基)巴比妥酸(戊巴比妥)、乙基-(1-甲基-4-苯基-4-哌啶-甲酰胺)那佐辛、苯哌利定、匹米诺定、福尔可定、3-甲基-2-苯基吗啡(芬美曲咪)、5-乙基-5-苯基巴比妥酸(司可巴比妥)、7-氯-5-苯基-1-(2-丙炔基)-1H-1,4-苯并二氮杂革-2(3H)-酮(匹那西洋)、 $\alpha$ -(2-哌啶基)二甲甲基-4-联哌啶-4-甲酰胺(哌替米特)、7-氯-1-(环丙基甲基)-5-苯基-1H-1,4-苯并二氮杂革-2(3H)-酮(普拉多、异丙哌替啶、丙氧芬、N-(1-甲基-2-哌啶子基乙基)-N-(2-吡啶基)丙酰胺(瑞芬太尼)、5-仲丁基-5-乙基巴比妥酸(仲丁巴比妥)、5-烯丙基-5-(1-甲基丁基)巴比妥酸(司可巴比妥)、N-(4-甲氧基甲基-1-(2-(2-噻吩基)乙基)-4-哌啶基)丙酰胺(舒芬太尼)、7-氯-2-羟基-甲基-5-苯基-1H-1,4-苯并二氮杂革-2-(3H)-酮(替马西洋)、7-氯-5-(1-环己烯基)-1-甲基-1H-1,4-苯并二氮杂革-2(3H)-酮(四氢西洋)、乙基-(2-二甲氨基-1-苯基-3-环己烷-1-甲酰胺)(替利定(顺式和反式))、曲马多、8-氯-6-(2-氯苯基)-1-甲基-4H-[1,2,4]三唑并[4,3-a][1,4]苯并二氮杂革(三唑仑)、5-(1-甲基丁基)-5-乙基巴比妥酸(乙哌比妥)、(1R,2R\*)-3-(3-二甲氨基-1-乙基-2-甲基-丙基)苯酚、(1R,2R,4S)-2-[二甲基氨基]甲基-4-(对-氟苯氧基)-1-(间-甲氧基苯基)环己醇,各自任选为相应的立体异构的化合物以及相应的衍生物,尤其是酯或醚的形式,并且全部为生理相容性的化合物,尤其是盐和溶剂化物。在一些实施方案中,药物可以药理学有效量存在于治疗组合物中。在一些实施方案中,药物可以约1wt%至约25wt%;约1wt%至约22wt%;约1wt%至约20wt%;约1wt%至约18wt%;约1wt%至约16wt%;约1wt%至约14wt%;约1wt%至约12wt%;约2wt%至约10wt%;约2wt%至约8wt%;约3wt%至约8wt%;约4wt%至约7wt%;约5wt%至约7wt%、或约6wt%至约7wt%的量存在于治疗组合物中。在一些实施方案中,药物可以约1wt%;约1.5wt%;约2wt%;约2.5wt%;约3wt%;约3.5wt%;约4wt%;约4.5wt%;约5wt%;约5.5wt%;约6wt%;约6.5wt%;约7wt%;约7.5wt%;约8wt%;约8.5wt%;约9wt%;约9.5wt%;约10wt%;约10.5wt%;约11wt%;约11.5wt%;约12wt%;约12.5wt%;约13wt%;约13.5wt%;约14wt%;约14.5wt%;约15wt%;约15.5wt%;约16wt%;约16.5wt%;约17wt%;约17.5wt%;约18wt%;约18.5wt%;约19wt%;约19.5wt%;约20wt%;约21wt%;约22wt%;约23wt%;约24wt%;或约25wt%的量存在于治疗组合物中。在一些实施方案中,药物可以约6.12wt%的量存在于治疗组合物中。



Nordazepam

## CINF 23

### Automating chemical structure and inhibition data extraction from patents: Text-mining approach

Francisco Costa, Ioannis Haldoupis, Jeffrey L. Nauss, jnauss77@yahoo.com, Andrew Hinton. Linguamatics, Cambridge, United Kingdom

Patents are a widely recognised valuable source of early structure-activity relationship (SAR) information; however, the manual efforts involved in extracting SAR data from patents is costly and exacerbated by the purposely obfuscated nature of patents. Technological advances in text mining of patents using natural language processing (NLP) approaches has recently resulted in the creation of efficient workflows for the extraction of SAR data. The workflow combines and integrates software capable of converting chemical text into structural representation with NLP software that recognises the wide variety of different ways inhibition data can

be described within the text of patents including biological assay data found inside tables. Linguamatics describes use of the I2E software in establishing a triage method that allows medicinal chemists to quickly assess the potential of a patent to be text mined for SAR data.

## CINF 24

### **BioAssay express: Creating and exploiting assay metadata**

**Philip Cheung**<sup>1</sup>, *philip.p.cheung@gmail.com*, **Alex Clark**<sup>1,2</sup>, **Janice Darlington**<sup>3</sup>. (1) *Research Informatics, Collaborative Drug Design, San Diego, Alabama, United States* (2) *Independent, Montreal, Quebec, Canada* (3) *Collaborative Drug Discovery, San Diego, California, United States*

The challenge of accurately characterizing bioassays is a real pain point for many drug discovery organizations. Research has shown that some organizations have legacy assay collections exceeding 20,000 protocols, the great majority of which are not accurately characterized. This problem is compounded by the fact that many new protocol registrations are still not following FAIR (Findability, Accessibility, Interoperability, and Reusability) Data principles.

BioAssay Express is a tool focused on transforming the traditional protocol description from an unstructured free form text into a well-curated data store based upon FAIR Data principles. By using well-defined annotations for assays, the tool enables precise ontology based searches without having to resort to imprecise keyword searches.

This talk explores a number of new important features designed to help scientists accelerate the drug discovery process. Some example use-cases include: enabling drug repositioning projects; improving SAR models; identifying appropriate machine learning data sets; fine-tuning integrative-omic pathways;

An aspirational goal for our team is to build a metadata schema based on semantic web vocabularies that is comprehensive to the extent that the text description becomes optional. One of the many possibilities is to take the initial prospective ELN entry for a bioassay protocol and feed it directly to an automated instrument. While there are many challenges involved in creating the ELN-to-robot loop, we will provide some insights into our collaborations with UCSF automation experts.

In summary, the ability to quickly and accurately search or analyze bioassay data (public or internal) is a rate limiting problem in drug discovery. We will present the latest developments toward removing this bottleneck.

## CINF 25

### **Building fast, robust, and reliable prediction models using very large biological data sets**

**Lutz Weber**<sup>1</sup>, *lutz.weber@ontochem.com*, **Hans-Joachim Boehm**<sup>2</sup>. (1) *IT, OntoChem, Stuttgart, Germany* (2) *Novomol, Loerrach, Baden, Germany*

Access to databases, patents, publications, webpages and other sources for biological activities of small molecules now allows the generation of very large sets of data with several billion knowledge triples. In our presentation, we will describe a new computational tool called EDEN that enables the conversion of very large data sets into computational models to predict the biological activity, DMPK and Tox related properties for any given molecule.

Using either 2D structures of active compounds and/ or the 3D protein structures of the target as input, EDEN generates a complex affinity fingerprint. This is then used to predict the activity of existing compounds and to design new compounds de novo with the desired biological activity.

We believe that EDEN delivers candidate compounds with very high accuracy and a low number of false positives because the computational model used to predict activity is more complete than other approaches. It uses both all “good bits” (required for binding) and all “bad bits” (detrimental to binding). This leads to very high enrichment factors. The program is very fast and can presently handle roughly 500 million compounds in one day on a standard notebook.

We will explain how EDEN identifies and addresses key challenges with big data sets such as noisy datasets

with many false positives or biased heterogeneous datasets with most of the data belonging to one chemical series.

Examples will be presented to illustrate some of the capabilities of EDEN.

## CINF 26

### **IUPAC and its role in the development of chemical nomenclature and structure representation**

**Richard Hartshorn**, *richard.hartshorn@canterbury.ac.nz*. University of Canterbury, Christchurch, New Zealand

*IUPAC (www.iupac.org) has a proud history in providing the essential tools for the application and communication of modern chemical knowledge. This has been particularly significant in the area of nomenclature and structure representation, where the “colour books” have had special significance.*

In this presentation I will provide a brief outline of the development of chemical nomenclature since the establishment of IUPAC in 1919. This will then lead to a discussion of changes in the ways that compounds are being presented in the literature and other documents, and what this means for the future of structure representation in a FAIR world of chemical data (FAIR = Findable, Accessible, Interoperable, Reuseable). Are the “colour books” as important as they used to be, and do we need formal nomenclature anymore?

As Secretary General of IUPAC, I have significant input into the future direction of IUPAC and its work. I will present my perspective on what the future holds for the organisation, and for you as nomenclaturists.

There will be some history, some explanation, some crystal ball-gazing, and a few long names (merely because I have to play up to expectations of someone still involved in nomenclature development – at least a little bit).

## CINF 27

### **IUPAC brief guides on nomenclature: Summary of the key nomenclature principles addressed in IUPAC colored books**

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

Over the past several years there have been IUPAC projects focused on the development of Brief Guides for Polymer, Inorganic and Organic chemistry. These brief guides capture the key nomenclature principles in the Purple, Red and Blue books. As a result of the very thoughtful approach taken to condense very large nomenclature resources into 3-4 pages, the brief guides are an excellent resource for both practicing scientists and members of the chemical education community when it comes to learn about, using or teaching nomenclature. In this talk I will provide an overview of the brief guides on polymer and inorganic nomenclature and how these can be utilized as resources, particularly in a classroom setting.

## CINF 28

### **Evolution of CAS nomenclature: Past, present, and future**

**Molly A. Strausbaugh**, *mstrausbaugh@cas.org*. CAS, Columbus, Ohio, United States

With over 110 years' experience curating the world's scientific literature, CAS has a long history of leadership in chemical substance nomenclature. Historically, CAS developed systematic names for substances for the purpose of a printed index. That index, *Chemical Abstracts*, empowered scientists to search literature by chemical name for the first time by ensuring a unique and consistent name for each discrete substance. However, as technology and utility of chemical nomenclature evolve, CAS's nomenclature and solutions are evolving as well to overcome new challenges. This talk will discuss key drivers and considerations for advances in CAS's nomenclature over the years, as well as emerging trends and technologies such as chemical information being embedded in text strings and big data technologies that are driving new nomenclature opportunities and solutions.



## CINF 29

### Updating the Braille Code of Chemical Notation 1997

*Philip Verhalen, philip.verhalen@gmail.com. Panola College, Scottsville, Texas, United States*

In 2017, the Braille Association of North America (BANA) asked the American Chemical Society (ACS) to help update the Braille Code of Chemical Notation. Their request stemmed from the differences in how individual textbook authors and editors would represent the same chemical concept. Differences in the ways that chemists represent the same concept made translation into Braille by non-chemists difficult. The ACS Nomenclature, Terminology and Symbols Committee worked with the Braille Association Chemistry Subcommittee to provide recommendations for updates and corrections to the 1997 code. The committee has completed its review in partnership with BANA. The suggested and updates there were developed by the working group are being submitted to BANA.

## CINF 30

### Carbon nanotube nomenclature: Challenges of naming emerging materials

*Elisabeth Mansfield, elisabethmansfield@gmail.com. Materials Reliability Division-853, NIST, Boulder, Colorado, United States*

Carbon nanotubes (CNTs) were first reported in a seminal Nature manuscript by Iijima. Since that 1991 publication, CNT reports have exploded, with new approaches to synthesis, opportunities for applications and unique properties reported consistently. Nomenclature of carbon nanotubes, however, has not progressed in the same fashion. A few individuals broached the subject but in general the terminology and nomenclature used varied publication-to-publication. The International Standards Organization (ISO) group on Nanomaterials (TC 229) undertook the terminology aspect, but CNT terminology had not yet been developed until an IUPAC project began in 2013. Approaches to nomenclature of CNTs is presented, along with the challenges of naming a material with varied dimensions, composition and structure.

## CINF 31

### Chemical nomenclature from books to computers: ACD/Name and IUPAC Division VIII

*Andrey Yerin, erin@acdlabs.ru. ACD/Labs, Toronto, Ontario, Canada*

The IUPAC, responsible for nomenclature development and celebrating its centennial this year, has published several volumes of specific rules. But in over 200 years of development, the nomenclature has grown too complex to learn and apply.

Algorithmic name generation entered the market in the early 1990s, with ACD/Labs among the first software providers with ACD/Name. The real challenge for programming was taking the multitude of nomenclature rules and converting them into computer algorithms. Since the early stages of our software's development, ACD/Labs has been invited to work with the IUPAC Commission on the Nomenclature of Organic Chemistry and later the Chemical Nomenclature and Structure Representation Division. This involvement in nomenclature projects allowed the development team behind Name to learn nomenclature in greater detail, and helped turn those huge printed volumes into reliable algorithms of name generation.

At the same time our work developing nomenclature algorithms allowed us to detect areas that lacked the necessary criteria, and propose procedures that were eventually included in the current IUPAC recommendations. With the implementation of the Preferred IUPAC Name (PIN) concept in Name algorithms, a number of name errors were identified in the 2013 version of the IUPAC Blue Book. Corrections to these errors are expected to be published in the Errata for this book; once again proving that even nomenclature experts make mistakes, and can benefit from algorithmic name generation.

The past 25 years of ACD/Name development showcases the mutual benefits for both chemical nomenclature as a field of study, and software developers alike. While name generation tools from several vendors are now

available and heavily used for name generation in various electronic media, algorithmic nomenclature development is far from complete, and requires further collaboration of nomenclature bodies and software vendors to ensure high quality names for all classes of chemical substances.

## CINF 32

### **IUPAC, nomenclature, and chemical representation: From the perspective of a worldwide structural database**

**Matthew P. Lightfoot**, *lightfoot@ccdc.cam.ac.uk*, Ian Bruno, Clare Tovee, Suzanna Ward, Seth Wiggin. Cambridge Crystallographic Data Centre, Cambridge, United Kingdom

Over the last 100 years IUPAC has had a critical role in the creation and the evolution of the common language used in chemistry. This presentation will reflect on the part IUPAC and developments in nomenclature and representation have played in the creation of the Cambridge Structural Database (CSD), a resource of 1 million small molecule crystal structures.

Initially, an important part of CSD creation was having a standardised naming system and for this we relied heavily on IUPAC naming conventions. As the CSD has evolved into a modern database for search and analysis, chemical names still provide an important way for users to find data and learn from the structure. Until about 10 years ago we primarily used the IUPAC nomenclature books to facilitate this; more recently we have integrated chemical naming software which itself heavily relies on IUPAC conventions.

The number and complexity of metal-organic structures in the CSD has risen sharply over the years and we will describe the challenges that this has given us with providing standardised chemical representations. Amongst the metal-organic structures in the CSD are an increasing number of MOF structures for which it is useful to have a clear indication of topology. This has seen us contribute to the IUPAC task force looking at topology representations and a PhD research project in this area.

It is not just metal-organic structures that present challenges for providing consistent naming, even simple organics can be complex. For example, one current challenge is how to name different forms of a structure, known as polymorphs, when the existence of other polymorphs is unknown.

We will describe how nomenclature and representation conventions are used within the CSD, areas where challenges still remain and how reliable nomenclature and representation helps enable researchers to gain new insights from structural data.

## CINF 33

### **Chemical representation: Toolbox for human and machine collaboration**

**Leah R. McEwen**<sup>1,2</sup>, *lrm1@cornell.edu*, Evan Hepler-Smith<sup>3</sup>. (1) Clark Library, Cornell University, Ithaca, New York, United States (2) Committee on Publications and Cheminformatics Data Standards, International Union of Pure and Applied Chemistry (IUPAC), Research Triangle Park, North Carolina, United States (3) History Department, Boston College, Boston, Massachusetts, United States

Chemical representation is a cornerstone of communication in the chemical sciences. IUPAC, in collaboration with many chemical institutions, has been engaged in refining this practice for over 100 years. From the very beginning it was evident that the "language" of chemical representation was important not only as a tool used by chemists to interpret and communicate their own research, but also as a systematic organizing principle for the entire field's "knowledge space": the accumulated results of global chemistry. Using chemical structures as an indexing motif, practicing chemists have been able to conduct sophisticated searches in the literature and identify gaps in the state of the art. Little did our venerable early colleagues realize the enormous expansion of analysis that would become possible through computing and cheminformatics based on the critical underlying framework of chemical representation. This is all made possible through active partnership of humans and machines using a number of approaches towards a common goal of unambiguous communication about the structural motifs of molecules and their associated properties. Ultimately chemical representation and informatics techniques are tools in the toolbox of the chemical sciences. This talk will advocate for the role of machines in helping us organize, communicate and analyze the chemical space, and emphasize the critical need for skilled chemists to continue to develop and apply these tools to chemical problems.

## CINF 34

### Helping students stand out in the academic job market

**Robert J. Gilliard**, *rjg8s@virginia.edu*. Dept of Chemistry, University of Georgia, Athens, Georgia, United States

For STEM students pursuing doctorate degrees, the ultimate goal is often a tenure-track faculty position. Current faculty typically mentor their students by sharing their knowledge and experiences in basic research, journal publication and grant funding. But as competition for coveted faculty positions continues to grow, graduate students must set themselves apart from their peers by developing soft skills and by securing leadership roles and accolades outside of the classroom. This talk will explore the various ways that faculty can help their students stand out in the academic job market by providing unique opportunities to collaborate, be recognized for scientific achievement and network beyond their institutional circle of influence.

## CINF 35

### Connecting the dots across academia and industry to ensure skill alignment

**Matthew Grandbois**, *grandboismatthew@gmail.com*. ACS Younger Chemists Committee, Boston, Massachusetts, United States

The career opportunities for STEM graduates are endless and expand beyond the lab. Are academics and their commercial counterparts working together in a way that make the opportunity of a career beyond the lab a possibility. Landing in the right spot requires an investment in skills beyond the technical acumen of the courses and research. Success for scientific students demands the ability to make connections and communicate. Should students and early-career professionals have to pursue these outside of their organization or are we preparing these professionals with the right skills to success not only to their first job but each step afterward?

## CINF 36

### Creating digital learning objects for chemistry

**Yulia Sevryugina**, *yulias@umich.edu*. University of Michigan, Ann Arbor, Michigan, United States

The Chemistry Librarian and the Digital Education Librarian will present their collaborative work on creating digital learning objects (DLOs) for teaching information literacy to chemistry students. One of the emerging primary needs for academic libraries is to provide students with subject specialized information literacy instruction. This demand creates an increasing teaching workload on liaison librarians. To sustain this workload while providing other essential university services, we considered implementing DLOs for teaching specialized information literacy instruction to chemistry students. DLOs have been demonstrated to provide effective, flexible, economical, facilitated, and interactive training to students. The goal of our DLO is to enhance students' information literacy skills by offering a simple and interactive tool that will assist students in preparing a bibliography for a scientific paper. Writing experimental reports or literature reviews is a common assignment for many STEM classes and creating a properly formatted bibliography is an important component of this assignment. Undergraduate students are generally unfamiliar with the concepts such as writing a bibliography, or finding the scholarly sources, or proper referencing someone else's work and our DLO addresses this gap in students' literacy skills. In our DLO, we adopt the formatting style of the American Chemical Society (ACS). Our DLO also covers general topics such as defining scholarly sources and importance of citing someone else's work. In this presentation, we will demonstrate our collaboration strategies as well as various aspects we considered while working on DLO's design. Those include selection of the most suitable module building software, accessibility considerations to meet the library's goals of providing diversity, equity and inclusion (DE&I) for all users, creative uses of technology to assess students' learning, and successful approaches to inter-departmental collaboration. We will share our strategies regarding the module's distribution and pilot testing. This presentation will be of interest to a broad audience that includes liaison librarians, instructors for

undergraduate and graduate chemistry courses, instructional designers, instruction librarians, online learning librarians and instructors, and any teacher or student.

### **CINF 37**

#### **RA21: Secure, seamless access for research**

*Ralph Youngen, r\_youngen@acs.org. ACS, Washington, District of Columbia, United States*

Secure and seamless identity authentication is essential for today's geographically-dispersed research collaboration. The RA21 initiative (<http://ra21.org>) recently published a NISO Recommended Practice, which calls for the use of federated identity management to provide a secure and streamlined user experience for access to scholarly information resources and research collaboration services. RA21 began in 2016 as a joint initiative sponsored by NISO and the International Association of STM Publishers. Since that time, RA21 has worked collaboratively with stakeholders from academic institutions, corporations, identity federations, and scholarly publishers to conduct pilots and test prototypes with practicing researchers. This session will discuss how the RA21 recommendations will improve student success by providing more seamless access to scholarly information resources while also striking the right balance with potential privacy concerns.

### **CINF 38**

#### **Experiences in scientific information literacy education**

*Jiuming Ji, jjm@ecust.edu.cn. Library, East China University of Science and Technology, Shanghai, China*

Since 1979, East China University of Science and Technology (ECUST) Library has been engaged in science information literacy education for nearly 40 years, which has played an important role in the success of students. With the development of information technology and the demand of industry for innovative students, we have made some changes and won some awards. To the credit course *literature retrieval* handled by librarians, changes were from teaching mode and content of learning materials. For example, teaching by a self-learning computer systems, students are guided not only to learn literature searching skills but also to learn knowledge discovery methods. We also endeavored to find ways for information literacy educations outside credit course systems and library buildings. Since September 2009, we have hosted a series of video competitions on information literacy. In March 2018, we launched the MOOC version of our literature retrieval course on the platform of China University MOOC. Since then, more than 40,000 college students enrolled in the course. Some librarians are mentors of University Students Research Experiments and Innovation Practices (USRP) programs granted by educational authorities of government or our school. We are also involved in our school's short-term practical curriculum system (STHC), and our research-oriented literature analyzing courses have won the favor of students. In this poster, list of research-oriented literature analyzing course titles and works of students will be showed.

### **CINF 39**

#### **Identifying the different definitions of student success between young scientists, faculty and administration in academia and hiring managers in industry**

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

CAS conducted a survey to determine how student success is defined by the various parties involved in the journey from academia to career placement. The survey results will be shared to determine how we work together to collaborate for success from the classroom to the boardroom. This talk will highlight any areas of disconnect between the definitions and identify ways we can work to ensure the success of our students.

### **CINF 40**

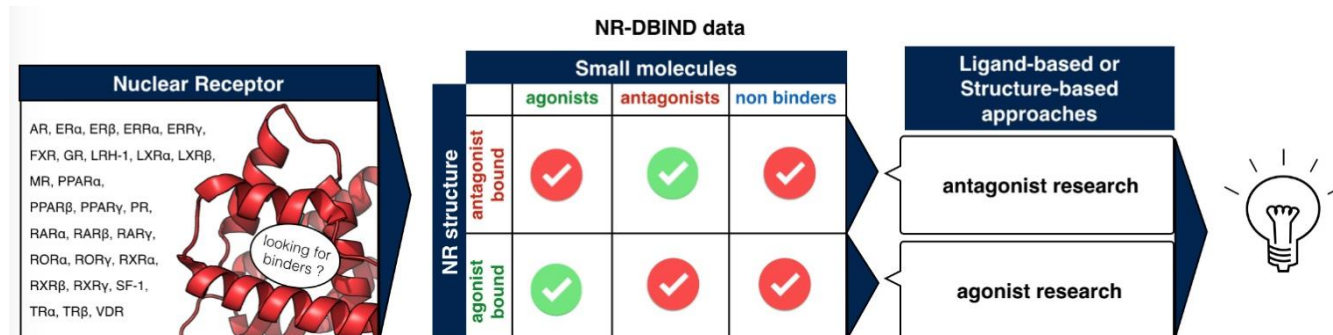
## Importance of inactive data in models: Application to virtual screening and example of nuclear receptors

*Manon Réau, manon.reau@lecnam.net, Nathalie Lagarde, Matthieu Montes. CNAM, Paris, France*

Virtual screening (VS) has become strongly established in the drug discovery process to identify hit(s), i.e. molecules able to bind and modulate the activity of their target, as well as for off-targets and endocrine disruptors prediction. Many ligand- and structure-based computer aided drug design (CADD) methods have emerged to propose strategies depending on the available information. Still, we observe a gap between the retrospective prediction performance of the CADD methods on reference data and their prospective performance on new targets. Two reasons may explain this observation: either the tools are not sensitive enough, or the reference data used for evaluation are not representative of real-life cases.

This introduces one aspect that has not been deeply considered yet in *in silico* drug design : the importance of inactive data in model construction and evaluation. Indeed, VS tools are evaluated and calibrated on standard benchmarking databases composed of an active and an inactive set. However, due to the lack of inactive data, decoy compounds, i.e. molecules assumed to be inactive, are used instead. Despite the efforts deployed in the decoys selection optimization, some inherent biases are observed. For instance, the structural dissimilarity with known active compounds imposed in the decoys selection does not expose to the activity gap frequently observed between highly similar molecules in real-life cases.

Herein, we present the first database including inactive data that can be used for model evaluation and construction : the NR-DBIND ([www.nr-dbind.drug-design.fr](http://www.nr-dbind.drug-design.fr)). The NR-DBIND is dedicated to the nuclear receptors (NR) family widely studied for therapeutic and public health concerns. It contains ~15000 manually reviewed binding data for 28 NRs including ~1500 inactive data mentioned in the literature and pharmacological profile annotation extracted from activity assays. The NR-DBIND constitutes a robust basis that should facilitate rational identification NR-modulators. Still, it contains a frustrating ratio of active/inactive compounds that reflects the publication bias. We believe that the only way to truly improve VS performances belongs to the medicinal chemists and pharmacologists community: the more inactive data are published, the more models performance is improved.



## CINF 41

### Crystal-structure prediction via basin-hopping global optimisation employing tiny periodic simulation cells and multipole expansion

*Christian Burnham, Pralok Samanta, Mohammad Reza Ghaani, mohammad.ghaani@ucd.ie, Niall English. School of Chemical and Bioprocess Engineering, University college Dublin, Dublin, Ireland*

A crystal-structure prediction (CSP) suite of algorithms and approaches for use under periodic-boundary conditions with empirical rigid models is presented, which employs (i) unrestricted cutoff radii for the real-space interactions, thus allowing the treatment of even very small unit cells, (ii) a global-optimisation algorithm based on the basin-hopping method, and (iii) exploiting the multipole expansion with a novel expression in the form of



a graphical and spherical-harmonics framework. These algorithms are applied to the phase prediction for ice polymorphs, in terms of lower-lying enthalpies over a broad pressure range, with a popular empirical potential, and high-quality electronic-structure methods are then used to re-rank the energetic stability, with reasonable, although imperfect, accord between empirical models and electronic-structure. Structure-factor and space-group determination are also presented, alongside with the outlook for finite-temperature determination of energies. Scope for extension of the present CSP methods for application towards 'floppier' molecules with substantially greater intramolecular degrees of freedom will be presented, alongside several case-study examples of interest to solid-state pharmaceuticals and organic chemistry.

## CINF 42

### CDD vault: Complexity simplified

**Janice Darlington**<sup>1</sup>, [jdarlington@collaboratedrug.com](mailto:jdarlington@collaboratedrug.com), Whitney W. Smith<sup>1</sup>, Barry A. Bunin<sup>2</sup>. (1) CDD, San Diego, California, United States (2) CDD, Belmont, California, United States

CDD Vault® is a platform that provides a hosted database solution for secure management and sharing of chemical and biological data. Researchers can organize chemical structures and biological study data, and securely collaborate with internal or external partners through a web interface. CDD Vault is differentiated by ease-of-use and superior collaborative data sharing workflows. Within the CDD Vault software, Activity & Registration, Visualization, Electronic Laboratory Notebook (ELN) and Inventory are well integrated for handling the majority of private drug discovery data requirements.

CDD Vault allows scientists to more effectively handle chemical complexity (registration, stereochemistry, mixtures, batches) and biological complexity (enzyme, phenotypic, cell, animals, IC50, Z/Z', and metadata), as well as natural workflows for secure collaborations.

CDD Public is an open database that can be mined through CDD Vault and is freely accessible to scientists at no charge. Importantly, it contains negative data which makes it valuable for those involved in predictive model development.

CDD Vault has been utilized as the central platform for multi-national collaborations including NIH Blueprint (11 organizations), Bill & Melinda Gates foundations (BMGF) Tuberculosis (TB) collaborations (14 organizations including 7 big pharma), and More Medications for Tuberculosis (MM4TB) (25 organizations including 2 big pharma). The search, analysis, and data visualization capabilities of CDD Vault will be presented.

## CINF 43

### Uncertainty estimation of individual QSAR predictions using multiple sources of information

**Christina Founti**<sup>1</sup>, [cmfounti1@sheffield.ac.uk](mailto:cmfounti1@sheffield.ac.uk), Val J. Gillet<sup>2</sup>, Jonathan Vessey<sup>3</sup>. (1) Information School, University of Sheffield, Sheffield, United Kingdom (2) Information School, Regent Court, University of Sheffield, Sheffield, United Kingdom (3) Lhasa Limited, Leeds, United Kingdom

Recognised practices for the development of QSAR models highlight the importance of reporting on the reliability of the models' output. However, this is not a trivial task for all modelling techniques, particularly in the case of regression algorithms where the accuracy of individual predictions needs to be assessed. This has been previously addressed by adding an additional step to the QSAR pipeline for the development of error models using machine learning algorithms. Although this approach is far from ideal, as it introduces further complexity to the workflow, it facilitates the use of information that is unseen by the QSAR model to estimate prediction error. However, the main challenge is that the performance of error models is, generally, poor.

In this work a consensus approach for the estimation of uncertainty in ADME predictions is investigated, which includes the use of variables from different applicability domain (AD) methods. The conformal prediction (CP) framework for uncertainty estimation facilitates the comparison of methods that produce estimates in different scales. Standard reliability indices of compounds based on AD were converted to compound-specific uncertainty estimates, which in CP are represented as prediction regions. The results were first assessed against the prediction intervals obtained from error models and sampling techniques and then against the uncertainty of the

assay. The variation of informative, prediction intervals was evaluated for each compound to obtain a final uncertainty estimate for each individual prediction.

## CINF 44

### Medicinal chemistry based measure of R group similarity

*Noel O'Boyle, baoilleach@gmail.com, Roger A. Sayle. NextMove Software, Cambridge, United Kingdom*

Molecular similarity is one of the most central concepts in chemoinformatics. Typical measures of molecular similarity (such as the Tanimoto coefficient of binary fingerprints) are used for tasks such as similarity search, distinguishing similar molecules from dissimilar (e.g. identifying actives in a virtual screen), measuring the diversity of a dataset or selecting a diverse subset. While the measurement of R group similarity is conceptually the same as for whole molecules, in practice existing methods for measuring molecular similarity perform poorly.

Improved methods to measure R group similarity are of particular importance in the context of a medicinal chemistry project. These often proceed by changing one R group at a time, advancing through matched pairs. Given an appropriate measure of R group similarity, it should be possible to suggest relevant modifications or identify gaps in the project data that should be filled. In a computational context, R group enumeration could be used to generate relevant candidate molecules for virtual screening or purchase.

In medicinal chemistry, the term bioisosteric replacement refers to a substitution that retains broadly similar biological properties. However, there is a need to go beyond the concept of bioisosteres/non-bioisosteres to handle levels of similarity; for example, chloro is regarded as more similar to fluoro than to amino. Clearly, in this context, the extent to which two R groups are similar is not just (or not even) a question of shared substructures. Previous approaches to this problem include the use of R group descriptors by Holliday et al which maps atom-based descriptor values onto a vector by distance from the attachment point, and the use of reduced graphs to encode bioisosteric equivalences by Birchall et al.

We propose to use co-occurrence in medicinal chemistry project data to derive a measure of R group similarity. We will use two distinct sources for these data, both in the public domain. The first is the ChEMBL database, which contains assay data extracted from a range of medicinal chemistry journals. The other source is the US patent literature, which provides text and ChemDraw sketches from which a large amount of medicinal chemistry data can be extracted. While large-scale mining of medicinal chemistry data has previously been used to detect bioisosteres, to our knowledge this is the first time it has been used to develop a method to measure R group similarity.

## CINF 45

### Systematic pipeline for automated structure-based molecular design: Beyond the static picture of hepatic organic anion transporting polypeptides

*Alzbeta Tuerkova, alzbeta.tuerkova@univie.ac.at, Barbara Zdrazil. Department of Pharmaceutical Chemistry, University of Vienna, Vienna, Austria*

Uptake transporters belonging to the solute carrier (SLC) family are playing a pivotal role in the development of new drugs, mainly due to their involvement in drug-drug interactions, adverse drug effects and toxicity. Therefore, structure-based modeling of SLC transporters can provide useful insights into binding events and resultant models can be applied for screening purposes. However, those efforts are limited by the lack of experimental 3D structure for the majority of SLC transporters, as well as their promiscuous nature with respect to ligand recognition.

State-of-the-art docking protocols do not usually account for protein flexibility which is being recognized as one of the crucial aspects for target-ligand recognition. Last but not least, an automated workflow for transporter modeling is very handy if the modeling procedure has to be repeated in case of newly released crystal structures or new results from biochemical studies (e.g. mutational studies). Here, we present a (semi)automatic modeling pipeline which provides a promising strategy for structure-based molecular modeling including iterative design cycles. First, structural templates are detected on basis of fold-recognition methods. The initially detected

template(s) is/are used to search the PDB database for analogous protein(s) with shared 3D fold. Afterwards, normal modes are calculated for identified proteins in order to explore their dynamics-based phylogeny, which can inform about possible additional templates suitable for homology modeling. Ensemble docking into multiple transporter conformations is then performed, since this presents a useful approach for unraveling binding mode hypotheses. For this purpose, an extensive conformational sampling of pre-selected templates is done on basis of normal modes. Distinct protein conformations are subsequently used for building structural comparative models of respective (SLC) transporter being captured in different functional states. Afterwards, large-scale docking screens into distinct structural models is performed to elucidate ligand binding events. Using the automatic workflow for generating binding mode hypotheses for a set of compounds with steroidal scaffold binding to three hepatic Organic Anion Transporting Polypeptides (OATP1B1, OATP1B3, and OATP2B1) delivered promising results which are currently under experimental evaluation. The protocol can be re-used for any target or off-target of pharmaceutical interest in the future.

## CINF 46

### Public database supporting evidence-based exposomics

**Risa R. Sayre**<sup>1,2,3</sup>, [sayre.risa@epa.gov](mailto:sayre.risa@epa.gov), **John Wambaugh**<sup>1</sup>, **Katherine Phillips**<sup>4</sup>, **Antony J. Williams**<sup>5</sup>, **Christopher Grulke**<sup>6</sup>. (1) National Center for Computational Toxicology, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, United States (2) Environmental Sciences & Engineering, University of North Carolina - Chapel Hill, Chapel Hill, North Carolina, United States (3) Oak Ridge Institute for Science and Education, Oak Ridge, Tennessee, United States (4) National Exposure Research Laboratory, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, United States (5) National Center for Computational Toxicology, Environmental Protection Agency, Wake Forest, North Carolina, United States (6) National Center of Computational Toxicology, US EPA, New Hill, North Carolina, United States

To support identification of likely sources of chemicals found in biological media through non-targeted/suspect screening mass spectrometry analysis, our project adds substance relationships between chemicals and their transformation products to the CompTox Chemicals Dashboard. We propose five categories for substances found in biomonitoring samples: (1) endogenous human metabolome, (2a) exogenous nutrients, (2b) markers of exposure to exogenous nutrients, (3a) xenobiotics, and (3b) markers of exposure to xenobiotics. Some compounds, such as formaldehyde, can appear in more than one category. Our effort is unique in that we include only *in vivo* empirically observed relationships, rather than products formed *in vitro*, in different species, or predicted based on pathways. Restriction to observable transformations will also include products formed by biotic, abiotic, or complex processes that have not yet been identified as relevant pathways for toxicology. We have curated a set of over 10,000 observations of 3a/3b chemical pairs and contextual metadata (such as analytical method) from databases and literature. Databases were filtered to only include empirical pairs and literature mappings were identified with natural language processing and manually verified. Description of experiments for improving the machine learning workflow to identify and extract this data will be the focus of this poster. Substance relationship mappings for curated substances are visible in the CompTox Chemicals Dashboard as a publicly available resource for top-down exposomics, in which biomarker analytes measured are used to infer association with exposure agents. These mappings also allow for the development of exposure estimates based on dose levels demonstrated to yield a detectable amount of product, and analysis of toxicity based on a greater number of relevant species.

## CINF 47

### Coupling the 1D, 2D, and 3D data worlds to facilitate drug discovery

**Daniel F. Ortwine**, [ortwine.daniel@gene.com](mailto:ortwine.daniel@gene.com). Genentech, South San Francisco, California, United States

The explosion of large databases of synthesizable molecules, searching technology, machine learning/AI capability, and high level scientific calculations such as free energy calculations is magnifying the challenge of organizing, analyzing, and presenting results in a unified fashion to inform drug design. Increases in compute power mean robust results can be delivered in near real time in many cases. Vendors are now stepping up to the plate to offer integrated data display solutions, but gaps remain. If desktop tools placed in the hands of medicinal chemists are to be routinely used, they must offer an intuitive interface that facilitates data exploration



with minimal software fatigue. We have adopted a strategy of placing larger scale calculations behind the scenes, callable from a single front end, then delivering results via 1D or 2D displays that are 'hot wired' directly to a 3D desktop modeling tool. Examples will be presented of the significant impact on therapeutic project progression the use of tools ranging from pKa calculations, 3D shape searching, docking, and QM-based torsion scans has had.

#### **CINF 48**

##### **Using knowledge graphs for prediction and visual hypothesis generation in drug discovery**

*David J. Wild, jm-acs@wild-ideas.org. Indiana Univ, Bloomington, Indiana, United States*

Building on prior developments in semantic and graph technologies, Knowledge Graphs (KGs) are increasingly being used to map together diverse, heterogeneous datasets in multiple domains. Algorithms based on KGs can be used to make predictions, and to extract patterns of insight. In this presentation, we will describe three kinds of knowledge graph that can be built (universal, domain and problem-specific), and will give examples of how link prediction and node embedding algorithms on top of domain and problem-specific graphs can be used as an alternate method of predicting drug-target activity (including off-target prediction) as well as identifying complex relationships between chemistry and biology that can be explored visually, and used to build hypotheses for biological effects of compounds, including off-target effects, adverse events, and phenotypic deconvolution.

#### **CINF 49**

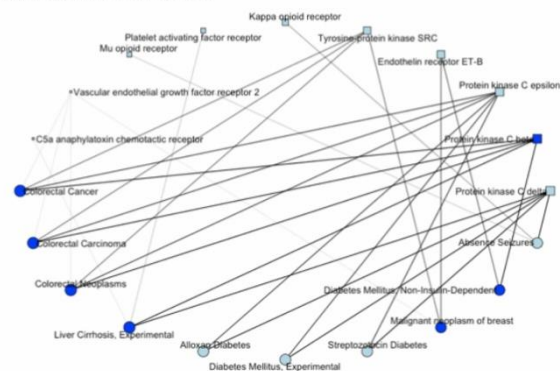
##### **Visualizing relationships between protein targets, GO annotations and diseases via dynamic network representations**

*Barbara Zdrazil<sup>1</sup>, barbara.zdrazil@univie.ac.at, Lars Richter<sup>1</sup>, Nathan Brown<sup>2</sup>. (1) Department of Pharmaceutical Chemistry, University of Vienna, Vienna, Austria (2) BenevolentAI, London, United Kingdom*

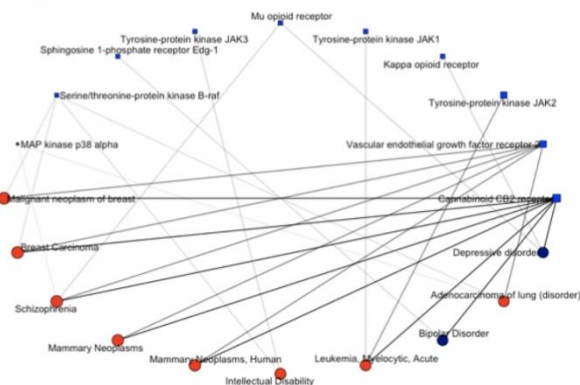
A deeper understanding of preclinical data can give crucial input to make firm decisions at earlier stages of the drug discovery pipeline since it is reflecting trends in research attention that the community is following over the years. It might help researchers to base their research endeavors on evidence-based criteria and direct their interest towards understudied proteins or core scaffolds. In this context, an emerging application of protein target indexing and characterization in drug discovery is the desire to capture the past, current, and future degree of research interest in main drug target families. Up to now, explorative studies on research attention across target families have been carried out mainly from a static point of view with only a few analyses including the time dimension.

In this study we were tracking compound-target associations over time in a target family-wise manner which delivered a picture of trends in research attention that the drug discovery community was following over time. Further, target innovation trends were linked to biological/therapeutic innovation patterns by linking targets to Gene Ontology (GO) annotations and disease annotations from DisGeNET. Inspecting the connectivity of protein targets with GO biological process annotations as well as diseases in network representations for different time periods revealed interesting trends in drug discovery. E.g. in cancer research, interest for some cancer types (e.g. prostate cancer) shifted from primarily research on targets of the EGFR family (mainly ErbB1) to research on Janus kinases, B-RAF and hepatocyte growth factor receptor. In addition, targets being involved in "immune system processes" have experienced a tremendous increasing attention over the last 20 years, both for kinases and GPCRs. Resultant dynamic networks visualizing associated target-disease annotations for different time periods, revealed a shifted focus from targets related to cancer and diabetes research to additional interest in targets associated with behavioral and neurodevelopmental diseases (such as schizophrenia and intellectual disability).

Immune System Process: 1995-2001



Immune System Process: 2009-2016



## CINF 50

### How a visual vocabulary defines what you see in your data

**Rajarshi Guha**, [rajarshi.guha@gmail.com](mailto:rajarshi.guha@gmail.com). Vertex Pharmaceuticals, Boston, Massachusetts, United States

Visualization aims to provide a specific view of your data, with the hope that such a view leads to some form of understanding or insight. However, a given data set can be visualized in multiple ways, with each way possibly highlighting different aspects. When a tool (or a method) provides a single view of the data, there is the possibility that you might end up with a skewed view of what the data represents and thus make erroneous or incomplete conclusions. This talk will highlight how it is important to examine data in multiple ways, using datasets from high throughput screening and chemical library design as examples. First, I will highlight some well known examples of visualization methods that can hide important details of the underlying data such as barcharts, and methods such as binning that can also hide important distributional details. Second I will describe how alternate representations can provide a different view of the data. In particular I will describe a network based representation of sets of molecules, that can be used to compare libraries. The use of such a representation allows us to consider library contents and compare contents of two libraries using network metrics, thus providing an alternative to traditional similarity based methods. I will conclude with a summary of best practices that helps one maximize their understanding of their data.

## CINF 51

### Molecular viz: I feel the need...the need for speed...and usability

**Jonas Boström**, [jonas.bostrom@astrazeneca.com](mailto:jonas.bostrom@astrazeneca.com). AstraZeneca, Mölndal, Sweden

In real-life everything is expected to be easy-to-use and happen instantaneous. Everything must flow. Think Google searches, text-messaging, video calling, social media feeds, etc. Drug discovery is starting to catch on, and more things are becoming automated, with computers doing much more for us than in the past. In this talk innovative computational approaches allowing new ways of working while making speedier progress will be discussed. Examples include Google-like ultrafast virtual screening capabilities, SAR analysis, Machine-learning and AI methods designing molecules with optimal properties as well as using immersive technologies for education and to gain a deeper understanding of the drugs we are trying to develop. A particular focus will be on my obsession on the need for speed and user-friendly molecular visualization.

## CINF 52

## Is virtual reality useful for visualizing and analyzing molecular structures?

**Thomas E. Ferrin**, *tef@cgl.ucsf.edu*. *Pharmaceutical Chemistry, University of California, San Francisco, California, United States*

Recent advances in instrumentation coupled with high-throughput methods have resulted in an unprecedented number of structural models available to researchers, including proteins, DNA/RNA, and their complexes with metals and small molecules. An indicator of this trend is the fact that the PDB archive surpassed 150,000 entries in 2019, compared to 10,000 entries in 2000. Moreover, there has also been significant growth in the size and complexity of these structural models, including the determination of many large multiprotein complexes. Together, this rapid expansion means increased challenges in our ability to interactively visualize and analyze structures of molecules, molecular assemblies, and protein sequence-structure relationships. Yet visualization and analysis are critical for addressing important and highly relevant biomedical problems such as identifying the molecular bases of disease, identifying targets for drug development, designing drugs, and engineering proteins with new functions. This invited presentation will focus on some of the current challenges and potential solutions in the visualization of molecules and molecular assemblies, and specifically how virtual reality, with its wide field of view, head tracking for better perception of molecular architectures, and 6-degree-of-freedom hand controllers for object manipulation is being used in drug binding studies and building accurate atomic models in electron microscopy and x-ray density maps.

### CINF 53

#### Chemical structure standardization and synonym filtering in PubChem

**Sunghwan Kim**, *kimsungh@ncbi.nlm.nih.gov*, *Paul Thiessen, Qingliang Li, Bo Yu, Evan Bolton*. *National Library of Medicine, National Institutes of Health, Bethesda, Maryland, United States*

PubChem (<https://pubchem.ncbi.nlm.nih.gov>) is a public chemical data repository that provides information on various chemical entities, including small molecules, siRNA, miRNA, peptides, lipids, carbohydrates, chemically modified biologics, etc. One of the most commonly requested tasks in PubChem is to search for a compound by chemical name (also commonly called “chemical synonym”). PubChem performs this task by looking up chemical synonym-structure associations provided by individual depositors to PubChem. These name-structure associations are used to create links between chemicals and Medical Subject Headings (MeSH) terms, which in turn are used to generate associations between chemicals and PubMed articles. The accuracy of these depositor-provided synonym-structure associations is dependent upon two important quality control methods used in PubChem: (1) chemical structure standardization and (2) synonym filtering based on crowd voting. In this presentation, we will discuss the two quality control methods and their effects on the chemical synonym-structure associations.

### CINF 54

#### Challenges in chemical registration system migrations, and how to deal with them

**Gerd Blanke**, *gerd.blanke@structurependium.com*. *StructurePendium Technologies GmbH, Essen, Germany*

Chemical registration systems have been in use in the biopharma, agro and chemical industries for many years in order to record a company's small molecule assets, determine their novelty, and ensure adherence to chemical business rules. However, many of these system are reaching the end of their life as they are not able to adequately cope with newer challenges such as a wider range of chemical modalities (e.g. large peptides, therapeutic antibodies, advanced materials), enhanced stereochemical representation, much higher volumes of data, collaborations, and integration of heterogeneous chemical data resulting out of company mergers and acquisitions.

Consequently, many companies are investigating a replacement of their legacy chemical registration system, or are already actively replacing it. When moving to the next generation chemical registration system correct migration of legacy chemical data poses a particular challenge, both in the sense of formal and technical

correctness, and in the 'semantic' sense of which actual compound is meant by a given compound record in the legacy system.

In our presentation we will discuss the specifics of common chemical database technologies employed in chemical registration systems as well as their differences, highlight the caveats this presents for chemical database and registration system migrations, share experiences gained from a variety of migration projects, and provide guidance how to overcome the difficulties.

## CINF 55

### **Making a hash of it: Advantage of selectively leaving out structural information**

*Noel O'Boyle, baoilleach@gmail.com, Roger A. Sayle. NextMove Software, Cambridge, United Kingdom*

One of the golden rules of cheminformatics is that tools should preserve chemical information exactly as provided by the user. Despite this, for some applications it may be useful to derive from the chemical structure a representation that discards or normalizes some of that information. We can refer to these as a molecular hash (or 'molhash') by analogy with hash functions from computer science. The use of molecular hashes presents an efficient route to solving many problems in cheminformatics which otherwise require exhaustive enumeration.

The most well-known molecular hash is the molecular formula, e.g. C<sub>9</sub>H<sub>8</sub>O<sub>4</sub> for aspirin. This is a representation of the structure that discards connectivity information and stereochemistry, but is sufficient to calculate the molecular weight. If particular rules are followed for the generation of the formula (e.g. by ordering the elements in Hill order) then the molecular formula is a convenient way to sort chemical structures so that they can easily be looked up; the Merck index has a formula index, for example. As the molecular formula ignores bond information, it is invariant to differences in bond representation common among inorganic complexes. Furthermore, by collating molecules based on their molecular formula, it is possible to identify isomers, which may be constitutional isomers or stereoisomers.

The example of the molecular formula illustrates that even the simplest of molecular hashes can be useful in identifying molecules that are structurally related in some way. This presentation will illustrate the application of molecular hashes to finding stereoisomers, alternate resonance forms, matched pairs, tautomers, regioisomers, redox pairs, mesomers, molecules that differ by a single-atom substitution, molecules with the same graph layout, and related peptides. We will discuss the concept of a hierarchical relationship between hashes, which leads to the concept of parent forms, and how calculation of slow hashes can be avoided by first checking parent forms.

## CINF 56

### **Crafting persistent identifiers and structure-based representations in DSSTox as surrogates for chemical names to better support interoperability in computational environments**

*Christopher Grulke, grulke.chris@epa.gov, Ann Richard, Antony J. Williams. National Center of Computational Toxicology, US EPA, New Hill, North Carolina, United States*

Nomenclature has been key to the conveyance of chemically associated information between scientists for over a century, as well as in the unstructured data environments that existed prior to the development of large-scale chemical databases. EPA's National Center of Computational Toxicology focuses on the collection and aggregation of hazard, exposure, and persistence data linked to chemicals to support environmental risk assessment. For this purpose, unique and intransient identifiers and structures provide more useful representations of chemical substances than human interpretable (also variable, error-prone and malleable) names. Unique DSSTox substance identifiers (DTXSIDs) and substance-list record identifiers (DTXRIDs) provide a simple way to separately manage chemical information associated with a particular sample or data source from a generic substance representation capable of aggregating data from many sources. When assignable, a chemical structure provides an unambiguous and information-rich representation of a substance that is universally interpretable by chemists and not subject to the permutations of chemical names. A unique DSSTox structure-identifier (DTXCID), in turn, provides an efficient indexing of a structure. Despite the value of such indexing, chemical names are, and will remain in widespread use as chemical currency by the public and

across scientific and regulatory domains. As such, names will continue to serve as primary linkages to source data and, thus, will continue to play an essential role in ensuring the accuracy of structure assignments, as well as for indexing in cases where structures cannot be assigned. As we extend our structure and substance storage methods to better document partially and ill-defined chemistry, the importance of the assigned name within our databases will likely wane, but it currently serves as the most heavily weighted source identifier when attempting to resolve a DTXRID to a DTXSID. *This abstract does not necessarily represent the views or policies of the U.S. Environmental Protection Agency.*

## CINF 57

### Classification of reactions by type or name

**Guenter Grethe**<sup>1</sup>, [ggrethe@att.net](mailto:ggrethe@att.net), **Josef Eiblmaier**<sup>2</sup>, **Hans Kraut**<sup>3</sup>, **Dagmar Kunzman**<sup>4</sup>, **Peter Loew**<sup>2</sup>. (1) Self-employed, Poway, California, United States (2) InfoChem, Muenchen, Germany

Chemists in research or development need access to large volumes of data to find the best answers to synthetic problems. Most of these are related to the preparation of new entities or to improving existing syntheses. Most frequently, searches over a wide range of reaction databases are carried out successfully using structure-based searches, but the formulation of an efficient query is sometimes difficult, particularly for novices. However, most chemists are very familiar with named reactions, such as Diels-Alder, Michael, etc. We will discuss a system in which a hierarchical index of name reactions is generated by a combination of RSS searches that are processed and added to a database. With over 700 well-known named reactions in the literature, the resulting taxonomy allows for simple text-based searches complimentary to structural searches. The taxonomy is based on seven main reaction mechanisms and can be extended and adopted for specific properties of chemical reactions like ring closures or multistep reactions. Large reaction data sets can be processed and clustered automatically for analysis in big data and machine learning projects.

## CINF 58

### UDM: Enabling exchange of comprehensive reaction information

**Frederik van den Broek**<sup>2</sup>, [f.broek@elsevier.com](mailto:f.broek@elsevier.com), **Gerd Blanke**<sup>1</sup>. (1) StructurePendium Technologies GmbH, Essen, Germany (2) Elsevier, Amsterdam, Netherlands

The first edition of the Beilstein Handbook of Organic Chemistry was published nearly 140 years ago. Electronic laboratory notebooks have been in use in chemistry for almost 20 years. And the life science industry still doesn't have a well-defined way of capturing and exchanging information about chemical reactions and relies on imprecise or vendor-specific data formats. Without a common language and structure to describe experiments, data integration is unnecessarily expensive and a significant part of published data has not been readily available for processing or analysis.

The Unified Data Model (UDM) project team aims to improve the situation. UDM is a collective effort of vendors and life science organizations to create an open, extendable and freely available reference model and data format for exchange of experimental information about compound synthesis and testing. Run under the umbrella of the Pistoia Alliance, the project team has published two releases of the UDM data format and it is expected that the model will continue to be improved as demand stipulates working with the Pistoia FAIR data implementation by industry community.

In our presentation we will discuss some specifics of the UDM and how the UDM is used to simplify the data exchange.

## CINF 59

### Reimagining IUPAC recommendations as a chemical ontology for semantic chemistry

**Stuart J. Chalk**, [schalk@unf.edu](mailto:schalk@unf.edu). Department of Chemistry, University of North Florida, Jacksonville, Florida,



## United States

A major activity of the IUPAC divisions is the development and publication of recommended terms in chemistry, through Pure and Applied Chemistry articles. Through the years this process has supported chemistry by formally defining concepts, creating a common language of chemistry - for humans. However, this rich knowledge base is currently not accessible to machines.

The digitization of recommended terms is therefore needed in order to represent chemical concepts for the semantic annotation of data. Recently, a new IUPAC project has been approved to move toward this goal. This project is the second phase of a multiphase project that started with the reinvention of the IUPAC Gold Book (<https://goldbook.iupac.org>). The IUPAC Gold Book (also called the Compendium of Chemical Terminology) is an aggregation of a small selection of terms (~7000) from IUPAC recommendations originally published in 1997 (hardback) and online since 2006. The new version of the Gold Book has made the terms available to machines via a REST API.

This talk will focus on the new project (Phase 2) to move the term definitions into a more formal ontological representation of knowledge in chemistry. A nominal ontological representation of recommended terms in chemistry will be shown and the process of migrating existing terms into this format will be discussed. Additionally, a prototype online system to allow divisions to manage this process will be presented.

### CINF 60

#### **Publishing FAIR spectral data and chemical structures: Report from the NSF workshop in Orlando**

**Leah R. McEwen**<sup>1</sup>, [lrn1@cornell.edu](mailto:lrn1@cornell.edu), **Vincent F. Scalfani**<sup>2</sup>, [vincent.scalfani@gmail.com](mailto:vincent.scalfani@gmail.com). (1) Clark Library, Cornell University, Ithaca, New York, United States (2) University Libraries, University of Alabama, Tuscaloosa, Alabama, United States

Spectral data and chemical structures are most often published as static figures embedded within publisher formatted documents (e.g. PDF). This practice has greatly limited both reuse and discovery of chemical data. The chemical community has recognized this limitation and, as a result, there is wide interest in enhanced sharing of spectral data and chemical structures in the form of machine-readable files. However, there is a lack of guidance for authors specifying how to create, describe, and package machine-readable chemical data alongside publication submissions. The NSF funded a workshop to bring together a wide variety of stakeholders in the chemistry community, including researchers, database providers, publishers, and librarians to map workflow models for preparing and publishing FAIR chemical spectral data and chemical structures. Workshop goals focused on formulating strategies based on what can be accomplished with current resources. Breakout discussions included proposed pilot workflows, metadata, value propositions for stakeholders, and community support and engagement. A full written report will be available in fall 2019 and will include summaries of the breakout discussion outcomes and use cases/value propositions to further engage the research community and other stakeholders.

### CINF 61

#### **SciWalker: Comprehensive ontology-based chemical search**

**Lutz Weber**<sup>1</sup>, [lutz.weber@ontochem.com](mailto:lutz.weber@ontochem.com), **Claudia Bobach**<sup>1</sup>, **Felix Berthelmann**<sup>1</sup>, **Timo Boehme**<sup>1</sup>, **Stephen Boyer**<sup>2</sup>, **Matthias Irmer**<sup>1</sup>, **Konstantin Kruse**<sup>1</sup>, **Ulf Laube**<sup>1</sup>, **Joachim Ludwig**<sup>1</sup>, **Anett Pueschel**<sup>1</sup>, **Christoph Ruttkies**<sup>1</sup>, **Ian Wetherbee**<sup>3</sup>. (1) OntoChem IT Solutions, Halle, Germany (2) Collabra Inc, San Jose, California, United States (3) Google, Mountain View, California, United States

SciWalker is a powerful new search resource that provides user-friendly access to massive amounts of scientific information. Its capabilities include accessing and analyzing public and private domain content by using ontologies to normalize knowledge concepts. SciWalker annotates and indexes large volumes of scientific and technical documents leveraging a registration

system that assigns ontology concept identifiers (OCID) that provide further value to the curated data. SciWalker integrates structure-based chemistry ontology with other ontologies resulting in improved associations and classifications of scientific content. The OCID compound and sequence registries, as well as other curated data and ontologies, are being made available publicly as Google BigQuery tables in the “SciWalker Open Data” project. This furthers the concept of an eChemistry counterpart of eScience with a computationally intensive environment carried out in the BigQuery highly distributed network. The grid computing and federated collaboration capabilities enabled by the combined power of SciWalker and Google BigQuery will allow researchers to harness the immense potential of publicly available databases.

## CINF 62

### **RDKit: Open-source cheminformatics from machine learning to chemical registration**

**Gregory Landrum**<sup>1,2</sup>, *greg.landrum@gmail.com*. (1) KNIME AG, Basel, Switzerland (2) T5 Informatics GmbH, Basel, BS, Switzerland

The RDKit is an open-source cheminformatics toolkit with a business-friendly license. The toolkit has been open source for more than 13 years, and provides a broad range of cheminformatic functionality ranging from descriptor calculation through molecule standardization and to chemical reactions. The toolkit is useable from the C++, Python, Java, and C# programming languages as well as from within the KNIME Analytics Platform and the PostgreSQL relational database system. It has been integrated into a number of other open-source software projects and multiple commercial software tools.

In this presentation I will provide an overview of the RDKit's capabilities, talk about its use in other open-source projects and commercial software, highlight some of the interesting functionality introduced in the most recent releases, and close with a personal perspective on the history of open-source software in our field and some of the challenges (and rewards!) of running an active open-source project.

## CINF 63

### **How the RCDK enables open source cheminformatics in R: From fingerprints to mass spectra**

**Rajarshi Guha**<sup>1</sup>, *rajarshi.guha@gmail.com*, Emma L. Schymanski<sup>2</sup>, Tobias Schulze<sup>3</sup>, Michael A. Stravs<sup>4</sup>. (1) Vertex Pharmaceuticals, Boston, Massachusetts, United States (2) Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Luxembourg, Luxembourg (3) Helmholtz Centre Env Res, Leipzig, Germany (4) Swiss Federal Institute of Science and Technology, Dübendorf, Switzerland

The CDK is a Java library for cheminformatics and is employed in a variety of applications. Given that many problems in cheminformatics involve analysis of collections of molecules, accessing cheminformatics in a statistical modelling environment enables flexible modelling and analysis of chemical information. R is an open source statistical modelling and programming environment and provides extensive support for data munging and machine learning, but does not support cheminformatics natively. As a result, a number of packages have been developed to provide such capabilities within the R environment. In this talk we discuss the rcdk, an R package that makes the CDK accessible from within R. After a brief discussion of the design of the rcdk [1], we briefly survey the ecosystem of packages that has grown up around rcdk to enable chemical data science within R. We then describe how these efforts have fed into larger, originally third-party R-based projects that require cheminformatics functionality, using the example of RMassBank [2]. Building on the rcdk functionality, the automated cheminformatic and spectral curation workflow has enabled the release of over 16,000 open mass spectra to the open spectral library MassBank [3] - which itself surfaces the structural information to the public using CDK libraries. The seamless integration between rcdk and the CDK enables consistent cheminformatics functionality for end users and developers alike. These once independent projects have now grown into a (spontaneous) geographically distributed, collaborative development between the CDK, rcdk, RMassBank and MassBank teams, driven mainly by communications via Github, showing that the ability to combine open source projects leads to products that are more than the sum of their parts.

## CINF 64

## Applying commonly overlooked corrections to DFT frequency calculations with GoodVibes

**Guilian Luchini**<sup>1</sup>, *Guilian.Luchini@colostate.edu*, **Robert S. Paton**<sup>1,2</sup>. (1) Chemistry, Colorado State University, Fort Collins, Colorado, United States (2) Chemistry Research Laboratory, University of Oxford, Oxford, United Kingdom

Developments in electronic structure theory have led to the widespread implementation and availability of approximate quantum chemical methods approaching accuracies within 1 kcal/mol of expensive ab initio calculations. These approaches are increasingly used to understand, validate and predict the outcomes of laboratory experiments. However, a direct comparison against experiment requires additional approximations to be made about the system under study such as the rigid-rotor harmonic oscillator (RRHO) description, whose accuracy can seriously undermine even the most elaborate computation. The effects of vibrational anharmonicity, multiple accessible conformations, molecular symmetry and the breakdown of the RRHO description collectively influence predictions of thermochemistry and kinetics.

We have developed an easy-to-use Python package that allows chemists to confidently process computational data obtained from DFT calculations. Corrections to thermodynamic values that otherwise would be unaccounted for are automated. These are universal considerations unrestricted to a particular software package. From the information provided by the Hessian matrix of second energy derivatives, thermodynamic values can be quickly obtained implementing a variety of scaling corrections to address limitations in the descriptions of low-frequency vibrations, at any temperature/pressure required. GoodVibes also has useful features such as creating output files of molecular Cartesian coordinates, graphing a reaction profile from relative energy values, or linking single point energy calculations to frequency calculations. This reliable and fast automated correction method is available on GitHub as an open source Python package and is readily installed and accessible to use by computational non-experts.

### CINF 65

## Analysis of the acid/base profile of natural products as starting points of epigenetic drug discovery

**Marisa G. Santibanez-Moran**<sup>1</sup>, *marisagsantibanez@gmail.com*, **J. Jesús Naveja**<sup>1,2</sup>, **B. Angélica Pilon-Jiménez**<sup>1</sup>, **Mariel P. Rico-Hidalgo**<sup>1</sup>, **David T. Manallack**<sup>3</sup>, **Jose L. Medina-Franco**<sup>1</sup>. (1) Department of pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Mexico City, Mexico (2) PECEM, Faculty of Medicine, Universidad Nacional Autónoma de México, Mexico City, Mexico (3) Monash Institute of Pharmaceutical Sciences, Monash University, Parkville, Victoria, Australia

The molecular acid/base profile has a significant influence on the ADMETox properties of molecules. Acidic and basic functional groups and their charge state are crucial in the target-molecule interactions as much as in pharmacokinetics, and toxicity. In this presentation, we will discuss the acid/base profile of molecular databases of natural products from different sources. The profile is compared to the acid/base properties of approved drugs, food chemicals, and semisynthetic compounds. Additionally, the profile of a molecular database with reported activity against epigenetic targets will be analyzed for finding overlaps with natural products and further advance the epigenetic drug discovery based on natural sources.

### CINF 66

## Pharos: Open-source target illumination platform

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

Pharos is a web resource that originated from the NIH "Illuminating the Druggable Genome" (IDG) program to characterize the understudied regions of the druggable genome. These so-called "dark" regions are important because they have potential druggable opportunities toward a wide range of disease areas. Part of the IDG mandate is to move from simply gathering and displaying target data to generating data about understudied



targets and therefore contribute to the illumination of dark targets. It is important that Pharos not only be able to display current data available, but to also be extendable and configurable to adapt to a variety of new data that is generated. The current version of Pharos collates datasets ranging from expression data to publications. The newest update of Pharos is a ground up UI rewrite designed to speed up the responsiveness of the site, as well as display newly generated data from our collaborators, ranging from data (imaging, phenotype, etc) to physical resources (knockout mice, cell lines, antibodies, etc). This update allows Pharos to be adapted to a wide range of data types, and positions Pharos to be a resource not only based on the dissemination of static target information, but to also produce resources and advance dark target knowledge which can then be utilized to further prioritize targets for drug discovery.

PHAROS Targets Diseases Ligands Topics API About FAQ search

Tchem **HTT** Huntingtin

May play a role in microtubule-mediated transport or vesicle function. Huntingtin is a disease gene linked to Huntington's disease, a neurodegenerative disorder characterized by loss of striatal neurons. This is thought to be caused by an expanded, unstable trinucleotide repeat in the huntingtin gene, which translates as a polyglutamine repeat in the protein product. A fairly broad range of trinucleotide repeats (9-35) has been identified in normal controls, and repeat numbers in excess of 40 have been described as pathological. The huntingtin locus is large, spanning 180 kb and consisting of 67 exons. The huntingtin gene is widely expressed and is required for normal development. It is expressed as 2 alternatively polyadenylated forms displaying different relative abundance in various fetal and adult tissues. The larger transcript is approximately 13.7 kb and is expressed predominantly in adult and fetal brain whereas the smaller transcript of approximately 10.3 kb is more widely expr ...[more](#)

Targets / Huntingtin

Protein Summary  
 IDG Development Level Summary  
 Associated Ligands  
 Disease Associations by Source  
 Target Expression Data  
 Protein to Protein Interactions  
 Publication Information  
 Sequence Details

**Protein Summary**

UniProt Accession IDs  
**P42858 Q9UQB7**

Gene Name  
**HTT**

Ensembl ID  
 ENST00000355072 ENSP00000347184  
 ENSG00000197386

Symbol  
 HD IT15 LOMARS

Illumination Graph

Knowledge Table

Most Knowledge About	Knowledge Value (0 to 1 scale)
biological process	1.00
hub protein	0.99
interacting protein	0.96
disease perturbation	0.95
biological term	0.92

IDG Development Level Summary

Representative target page with associated ligands.

## CINF 67

### Using open data, services, and source software to deliver the EPA CompTox Chemicals Dashboard

**Antony J. Williams<sup>1</sup>**, [tony27587@gmail.com](mailto:tony27587@gmail.com), **Christopher Grulke<sup>2</sup>**, **Kamel Mansouri<sup>3</sup>**, **Jeremy Dunne<sup>1</sup>**, **Jeff Edwards<sup>1</sup>**. (1) National Center for Computational Toxicology, Environmental Protection Agency, Wake Forest, North Carolina, United States (2) National Center of Computational Toxicology, US EPA, New Hill, North Carolina, United States (3) Integrated Laboratory Systems, Inc., Research Triangle Park, North Carolina, United States

The US EPA CompTox Chemicals Dashboard website provides access to various data types associated with ~900,000 chemical substances and supports the needs of the National Center for Computational Toxicology. The dashboard both consumes data, models and open source code from open as well as delivering data and services back to the community. The dashboard offers access to various types of aggregated and integrated chemistry, biology and toxicology data. The dashboard provides web-based access to data hosted in multiple databases, integrates to an underlying chemical registration system and utilizes both commercial and open QSAR/QSPR models to deliver predicted data for the chemicals. Some of the open source software that we utilize is used for InChI generation, for structure drawing and for real time prediction of both toxicity and

physicochemical endpoints. This presentation will provide an overview of the dashboard, review our usage of open source code to deliver the web application, and discuss our present and planned contributions to open science. *This abstract does not necessarily represent the views or policies of the U.S. Environmental Protection Agency.*

#### **CINF 68**

##### **Facilitating community-based chemical curation by providing an open source version of the DSSTox chemical and list registration software that supports the EPA CompTox Chemicals Dashboard**

**Christopher Grulke**, *grulke.chris@epa.gov*, Antony J. Williams, Amar Singh, Jeremy Dunne, Jeff Edwards, Ann Richard. National Center of Computational Toxicology, US EPA, New Hill, North Carolina, United States

The Distributed Structure Searchable Toxicity (DSSTox) database serves as the chemical substance foothold that allows for the collection, integration, and surfacing of data in US-EPA's CompTox Chemicals Dashboard (<https://comptox.epa.gov/dashboard>). Whereas the DSSTox project has always had as a primary focus the sharing of data linked to curated chemical information, the tools built to support our data-structure curation process have not been sufficiently documented and bug-free to provide to the community. As a result, this closed source, open data model has prevented others from being able to directly employ our processes for curating and managing their own sets of chemistry, thus inhibiting a larger community-based curation effort. Hence, we have launched a Chemical Registration Open Source project to provide a public version of the application that has supported DSSTox chemical registration for the past 4 years. This project is crafting a set of publicly available microservices for integrating cheminformatics support functions using Open Source toolkits. These components will provide the underpinning for a user interface to support curator tasks that will handle all aspects of structure normalization, substance mapping, conflict-resolution, substance registration and storage. Database population functions allowing a new user to initialize the database with publicly downloadable DSSTox content will be included. This first step in publishing an Open Source application that will allow users to incorporate and manage the DSSTox database and associated data is expected to be augmented with additional capabilities covering our other data domains being surfaced on the Comptox Chemicals Dashboard. The goal will be to enable efficient integration of data resources to facilitate the management and transfer of environmentally relevant data between stakeholders. This presentation will provide an overview of the Chemical Registration Open Source project and status update for delivering enhanced solutions to the community. *This abstract does not necessarily represent the views or policies of the U.S. Environmental Protection Agency.*

#### **CINF 69**

##### **Matched molecular pair (MMP) and matched molecular series (MMS) visualizations for drug discovery**

**Christopher Keefer**, *cekeefe@gmail.com*. Pfizer Inc., North Stonington, Connecticut, United States

The use of Matched Molecular Pairs has grown ever since the advent of algorithms for their fast computation for large screening databases. They are attractive to medicinal chemists since they are based on easy to understand structural changes. Matched Molecular Pair Analysis (MMPA) has been successfully utilized for both prediction and design idea generation. An extension of MMPA is MMSA where multiple series of MMPs are analyzed. One of the challenges in MMPA and MMSA is how to distill the vast amount of data and present the most salient information in a clear and concise manner that facilitates medicinal chemistry decision making. This talk will demonstrate several approaches we have taken to address the issue of MMP and MMS visualization. Examples include MMP data change summarization, contextualization, and drilldown information. It will also cover the visualization of MMS SAR Networks, the concept of cross MMS SAR Networks and their visualizations, and the mapping of MMP transforms for hot spot analysis.

#### **CINF 70**

##### **Using DOCK and ZINC to visualize ultra-large chemical libraries**

**John J. Irwin**<sup>1</sup>, *jjj@cgl.ucsf.edu*, Brian Shoichet<sup>2</sup>, Lyu Jiankun<sup>1</sup>, Trent E. Balius<sup>3</sup>, Roger A. Sayle<sup>4</sup>, Isha Singh<sup>1</sup>,

Anat Levit<sup>1</sup>, Yurii Moroz<sup>5</sup>, Matthew O'Meara<sup>1</sup>, Chinzorig Dandarchuluun<sup>1</sup>, Benjamin Wong<sup>1</sup>, Jennifer Young<sup>1</sup>, Khanh Tang<sup>1</sup>. (1) Pharmaceutical Chemistry, University of California San Francisco, San Rafael, California, United States (2) Univ of Calif San Fran, San Francisco, California, United States (3) Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, California, United States (4) NextMove Software, Cambridge, United Kingdom (5) Chem-Space, Riga, Latvia

By the time of this meeting, we expect ZINC will contain more than 2.5 billion commercially available, rule-of-4 (Ro4), biologically relevant compounds that can be purchased and tested within two months. With the help of commercial compound suppliers, we expect purchasable chemical space to continue to grow sharply for the foreseeable future. By summer 2020, we plan to have over 1 billion Ro4 3D molecules ready for large scale docking.

We are focusing on two approaches to access this new and growing chemical space. First, molecular docking is a pragmatic approach to use protein structure to discover new small molecules to modulate protein activity. Successful projects against over 100 targets have now been reported in the literature. Our recent work on AmpC beta-lactamase and Dopamine D4 (Lyu, Nature, 2019) showed that docking screens of large libraries can find novel chemotypes and compounds with high affinity.

A second approach to ligand discovery is analog-by-catalog. On average, compounds in ZINC have 20 close analogs, often allowing for rapid exploration of SAR around many hits when they are found. A key problem is being able to search the database in real time as the database grows. We have recently begun using Arthor and Smallworld by Nextmove Software to provide rapid public searching capabilities for ZINC.

Together, we use these two approaches to visualize and prioritize the parts of purchasable chemical space that are of interest for our target.

The scalability of our approaches to the planned future growth of purchasable chemical space will be discussed.

## CINF 71

### **Emerging AI and machine learning approaches for designing novel chemicals and materials with the desired properties**

Maria Popova<sup>1</sup>, Olexandr Isayev<sup>2</sup>, **Alexander Tropsha**<sup>1</sup>, alex\_tropsha@unc.edu. (1) Univ of North Carolina, Chapel Hill, North Carolina, United States (2) UNC Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States

We will present novel approaches to representing, visualizing, and analyzing chemicals and materials as well as insights from model interpretation to enable accelerated discovery of novel functional chemicals and materials. Recently, we and others have developed novel structural and electronic materials descriptors that can be used to visualize the materials space (materials cartogram) and identify regions of materials with specific compositions and properties [1], as well as develop Quantitative Materials Structure-Property Relationships (QMSPR) models. Using the data from the AFLOW repository of high-throughput *ab-initio* calculations for inorganic molecules, we have generated QMSPR models to predict critical material properties such as metal/insulator classification, bulk modulus, Fermi energy, and band gap energy [2]. As an experimental proof-of-concept, we have employed the QMSPR approach to identify a novel photocathode material for dye-sensitized solar cells (DSSCs) [3]. In parallel, we and others have developed a novel computational strategy based on deep and reinforcement learning techniques for de-novo design of molecules with desired properties. Our strategy integrates two deep neural networks – generative and predictive – that are trained separately but employed jointly, with an added component of reinforcement learning, to generate novel chemical structures with the desired properties. In the proof-of-concept study [4], we have employed this strategy (termed Reinforcement Learning for Structure Evolution, or ReLeaSE) to design chemical libraries biased toward compounds with specific ranges of physical properties, such as melting point and hydrophobicity, as well as to develop novel compounds selective against specific targets such as kinase inhibitors. I will discuss recent advances in the ReLeaSE technology including visualization of model outcomes and talk about opportunities for combining ReLeaSE with property filters and robotic chemistry to accelerate practical design and discovery of novel chemical entities with the desired properties.

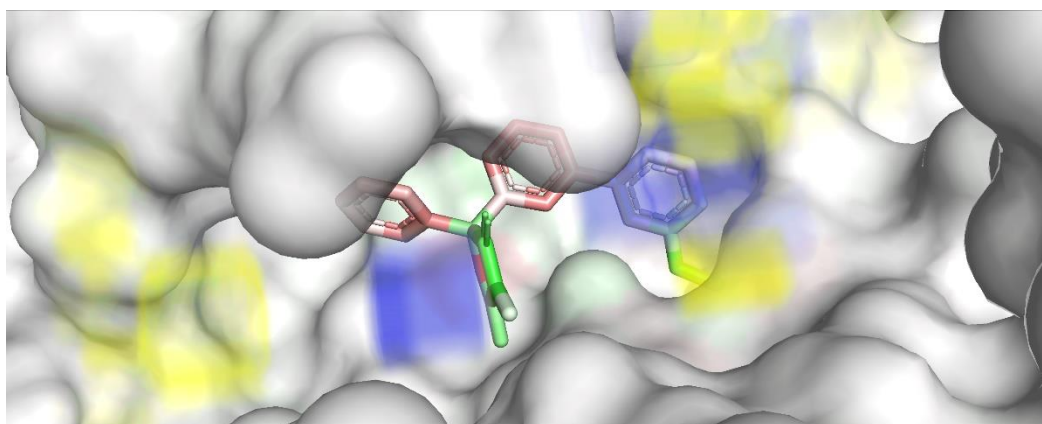
## CINF 72

### Visualizing structure-based deep learning scoring functions for protein-ligand interactions

**David Koes**, *dcoes@pitt.edu*. Computational and Systems Biology, University of Pittsburgh, Pittsburgh, Pennsylvania, United States

Protein-ligand scoring is an important step in a structure-based drug design pipeline since a correct binding pose and predicting the binding affinity of a protein-ligand complex enables effective virtual screening. Deep neural networks are a particularly promising approach for learning how to more effectively score protein-ligand interactions from the growing amount of cheminformatic and structural data.

We will describe several approaches for visualizing how to decompose grid-based convolutional neural network scoring functions into human interpretable visualizations. We will demonstrate how such visualizations can be used (or not) to improve the training of the network and potentially guide medicinal chemistry.



## CINF 73

### Cheminformatics-powered visualization methods of complex multidimensional SAR data

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

The amount of chemical biological data accessible to medicinal chemists and modelers is skyrocketing. In fact, the continuous generation of experimental data points via high throughput/content screening along with the rise of fully automated chemical synthesis platforms will even accelerate the data revolution in drug discovery. As a result, the actual integration, visualization, and modeling of such *Big Chemical Data* represent increasingly complex tasks to accomplish, even for experts in cheminformatics. Meanwhile, medicinal chemists are demanding easy-to-use tools that give them a direct access and visual representation to critical information to guide the rational molecular design. In particular, needed are methods and software for representing structure-activity relationships (SAR) for chemical series tested in multi-dimensional experimental assays. Herein, we recapitulate several methods and associated software allowing the visualization of complex SAR data. Notably, we will discuss: (1) the use of ligand- and structure-based graphs to represent complex associations between targets, chemicals, and experimental activities; (2) the complementarity of 2D/3D/MD descriptors to build ligand-based radial plots and circular dendrograms to rapidly establish SAR and solve activity cliffs; (3) the emergence of web-based services such as ChemMaps.com to easily browse and navigate the chemical space of drugs and drug candidates; (4) 3D printing to physically print complex protein-ligand or protein-protein interfaces; and (5)

the rapid rise of augmented/virtual reality to visualize SAR data in a fully immersive virtual environment. At last, we will discuss the future of SAR data visualization with some suggestions for next-generation methods and tools.

## CINF 74

### **Data visualization for compound library enhancement: Application of artificial intelligence algorithms from computer chess**

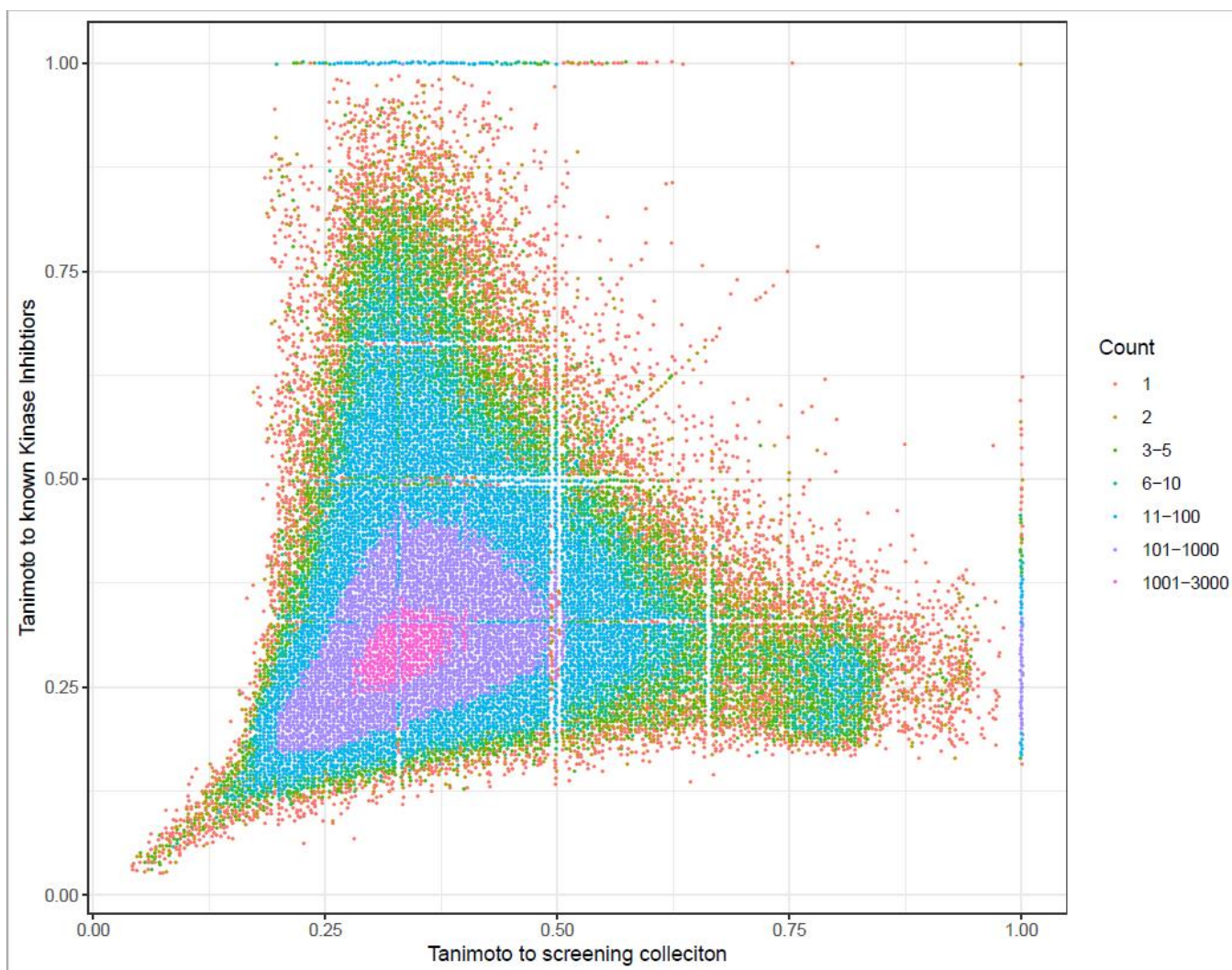
**Roger A. Sayle**<sup>1</sup>, [roger@nextmovesoftware.com](mailto:roger@nextmovesoftware.com), Noel O'Boyle<sup>1</sup>, Nicolas Zorn<sup>2</sup>, Roman Affentranger<sup>2</sup>. (1) NextMove Software, Cambridge, United Kingdom (2) Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd, Basel, Switzerland

Real-world diversity selection is complicated by the existence of the current screening collection of several million compounds (that gets depleted over time and may no longer be available or optimally desirable) and the desire to sample different regions of chemical space with different densities. For example as much of chemical space as possible should be covered by cheap compounds, and more expensive compounds only used to fill any remaining interstitial voids.

The MaxMin algorithm is frequently used in cheminformatics to pick diverse sets of compounds. We recently developed a significant improvement to MaxMin selection (contributed to RDKit) reducing the number of Tanimoto comparisons required, thereby significantly increasing the size of data sets that can be processed; A major source of this improvement is alpha-beta pruning, an AI technique more commonly encountered in chess and Go playing programs.

This talk describes diversity selection as a multi-objective optimization; without constraints diversity picking tends to initially select “wacky” molecules. Hence we formulate diversity selection as an operation over three compound sets; selecting from available compounds those that are most dissimilar to an existing in-house screening collection, but are maximally similar to a reference set of desirable compounds. The goal is not to select the most diverse compounds in the entirety of chemical space, but to sample from within a constrained “drug-like” space. This approach can be visualized as a scatter plot with novelty (similarity to the current collection) on the X-axis, and desirability (similarity to prototype ideal compounds) on the Y-axis. That alpha-beta pruning can be applied to one axis and not the other, leads to an interesting asymmetry, but enables the efficient identification of compounds on (or near) the Pareto frontier (i.e. those to be considered for purchasing) in a fraction of the computational effort previously required.





## CINF 75

### Reaction InChI (RInChI): Present and future

**Gerd Blanke**<sup>1</sup>, [gerd.blanke@structurependium.com](mailto:gerd.blanke@structurependium.com), **Jonathan M. Goodman**<sup>2</sup>, **Guenter Grethe**<sup>3</sup>, **Hans Kraut**<sup>4</sup>. (1) StructurePendium Technologies GmbH, Essen, Germany (2) Dept of Chemistry, Cambridge, United Kingdom (3) Self-employed, Alameda, California, United States (4) Infochem GmbH, Munich, Germany

The International Chemical Identifier for Reactions (RInChI) provides a vendor-neutral, machine-readable string representing chemical reactions. The prototype of RInChI was first released in 2011 and Version 1.00 has been downloadable since 2017. RInChI is starting to be used in databases and cheminformatics software packages, drawing tools are beginning to provide the calculations of RInChIs from reaction depictions and publishers are going ahead with integrating the RInChI into their web pages. We will discuss what has been achieved thus far and plans for the future. This will include adding data to auxiliary information layers, including the ability to assign atom-atom mapping information and chemical process details, such as reaction temperature and yield. Information like this can be used to optimize reaction pathways leading to high levels of automation of smart chemical syntheses. The planned extensions will incorporate future developments of the InChI standard.

## CINF 76

### Chemical mixtures: File format, open source tools, example data, and mixtures InChI derivative

**Alex Clark**<sup>1</sup>, [aclark.xyz@gmail.com](mailto:aclark.xyz@gmail.com), Philip Cheung<sup>1</sup>, Janice Darlington<sup>1</sup>, Leah R. McEwer<sup>2</sup>. (1) R&D, Collaborative Drug Discovery, Burlingame, California, United States (2) Clark Library, Cornell University, Ithaca, New York, United States

We will present our work on defining a common datastructure that captures information about mixtures of chemicals, which has been conspicuously absent from the repertoire of common file formats used for cheminformatics. Our "Mixfile" is an analog of the omnipresent "Molfile", and represents mixed substances in a hierarchical format that captures structure, name, quantity and other metadata in a way that conveniently represents order of mixing, and gracefully handles real world uncertainties. The Mixfile format has been designed in collaboration with the IUPAC working group to be used as a suitable input source for the Mixtures InChI (MInChI) notation. Having tools and data in Mixfile/MInChI form will open up mixture data to cheminformatics methods in the same way as Molfile/InChI has, allowing content to be accurately rendered, searched, linked and analyzed with all manner of derived algorithms.

One of the main problems with proposing a machine readable representation for mixtures is that all existing data is represented either as text or is stored in custom data formats that are not suitable for broad adoption. In order to bootstrap the data content problem we are developing a text extraction procedure for use on sources such as catalogs or organisation-wide inventories. We have tested this procedure by generating thousands of mixtures from publicly available catalog descriptions, and made the results available to the community. We have also created a graphical editor and manipulation/rendering libraries, which is also open source. The editor can be run as a desktop app or integrated into a web page. We will describe our progress toward integrating this new datatype, and the corresponding importing & editing tools, into the CDD Vault Electronic Lab Notebook.

## CINF 77

### Organometallics: InChIng forwards to better representations and happier chemists

**Ian Bruno**<sup>1</sup>, [bruno@ccdc.cam.ac.uk](mailto:bruno@ccdc.cam.ac.uk), Colin Batchelor<sup>4</sup>, Jonathan M. Goodman<sup>2</sup>, Gerd Blanke<sup>3</sup>. (1) Cambridge Crystallographic Data Centre, Cambridge, United Kingdom (2) Dept of Chemistry, Cambridge, United Kingdom (3) StructurePendium Technologies GmbH, Essen, Germany (4) Royal Society of Chemistry, Cambridge, United Kingdom

The standard InChI does not currently provide a very effective description of organometallic compounds, and this is restricting its use. Key user groups for the InChI include synthetic organic chemists, who may be disappointed to discover that widely-used chemicals, such as Grignard reagents and the Grubbs catalyst, can only be described in a rather approximate way and that InChI is not an ideal identifier for these important molecules. Furthermore, over half of the compounds represented in resources such as the Cambridge Structural Database (CSD) contain a metal and as such cannot be reliably represented by an InChI. This limits the extent to which the CSD can be interlinked with other resources of chemical data. The InChI Trust is exploring ways to satisfy this user requirement, and has put out an RFP for the implementation of an extra InChI layer which will make the InChI even more powerful and generally applicable. We will outline the technical strategy for doing this, the steps which will need to be taken building on this first step, and the testing and validation strategy, which will be focussed on the organometallic compounds held in the CSD. Whilst it is likely that edge-cases will remain, an organometallic InChI protocol which satisfies the majority of practical use cases will be an important step forward for the InChI.

## CINF 78

### Names for structural variability: Alkanes from maximum efficiency to the limits of existence

**Jonathan M. Goodman**, [jmg11@cam.ac.uk](mailto:jmg11@cam.ac.uk). Dept of Chemistry, Cambridge, United Kingdom



How can variable structures best be encoded in an InChI, so that the data is compact, useful and canonical? The general question is a very challenging one, but answers for restricted groups of molecules are more accessible and provide stepping-stones towards a full analysis. A process for generating compact, canonical identifiers for groups of alkanes ( $C_nH_{2n+2}$ ) has been developed and tools to investigate their properties are being prototyped. For short lists of molecules, this representation provides little advantage over enumeration. For longer lists, this may provide a rapid, efficient and compact way of handling groups of structures. Is a molecule contained within a canonical list? What molecules do two lists have in common? How can two lists be combined into a new canonical list? Answers to these questions, even for subsets of molecular space, provide a step towards compact, canonical, InChI-based representations of Markush structures.

## CINF 79

### IUPAC SMILES+ specification: Proposed community effort to advance interoperability of the SMILES chemical structure representation

**Vincent F. Scalfani**<sup>1</sup>, [vincent.scalfani@gmail.com](mailto:vincent.scalfani@gmail.com), Leah R. McEwen<sup>2</sup>, Christopher Grulke<sup>3</sup>, Evan Bolton<sup>4</sup>, Gregory Landrum<sup>5</sup>, Helen Cooke<sup>12</sup>, Issaku Yamada<sup>6</sup>, John J. Irwin<sup>7</sup>, Jose L. Medina-Franco<sup>8</sup>, Miguel Q. Olozába<sup>9</sup>, Oliver Koehler<sup>10</sup>, Susan Richardson<sup>11</sup>. (1) University Libraries, University of Alabama, Tuscaloosa, Alabama, United States (2) Clark Library, Cornell University, Ithaca, New York, United States (3) National Center of Computational Toxicology, US EPA, New Hill, North Carolina, United States (4) National Center for Biotechnology Information, Bethesda, Maryland, United States (5) T5 Informatics GmbH, Basel, Switzerland (6) The Noguchi Institute, Tokyo, Japan (7) Pharmaceutical Chemistry, University of California San Francisco, San Rafael, California, United States (8) Universidad Nacional Autónoma de México, Scottsdale, Arizona, United States (9) Universidad de Granada, Granada, Spain (10) German National Library of Science and Technology, Hannover, Germany (11) Royal Society of Chemistry, Cambridge, United Kingdom (12) Nantwich Museum, Cheshire, United Kingdom

The IUPAC International Chemical Identifier (InChI) and Simplified Molecular Input Line Entry System (SMILES) are the two most important and commonly used line notations today. The InChI is a chemical identifier, while SMILES is chemical representation format. SMILES and InChI are complementary to each other and serve different use-cases within the chemical information and cheminformatics communities. InChI is well-documented and standardized through IUPAC. In contrast, there is no up-to-date specification documentation for SMILES, and this has led to interoperability issues between cheminformatics toolkits, greatly affecting the accurate exchange of chemical information globally. This presentation will discuss the current status of SMILES documentation, interoperability, and our efforts to establish an IUPAC SMILES+ specification and community forum for ongoing SMILES development. *This abstract does not necessarily represent the views or policies of the U.S. Environmental Protection Agency.*

## CINF 80

### InChI open education resource (OER)

**Robert E. Belford**<sup>1</sup>, [rebelford@ualr.edu](mailto:rebelford@ualr.edu), Ehren C. Bucholtz<sup>2</sup>, Steven P. Wathen<sup>3</sup>, Martin A. Walker<sup>4</sup>, Jordi Cuadros<sup>5</sup>, Tanya Gupta<sup>6</sup>, Nathan Brown<sup>7</sup>, Vincent F. Scalfani<sup>8</sup>. (1) Univ of Arkansas at Little Rck, Little Rock, Arkansas, United States (2) Basic Sciences, St. Louis College of Pharmacy, St. Louis, Missouri, United States (3) Chemistry, Siena Heights University, Adrian, Michigan, United States (4) SUNY Potsdam, Potsdam, New York, United States (5) Departament Estadística Aplicada, IQS, Barcelona, Spain (6) Chemistry, South Dakota State University, Brookings, South Dakota, United States (7) The Institute of Cancer Research, Sutton, United Kingdom (8) University Libraries, University of Alabama, Tuscaloosa, Alabama, United States

This presentation will describe an Open Education Resource (OER) designed to help chemists, educators and students find information on, and applications of, the IUPAC International Chemical Identifier (InChI). InChI is a machine readable semantic identifier based on layered line notation for the representation of chemical structures that was developed by IUPAC and NIST. InChI is an open and freely available identifier, with the standard InChI and its correlative hashed key arguably being the new nomenclature of the digital area, enabling a wide variety of 21<sup>st</sup> century semantic web applications and activities that should be utilized in chemical education and the practice of chemistry.

Usage of the World Wide Web has become ubiquitous by practicing chemists in the pursuit of science, and yet few take advantage of semantic features that advances like InChI enable, instead navigating the web the same way they would a book, browsing from one webpage to another. In 2017 the InChI Trust initiated a working group to tackle issues related to the adoption of InChI by the greater practicing community of chemists, and the InChI OER is the first project of this working group.

Using the InChI OER, chemists, educators and students can find resources on InChI, including downloadable classroom and reference materials such as modifiable documents and spreadsheets. This presentation will describe the InChI Open Education Resource designed to help chemists, educators and students find information about and applications based on the InChI standard.

## CINF 81

### Keeping up the momentum: Brief report from the InChI San Diego workshop

**Raymond J. Boucher**<sup>1,6</sup>, [rboucher@wiley.com](mailto:rboucher@wiley.com), **Richard Kidd**<sup>2,6</sup>, [kiddr@rsc.org](mailto:kiddr@rsc.org), **Ian Bruno**<sup>3,6</sup>, **Stephen R. Heller**<sup>4,6,7</sup>, **Leah R. McEwen**<sup>5,7,6</sup>. (1) John Wiley and Sons Ltd, Chichester, United Kingdom (2) Royal Soc of Chem T Graham Hse, Cambridge, United Kingdom (3) Cambridge Crystallographic Data Centre, Cambridge, United Kingdom (4) Retired, Silver Spring, Maryland, United States (5) Clark Library, Cornell University, Ithaca, New York, United States (6) InChI Trust, Cambridge, United Kingdom (7) International Union of Pure and Applied Chemistry (IUPAC), Research Triangle Park, North Carolina, United States

We will report on the InChI San Diego workshop, adding issues of broad community interest, and discuss future activities supported by IUPAC and the InChI Trust.

## CINF 82

### Fast and accurate interatomic potential models by genetic programming

**Alberto Hernandez**, **Adarsh Balasubramanian**, **Fenglin Yuan**, **Simon Mason**, **Tim Mueller**, [tmueller@jhu.edu](mailto:tmueller@jhu.edu). Johns Hopkins University, Baltimore, Maryland, United States

In recent years there has been great progress in the use of machine learning algorithms to develop interatomic potential models. Potential models developed using machine learning are typically orders of magnitude faster than density functional theory but also orders of magnitude slower than physics-derived models such as the embedded atom method. We demonstrate that machine learning, in the form of genetic programming, can be used to develop accurate and transferable many-body potential models that are as fast as the embedded atom method, making them suitable to model materials on extreme time and length scales. The key to our approach is to explore a hypothesis space of models based on fundamental physical principles and to select models from this hypothesis space based on their accuracy, speed, and simplicity. We demonstrate our approach by developing fast and accurate interatomic potential models for copper that generalize well to properties they were not trained on. Our approach requires relatively small sets of training data, making it possible to generate training data using highly accurate methods at a reasonable computational cost.

## CINF 83

### Accelerating design of inorganic materials with machine learning and AI

**Olexandr Isayev**, [olexandr@olexandrisayev.com](mailto:olexandr@olexandrisayev.com). UNC Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States

Historically, materials discovery is driven by a laborious trial-and-error process. The growth of materials databases and emerging informatics approaches finally offer the opportunity to transform this practice into data- and knowledge-driven rational design—accelerating discovery of novel materials exhibiting desired properties. The Materials Genome Initiative (MGI) has transformed Materials Science into a data-rich discipline. These developments open exciting opportunities for knowledge discovery in materials databases using informatics

approaches to inform the rational design of novel materials with the desired physical and chemical properties. Statistical and data mining approaches have been successfully employed in both chemistry and biology leading to the development of cheminformatics and bioinformatics, respectively. However, until recently their application in materials science has been limited due to the lack of sufficient body of data, methods and reliable infrastructure.

In this work we showcase a pilot materials informatics platform capable of (i) instantaneously query and retrieve the necessary material information in the form of web app and RESTful API; (ii) identify, visualize and study important data patterns, and (iii) generate experimentally-testable hypotheses by building predictive Machine Learning (ML) models based on materials fingerprints and descriptors. Our computational approach relies on cheminformatics methodologies that one of our groups has developed and employed successfully to enable rational design of organic compounds with desired properties (e.g., drug candidates). By using data from the AFLOW repository ([www.aflow.org](http://www.aflow.org)) for high-throughput ab-initio calculations, we have generated ML models to predict many critical material properties like superconductivity, Debye temperature, Seebeck coefficient, bulk modulus, and band gap energy.

#### **CINF 84**

##### **Deep learning from crystallographic representations of periodic systems**

*Phillip M. Maffettone*, [phillip.maffettone@liverpool.ac.uk](mailto:phillip.maffettone@liverpool.ac.uk), *Andrew I. Cooper*. Chemistry, University of Liverpool, Liverpool, United Kingdom

While significant advances have been made in accelerating chemical discovery with machine learning, the use of these methods for crystalline materials usually requires restraints on the input structure or manually constructed feature vectors. The arbitrary size and periodicity of crystalline systems pose challenges as these systems need to be represented with a fixed dimensionality and retain translational, rotational, and permutation invariance. We present the use of crystallographically inspired transformations that are amenable to deep learning algorithms. By calculating a modified structure factor, a suite of representations are developed that are of fixed size irrelevant of the input dimensionality. The ability for these representations to capture the periodic structural information is demonstrated through classification and regression problems related to crystalline materials. This approach has direct impact in crystal structure prediction and the development of energy-structure-function maps for autonomous discovery.

#### **CINF 85**

##### **Application of machine learning tools for the analysis of combinatorial libraries of all metal-oxides photovoltaic cells**

*Hanoach Senderowitz*<sup>1</sup>, [hsenderowitz@gmail.com](mailto:hsenderowitz@gmail.com), *Abraham Yosipof*<sup>2</sup>, *Omer Kaspi*<sup>1</sup>. (1) Chemistry, Bar Ilan University, Ramat Gan, Israel (2) Information Systems, College of Law & Business, Ramat Gan, Israel

Growth in energy demands, coupled with the need for clean energy, are likely to make solar cells an important part of future energy resources. In particular, cells entirely made of metal oxides (MOs) have the potential to provide clean and affordable energy if their power conversion efficiencies are improved. Such improvements require the development of new MOs which could benefit from combining combinatorial material sciences for producing solar cells libraries with machine learning approaches to analyze the resulting libraries and direct synthesis efforts. In this work we present the application of several machine learning tools to the analysis of multiple MO-based solar cell libraries. First we present several dimensionality reduction methods for the visualization of the photovoltaic (PV) space and discuss their performance in terms of their ability to segregate libraries made of different MOs. Next we present a unified workflow for the derivation of predictive QSAR models for key PV properties including short circuit current, open circuit voltage, and the internal quantum efficiency. The workflow is composed of several components including library characterization, library visualization, removal of outliers and model derivation using several algorithms such as  $k$  nearest neighbors ( $k$ NN), genetic programming (GA) and RANSAC. Our results demonstrate that QSAR models with good prediction statistics could be developed and that these models highlight important factors affecting these properties in accord with experimental findings. The resulting models are therefore suitable for designing better

solar cells. Finally we highlight the importance of collaborating with experimentalists to provide physics/chemistry based insight to the observed trends and to capitalize on the results.

## CINF 86

### **Database of low-energy cluster structures for atomically precise nanoclusters across the periodic table calculated using density functional theory**

**Peter Lile**<sup>1</sup>, *plile1@jhu.edu*, **Tim Mueller**<sup>2</sup>. (1) *Materials Science and Engineering, Johns Hopkins University, Baltimore, Maryland, United States* (2) *Johns Hopkins University, Baltimore, Maryland, United States*

The chemical and structural properties of atomically precise nanoclusters are of great interest in numerous applications, but the structures of the clusters can be difficult to predict. In this work, we present the largest database of atomic structures determined using ab-initio methods to date. We report the methodology used to rapidly evaluate the energies and relaxed structures for over 16,000 low energy cluster structures across 55 elements using density functional theory. The database has been validated against previous computational and experimental data where available. Patterns in the data reveal insights into the chemical and structural relationships among the elements at the nanocluster scale. We describe how the database can be accessed for future studies and design of nanocluster materials.

## CINF 87

### **Self-assembly of metal-organic frameworks**

**Yamil J. Colon**<sup>1</sup>, *ycolon@u.northwestern.edu*, **Ashley Guo**<sup>2</sup>, **Lucas W. Antony**<sup>2</sup>, **Kyle Hoffmann**<sup>2</sup>, **Juan J. De Pablo**<sup>2</sup>. (1) *Chemical and Biomolecular Engineering, University of Notre Dame, Notre Dame, Indiana, United States* (2) *Institute for Molecular Engineering, University of Chicago, Chicago, Illinois, United States*

Algorithms for the large-scale generation and high-throughput screening of metal-organic frameworks (MOFs) have been instrumental in the computationally-guided discovery and design of these materials, especially for gas storage and separations. These efforts have also highlighted an important issue: the synthetic feasibility of promising materials that are computationally discovered. Modeling the self-assembly can aid in answering that question. For this study, we focus on the self-assembly of MOF-5.

We employ a combination of enhanced sampling techniques to study MOF self-assembly for different scenarios and system sizes. We model a MOF-5 nanoparticle of either a single unit cell or four unit cells in size, in explicit solvent. We perform replica exchange with solute scaling (REST2) simulations and calculate the free energy profile using the average distance and average angle between nodes as collective variables. The free energy minima observed in these systems are close to the values in the experimental unit cell, therefore validating that the chosen model represents at least a metastable state.

To model self-assembly, we perform finite temperature string (FTS) method calculations, using the average distance and angle between nodes and the total coordination between nodes and linkers as collective variables. The process was modeled for both single and four unit cells systems from two starting points: fully disassembled and amorphous. The self-assembly starting from the disassembled state was found to be downhill in free energy for both the single unit cell and four unit cells systems. In contrast, the self-assembly starting from the amorphous state contains free energy barriers on the way to the final MOF structure. Finally, we discuss the mechanism of self-assembly of MOF-5 for the scenarios and system sizes considered.

## CINF 88

### **Accelerated discovery of high-refractive-index polyimides via *First-Principles* materials modeling and informatics**

**Johannes Hachmann**, *hachmann@buffalo.edu*. *Dept of Chemical and Biological Engineering, University at Buffalo, SUNY, Buffalo, New York, United States*

We present a high-throughput computational study to identify novel polyimides (PIs) with exceptional refractive

index (RI) values for use as optic or optoelectronic materials. Our study utilizes an RI prediction protocol based on a combination of first-principles and data modeling developed in previous work, which we employ on a large-scale PI candidate library generated with the *ChemLG* code. We deploy the virtual screening software *ChemHTPS* to automate the assessment of this extensive pool of PI structures in order to determine the performance potential of each candidate. This rapid and efficient approach yields a number of highly promising leads compounds. Using the data mining and machine learning program package *ChemML*, we perform a materials informatics analysis of the top candidates, e.g., with respect to prevalent structural features and feature combinations that distinguish them from less promising ones. In particular, we explore the utility of various strategies that introduce highly polarizable moieties into the PI backbone to increase its RI yield. The derived insights provide a foundation for rational and targeted design that goes beyond traditional trial-and-error searches.

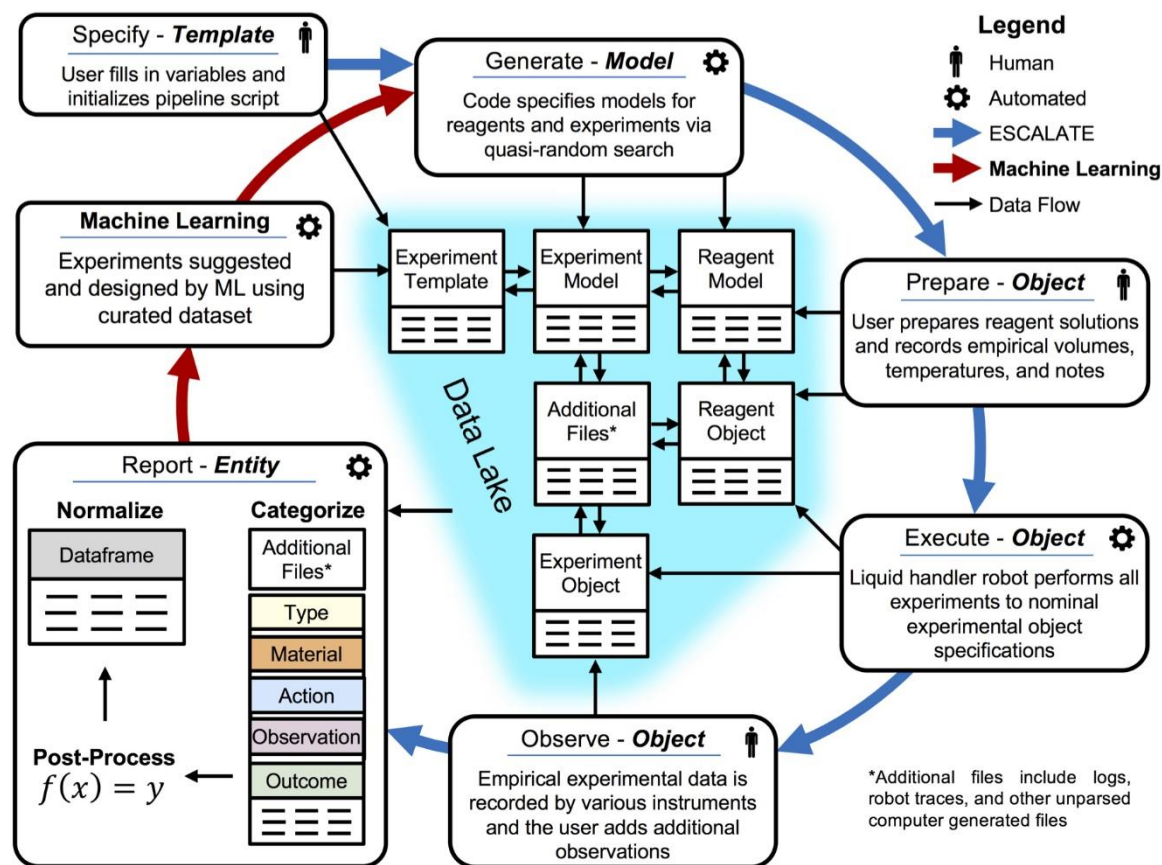
## CINF 89

### **Experiment specification, capture and laboratory automation technology (ESCALATE): Software pipeline for automated chemical experimentation and data management, with application to metal halide perovskite discovery**

*Joshua Schrier, jschrier@fordham.edu. Chemistry, Fordham University, The Bronx, New York, United States*

Applying artificial intelligence to materials research requires abundant curated experimental data and the ability for algorithms to request new experiments. **ESCALATE** (Experiment Specification, Capture and Laboratory Automation Technology) is an ontological framework and open-source software package that solves this problem by providing an abstraction layer for human- and machine-readable experiment specification, comprehensive and extensible (meta-)data capture, and structured data reporting. **ESCALATE** simplifies the initial data collection process, and its reporting and experiment generation mechanisms simplify machine learning integration. In this way it helps turn automated and semi-automated laboratory experimentation into a subroutine that can be called by external programmers, filling an unmet need that is between that of traditional electronic laboratory notebook (ELN) software and more heavy-duty automation solutions. In this talk, I will discuss "lessons learned" from an initial **ESCALATE** implementation for metal halide perovskite discovery, which has been used to perform thousands of algorithmically-controlled experiments at multiple laboratory sites, with participants across the country.





Schematic of the ESCALATE workflow.

## CINF 90

### Standardization of structural representation of polymers used in medicinal products

**Yulia Borodina**<sup>1</sup>, [yulia.borodina@fda.hhs.gov](mailto:yulia.borodina@fda.hhs.gov), **Igor Filippov**<sup>2</sup>, **Tyler Peryea**<sup>3</sup>, **Yuri Pevzner**<sup>3</sup>. (1) FDA, Silver Spring, Maryland, United States (2) VIF Innovations, LLC, Rockville, Maryland, United States (3) Chickasaw Nation Industries, Rockville, Maryland, United States

The US FDA registers ingredients in medicinal products and assigns Unique INgredient Identifiers (UNIs) that must be used by companies marketing their products in the US. Many inactive ingredients (excipients) receiving UNIs are polydisperse materials such as polymers. Despite the registration of polymers complying with the ISO 11238 standard, multiple alternative representations are possible because of the generality of the standard on the one hand and the lack of chemoinformatical standards for description of polydisperse materials on the other. Here we describe an approach for standardization of the structural representation of polymers for data identification and exchange purposes. The approach is based on the InChI algorithm and is used to populate substance indexing files that accompany product labels published by FDA on the DailyMed website.

## CINF 91

### Monitoring progress in lead optimization

**Jürgen Bajorath**<sup>1,2</sup>, [bajorath@bit.uni-bonn.de](mailto:bajorath@bit.uni-bonn.de). (1) Life Science Informatics, University of Bonn, B-IT, Bonn, Germany (2) Biological Structure, University of Washington, Seattle, Washington, United States

It is generally difficult to estimate the odds of success in hit-to-lead and lead optimization projects. In medicinal chemistry, decisions to continue or discontinue individual series are typically made on the basis of subjective judgment. A computational method termed Compound Optimization MOnitor (COMO) is presented that helps to determine if further optimization progress can be expected for a given series or if sufficient numbers of analogs have been generated. The approach integrates the assessment of chemical saturation of analog series and SAR progression. Chemical space distributions of existing analogs and virtual candidates are compared and neighborhoods of analogs are systematically analyzed. Chemical saturation and SAR progression scores are calculated, which make it possible to predict the potential of series for further chemical optimization. COMO also includes a compound design component. Virtual analogs generated for chemical saturation analysis provide a pool of candidates for synthesis. Machine learning enables the prediction of active analogs to further advance series prioritized by COMO. The methodology is easily expandable to include multiple optimization relevant properties.

## CINF 92

### Electronic-structure informatics using 3D descriptors of molecules

*Manabu Sugimoto, sugimoto@kumamoto-u.ac.jp. Faculty of Advanced Science and Technology, Kumamoto University, Kumamoto, Japan*

In this presentation, we will report our newly developed quantum chemical descriptors of molecules and their applications to cheminformatics. The new descriptors are designed to describe three-dimensional features obtained in quantum chemistry calculations.

As is well known, chemical properties of molecules are tightly related to their electronic structures. This fact justifies the cheminformatics based on quantum chemistry calculations. The present author and his collaborators have been challenging to define and apply his own electronic descriptors through considerations on mechanistic aspects of the chemical phenomena and the related physical modellings. So far, most of the descriptors that we have been applying correspond to spectroscopic features of molecules. This set of descriptors seems to have limitation or weakness in describing three-dimensional features related to molecular recognition. Herein we will suggest three dimensional descriptors representing topological features of interaction energy surfaces and molecular orbitals. We will also suggest coarse-grained descriptions for three-dimensional features of molecules for efficient cheminformatics modeling.

## CINF 93

### Fast evaluation of potential synthesis routes using DFT calculations on the basis of Transition State Data base (TSDB)

*Kenji Hori, kenji2969@gmail.com. Materials Engineering, Yamaguchi University, UBe, Yamaguchi, Japan*

There is a CREST project of JST which consists of four groups; the first makes a very large scale library (VLSVL) of drug candidate molecules; the second creates a prediction model which screens candidates in VLSVL and picks up potential molecules; the third is concerning with chemical process and the last is our project. We are constructing a data base, called TSDB and QMRDB, which are used for analyzing reaction mechanisms to synthesize many candidate molecules in a short time. It is because the existence of transition states is the key for the reaction to proceed. For this purpose, we developed a cloud system managing the data bases as well as theoretical calculations. Two programs were also created; one is an interface between the cloud system and windows terminals and the other makes input files for Gaussian calculations based on search results of TSDB. In the present talk, we will show the summary of the TSDB system and some results of reaction mechanism analyses for synthesizing drug candidate molecules for inhibiting the PME-1 protein.





## CINF 94

### Development using materials informatics in Japanese companies

**Yukihiro Uchi**, *uchi.yb@om.asahi-kasei.co.jp*. ASAHI KASEI CORPORATION, Chiyoda-ku, Tokyo, Japan

The impact of the Material Genome Initiative, which began in the United States in 2011, on Japan is significant, and similar efforts have been taken in Japan. Materials Informatics is one of the leading initiatives, and the Japanese industry has just begun to include Materials Informatics in product development.

Materials Informatics is perceived as a new approach, but in Japan, Prof. Funatsu has long been working on the application of chemoinformatics methods to material design and chemical process control, so the foundation of materials informatics was already done.

Asahi Kasei Co. is a comprehensive chemical manufacturer in Japan engaged in businesses such as chemicals, textiles, housing, building materials, electronics, pharmaceuticals, and medical care. Asahi Kasei Co. has also received some advice from Prof. Funatsu and has used the results of his research.

In this presentation, we will introduce some of the material development and product development efforts using informatics technology. In materials development, for example, 10 types have been selected from 80 types of raw materials, which have been advanced based on human intuition and experience so far, and there are innumerable combinations to determine the ratio of their amounts, and in the past. We have found the possibility to reduce the number of trials and errors by using informatics technology.

## CINF 95

### Prediction and control of vacuum deposition process by data-driven method

**Yuya Takeda**, *takedakinako@yahoo.co.jp*, **Yoichi Zushi**, *Yoichi.Zushi@kaneka.co.jp*, Takahiro Ogushi, Eiji Kuribe. Kaneka.co, Settu-city, Osaka-Pref, Japan

In Manufacturing processes, there are some process variables which are difficult to be measured. Soft sensor was developed in order to predict behavior of process variables and to control overall process. In thin-film PV (photovoltaic), layers are deposited by CVD (Chemical Vapor Deposition) process on the glass. There are some relations between deposition temperature of glass and efficiency of PV, but the temperature of glass is difficult to be measured directly in vacuum process. To predict the temperature of glass and to control the efficiency, we developed a soft sensor in CVD process. In OLED (Organic Light Emitting Diode) lighting device, there are quality specifications such as brightness, driving voltage and color etc. But they have a trade-off relationship, so it is difficult to decide operating condition. To control and improve the quality of OLED lighting devices, we

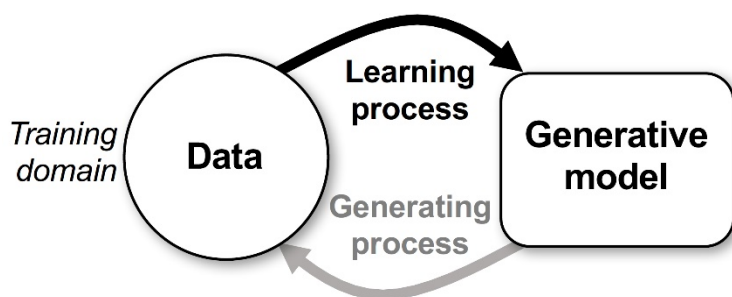
developed the method to decide operating condition.

## CINF 96

### Designing synthesizable bioactive compounds with chemistry-savvy machine intelligence

**Gisbert Schneider**<sup>1</sup>, [gisbert@ethz.ch](mailto:gisbert@ethz.ch), **Daniel Merk**<sup>1</sup>, **Francesca Grisoni**<sup>1</sup>, **Alexander Button**<sup>2</sup>, **Lukas Friedrich**<sup>2</sup>, **Jan A. Hiss**<sup>1</sup>, **Petra Schneider**<sup>2</sup>. (1) Department of Chemistry and Applied Biosciences, RETHINK, ETH Zurich, Zurich, Switzerland (2) Chemistry and Applied Biosciences, IPW, ETH Zurich, Zurich, Switzerland

Generative machine learning, which aims to find the distribution in the training data set to generate new samples, can take the role of a chemist not only in the formulation of testable hypotheses, but also in the creative aspect, in the assembly of innovative molecules. We have implemented and challenged 'chemistry-savvy' deep learning models in prospective molecular design projects that aimed to obtain synthetically easily accessible new chemical entities. In the first study, recurrent networks were trained with structures of known synthetically accessible, druglike compounds. By applying transfer learning, the learned feature distributions were biased towards certain pharmacologically desired endpoints. The computationally generated de novo designs were subsequently prioritized, chemically synthesized and biochemically tested for the predicted activities with high success rates. At this point, human expert knowledge turned out to be beneficial for selecting synthetically accessible chemical designs. In the second study, we developed a novel virtual synthetic assembly method that combines a rule-based approach with a neural network trained on successful synthetic routes described in chemical patent literature. This unique combination enabled a balance between ligand-similarity based generation of innovative compounds by scaffold hopping and forward-synthetic feasibility of the designs. In a prospective proof-of-concept application, the software successfully produced sets of de novo designs for four approved drugs that were in agreement with the desired structural and physicochemical properties. Target prediction indicated more than 50% of the computer-generated molecules as biologically active. Selected computer-generated compounds were successfully synthesized in accordance with the synthetic route proposed by this method. If successful in the long run, these concepts will combine a continuously learning machine intelligence with the synthesis and testing of pharmacologically relevant chemical matter. Such an envisaged automated drug design engine may not only imitate but exceed human decision making as a core aspect of the drug discovery process.



## CINF 97

### Activity landscape and its application to molecular design

**Kiyoshi Hasegawa**, [hasegawakiy@chugai-pharm.co.jp](mailto:hasegawakiy@chugai-pharm.co.jp). Chemistry, Chugai Pharma, Kamakura, Kanagawa, Japan

Activity landscape is useful technique to visualize chemical space with the desired molecular profiles. We have applied this technique to clearance data against mouse, rat and human in order to select the stable lead

compounds from huge high-throughput cluster hits. Furthermore, we have invented, so called, the atom-coloring method to clarify which molecular parts are susceptible to clearance. This information is important for designing the stable compounds. Another topic is the gap between the enzyme and cell activities. This phenomenon is often encountered in drug discovery. That is, the cell activity of the molecule is not high even though the enzyme activity is high. Comparing two activity landscapes of the enzyme and cell activities, we can investigate which molecular skeleton is a promising target for next lead optimization. Because we can easily detect the promising chemical space, we can design the libraries to fill the chemical space.

## **CINF 98**

### **Data-driven drug discovery and medical treatment by machine learning**

*Yoshihiro Yamanishi, yoshihiro.yamanishi@gmail.com. Kyushu Institute of Technology, Iizuka, Japan*

Drug repositioning is an efficient strategy for drug development and medical treatment, and it has received remarkable attention in pharmaceutical and medical science. The drug repositioning approach can increase the success rate of drug development and to reduce the cost in terms of time, risk, and expenditure. In this study, we developed novel machine learning methods for automatic drug repositioning in order to predict unknown indications of known drugs or drug candidate compounds. The prediction is performed based on the analysis of various large-scale omics data of drugs, compounds, genes, proteins, and diseases in a framework of multi-task learning. Our results show that the proposed method outperforms previous methods in terms of accuracy, applicability, and interpretability. We performed a comprehensive prediction of new indications of all approved drugs and bioactive compounds for a wide range of diseases defined in the International Classification of Diseases. We show several biologically meaningful examples of newly predicted drug indications for cancers and neurodegenerative diseases. The proposed methods are expected to be useful for various applications in drug discovery and medical treatment.

## **CINF 99**

### **Development of data driven chemistry in chemistry and chemical engineering**

*Kimito Funatsu, funatsu@chemsys.t.u-tokyo.ac.jp. Univ Tokyo Dept Chem Sys Eng, Tokyo, Japan*

Cheminformatics has been applied to various kind of area of chemistry, molecular design, materials design, organic synthesis design, structure elucidation and process control. In this lecture, I will present overview of these applications during my research life.

## **CINF 100**

### **Application of extended reality (XR) technologies in the academic library to support innovative research and instruction in the physical sciences and engineering disciplines**

*Elisandro Cabada, cabada@illinois.edu, Mary C. Schlembach. Innovation, Discovery, DDesign, and DAta (IDEA) Lab, Grainger Engineering Library Information Center, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States*

As Extended Reality (XR) technologies (Augmented Reality, Virtual Reality, and Mixed Reality) have become more affordable, scalable, and adaptable, they are increasingly being integrated into research and instruction. These technologies *extend* the lab and classroom by providing faculty and students with virtual environments that can generate discipline-specific simulations, whereby making the learning experience more visceral, engaging, and immersive. In the Innovation, Discovery, DDesign, and DAta (IDEA) Lab in the Grainger Engineering Library, library faculty and staff have taken a multifaceted approach to supporting research and instruction with XR technologies. Along with providing access to space and XR equipment, the IDEA Lab is developing XR programming capabilities. With this growing domain expertise, the library is able to offer consultative services to help faculty and staff integrate these emerging technologies into their research and instruction. By building extensible and portable web tools for the depositing and accessing of campus generated

XR content, the IDEA Lab will provide campus with a platform for the archiving and distribution of locally sourced XR software and programs. As a node in the campus-wide VR@Illinois group, Siebel Center for Design, and Health Maker Lab networks, the IDEA Lab serves as a nexus of emerging technologies and interdisciplinary innovation. By leveraging these relationships with campus partners, the Grainger Engineering Library and the Chemistry Library are positioned to apply XR technologies to serve the research and instruction needs of the Grainger College of Engineering, School of Chemical Sciences, and the Carle Illinois College of Medicine.

## CINF 101

### Deploying a VR workstation and molecular visualization at Caltech library

**Thomas E. Morrell**, *tmorrell@caltech.edu*, **Donna Wrublewski**, *Caltech Library MC 1-43, California Institute of Technology, Pasadena, California, United States*

Virtual Reality (VR) technology has great potential to improve molecular visualization, among other valuable applications. The Caltech Library has made VR available to campus via a workstation that is available to all campus users. The workstation is located in a reservable conference room, and users can check out the key to the locked cart that contains a HTC Vive headset. The workstation supports both commercial software via Steam and custom development of applications. Molecular visualization of proteins is used as a custom development example: visualizing a protein in VMD, adding it to an environment in Unity, and then looking at the protein in VR. This talk will discuss the hardware and software choices made for the Library setup, the logistics of running a VR system at a campus library, successful outreach strategies, and how chemistry applications are a key component of VR services.

## CINF 102

### Librarians and extended reality: Enhancing teaching and learning in the chemical sciences

**Samuel Putnam**, *srputnam@ufl.edu*, **Michelle M. Nolan**, *michellenolan@ufl.edu*, **Ernie Williams**, *aaronwilliams@ufl.edu*. *Marston Science Library, University of Florida, Gainesville, Florida, United States*

As extended reality (XR) technology has become more accessible, academic communities have begun experimenting with various platforms and applications to integrate XR into their pedagogical practices. Academic libraries specifically have embraced these new technologies by creating spaces for instructors and students to explore and create XR experiences. Libraries have curated these spaces as veritable laboratories for instructors and students to test and prototype without the commitment to acquiring the requisite hardware, software, and expertise. Librarians at Marston Science Library at the University of Florida created Made@UF, a development and exploration space for XR. In Made@UF, library workers consult faculty interested in implementing XR experiences into their courses, host classes for small XR sessions, and facilitate asynchronous sessions for course work.

In chemistry, XR's allure promises the potential for exploring atomic structures, practicing laboratory techniques, and modeling reactions without the dangers or cost committed to real-time bench work. In addition to traditional teaching laboratory experiments, XR can be used to perform synthetic protocols in a truly environmentally sound way. Technologies such as PyMol enable 3D models to be imported and explored in virtual environments with relative ease. As technology improves, immersive experiences are expanding beyond simple manipulation. As librarians enhance their expertise in the area of XR, chemistry faculty are coming to the libraries to foster partnerships focused on improving teaching and learning in the chemical sciences. Libraries can enhance teaching and learning in the chemical sciences by evolving beyond traditional collections based on text and building multimodal collections including XR experiences.

## CINF 103

### Using XR to teach about chemical lab safety

**Shalini Ramachandran**<sup>1</sup>, *shalinir@usc.edu*, **Rebecca Broyer**<sup>2</sup>, *rbroyer@usc.edu*, **Steven Cutchin**<sup>3</sup>, *stevencutchin@boisestate.edu*, **Sheree Fu**<sup>4</sup>, *sfu7@calstatela.edu*. (1) Libraries, University of Southern California, Los Angeles, California, United States (2) Chemistry, University of Southern California, Los Angeles, California, United States (3) Computer Science, Boise State University, Boise, Idaho, United States (4) Library, Cal State Los Angeles, Los Angeles, California, United States

In recent years, immersive technology tools have burgeoned. After the release of the affordable Oculus Go headset and the Merge Cube, there has been increasing use of Virtual and Augmented Reality in classrooms. Libraries have also taken interest in these emerging trends. In 2018, a partnership between the virtual reality company HTC VIVE and the California and Nevada state libraries, deployed immersive technology systems in 100 public libraries throughout California and 11 public libraries in Nevada. While the integration of Virtual and Augmented Reality (VR and AR) and Mixed Reality (MR), (the combination of which we will refer to as Extended Reality or XR), in K-12 settings, public libraries, and museums has been more widespread, academic libraries have not been as prolific in adopting immersive technologies. However, that may be changing with the launch of virtual lab simulations by Labster and HoloLab Champions and the VR app Nanome, which can be used to virtually manipulate chemicals and proteins. In our talk, we will discuss how XR can be used to provide instruction to students about lab safety. This summer, we plan to use VR technology to demonstrate experiments such as synthesis of gun cotton or hydrolysis of tert-butyl chloride, and a visualization of how a dye or fluorophore is spread from gloves. Our approach to XR is experimental and explorative at this stage, but our goal is to create a toolkit of resources that can be used for lab safety. We bring our diverse areas of expertise in chemistry instruction, VR development, and library innovation to our project. Specifically, our vision is to provide XR teaching and learning tools for science and engineering students via the academic library space.

#### **CINF 104**

##### **Digital collections at Cal Poly Pomona and the California State University campuses**

**Jodye Selco**, *jiselco@cpp.edu*. CeMAST, Cal Poly Pomona, Whittier, California, United States

Cal Poly Pomona, one of the 23 California State University (CSU) campuses, has both a repository of multi-media learning objects hosted on a dedicated server (discovery layer with user interfaces) and a preservation collection housed through the library. The multi-media learning objects are developed collaboratively by faculty, instructional designers, and programmers. Each multi-media learning object is accessible; the accessibility is either designed in from the beginning of the project or a separate program designed later. The library collection, called "Bronco Scholar", is one space within the CSU system's digital collections to collect and curate multiple types of data from data sets for learning and research data sets to digitized materials such as books, journals, and images. Both of these digital collections are helping the university build open courseware. The CSU also supports MERLOT.org which is a free, curated online learning collection that supports materials and content creation tools which is led by an international community of educators, learners, and researchers. MERLOT has adopted the intellectual property (IP) protection policies of the consortium, Creative Commons ([creativecommons.org](http://creativecommons.org)). Students use the multi-media learning objects in courses to help them visualize content and practice using this knowledge to solve problems.

#### **CINF 105**

##### **Discovery of novel inhibitors of human galactokinase by virtual screening**

**Min Shen**, *shenmin@mail.nih.gov*, *Xin Hu*. NCATS/NIH, Rockville, Maryland, United States

Classic Galactosemia is a potentially lethal autosomal recessive metabolic disorder caused by deficient galactose-1-phosphate uridylyltransferase (GALT) that results in the buildup of galactose-1-phosphate (gal-1-p) in cells. Galactokinase (GALK1) is the enzyme responsible for converting galactose into gal-1-p. A pharmacological inhibitor of GALK1 is hypothesized to be therapeutic strategy for treating galactosemia by reducing production of gal-1-p. In this study, we report the discovery of novel series of GALK1 inhibitors by structure-based virtual screening (VS). Followed by an extensive structural modeling and binding mode analysis of the active compounds identified from quantitative high-throughput screen (qHTS), we developed an efficient



pharmacophore-based VS approach and applied for a large-scale *in silico* database screening. Out of 230,000 compounds virtually screened, 350 compounds were cherry-picked based on multi-factor prioritization procedure, and 75 representing a diversity of chemotypes exhibited inhibitory activity in GALK1 biochemical assay. Furthermore, a phenylsulfonamide series with excellent *in vitro* ADME properties was selected for downstream characterization and demonstrated its ability to lower gal-1-p in primary patient fibroblasts. The compounds described herein should provide a starting point for further development of drug candidates for the GALK1 modulation in the Classic Galactosemia.

## CINF 106

### Measuring R group similarity using medicinal chemistry data

**Noel O'Boyle**, *baoillean@gmail.com*, Roger A. Sayle. NextMove Software, Cambridge, United Kingdom

Molecular similarity is one of the most central concepts in chemoinformatics. Typical measures of molecular similarity (such as the Tanimoto coefficient of binary fingerprints) are used for tasks such as similarity search, distinguishing similar molecules from dissimilar (e.g. identifying actives in a virtual screen), measuring the diversity of a dataset or selecting a diverse subset. While the measurement of R group similarity is conceptually the same as for whole molecules, in practice existing methods for measuring molecular similarity perform poorly.

Improved methods to measure R group similarity are of particular importance in the context of a medicinal chemistry project. These often proceed by changing one R group at a time, advancing through matched pairs. Given an appropriate measure of R group similarity, it should be possible to suggest relevant modifications or identify gaps in the project data that should be filled. In a computational context, R group enumeration could be used to generate relevant candidate molecules for virtual screening or purchase.

In medicinal chemistry, the term bioisosteric replacement refers to a substitution that retains broadly similar biological properties. However, there is a need to go beyond the concept of bioisosteres/non-bioisosteres to handle levels of similarity; for example, chloro is regarded as more similar to fluoro than to amino. Clearly, in this context, the extent to which two R groups are similar is not just (or not even) a question of shared substructures. Previous approaches to this problem include the use of R group descriptors by Holliday et al which maps atom-based descriptor values onto a vector by distance from the attachment point, and the use of reduced graphs to encode bioisosteric equivalences by Birchall et al.

We propose to use co-occurrence in medicinal chemistry project data to derive a measure of R group similarity. We will use two distinct sources for these data, both in the public domain. The first is the ChEMBL database, which contains assay data extracted from a range of medicinal chemistry journals. The other source is the US patent literature, which provides text and ChemDraw sketches from which a large amount of medicinal chemistry data can be extracted. While large-scale mining of medicinal chemistry data has previously been used to detect bioisosteres, to our knowledge this is the first time it has been used to develop a method to measure R group similarity.

## CINF 107

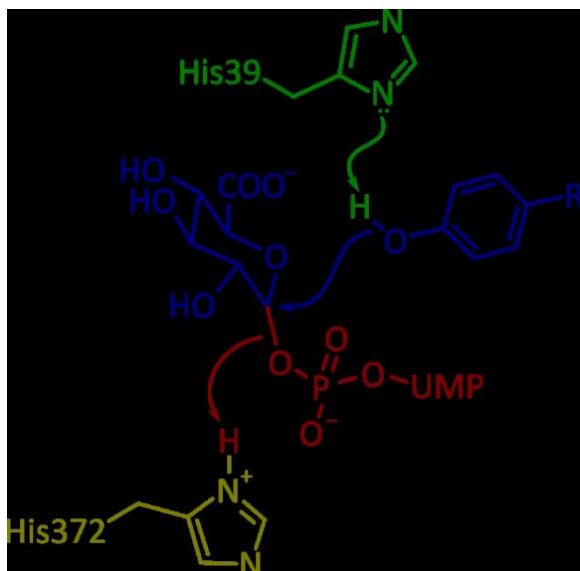
### Mechanism and prediction of UGT metabolism

**Mario Öeren**<sup>1</sup>, *mario@optibrium.com*, Peter Hunt<sup>1</sup>, David J. Ponting<sup>2</sup>, Matthew Segall<sup>1</sup>. (1) R&D, Optibrium Limited, Cambridge, Cambridgeshire, United Kingdom (2) Lhasa Limited, Leeds, Yorkshire, United Kingdom

Understanding the metabolism of xenobiotics is an important aspect of the drug design process. UDP-glucuronosyltransferases (UGTs) are major contributors to phase II metabolism, aiding in the formation of soluble metabolites by conjugation of a polar glucuronic acid moiety to their substrates. The usual UGT-mediated reaction is *O*-glucuronidation on hydroxyl or carboxyl groups with some overlap in the substrate specificity for the various UGT isoforms. An exception is UGT1A4 where *N*-glucuronidation is predominant. We will present an elucidation of the conjugation mechanism of UGTs, based on detailed Density Functional Theory calculations and validated with experimental data. Using these results, the reactivity of potential sites of metabolism can be estimated, based on the activation energy ( $E_a$ ) of the rate limiting step of the conjugation

reaction.

Random forest models were trained with high-quality regioselectivity data for the major drug-metabolising isoforms of UGT. These combine the  $E_a$  of each potential site of metabolism with steric and orientation descriptors that capture the influence of the protein environment of each isoform on the sites of metabolism. The resulting models predict if a compound is likely to be metabolised by UGTs, as well as identifying the likely isoforms responsible for the metabolism.



## CINF 108

### Signals lead discovery as the Corteva Cheminformatics Workbench

**Dirk Tomandl**<sup>1</sup>, [DTomandl@Dow.com](mailto:DTomandl@Dow.com), **Scott Smith**<sup>1</sup>, **Jeremy Wilmo**<sup>2</sup>. (1) Data Science and Cheminformatics, Corteva Agriscience, Indianapolis, Indiana, United States (2) Discovery Chemistry, Corteva Agriscience, Indianapolis, Indiana, United States

This presentation describes the Corteva journey that started with a raw "Cheminformatics Workbench" vision. In collaboration with PerkinElmer, Signals Lead Discovery as the incarnation of the Workbench took shape. We will describe our insights into what made this collaboration across companies successful and how Spotfire and Signals have the potential to impact crop protection research. An outlook on future plans and where we see the potential of the Signals technology stack for crop protection and pharmaceutical research concludes the presentation.

## CINF 109

### Probing allosteric modulators of AMP-activated protein kinase

**Xin Hu**, [hux61@mail.nih.gov](mailto:hux61@mail.nih.gov), **Juan J. Marugan**, **Wei Zheng**. NCATS, Rockville, Maryland, United States

AMP-activated protein kinase (AMPK) is a key regulator of energy homeostasis and has been implicated in many human diseases including metabolic disorder and cancer. As a potential target of Niemann-pick disease type C (NPC), we have previously reported that methyl- $\beta$ -cyclodextrin restored impaired autophagy flux in NPC cells through activation of AMPK. In this study, we first probed the binding interactions of M $\beta$ CD and various



AMPK activators bound to the allosteric binding site at different AMPK subunit. The selectivity of allosteric binding was investigated through structure-based pharmacophore modeling and MD simulations. Finally, we performed ligand- and structure-based virtual screening and have identified a number of novel small molecules with activities in the AMPK-activation assay and cell-based NPC phenotypic assay. The compounds provided a starting point for further development of drug candidates for NPC therapeutics.

#### **CINF 110**

##### **Underlying scientific evidence discovery for FDA orphan drug designations from the GARD integrative knowledge graph: Towards drug discovery for rare diseases**

*Qian Zhu, qian.zhu@nih.gov, Dac-Trung Nguyen, Noel Southall. NCATS, NIH, Rockville, Maryland, United States*

While researchers have made great progress in learning more rare diseases and developing treatments for those rare diseases, the exact cause of many rare diseases is still unknown, consequently most rare diseases still have no treatments yet. The FDA Office of Orphan Products Development (OOPD) evaluates scientific and clinical data submissions from sponsors to identify and designate products as promising for rare diseases and to further advance scientific development of such promising medical products. FDA orphan drug designations as an incredible orphan drug resource is publicly accessible to support scientific research and clinical practice. To better understand the rationale directing the drug approval towards evidence-based drug discovery, we propose to assess the designations by identifying and analyzing their underlying scientific evidence from the GARD integrative knowledge graph, developed in our previous study. The Genetic and Rare Diseases Information Center (GARD) was established in 2002 to provide up-to-date information for approximately 7,000 genetic and rare diseases, and GARD is currently managed by the Office of Rare Diseases Research (ORDR) within the National Center for Advancing Translational Sciences (NCATS). The GARD integrative knowledge graph includes diverse types of data, not only drug relevant data, such as, the FDA orphan drug designations and Inxight Drugs, but also disease and gene related data, including GARD, Orphanet, OMIM, Disease Ontology, GHR, GO, Ontology of Genes & Genomes, and etc. In this study, we will systematically explore the GARD integrative knowledge graph to collect and analyze underlying scientific evidence for those approved designations. Subsequently we will discover potential new usages of approved drugs for rare diseases by learning patterns derived from the above step.

#### **CINF 111**

##### **Explore, exploit, and extrapolate: How AI-driven SAR navigation facilitates lead optimisation in drug discovery**

*David Marcus, david.x.marcus@gsk.com, Chris Luscombe, Stephen Pickett, Stefan Senger, Darren Green. Data and Computational Sciences, GlaxoSmithKline, Stevenage, United Kingdom*

Small molecule drug discovery involves a complex multi-parameter optimisation process with cycles of design, make and test to establish a desired compound profile. Within this context, machine learning methods, experimental design and de-novo structure generation have all found a place to facilitate and accelerate lead optimisation. However, they have tended to be used in a reactive manner, to address problems posed by program teams rather than as a continual and proactive process. In this presentation we will describe how data-driven cheminformatics based AI methods has the potential to automate parts of the lead optimisation process which historically has been a very time consuming task. This strategy adds the ability to explore and exploit multiple paths within chemical space and suggest structural modifications that will gain better understanding of the SAR problem at hand. In addition, by continually improving computational models we can extrapolate to new regions of chemical space and suggest novel compounds. The implications of automation for the human-machine interface will be explored and illustrated with examples from BRADSHAW, GSK's experimental automated design environment.

#### **CINF 112**

## AI-driven drug design across the discovery spectrum: Case studies

*John H. Griffin, john@numerate.com. Numerate, Inc., Atherton, California, United States*

Breakthroughs in machine learning theory and practice, coupled with ready access to cloud based supercomputing resources and ever-increasing amounts of experimental data, are enabling truly AI centric processes for small molecule drug design wherein predictive models successfully substitute for laboratory assays throughout the Discovery critical path. This presentation will describe how diverse machine learning techniques, ranging from multidimensional and multitask boosting to deep neural networks, can extract accurate, scaffold independent, ligand based predictive models for important phenomena: target binding, functional activity, selectivity, PK/ADME properties, and toxicity. Applications of these models will be illustrated with examples from therapeutic programs and discussed in terms of their potential to enhance success/reduce attrition in drug discovery.

### CINF 113

#### What compound to synthesize next? How machine learning and artificial intelligence impact compound optimization

*Daniel Kuhn<sup>2</sup>, daniel.kuhn@merckgroup.com, Kristina Preuer<sup>1</sup>, Michael Krug<sup>2</sup>, Günter Klambauer<sup>1</sup>, Sepp Hochreiter<sup>1</sup>, Friedrich Rippmann<sup>2</sup>. (1) LIT AI Lab & Institute for Machine Learning, Johannes Kepler University, Linz, Austria (2) Computational Chemistry & Biology, Merck KGaA, Darmstadt, Germany*

Machine Learning and Artificial Intelligence techniques are increasingly used in ongoing drug discovery projects at Merck KGaA, Darmstadt, Germany. We present how Machine Learning and Deep Learning models are utilized for target identification and optimization of protein kinase inhibitors with respect to potency, selectivity and ADME profile (e.g. microsomal stability, permeability, and brain penetration). For prediction of kinase selectivity, we employ more than 270 deep learning models, which help medicinal chemists in the selection of the right molecules to synthesize. We will also share lessons learned from analysis of large protein kinase profiling datasets. Finally, three challenges for predictive modelling in drug discovery projects will be addressed: a) how the quality and applicability of predictive models can be assessed, b) how models are interpreted, and chemists can be guided by mapping information from Deep Networks back onto molecular structures and c) how to best operationally integrate predictive models into project work to maximize impact.

### CINF 114

#### Pretraining deep learning molecular representations for property prediction

*Bowen Liu, liubowen@stanford.edu, Weihua Hu, Jure Leskovec, Percy Liang, Vijay S. Pande. Stanford University, Stanford, California, United States*

Predicting molecular properties is an essential task in computational drug discovery. Deep learning methods have recently shown excellent progress in improving ligand based virtual screening performance. However, a significant problem observed by multiple researchers is that these algorithms do not generalize well in prospective settings and instead overfit to training data. This is due to two main challenges: 1. most chemical datasets are small and there is significant class imbalance. 2. when these models are used in a prospective setting, the molecules being tested are often structurally different from molecules seen during training. To close this generalization gap, we introduce a pre-training framework for graph neural networks (GNNs), inspired by recent successes in computer vision and natural language processing. We achieve this by performing a series of unsupervised and supervised pre-training auxiliary tasks on large public datasets. The pre-trained models can be subsequently fine-tuned for downstream tasks. Our pre-trained models show significant improvements over models trained from scratch, and consistently outperform several strong baselines over a wide range of benchmark datasets.

### CINF 115

## Modeling protein flexibility with conformational sampling improves ligand pose and bioactivity prediction

**Kate A. Stafford**, *kate@atomwise.com*, Jon Sorenson, Izhar Wallach. Atomwise, San Francisco, California, United States

Accurate prediction of ligand pose in protein-ligand complexes is a challenging problem. It is expected that the quality of pose prediction will affect performance in predicting other properties such as binding affinity based on a modeled complex. Many proteins of biological interest are highly dynamic, occupying multiple substates in solution that are functionally relevant and may preferentially bind different classes of ligands. However, most protein-ligand docking strategies employ a rigid model of the protein receptor structure, as a fully flexible representation of the protein is generally computationally prohibitive for large-scale virtual screening. We describe a protein conformational sampling pipeline that enables access to higher-quality ligand poses and successful identification of these poses from the resulting ensemble of protein-ligand complex models. We then use this data to explore the effects of higher-quality ligand poses on bioactivity prediction using AtomNet, a three-dimensional convolutional neural network for structure-based virtual screening. These considerations are especially important in predicting the ligand binding modes of compounds that bind kinases and GPCRs---two prominent protein classes for pharmaceutical development. Our approach to modeling receptor flexibility strikes a balance between relying on a single experimentally determined protein structure and full modeling of the receptor structure with molecular simulations.

### CINF 116

#### Machine learning for the discovery of $\alpha_v\beta_6$ integrin antagonists

**Jonathan D. Hirst**<sup>1</sup>, *jonathan.hirst@nottingham.ac.uk*, Steven Oatley<sup>1</sup>, Ellen Guest<sup>1</sup>, Thomas Gaertner<sup>2</sup>, Simon J. MacDonald<sup>3</sup>. (1) School of Chemistry, University of Nottingham, Nottingham, United Kingdom (2) School of Computer Science, University of Nottingham, Nottingham, United Kingdom (3) GlaxoSmithKline, Stevenage, United Kingdom

Chemical space is too large to enumerate explicitly. Thus, it represents a so-called intensionally defined design space. Search strategies for intensionally designed spaces are a current area of interest in machine learning. In the context of a drug design problem, we have investigated the application of a data-driven adaptive Markov chain approach, where the acceptance probability is given by a probabilistic surrogate of the target property, modelled with a maximum entropy conditional model. We have applied the approach to a lead development search for an antagonist of an alpha-v integrin, using a molecular docking score as the optimisation function. The RGD integrin receptors are thought to play a key role in fibrosis. Antagonism of alpha-V-beta-6 is one promising avenue for the development of a novel therapeutic treatment and some success has been reported in discovering compounds with significant activity against alpha-V-beta-6 and physicochemical properties commensurate with oral bioavailability. Molecular docking was performed using OpenEye FRED, which uses a rigid ligand approach, where a large number of conformations are generated and each of those are docked successively. We have discovered compounds with greater predicted activity than compounds found in our previous work by employing two strategies. We have substantially increased the search space explored, to the order of  $\sim 10^{20}$  possible compounds, and we have considered receptor flexibility by docking to an ensemble of snapshots from molecular dynamics simulations.

### CINF 117

#### One million crystal structures in the CSD: Cause for celebration, cause for consideration

**Robin Taylor**, *robin@justmagnolia.co.uk*, Ian Bruno. Cambridge Crystallographic Data Centre, Cambridge, United Kingdom

By the time we meet in San Diego, the Cambridge Structural Database (CSD) will contain over a million crystal structures of organic and metal-organic compounds. The uses that have been made of the database are manifold, but may broadly be divided into three categories: (a) fundamental research, the results of which have

often been of seminal importance; (b) knowledge-driven software developments; (c) commercial or commercially-relevant applications. Primary among the latter is the use of the CSD to assist the invention of new pharmaceuticals and agrochemicals. Promising new applications are mainly in the field of novel materials, e.g. for gas storage and separation, electronics, thin films. Of course, the CSD has its share of challenges and uncertainties. One of particular interest is the applicability (or otherwise) of small-molecule crystal structures to *in vivo* environments. Another is the ongoing need to maintain high data quality: despite the arrival of big data and machine learning, the basic principle of garbage in, garbage out is still with us. But these cautionary points must be seen in context. The undeniable fact is that the CSD has been a great success, fulfilling the vision of its founders and showing what can be achieved when a scientific community works together to collate and exploit the results of its endeavors.

## **CINF 118**

### **Leveraging the CSD's one million structures in course-based undergraduate research experience**

**Heba Abourahma**<sup>1</sup>, *abourahm@tcnj.edu*, **Amy Sarjeant**<sup>2</sup>. (1) Chem Dept, The College of New Jersey, Ewing, New Jersey, United States (2) Cambridge Crystallographic Data Centre, Piscataway, New Jersey, United States

The Cambridge Structural Database provides a wealth of knowledge for informatics research. At The College of New Jersey (TCNJ), a primarily undergraduate institution, the CSD is used in multiple courses throughout the curriculum. In second-year courses the CSD is used to help students appreciate geometrical features of small organic molecules. In an upper-level, special topics course the CSD is used in a Course-based Undergraduate Research Experience (CURE) that involves identifying metal-organic framework candidates suitable for catalysis. This presentation will demonstrate how the use of the CSD can enhance undergraduates' learning experience.

## **CINF 119**

### **Use of the Cambridge Structural Database in the undergraduate chemistry curriculum**

**Arun T. Royappa**, *royappa@uwf.edu*. Chemistry, University of West Florida, Pensacola, Florida, United States

Crystallography has provided more information on molecular structure, bonding and geometry than any other scientific technique, and the Cambridge Structural Database (CSD) is the world's largest repository of crystal structures. As such, the CSD is a valuable resource in several different courses across the undergraduate Chemistry curriculum at the University of West Florida. In this presentation, the use of the CSD at a Primarily Undergraduate Institution will be described. In brief, the CSD is integrated not only into our research programs but also into lecture and laboratory courses, e.g., in extended exercises and as a source for obtaining important molecular structural parameters.

## **CINF 120**

### **Examining research data through a crystal lens: Teaching students about primary data, data representation, and data management using crystal structure databases**

**Judith N. Currano**, *currano@pobox.upenn.edu*. Chemistry Library, University of Pennsylvania, Jenkintown, Pennsylvania, United States

The recent move of funding agencies towards requiring researchers to deposit primary data alongside publications of funded research has increased the amount of primary data available to individual researchers, and a smart scientist uses all available information! When faced with databases of structures and properties, though, students are not always certain as to which contain primary data and which are presenting data previously published in other sources. This presentation focuses on the use of curated crystallography databases in a classroom setting and examines three different concepts that they can help present to students. First, since they are primary data repositories that contain the actual endproducts of a chemist's research, rather than indexer-interpreted data, they can be used as a case study in data curation and management. Second, the

fact that they are curated can be used to initiate a conversation about critically evaluated data, data trustworthiness, and the ways in which providing access to one's primary data encourages reproducibility and discourages fraud. Finally, they offer an opportunity to examine structural representation and techniques of structure-based searching in a classroom setting.

## CINF 121

### **Materials genome approach to functional materials discover using the CSD**

**Kimberley R. Cousins**, *kcousins@csusb.edu*, **Sarah B. Rodriguez**, *006013497@coyote.csusb.edu*. *Chemistry Biochemistry, California State University, San Bernardino, California, United States*

The materials genome initiative seeks to discover novel materials for tomorrow's technology. At the CSUSB Center for Advanced, Functional Materials we harness "UI" (undergraduate intelligence) and the CSD to find functional materials demonstrating piezoelectric and ferroelectric properties among existing, known crystal structures as well as new materials suggested by features of existing structures. In this presentation, we will describe our multiple approaches for structure discovery using the CSD to suggest known and new materials and will highlight several successful functional materials uncovered by this approach. Our team has used first principles (DFT) calculations and experimental methods for property characterization, and we will share some results of the materials that not yet published. Among known crystal structures we have uncovered and characterized a highly responsive non-linear optical material, and a novel electronic ferroelectric/piezoelectric co-crystal. In addition, two novel materials have been crystallized based on our investigation of known hydrogen bonding co-crystals in the CDS. While automation would undoubtedly speed the screening process, our manual investigations provide new insight into structure/function, as well as a highly valuable educational experience.

## CINF 122

### **Building a collection of metal–organic frameworks in the Cambridge Structural Database for materials discovery**

**David Fairen-Jimenez**, *df334@cam.ac.uk*. *Dept. of Chemical Engineering & Biotechnology, University of Cambridge, Cambridge, United Kingdom*

The building-block approach to the synthesis of metal-organic frameworks (MOFs) has opened the possibility to synthesise a virtually infinite number. This creates exciting opportunities, but also raise the question of how to identify and classify MOFs among the plethora of existing crystal structures – this is particularly difficult as the definition of MOFs is still under debate. At the same time experimental trial-and-error discovery of MOFs is not fast enough and therefore new methods accessible not only to computational researchers but mainly to experimentalists need to be developed.

In collaboration with the Cambridge Crystallographic Data Centre, we have developed a curated database containing all the MOFs deposited in the Cambridge Structural Database (CSD). This initiative provides the MOF community with tools to extract their desired structures from the pool of CSD crystalline structures and to visualise their data of interest. In order to extract the desired MOF structures, we developed a number of "look-for-MOF" criteria. We also developed new capabilities to enable researchers to browse and look for MOF families based on metal-clusters, chirality, surface chemistry (functional groups) and pore and network dimensionality. This has resulted in a regularly updated CSD-MOF subset of ca. 90,000 structures to date. This subset can also further evolve depending on changes to the definition of a MOF, allowing users to match the criteria relevant to their area of interest. We also offer the possibility of visualizing the MOF landscape of properties on an interactive webpage (<https://tinyurl.com/CSD-MOF-vis>).

We also have demonstrated the power of the CSD-MOF subset for computational high-throughput screening (HTS), where we analyzed their performance in different applications. We have completed the full cycle from the screening of MOFs to the identification and synthesis of optimal materials. Our studies delimit the relationships between structural properties and gas adsorption performance in dynamic 5D representations, allowing the final user to select any material analyzed. Furthermore, as the computational cost of the molecular simulations needed in HTS is still too high, we show how the integration of machine learning algorithms allows predicting

adsorption performance not only of existing materials but also future ones. All in all, we believe the combination of developed tools will greatly enhance the HTS MOFs discovery for multiple applications.

## CINF 123

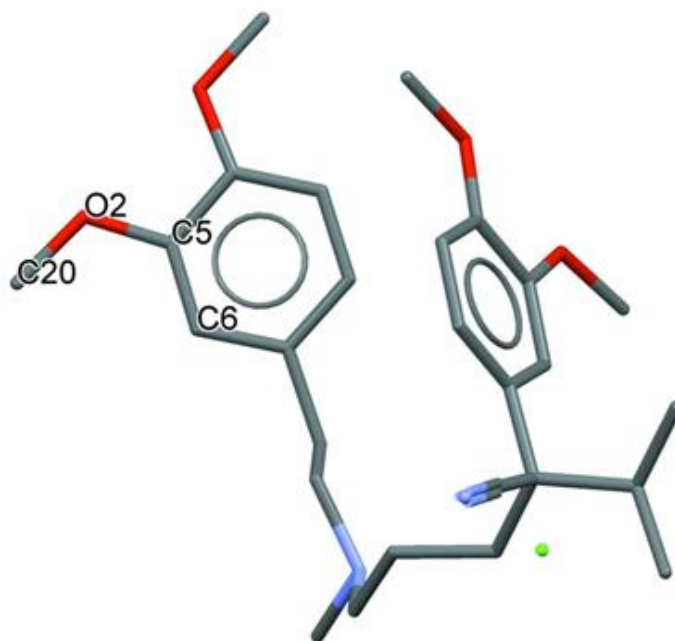
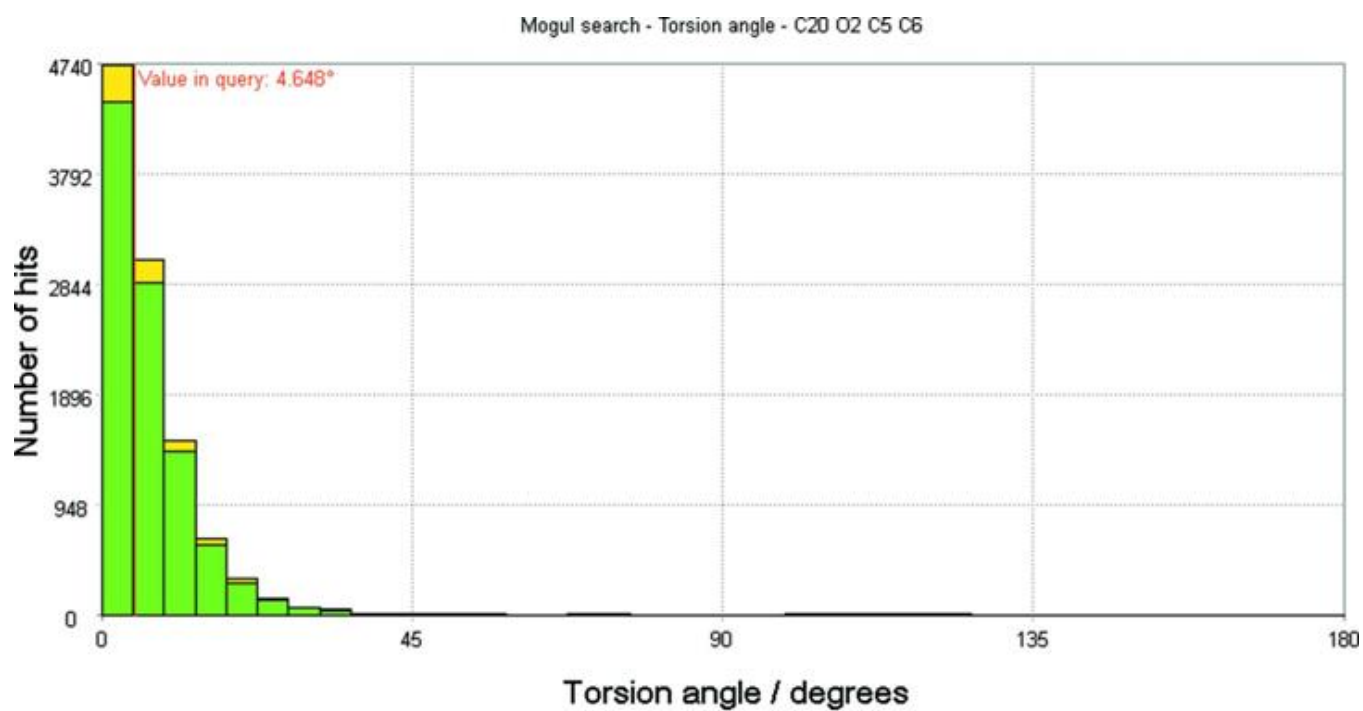
### **Improved crystal structure determination from powder diffraction data using the Cambridge Structural Database system**

**Kenneth Shankland**<sup>1</sup>, *k.shankland@reading.ac.uk*, **Elena Kabova**<sup>1</sup>, **Jason Cole**<sup>2</sup>. (1) Pharmacy, University of Reading, Reading, Berkshire, United Kingdom (2) CCDC, Cambridge, United Kingdom

Crystal structure determination from powder diffraction data (SDPD) using global optimisation methods relies heavily on the input of a reasonably accurate starting 3D model of the molecule under study, such that the structure determination involves only the optimisation of molecular position and orientation within the unit cell, plus those torsion angles whose values are not known in advance. The CSD is excellent resource in this regard, but its utility in SDPD goes well beyond simply providing closely-related fragments from which models of the molecule under study can be constructed.

This presentation will describe the close link between the DASH structure-solving program and the CSD in the context of SDPD, showing how use of the Mogul library can significantly improve the chances of solving challenging structures with large numbers of torsional degrees of freedom. Furthermore, the use of CSD Conformer Generator in model construction, and the use of Mogul in input model validation will be described.





The Mogul-derived distribution (top) for the C20—O2—C5—C6 torsion angle of verapamil hydrochloride (bottom).

**CINF 124**

**Million opportunities: Using the CSD to design color changing molecular switches**



**Paul R. Raithby**, *p.r.raithby@bath.ac.uk*. Department of Chemistry, University of Bath, Bath, United Kingdom

One of the key areas of development in modern structural chemistry is the design, synthesis and fabrication of new “smart” materials with targeted properties or functions. A particular area of importance is solid-state materials that change color when exposed to external agencies such as light, changes in pressure or temperature, or in the presence of volatile organic compounds or toxic gases. The Cambridge Structural Database (CSD) now has a 1,000,000 structures, and through “knowledge mining” using the structural information in this database linked with spectroscopic data it is possible to identify some of the structural changes that are necessary to induce a change in the color of the material, and design reversibility into the switching process for the material, and influence the time-scale over which the color change occurs. With the aid of the CSD we have designed rapidly reversible color switching platinum(II) complexes that change in the presence of low concentrations of water or methanol, and organic thermochromic cocrystals. We now describe the factors that favor the formation of solid-state linkage isomers that change color when exposed to light of a specific wavelength, and describe the design developments that we have used to control the temperature at which the changes occur.

## CINF 125

### Biologics information in PubChem

**Jian Zhang**<sup>1</sup>, *jjzhang@ncbi.nlm.nih.gov*, **Paul Thiessen**<sup>1</sup>, **Tiejun Cheng**<sup>1</sup>, **Ben Shoemaker**<sup>1</sup>, **Evan Bolton**<sup>1</sup>, **Noel O'Boyle**<sup>2</sup>, **Roger A. Sayle**<sup>2</sup>. (1) NLM, National Institutes of Health, Bethesda, Maryland, United States (2) NextMove Software, Cambridge, United Kingdom

PubChem provides a wealth of chemical information from deposited substances to standardized compounds, from bioactivity data to patent information. Beyond the small molecules, PubChem also contains a large amount of information for larger molecules known as biologics which includes various biopolymers such as glycans, amino acids, nucleic acids, and lipids. For these biologics, all atom/bond approaches struggle to convey meaning to human viewers. To help users, PubChem uses biopolymer-based line notations and descriptions, including PLN, HELM, condensed IUPAC, LINUCS, and SNFG depictions. In this presentation, we will discuss how PubChem processes large molecule structures, how standards are harnessed, and opportunities for future improvements.

## CINF 126

### Current progress in HELM representation, integration, and data migration

**David Deng**<sup>1</sup>, *dengw2@gmail.com*, **Tony Yuan**<sup>1</sup>, **Jinbo Lee**<sup>2</sup>, **Rajeev Hotchandani**<sup>1</sup>. (1) Scilligence, Cambridge, Massachusetts, United States (2) Scilligence Corporation, Cambridge, Massachusetts, United States

Since its creation by Pfizer scientists and introduction by the Pistoia Alliance, HELM (Hierarchical Editing Language for Macromolecules) notation has been increasingly adopted as an industry standard for biologic sequence representation and registration. Scientists can now save their biologic sequences in an exchangeable format rather than that of their own organization. By providing a free-to-use sketcher, the release of an open source HELM web editor has helped expand HELM adoption.

Various informatics vendors have added HELM notation support, often additionally including monomer library management. HELM has the ability and flexibility to represent various types of complex macromolecules. Given the variety of custom notations in use, many nuanced specifications are required for production use. With the adoption of HELM notation in its Chem and Bioinformatics products, Scilligence has been assisting its adoption and has encountered a great deal of “real-world” scientific data. In this presentation, our progress with the HELM representation, data integration, and our efforts to ease incorporation of legacy data will be summarized.

## CINF 127

### Representational and algorithmic challenges in biologic informatics 2019

**Roger A. Sayle**, *roger@nextmovesoftware.com*, Noel O'Boyle. NextMove Software, Cambridge, United Kingdom

Recent years have seen a dramatic improvement in the technology and notations used to represent biologics, such as peptides, nucleic acids and antibodies. As of March 2019, the NCBI's PubChem database contains over 526K compounds annotated with a biological line notation. However there remain a number technical challenges and representational issues at the frontiers of what can currently be handled. This talk discusses some of these issues, including the challenges created by the differing variants of Pistoia's HELM notation, and the representation of non-standard nucleic acids, lipids, protein variants and PROTACs.

## CINF 128

### **Notation for identification of glycans contained in glycoproteins, glycolipids, and other biomolecular structure data**

**Issaku Yamada**<sup>1</sup>, *issaku@noguchi.or.jp*, Nobuaki Miura<sup>1</sup>, Shinichiro Tsuchiya<sup>2</sup>, Kiyoko F. Aoki-Kinoshita<sup>2</sup>. (1) The Noguchi Institute, Tokyo, Japan (2) Faculty of Science and Engineering, Soka University, Tokyo, Japan

The IUPAC International Chemical Identifier (InChI) is a well-known notation for chemical substances that can be applied to many types of chemical structures. Recently, InChI has been trying to cover a variety of substances such as mixtures, metal-complexes, tautomers, etc., which is very important in chemical fields. By using InChI/InChIKey, chemical databases can be easily developed and allows linking between many database resources. However, carbohydrate structures can not be covered by the InChI system at this time.

In the field of glycobiology, monosaccharides are considered the basic unit of carbohydrates, similar to nucleic acids of DNA and amino acids of proteins. However, carbohydrates are more complex than these linear biomolecules because of their branched structure, where multiple monosaccharides can attach glycosidic bonds to the same parent monosaccharide. Moreover, carbohydrate structures are difficult to determine distinctly. Various structures that can be considered carbohydrates are published in the literature, such as monosaccharide compositions, structures with undefined glycosidic bonds, and those having underdetermined capping units.

We have developed a new representation method focusing on carbohydrate structures called WURCS (Web3 Unique Representation Carbohydrate Structures). WURCS can represent all different kinds of carbohydrate structures, as described above, which can all be described as unique strings. Therefore, WURCS can be used to identify carbohydrate structures.

We have been developing the International Glycan Structure Repository, GlyTouCan, which assigns unique accession numbers to glycan structures that are published in the literature. WURCS and related tools which we have developed are used as the foundation of GlyTouCan. Because GlyTouCan is now being officially recognized as the standard repository for carbohydrate structures, WURCS will become an important notation for uniquely representing glycans in the future.

## CINF 129

### **Trends in biologics research and development: Analytic studies based on CAS-curated data**

**Cynthia Y. Liu**, *cliu@cas.org*, Yingzhu Li, Yi Deng. CAS, Columbus, Ohio, United States

Biologics or biopharmaceuticals have offered new and exciting options for the treatment of diseases previously considered incurable and are becoming one of the most promising categories of drugs in the 21<sup>st</sup> century. Given the intense and growing interest these drugs are receiving, CAS, a division of American Chemical Society, has conducted, in collaboration of National Science Library, the Chinese Academy of Science, a series of scientific analyses in order to help organizations stay abreast of the rapid development in this area. The big data analyses were based on 30 years' worth of published data on four major classes of biologics: antibodies, fusion protein, gene and cell therapy, and vaccines. About 250,000 substances and 500,000 published documents were extracted from CAS's curated substance collection and document database, respectively. Comprehensive analyses of global R&D trends were conducted from various perspectives that covered organization sources,

patent flow, molecular types, targets, disease indications, among others. Our analyses have revealed interesting findings and also identified both opportunities for rapid and new successes in biologics development and potential challenges.

### CINF 130

#### **Biosequence searching: How CAS is expanding workflow solutions for IP searchers and beyond**

**Robbie J. Walczak**, *randkwalczak@att.net. Chemical Abstracts Service, Hilliard, Ohio, United States*

The area of bioinformatic searching in biological research and development continues to open new avenues for innovation. Creating search solutions that efficiently meet new innovation needs, however, remains a challenge due to the overall complexity of existing biological sequence databases and the lack of comprehensiveness of desired information. Recently, CAS and Apteian have partnered to assist the Intellectual Property community perform more robust and comprehensive biosequence searching in order to discover areas of innovation opportunities. This partnership will increase workflow efficiency by uniting critical sequence sources for IP searching in a single interface familiar to intellectual property professionals, Apteian GenomeQuest. Using CAS Biosequences™, customers can search and identify structure detail and context within patents beyond traditional machine indexing. CAS scientists also curate exclusive information, which includes detailing uncommon chemically-modified sequences and sequence variants not found in any other database. Furthermore, CAS curation adds unique sequences from non-patent sources, expanding search comprehensiveness. We will discuss how this collaboration successfully met a significant customer concern, and examine how CAS Biosequences could potentially generate other bioinformatic workflow solutions with current and future indexing strategies.

### CINF 131

#### **Pesticide quantitative structure-greenhouse-activity relationship models**

**Dirk Tomandl**<sup>1</sup>, *DTomandl@Dow.com*, **Carla Klittich**<sup>4</sup>, **John Herbert**<sup>2</sup>, **Norbert M. Satchiv**<sup>2</sup>, **Dave Demeter**<sup>3</sup>. (1) *Data Science and Cheminformatics, Corteva Agriscience, Indianapolis, Indiana, United States* (2) *Discovery Biology, Corteva Agriscience, Indianapolis, Indiana, United States* (3) *Computational Chemistry and Modeling, Corteva Agriscience, Indianapolis, Indiana, United States* (4) *Retired - Discovery Biology, Corteva Agriscience, Indianapolis, Indiana, United States*

Crop Protection scientists at Corteva run a wide variety of biological assays to evaluate experimental analogs for bioactivity. These assays are typically whole-organism tests. As a result, the activity data are noisy and come from a multitude of biological mechanisms of action.

This work describes the development of QSAR models that predict the greenhouse activity of small organic molecules. These QSAR models are beginning to be used for Lead Generation and Lead Optimization. Model development included a complex data curation process, a descriptor selection process, model development and validation. Statistical validation data and preliminary comparisons with Deep Learning Networks will be presented.

### CINF 132

#### **Pore volumes and surface areas of metal-organic frameworks as descriptors for materials discovery**

**Austin Mroz**, *amroz@uoregon.edu*, **Christopher H. Hendon**. *Chemistry and Biochemistry, University of Oregon, Eugene, Oregon, United States*

Metal-organic frameworks (MOFs) are an emerging material in the energy storage device arena due to their nanoporous, crystalline structure, and tunability of the two basic building blocks: metal centers and organic linkages. The structural modularity of these materials renders them ideal for high-throughput materials design approaches that take advantage of the predictive power of machine learning (ML). Accuracy of ML models depends on the reliability of the labeled training data set. Therefore, improved data labels will have significant

impact on successful design and property prediction of MOFs. The chemical and physical properties of a MOF are derived from the pore, which is best described via the pore volume (PV) and surface area (SA). Current methods of calculating and measuring PV and SA rely on inert gas probe molecules, yielding only the accessible PV and SA, in addition to an inherent dependence on probe molecule identity. This is detrimental to ML models, which rely on accurate predictive features for training. We have increased the accuracy and efficiency of PV and SA calculations by taking advantage of density functional theory, which yields the electrostatic potential at a discrete number of volumetric pixels within the unit cell. Using a systematic sampling approach coupled with blob and Canny edge detection algorithms we successfully demonstrate the utility of this methodology for a data set of 100 MOFs. The presented methodology is independent of probe molecule identity, thus universalizing PV and SA calculations for all porous materials, thereby removing bias from the model.

### CINF 133

#### Artificial neural network-based approach to thermodynamic property estimation

**Ruben Van de Vijver**<sup>1</sup>, *Ruben.VandeVijver@UGent.be*, **Pieter Plehiers**<sup>1</sup>, **Pieter-Jan Verberckmoes**<sup>1</sup>, **Guy B. Marin**<sup>1</sup>, **Christian V. Stevens**<sup>2</sup>, **Kevin Van Geem**<sup>1</sup>. (1) *Laboratory for Chemical Technology, Ghent University, Ghent, Belgium* (2) *Department of Sustainable Organic Chemistry and Technology, Ghent University, Ghent, Belgium*

The optimization of various chemical processes, such as fuel combustion in engines and pyrolysis of hydrocarbons in a steam cracker, relies heavily on highly detailed kinetic models. Such models require accurate kinetic and thermodynamic data. Ideally, these are provided by high-level-of-theory *ab initio* calculations. However, for the time being, performing such calculations “on-the-fly” is prohibitively expensive. Combining existing data with group additivity is the current state-of-the-art for thermodynamic and kinetic property estimation. One major disadvantage of this approach is the large amount of expert-user intervention. Essentially, the task of this expert-user is recognizing the important (functional) groups that constitute the molecule and that contribute to its thermodynamic behavior.

Artificial neural networks have been successfully applied to several problems involving similar pattern- and element recognition tasks, especially in image processing and natural language processing. As neural networks have also been frequently applied in several QSPR problems, they are seen as a promising alternative for group additivity.

In this work, a generally applicable artificial neural network approach is presented. It allows the estimation of various thermodynamic properties such as enthalpy- and entropy of formation and specific heat capacities at different temperatures. The neural network is trained on a large number of C,H,N,O-containing hydrocarbons. While the current results do not yet achieve the desired chemical accuracy of group additive methods, the trends are captured well. With some further improvements on the molecular representation in the input, this approach can become competitive with group additivity as it is applicable to a wider range of hydrocarbons than any individual group additive method.

### CINF 134

#### Persistent homology for chemical applications: Story of birth and death

**Konstantinos D. Vogiatzis**<sup>1</sup>, *kvogiatz@utk.edu*, **Alan Cherne**<sup>2</sup>, **Justin K. Kirkland**<sup>3</sup>, **Vasileios Maroulas**<sup>2</sup>, **Cassie Putman Micucci**<sup>2</sup>, **Jacob Townsend**<sup>3</sup>. (1) *Department of Chemistry, University of Tennessee, Knoxville, Tennessee, United States* (2) *Department of Mathematics, University of Tennessee, Knoxville, Tennessee, United States* (3) *Department of Chemistry, University of Tennessee, Knoxville, Knoxville, Tennessee, United States*

We have developed a novel molecular fingerprinting method based on persistent homology that can encode the geometrical and electronic structure of molecules for chemical applications. We have demonstrated its applicability on two different test cases. The first is related to non-covalent interactions between functional groups of materials and small gas molecules for environmental applications. The second is related to biomimetic catalysis via heme and non-heme Fe(IV)-oxo sites for C-H activation. For both cases, quantum chemical calculations were performed on a small number of molecules (50-100) for the generation of meaningful data. We have used these data in order to train a statistical model that includes the new fingerprinting method and

machine learning algorithms. The trained models have been used for high-throughput virtual screening by predicting the properties of larger molecular databases (more than 100,000 entries) where quantum chemical data are not available.

### CINF 135

#### **ML models that both learn and teach chemistry via partitioning reactive trajectories by reaction product in phase space using support vector machines**

**Gianmarc Grazioli**<sup>1,2</sup>, *g.grazioli@uci.edu*, **Saswata Roy**<sup>2</sup>, **Carter T. Butts**<sup>1,3,4</sup>. (1) California Institute for Telecommunications and Information Technology (Calit2), University of California, Irvine, Irvine, California, United States (2) Dept. of Chemistry, University of California, Irvine, Irvine, California, United States (3) Department of Computer Science, University of California, Irvine, Irvine, California, United States (4) Departments of Sociology, Statistics, and Electrical Engineering and Computer Science, University of California, Irvine, Irvine, California, United States

The highly layered architectures featured in deep learning have paved the way for remarkable recent advances in a wide variety of fields. At the same time, the accuracy of these deep learning models often comes at a cost: it is typically difficult or impossible to determine how a given model works. Although this may not be a problem for some applications, there is a need for machine learning-driven approaches to molecular simulation whereby ML models can not only be called upon to make predictions, but also be interrogated to uncover chemical insights pertaining to why the model is able to make the predictions it makes. For instance, if an ML model, trained on phase space data from chemical simulations, is able to consistently predict the reaction product of a multi-reaction pathway chemical system long before the reaction occurs, then the model must necessarily have learned an approximation to the isocommittor surfaces that separate trajectories bound by distinct product states. However, obtaining chemical insight from this solution depends on being able to extract information on these surfaces from its often highly indirect representation within the ML model and summarize it in humanly accessible terms. Kernel learning systems such as support vector machines (SVMs) provide an excellent framework for developing models that not only learn latent predictive patterns in chemical data, and use them to make predictions, but also inform the user as to how the prediction was made. The reason SVMs and related methods excel in this capacity is that the source of the flexibility in their decision boundaries is the expression of the nonlinear target function in terms of a set of known basis functions defined by the choice of kernel and training data, making them susceptible to subsequent analysis. In particular, every prediction regarding e.g. a transition state between products can be expressed in terms of linear combinations of knowable functions of input features such as atomic coordinates and momenta, facilitating approximation in chemically intelligible terms. This presentation will include a theoretical description of our approach as well as a demonstration whereby the methodology is applied to ab-initio dynamics of the photodissociation of acetaldehyde.

### CINF 136

#### **Deep neural network model for MD-level packing density predictions and its application in the study of 1.5 million organic molecules**

**Johannes Hachmann**, *hachmann@buffalo.edu*. Dept of Chemical and Biological Engineering, University at Buffalo, SUNY, Buffalo, New York, United States

The process of developing new compounds and materials is increasingly driven by computational modeling and simulation, which allow us to characterize candidates before pursuing them in the laboratory. One of the non-trivial properties of interest for organic materials is their packing in the bulk, which is highly dependent on their molecular structure. By controlling the latter, we can realize materials with a desired density (as well as other target properties). Molecular dynamics simulations are a popular and reasonably accurate way to compute the bulk density of molecules, however, since these calculations are computationally intensive, they are not a practically viable option for high-throughput screening studies that assess material candidates at a massive scale. In this work, we employ machine learning to develop a data-derived prediction model that is an alternative to physics-based simulations, and we utilize it for the hyperscreening of 1.5 million small organic molecules as well as to gain insights into the relationship between structural makeup and packing density. We also use this study to analyze the learning rate of the employed neural network approach and gain empirical data on the



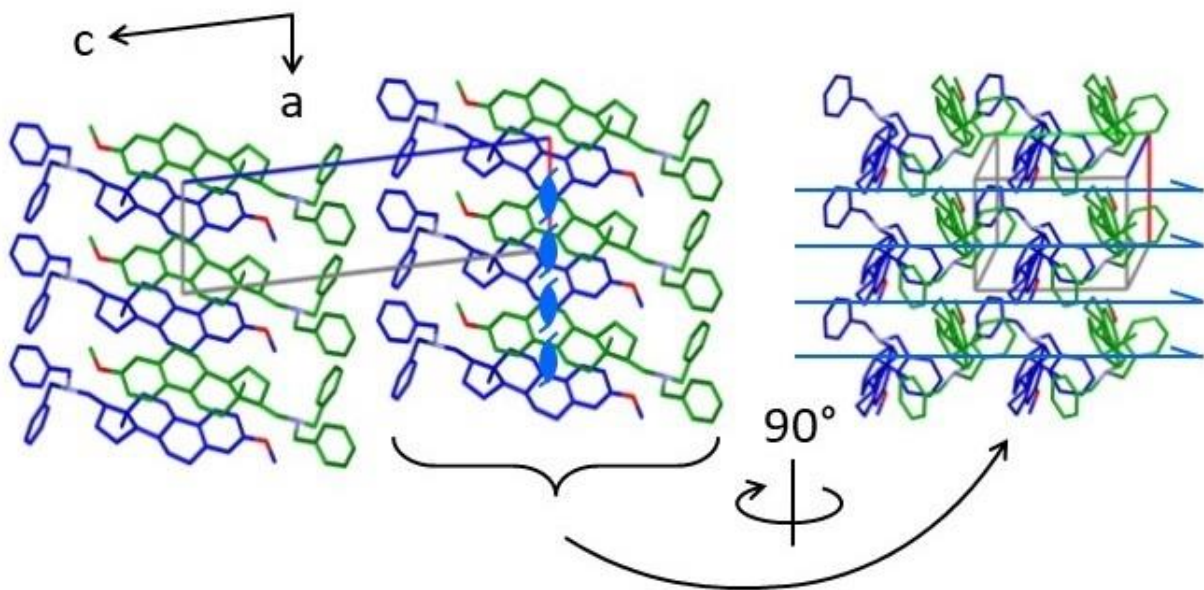
dependence of model performance and training data (size), which will inform future investigations.

### CINF 137

#### Pervasive approximate symmetry in $P1$ and high- $Z'$ organic crystals: Implications for crystal nucleation

**Carolyn P. Brock**, *cpbrock@uky.edu*. Chemistry, University of Kentucky, Lexington, Kentucky, United States

Data in the Cambridge Structural Database (*i.e.*, the CSD) for *ca.* 750 well determined ( $R \leq 0.075$ ) organic crystal structures having translational symmetry only and *ca.* 300 more with more than four independent formula units (*i.e.*,  $Z' > 4$ ) have been investigated in detail. Obvious approximate symmetry has been found in well over 50% of them. Approximate  $2_1$  screw axes, pseudoinversion centers, and pseudotranslations all occur frequently; approximate glides are less common. A program to find pseudotranslations (*i.e.*, modulations) has been written in collaboration with Robin Taylor of the CCDC. The program has been used to identify relationships between phases archived in the CSD and also to identify the probable basic cell in modulated structures. It seems likely that the crystal nucleus is often more symmetric than the macroscopic crystal. The number of structures having obvious layers has been a surprise; those layers often have good approximate symmetry but are stacked such that the overall 3-D symmetry is lower than the near 2-D symmetry. If interlayer forces are weaker than intralayer forces then deformations during the early stages of crystal growth are more likely between layers than within them.



Layers (001) in GURBIG ( $P1$ ,  $Z = 2$ ;  $\alpha, \beta, \gamma = 86.5, 83.1, 89.5^\circ$ ). There are approximate  $2_1$  screw axes along **b**.

### CINF 138

#### One million crystal structures: One million disappearing polymorphs waiting to happen?

**Julian Helfferich**, *julian.helfferich@avmatsim.eu*, *Jacco van de Streek*, *Marcus Neumann*. Avant-garde Materials Simulation GmbH, Freiburg, Germany

On rare occasions, a crystal structure that has been prepared for years is unexpectedly “superseded” by a

thermodynamically more stable polymorph, rendering the initial crystal structure almost impossible to obtain—a so-called “disappearing polymorph”. Famous cases included the drug compounds Ritonavir and Rotigotine: the thermodynamically more stable polymorph is substantially less soluble and required a costly reformulation before a product with comparable specifications could be manufactured again.

On a case-by-case basis, the probability of a kinetically hindered but thermodynamically stable polymorph having been missed in an experimental polymorph screen can be assessed by a computational crystal structure prediction study: the calculations do not suffer from kinetics and will in principle enumerate all possible crystal structures for a given compound; comparison against the experimental structures identifies any missed polymorphs. In a recent paper, we presented the results of 41 such computational crystal structure prediction studies to draw up statistics regarding the number of hidden Ritonavir cases. The main finding of the paper was that between 15 and 45% of all experimental crystal structures are in fact thermodynamically unstable.

The large spread in our estimate was due to the computational error in the calculation of the relative free energies of the predicted polymorphs. Even though the computational error is estimated to be as small as 0.5 kcal/mol, about half the size of the gold standard referred to as “chemical accuracy”, the experimental energy differences between polymorphs are also around the 0.5 kcal/mol mark, and to reduce the large spread in our estimate an even more accurate energy model was needed.

The shortcomings of our old energy model—neglect of temperature, relatively poor accuracy of the exchange part of the functional used in the DFT calculations, and the assumption that Van der Waals interactions are pairwise additive—are well documented, and in this contribution we will present the results of an energy model in which these shortcomings have been addressed. With the availability of more accurate relative lattice energies, we are able to narrow down our estimate of how many of the one million experimental crystal structures represent a thermodynamically unstable polymorph at ambient conditions.

## CINF 139

### What the Cambridge Structural Database tells us about hydrates

*Jen Werner, jw1701@georgetown.edu, Jennifer A. Swift. Department of Chemistry, Georgetown University, Washington, District of Columbia, United States*

A large fraction of organic molecules can crystallize as hydrates, but the structural features that predispose hydrate formation remain unclear. Back-end search methods were implemented using Python API (API = Application Programming Interface) to identify trends across the more than 35,000 organic hydrate structures in the Cambridge Structural Database (CSD). Pairs of hydrated and anhydrous structures were identified based on structure matching of the organic component using SMILES strings. Several features were then analyzed including packing density, symmetry, and the ratio of hydrogen bond donors and acceptors in order to identify trends. A mathematical approach was developed in parallel to classify each hydrate based on its topology as either a channel, zigzag, or isolated hydrate.

## CINF 140

### Energetics of co-crystal formation: Informing prediction through combining the database with large scale simulations and machine learning

*Graeme M. Day, g.m.day@soton.ac.uk, Christopher R. Taylor, David McDonagh, William Fyffe, Chris-Kriton Skylaris. Chemistry, University of Southampton, Southampton, United Kingdom*

A grand challenge in the area of computational chemistry is the prediction of crystal structures from first principles. The area of crystal structure prediction (CSP) has seen important advances in the past few years, both in the underlying theoretical models and algorithms and applications in, for example, pharmaceutical solid form screening and the discovery of functional materials. As we develop these methods, it is also crucial to develop rules for interpreting their results. For example, what is the relevant energetic range for observable polymorphism or what is the typical energetic driving force for co-crystallization? To answer these questions, we make use of the wealth of information in the Cambridge Structural Database, which we combine with large scale simulations. This presentation focusses on co-crystallization and the information that can be learned from lattice energy calculations on a large set of observed structures. The study investigates the range of co-crystallization driving force for observed co-crystals: the stability of a co-crystal with respect to the separate crystallization of its



pure components and the structural features that lead to stable co-crystals. We also present attempts to apply machine learning to predict the co-crystallization driving force from molecular features alone. In this work, we apply crystal structure prediction to supplement the dataset to which machine learning can be applied.

## CINF 141

### Using knowledge-based tools to evaluate solid-form design and risk assessment

**Bhupinder Sandhu**<sup>1</sup>, *bhupindersandhu90@ksu.edu*, **Christer B. Aakeroy**<sup>1</sup>, **Susan M. Reutzel Edens**<sup>2</sup>, **Amy Sarjeant**<sup>3</sup>, **Shyam Vyas**<sup>4</sup>. (1) Dept of Chemistry, Kansas State University, Manhattan, Kansas, United States (2) Eli Lilly Co, Indianapolis, Indiana, United States (3) Cambridge Crystallographic Data Centre, Piscataway, New Jersey, United States (4) Center for Integrative Proteomics Research,, Cambridge Crystallographic Data Centre, Union Beach, New Jersey, United States

In the pharmaceutical industry, the vast majority of drugs are delivered in a crystalline form. The performance of any solid form of an active pharmaceutical ingredient (API) depends on a range of factors such as aqueous solubility, chemical and physical stability, hygroscopicity, particle control, mechanical properties etc. Despite an extensive amount of research in this field, it is still difficult to predict, simply from the chemical structure, the number of crystal forms of a compound and how to control and prepare unknown crystal forms. The Cambridge Structural Database (CSD) has developed a knowledge-based model within the Mercury software which utilizes hydrogen-bond propensity, hydrogen-bond coordination and full interaction maps of a molecule. In this study, three case studies will be presented to determine whether these knowledge-based tools can guide us towards predicting the right intermolecular interactions in individual molecules as well as in multicomponent systems such co-crystals. The knowledge gained from this study can be applied to more flexible, larger drug-like molecules with higher complexity.

## CINF 142

### What did the CSD ever do for drug discovery?

**John Liebeschuetz**, *John.Liebeschuetz@astx.com*. Astex Pharmaceuticals, Cambridge, United Kingdom

The Cambridge Structural Database provides the ultimate reference for evaluating the geometry and interactions of a putative drug design bound to its protein target. Retrospectively we illustrate how use of this information has positively contributed to successful drug candidate discovery at Astex.

As the database grows and its data content increases, new questions may become answerable. For instance, as well as “What is the best geometry for this ligand-protein hydrogen bond?” might we also ask “How is the preferred interaction of ligand A with residue B, influenced by adjacent residue C?”? We look to how the next million structures might be used to help the drug designer answer this and other important questions.

## CINF 143

### Improved structure-based drug design with one million small molecule crystal structures

**Bernd Kuhn**, *bernd.kuhn@roche.com*. F. Hoffmann-La Roche, Basel, Switzerland

For efficient structure-based drug design it is crucial to quickly assess the relevance of molecular conformations and interactions generated in a computer model. A reality check can be performed by comparing the outcome of a simulation with the wealth of experimental information that is contained in the Cambridge Structural Database (CSD). We will show how data mining of the CSD can be used to generate practical guidelines about preferred small molecule conformations, intramolecular hydrogen bonding motifs, and intermolecular interaction preferences. Examples from small molecule drug discovery projects at Roche where this knowledge was purposefully used in the design process will be highlighted.

Another important use of CSD data is in virtual compound or fragment libraries of drug design software tools.

Prominent examples are ReCore for inhibitor scaffold replacement, TorsionAnalyzer for assessing small molecule ligand strain, and CSD-CrossMiner for pharmacophore searching in protein binding sites. We will present case studies where these tools were used prospectively to improve binding affinity and selectivity, and to generate alternative inhibitor scaffolds.

#### CINF 144

##### Computational database for first-row transition metals

**Kevin Basemann**, *kdbasema@iastate.edu*, Alex Leffel, Aaron D. Sadow, Theresa L. Windus. Iowa State University, Ames, Iowa, United States

Computational chemists have long utilized experimental databases of thermodynamic properties for benchmarking method to determine how well new models calculated measured properties. These databases have also been utilized in fitting force-field and semi-empirical models, and with current trends in machine learning they are becoming increasingly valuable. The existence of highly accurate measured thermodynamic properties of gas phase molecules, in particular those containing elements from hydrogen to neon, are intrinsically linked to the development of methods even today. Unfortunately, as one moves further down the periodic table fewer and fewer of these highly accurate gas-phase thermodynamic properties exist. Upon reaching the first row transition metals for example, the largest compiled database of experimental or high accuracy computational data is around only 100 examples many of which are exotic species and do not reflect the sort of compounds most often utilized in computational or experimental systems. We have developed a new database that increases the amount of high accuracy computational calculations for transition metal complexes by an order of magnitude. By utilizing the NWChem software packages optimally designed parallelization methods, with the correlation consistent composite approach for transition metals (ccCA-TM) developed by Angela Wilson et. al. we have obtained a database that can be utilized in the development of new computational methods including novel density functionals geared toward transition metals and new machine learned models for application in the computational chemistry community. In addition, the wealth of computational information will enable correlation of properties and molecular orbital trends across multiple metal systems. Varied formats for accessing the data collected throughout this process has been geared to consider the needs of both the machine learning and computational chemistry communities are being utilized.

#### CINF 145

##### Open chemistry: Democratizing web-based chemistry databases

**Marcus D. Hanwell**<sup>1</sup>, *marcus.hanwell@kitware.com*, Chris Harris<sup>1</sup>, Alessandro Genova<sup>1</sup>, Muammar El Khatib<sup>2</sup>, Mojtaba Haghightalari<sup>3</sup>, Johannes Hachmann<sup>3</sup>, Wibe Dejong<sup>2</sup>. (1) Scientific Computing, Kitware, Inc., Clifton Park, New York, United States (2) 50F1650, LBNL, Berkeley, California, United States (3) Dept of Chemical and Biological Engineering, University at Buffalo, SUNY, Buffalo, New York, United States

Web-based chemistry databases have long been maintained by a few large centralized providers, and offered to the community largely as read-only collections. Research groups and institutions had the option of putting together something bespoke for their data, contributing to some of the centralized repositories, or using ad-hoc data management techniques locally. These often entailed the ad-hoc use of the file system as a simple database, and relied on group knowledge to communicate data/results. As computational resources grow even for relatively small groups the importance of democratizing web-based chemistry databases grows for both internal use and sharing results generated by groups more widely.

The development of an open source web-based database with JupyterLab for computational chemistry will be described. The core is implemented in the Python programming language, using MongoDB for metadata storage, and file system abstractions for large data storage. It features a RESTful API, a modern HTML5 React-based single page web application, and extensions to JupyterLab. It can run a number of quantum codes such as NWChem, PSI4, and more under development in addition to trained machine learning models. The Jupyter notebook can trigger computational jobs, with automated ingestion of data into the web-based database. The single page application provides access to individual results, including 3D visualization using WebGL-based rendering techniques. Public data can be searched, and authenticated users can do more. It also integrates with

a number of web-based chemistry databases, and aims to provide another option for offering data in increasingly data-centric workflows on the web.

The screenshot displays a JupyterLab environment with a Python notebook titled "Single Point Energy Calculation". The notebook contains the following code and output:

```
The mol.energy() method is a specialized helper function that adds 'task': 'energy' to the input_parameters dictionary, and then calls the generic mol.calculate() method internally.
```

```
[4]: result = mol.energy(image_name, input_parameters)
```

```
[7]: result.orbitals.show(mo='homo', iso=0.005)
```

Below the code is a 3D visualization of a molecular orbital, showing a red and blue lobed structure. The interface also includes a "Geometry Optimization Calculation" section with the following code and output:

```
The mol.optimize() method is a specialized helper function that adds 'task': 'optimize' to the input_parameters dictionary, and then calls the generic mol.calculate() method internally.
```

```
[8]: result = mol.optimize(image_name, input_parameters)
```

```
[9]: result.orbitals.show(mo='lumo', iso=0.005)
```

At the bottom, there is a "Pending Calculations" table:

ID	Code	Version	Status	Logs
5c93976ea8f9c0001fbc9c6	NwChem	6.6	QUEUED	

JupyterLab interface capable of running jobs, querying web-based chemistry database, and editing the database.

## CINF 146

### KinaMetrix: Web resource to investigate kinase conformations and inhibitor space

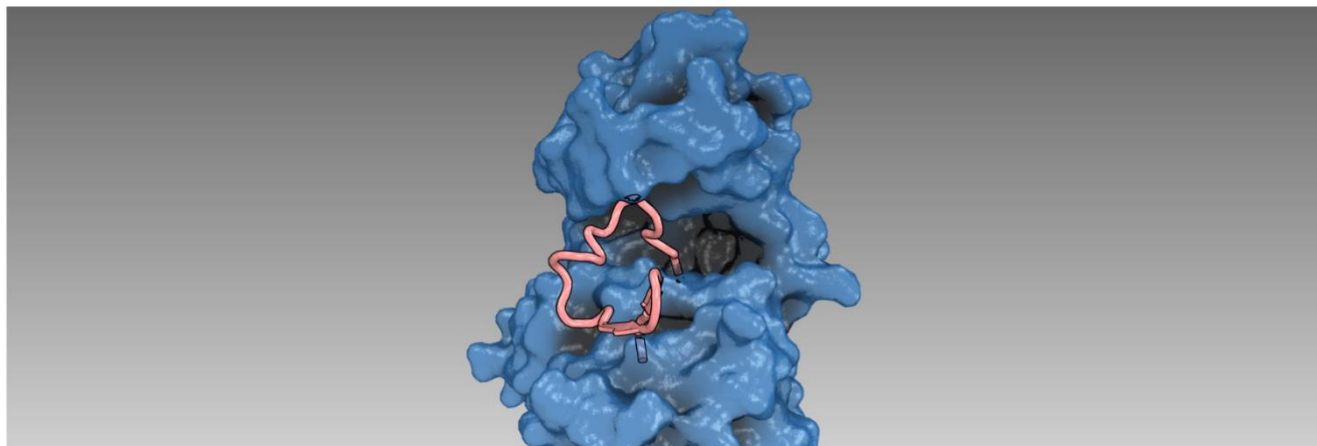
**Rayees Rahman**<sup>1</sup>, [rayees.rahman@icahn.mssm.edu](mailto:rayees.rahman@icahn.mssm.edu), **Peter M. Ung**<sup>2</sup>, **Avner Schlessinger**<sup>1</sup>. (1) Department of Pharmacological Sciences, Icahn School of Medicine, Floral Park, New York, United States (2) Pharmacology, Yale University, New Haven, Connecticut, United States

Protein kinases are among the most explored protein drug targets. Visualization of kinase conformations is critical for understanding structure–function relationship in this family and for developing chemically unique, conformation-specific small molecule drugs. We have developed *Kinformation*, a random forest classifier that annotates the conformation of over 3500 protein kinase structures in the Protein Data Bank. *Kinformation* was trained on structural descriptors derived from functionally important motifs to automatically categorize kinases into five major conformations with pharmacological relevance. Here we present KinaMetrix

(<http://KinaMetrix.com>), a web resource enabling researchers to investigate the protein kinase conformational space as well as a subset of kinase inhibitors that exhibit conformational specificity. KinaMetrix allows users to classify uploaded kinase structures, as well as to derive structural descriptors of protein kinases. Uploaded structures can then be compared to atomic structures of other kinases, enabling users to identify kinases that occupy a similar conformational space to their uploaded structure. Finally, KinaMetrix also serves as a repository for both small molecule substructures that are significantly associated with each conformation type, and for homology models of kinases in inactive conformations. We expect KinaMetrix to serve as a resource for researchers studying kinase structural biology or developing conformation-specific kinase inhibitors.



Kinases are dynamic proteins that can adopt several distinct conformational states.



That's why we built KinaMetrix.

A webserver to investigate kinase conformations and inhibitor space.

## CINF 147

### Structure-based search of chemical libraries with Pharmit

**David Koes**, [dcoes@pitt.edu](mailto:dcoes@pitt.edu). *Computational and Systems Biology, University of Pittsburgh, Pittsburgh, Pennsylvania, United States*

Pharmit (<http://pharmit.csb.pitt.edu>) is an open-source web server that provides interactive virtual screening of multiple chemical databases containing millions of compounds. Compounds can be searched for using pharmacophore and shape constraints and then energy minimized and ranked with respect to a receptor structure. We will describe the algorithmic underpinnings of Pharmit, describe its use, including describing successful prospective screens accomplished using Pharmit, and provide updates on its most recent developments.

## CINF 148

### Molecular malthusianism? Next three logs of the growth of purchasable chemical space

**John J. Irwin**, [jjj@cgl.ucsf.edu](mailto:jjj@cgl.ucsf.edu). *Pharmaceutical Chemistry, University of California San Francisco, San Rafael, California, United States*

Purchasable lead-like chemical space has grown from millions to billions in the last decade. Current growth suggests trillions will soon be only FedEx and a credit card (or PO) away. We will discuss how ZINC is struggling to adapt in this environment.

#### CINF 149

##### **Searching for similar reactions and molecules using the power of graph databases and the graph edit distance metric**

**Victorien Delannee**<sup>1</sup>, *victorien.delannee@nih.gov*, **Marc C. Nicklaus**<sup>2</sup>. (1) National Cancer Institute, Frederick, Maryland, United States (2) NCI-Frederick Bldg 376 RM 207, Natl Inst Health NCI Ft Detrick, Frederick, Maryland, United States

Looking for similar molecules and reactions in a database is often based on fingerprint methods where the Tanimoto scores between the query and the target are computed. In case of an exact match, the results can be relevant. However, they can be surprising and quite inaccurate when the objective is to find closely related molecules or reactions. In addition, these methods do not offer any flexibility regarding the result space returned. For example, it is not possible for the user to request only similar molecules and/or reactions that have fewer than two ring substitutions. Here, we present a new approach using the powerful graph database approach and the graph edit distance (GED) metric to search for similar molecules and reactions. The GED is a measure evaluating the number of operations (insertions, deletions, and substitutions) required to change one graph into another one. Calculating this GED between two graphs is usually a slow process. To address this, we implemented a fragmentation method shrinking the molecules. All fragments and their relationships are entered in the graph database allowing one to rebuild iteratively the original molecule. In addition, in order to find similar reactions, each reaction is transformed into a pseudo-molecule, i.e. a Condensed Graph of Reaction, which is also fragmented. Thus, in order to find similar molecules or reactions in the database, the query is fragmented by the same process and similar molecules or reactions are returned as a function of the GED. The procedure can be fully parameterized by the user (number of ring related operations, number of aliphatic chain/linker operations, number of atoms and bonds operations, substructure search with or without consideration of a threshold for number of operations, etc.). The strengths of this method are its accuracy and high flexibility.

#### CINF 150

##### **Helping chemists identify new opportunities during chemistry research: Building and turning high-quality data into actionable insights**

**Juergen Swienty Busch**, *juergen@swienty-busch.de*. Elsevier Information Systems GmbH, Bodingen, Germany

Reaxys and Reaxys Medicinal Chemistry are well known for their sophisticated data structure, which formed the basis of the UDM (Unified Data Model) published by Pistoia, and their rich and detailed experimental chemistry content sourced from journal articles and patents. This talk will discuss how Reaxys and Reaxys Medicinal Chemistry are being built and how chemists in industry and academia access the system and use the data for their daily research with a focus on drug discovery.

#### CINF 151

##### **Challenges and opportunities of delivering structural data on the web**

**Matthew P. Lightfoot**, *lightfoot@ccdc.cam.ac.uk*, **Ian Bruno**, **Suzanna Ward**. Cambridge Crystallographic Data Centre, Cambridge, United Kingdom

Thirty years on from the introduction of the World Wide Web, it is now taken for granted that data will be accessible on the web. The Cambridge Structural Database (CSD) contains 1 million crystal structures and has been in existence since 1965. Initially in book form, this soon evolved into electronic form allowing researchers to search and analyse the data through desktop software. In 2009 the Cambridge Crystallographic Data Centre

(CCDC) developed its first web interface to the CSD, WebCSD, to enable scientists to search the 470,000 structures available in the CSD at that time. Since then, not only has the database grown significantly but so too have the capabilities and complexities of our web-based resources.

This talk will highlight some of the tools we make available on the web, from Access Structures and WebCSD, to more complicated applications. We will look at work we have done to link and integrate to other data sources, ORCID records, raw data files, other data repositories and journal articles. We will go on to describe the importance of the quality control processes, validation systems and data integrity work we have in place to ensure that users can reuse, discover and ultimately trust the data we hold.

One of the hidden challenges we face in delivering data to the community is in making this data as freely accessible as possible while ensuring the value of our more advanced web-based resources and we will discuss some of the systems we have in place to ensure this.

We will conclude the talk by looking at how the web continues to provide many exciting opportunities for the delivery of data and how we are looking to exploit these to ensure that structural data is as accessible and discoverable as possible.

## **CINF 152**

### **Data at the Royal Society of Chemistry**

*Richard Kidd, kiddr@rsc.org. Royal Soc of Chem T Graham Hse, Cambridge, United Kingdom*

This presentation will focus on our efforts to build web-based databases to serve chemistry researchers in organic, analytical, natural products and reactions areas, as well as efforts to link our services to other platforms, such as journals publishing and APIs.

## **CINF 153**

### **PubChem: Improving access to chemical information on the web**

*Asta Gindulyte, mandroji@yahoo.com. National Center for Biotechnology Information, National Institutes of Health, Bethesda, Maryland, United States*

Since the launch in 2004, PubChem has become one of the most important chemical information websites for scientists, students, and the general public. Each month several million users worldwide access PubChem either interactively or programmatically. Given the complexity and amount of chemical information available in PubChem, however, there are many challenges in creating a website that provides users with good experience. This talk will describe recent efforts by PubChem team to take advantage of advancements in web technologies in order to create a streamlined, easier to use website, thus improving access to chemical information worldwide.

## **CINF 154**

### **Neural network potential for modeling radical reactions**

*Richard Messerly<sup>1</sup>, rames101@gmail.com, Peter St. John<sup>1</sup>, Adrian E. Roitberg<sup>3</sup>, Seonah Kim<sup>2</sup>. (1) Biosciences Center, National Renewable Energy Laboratory, Superior, Colorado, United States (2) National Bioenergy Center, National Renewable Energy Laboratory, Lakewood, Colorado, United States (3) Chemistry, University of Florida, Gainesville, Florida, United States*

Machine learning has become a ubiquitous tool in data science and is becoming increasingly common in fundamental computational chemistry. Of particular interest to this study are neural network potentials (NNPs). NNPs are trained to predict *ab initio* molecular energies and forces, similar to classical force fields, at a fraction of the computational cost of traditional quantum mechanical (QM) calculations. The recently developed family of NNPs known as the "Accurate Neural network engine for Molecular Energies" (ANAKIN-ME or ANI) is one of the most accurate NNPs for estimating conformational energies of compounds containing carbon, hydrogen, oxygen, and nitrogen (CHON). Because the previous version of ANI (ANI-1x) was trained to a data set



consisting of equilibrium and low-energy non-equilibrium closed-shell structures, however, ANI-1x is not well-suited for predicting bond dissociation energies and other high-energy reaction processes. This study applies active learning (query by committee) and transfer learning to retrain ANI-1x with an augmented data set. This work discusses three complementary approaches for automatically generating tens of thousands of additional structures that include radical species relevant to combustion chemistry. The retrained ANI potential (ANI-1rx) demonstrates considerable improvement in estimating bond dissociation energies and transition state structures. In conjunction with reactive molecular dynamics and minimum energy path sampling, ANI-1rx enables the rapid and accurate prediction of reaction mechanisms and activation energies necessary for traditional combustion kinetic modeling.

## CINF 155

### Predicting NMR in real-time through message-passing neural network

*Yanfei Guan<sup>1</sup>, yanfei@colostate.edu, Robert S. Paton<sup>2</sup>. (1) Chemistry, Colorado State University, Fort Collins, Colorado, United States (2) Chemistry Research Laboratory, University of Oxford, Oxford, United Kingdom*

Nuclear Magnetic Resonance (NMR) spectroscopy is a powerful technique used every day to elucidate the 3D structures of organic and biological molecules at atomic resolution. NMR spectra report detailed information on the local chemical environments of atoms in molecules which, in combination the experience and chemical intuition of experts, can be used to assign the connectivity and shape of molecules. Recent advances in computation have paved the way for the prediction of chemical shifts and the era of computer-automated structure elucidation is now within reach. However, existing computational methods are still very expensive and inherently low-throughput tools, which are not easily accessible to the broad chemistry community. To address these limitations, we have developed a deep learning model that leverages large theoretical and experimental databases. This model is able to predict the chemical shifts of organic and bioorganic molecules in a fully automated manner with high fidelity. Our new approach can be used to identify the structures of previously unidentified natural products and to correct human misassignments in real-time by non-experts. Access is provided through a product level web-app which has the potential to act as a “robot reviewer” for the chemical sciences.

## CINF 156

### Practical applications of deep learning to imputation of drug discovery data

*Thomas Whitehead<sup>2</sup>, Benedict Irwin<sup>1</sup>, Peter Hunt<sup>1</sup>, Matthew D. Segall<sup>1</sup>, matthew.d.segall@gmail.com, Gareth Conduit<sup>2,3</sup>. (1) Optibrium Limited, Cambridgeshire, United Kingdom (2) Intellegens Limited, Cambridge, United Kingdom (3) Cavendish Laboratory, University of Cambridge, Cambridge, United Kingdom*

We describe a novel deep learning method for data imputation that accepts both molecular descriptors and sparse experimental data as inputs to exploit the correlations between experimentally measured endpoints, as well as structure-activity relationships (SAR). The method can robustly estimate the confidence in each prediction and improves the accuracy of prediction over conventional quantitative SAR models. We describe practical applications to drug discovery, including pharma-scale collections comprised of over one million compounds and smaller, project-specific data sets. We illustrate how imputation of missing data, combined with the ability to focus on the most confident predictions, guides the selection of compounds and prioritization of experimental resources in hit-to-lead and lead optimization



## CINF 157

### Protein binding site fingerprinting for activity screening in machine learning

**Bastiaan Bergman**, *bastiaan.bergman@gmail.com*, Kate A. Stafford, Denzil Bernard, Stefan Schroedl. Atomwise, San Francisco, California, United States

Proteins can be characterized in several ways, schemes exist from very complete to more generalistic. On the one extreme, it is possible to give complete structural descriptions in terms of atoms and their 3-dimensional positions while on the other hand, one could look at just the amino acid sequence or even more broadly, use a protein family categorization method such as EC or PFAM. In practice the method one chooses to use is a compromise between computation efficiency and descriptiveness of the system. For activity screening using machine learning methods, the complete structural description can be too large, especially for small more simple methods such as decision tree based learning, drowning the features of importance with lots of structural information that does not have much impact in the binding process. We here propose a method for fingerprinting the binding site of proteins using molecular interaction fields expressed by only the amino acids exposed in the binding pocket. The methodology derives at a fixed length fingerprint bearing the binding site particularities, inspired by methods used for surface shape fingerprinting. This fingerprint can be used for similarity metrics as well as supervised and unsupervised learning and is not limited to comparisons with a reference protein.

## CINF 158

### Multiagent consensus equilibrium (MACE) for addressing the scaling challenges of computational chemistry

**James R. Ulcickas**<sup>1</sup>, *julcicka@purdue.edu*, Garth J. Simpson<sup>2</sup>. (1) Chemistry, Purdue University, West Lafayette, Indiana, United States (2) Purdue Univ Dept of Chemistry, West Lafayette, Indiana, United States

Investigations into the physical nature of complex phenomena increasingly depend upon the unified investigation of theoretical and experimental methods. However, as the size and complexity of these systems increases, quantum chemical simulation methods become increasingly costly, particularly in studies pertaining to extended solid state systems. Consequently, methods which enhance numerical accuracy while minimizing the corresponding computational scaling with respect to system size are highly desirable. Multi-agent consensus equilibrium is a criterion which may be used with convex sets of functions to fuse multiple sets of data, under the constraint that each agent describes the system in the same dimensionality.

In the presented work, MACE is utilized to fuse the output of multiple methods in computational chemistry to a single combined result. Preliminary studies integrate results computed from PM3, Hartree-Fock/6-31+G\*\*, MP2/6-31+G\*\*, B3LYP/6-31+G\*\*, and Hartree-Fock/cc-pVDZ methods, each computed in parallel. The MACE result represents the intersection of manifolds corresponding to each agent and consequently the recovered result is not equivalent to the arithmetic average of each individual model. Furthermore, the intrinsic uncertainty

within each agent (e.g., the slope of the potential energy surface in a geometry optimization) provides unsupervised feedback into the MACE criterion, dictating the equilibrium state. Preliminary MACE results optimizing the geometry of water and other small molecules demonstrating greater accuracy than gradient descent upon the individual agents will be presented. The capacity for integrating experimental measurements as agents and the greater applicability of the method discussed. Application of this data-science based approach to *ab initio* chemistry aims to improve the accuracy of quantum chemistry without increasing computation time beyond the agent with the largest scaling.

## CINF 159

### New approach to regression uncertainty analysis and applications to drug design

**Marvin Waldman**, *mwaldman3@san.rr.com*, Robert Clark. *Simulations Plus, Inc., San Diego, California, United States*

The ever-increasing use of *in silico* models in drug discovery has led to a concomitant need for assessing the reliability of predictions on an individual, rather than global, basis. Whether the models are used for estimating biophysical properties (logP, solubility), ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) characteristics, or biological activity, an ability to estimate the uncertainty (i.e., a confidence interval) in the prediction can help assess which compounds should be prioritized for synthesis, additional testing or property measurement. Recently, we developed an approach for estimating confidence in binary classification models composed of an ensemble of artificial neural networks (ANNE) based on the degree of concordance among the individual network predictions. We have now developed a similar treatment for ensemble-based regression models based on the variance of the individual network predictions. While the general idea of correlating prediction error with variance of multiple model predictions has been previously explored, our specific approach is new. It makes use of a general probabilistic analysis of the conditional error with respect to ensemble variance while employing generalized gamma distributions to model the underlying probability distributions. Applications of the method to various property models and how they can help assess model quality, aid in data curation, and drug discovery decision making will be discussed.

## CINF 160

### Insights from CSD crystallographic data applied to drug discovery

**Neysa Nevins**, *Neysa.2.Nevins@gsk.com*. UP1450, *GlaxoSmithKline, Collegeville, Pennsylvania, United States*

This will be a “tales from the trenches” style talk discussing examples of how Cambridge Structural Database (CSD) data has provided guidance for drug design in the context of hit to lead discovery. Examples will include tautomer insights, conformational analysis, and solubility assessment.

## CINF 161

### What fragment hit to follow and how? Using hotspots to prioritise chemistry resources

**Mihaela Smilova**<sup>1</sup>, **Peter Curran**<sup>2</sup>, **Chris Radoux**<sup>3</sup>, **Will Pitt**<sup>4</sup>, **Jason Cole**<sup>2</sup>, **Anthony Bradley**<sup>5</sup>, **Brian Marsden**<sup>1</sup>, *brian.marsden@sgc.ox.ac.uk*. (1) *University of Oxford, Oxford, United Kingdom* (2) *CCDC, Cambridge, United Kingdom* (3) *EBI, EMBL, Cambridge, United Kingdom* (4) *UCB, Slough, United Kingdom* (5) *Exscientia, Oxford, United Kingdom*

Fragment based drug discovery (FBDD) has established itself as a powerful tool for developing probe and drug candidates by rationally elaborating small chemical fragment hits into larger, optimised lead compounds. The use of X-ray crystallography technologies, such as XChem, as a medium-throughput screening tool for FBDD results in a wealth of structural data on low molecular weight molecules in complex with a protein target. Interpreting this data and distilling it into prioritised suggestions for the elaboration of fragment hits into leads with increased potency and selectivity for the target protein is currently a significant challenge to using this technique. Thus, computational methods to address this problem are in high demand.

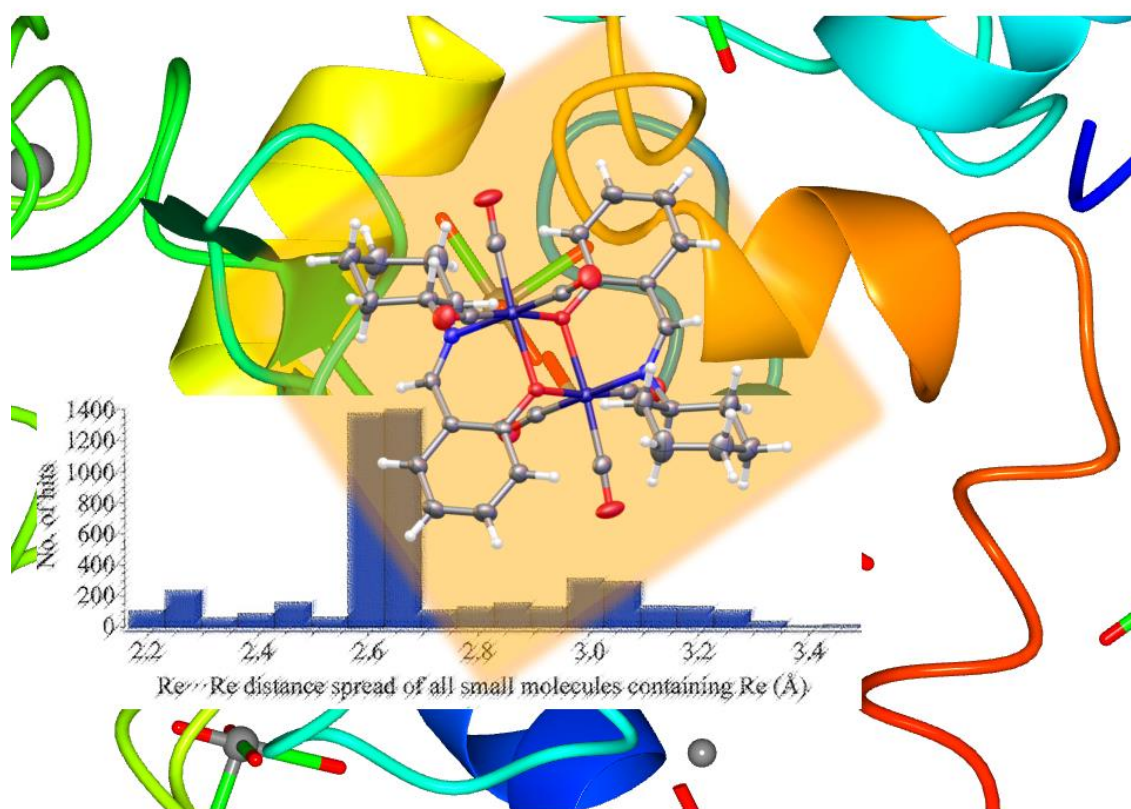
Introduced in 2016 by Radoux *et al.* based upon the CSD, fragment hotspot mapping highlights specific interactions a protein makes with a bound fragment. The highest scoring interactions predicted are often those that drive binding of the initial fragment; moderately scoring areas correspond to areas of the binding site into which initial fragment hits can be elaborated. Since crystallographic FBDD campaigns result in multiple structures of the same protein, often with evidence of intrinsic protein flexibility, we combine fragment hotspot maps calculated for each protein/fragment structure into an "ensemble map" for the protein target. Comparing ensemble maps of a target and a related off-target protein from the same protein family can then inform the prioritisation and onward elaboration of fragment hits into lead compounds with a tailored selectivity profile within a target's protein family. A set of diagnostics for assessing the quality of the method's predictions has been developed, based on ensembles of apo, fragment-bound, and selective inhibitor bound structures of proteins in two well-researched families: protein kinases and bromodomains.

## CINF 162

### Traversing interoperability: Drug development harnessing the CSD and PDB

*Alice Brink, brinka@ufs.ac.za. Chemistry, University of the Free State, Bloemfontein, South Africa*

Drug design, particularly the development of target specific radiopharmaceuticals which involves the selective receptor binding of a radioactive organometallic complex to a possible disease site involves multiple facets. Simple manipulation of the ligand system bound to the metal centre can significantly alter parameters such as steric and electronic character, chirality, stability, biological and hydro/lipophilicity properties. Our organometallic research utilising the group 7 transition metal triad of manganese, technetium and rhenium for nuclear medical imaging and therapeutic agents, includes the interactions with proteins using protein crystallography. This provides valuable structural information in a similar vein to fragment based drug discovery (FBDD). The domain of chemical versus biological crystallography has resulted in multiple discipline variations, such as incompatible software, data formatting and terminology. A key challenge which hinders research advancement is the lack of interoperability between chemical and biological crystallographic data. On the brink of the CSD's celebration of 1 Million crystal structures during 2019, this perspective will highlight the opportunities of harnessing the databases of the CSD and PDB, as well as the advantages of software which can convert organometallic small molecule structural data for use in protein refinement software. This multidiscipline approach to radiopharmaceutical development will include kinetic reactivity studies highlighting how subtle structural changes can significantly affect chemical reactivity – trends noticeable due the structural data made available through the CSD.



## CINF 163

### Semantic representation of CIF files: Mining crystal structures in the CSD

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

The Crystallographic Information File (CIF) format is a specification rich with data and metadata about a crystal structure. It has become an important part of the CCDC's curation of crystal structure information within in the CSD and a benchmark by which authors can and must submit their data to the CSD.

The CIF format is built upon a broad collection of data dictionaries overseen by the ICUR Committee for the Maintenance of the CIF Standard (COMCIFS). The open development/extension of the CIF specification coordinated by COMCIFS has allowed the crystallography community to trust CIF as an important mechanism to share crystallography data in a standard way.

This presentation will focus on the translation of the data dictionaries to ontological terms such that the data in CIF files can be represented semantically. As a result each piece of data and metadata can be represented as a Resource Description Framework (RDF) triple - statements that link a subject (s) using a predicate (p) to an object (o). CIF files from the CSD educational dataset were converted to JSON-LD files (a representation of RDF) and imported into a graph database. SPARQL queries were written to extract crystallographic features from the triple store data to evaluate the potential of the semantic representation of the CIF to elucidate additional information about crystal structures.

## CINF 164



## Learning from a database of a million crystalline materials

**Richard I. Cooper**, *richard.cooper@chem.ox.ac.uk. Chemistry, University of Oxford, Didcot, Oxon, United Kingdom*

There are many high impact applications of the knowledge in the CSD, such as geometry interaction statistics, and they have been used for chemical validation, development of knowledge-based potentials and scoring functions.

However, there are also interesting *one-sided questions*, which we don't *quite* have enough information to answer: e.g., which molecules will form stable crystalline structures? This question is implicit in knowledge-based structure prediction approaches, which score putative crystal structures by features that they share with known materials, and also explicit in statistical models to predict 'crystallizability'. In both cases, the CSD only has data from the successful results of crystal growth and diffraction analysis, but we have no record of any failed crystallization attempts.

An analogy of such one-sided problems is found in an analysis of the patterns of bullet holes in aircraft returning from missions during World War II. Engineers set out to find where to add armor to reduce loss of aircraft and crew. Perhaps counter-intuitively, the best places to add armor were those that had no recorded bullet damage; the patterns were not showing where planes were most likely to get shot, rather, where they could take a hit and still return home. To ignore the fact that there is information from the planes that did not return is to fall foul of survivor bias.

Negative results in classification problems such as crystallizability are particularly hard to obtain, because there is no clearly defined end point for a recrystallization experiment - there are always more and different experiments that can be attempted. I will present our approach to tackling these problems in predicting crystallizability, and discuss how we have progressed to using the CSD to help search for relationships between the molecular constituents of materials and other physical properties.

Even as the CSD starts to scratch the surface of recording the structures of molecular crystalline substances -- it has now  $10^6$  structures from an estimated  $10^{60}$  small molecule chemical space -- we should not worry that we will run out of materials to discover! Beyond  $10^{60}$  molecules the chemical space of multi-component crystals opens up a combinatorial hyperspace of new materials. I will present results from our research which can guide exploration of a combinatorial co-crystal experimental space using statistical analysis of an iteratively updated ML model in order to save time and resources.

## CINF 165

### From structure to crystallisation and manufacturing: Journey in applications of the CSD

**Chick C. Wilson**<sup>1,2</sup>, *C.C.Wilson@bath.ac.uk. (1) Department of Chemistry, University of Bath, Bath, United Kingdom (2) CMAC Future Manufacturing Hub, University of Bath, Bath, United Kingdom*

The reach of the Cambridge Structural Database (CSD) in structural systematics has been enormous; crystal engineering, for example, has benefited from this information in enabling the design of new solid-state products with predictable architectures. This essentially brings to centre stage the area of molecular (intermolecular interaction) recognition. Tools for recognising molecular complementarity have been developed, attempting to suggest the most likely pairs of molecules that will interact, for example in designing co-crystals and molecular complexes. This presentation, however, will focus on the evolution of applications of these tools in emerging contemporary application areas in molecular solid state, notably in pharmaceutical materials.

In this context, molecular complementarity approaches also encompass interactions that are not (yet) in the solid state, hence having potential application in driving choices in crystallisation science. The use of these principles in the design of crystallisation routes to desired solid forms, to identify possible additives that might influence critical aspects of the crystallisation process, will be discussed.

The use of additives can have a huge influence on the final product. They can contribute to polymorph



selectivity, allow access to elusive solid forms, or can be used to alter macroscopic morphology of the resulting product. Designing the optimal intermolecular interactions that will allow the correct choice of additive to influence crystallisation, without itself being incorporated in the final product, is a challenge. Structural informatics using the CSD have a critical role to play alongside other experimental and computational crystallisation and crystal growth methodologies. A number of examples of the use of additives in each of these areas will be highlighted.

The broader context of this approach will also be discussed. Crystallisation processes form a critical path in the production and manufacturing of many functional chemicals, including pharmaceuticals. Major research efforts across the world are seeking to optimise pharmaceuticals manufacturing, including by continuous manufacturing methods. This requires often sophisticated integrated workflows to be developed from candidate API molecule to final formulated product. Digital design is an increasingly important element of such workflows, complementing experimental approaches, and the role of the CSD in this context will be addressed.

## **CINF 166**

### **New frontiers beyond one million: New horizons for structural chemistry**

*Juergen Harter, jharter@ccdc.cam.ac.uk, Ian Bruno. Cambridge Crystallographic Data Centre, Cambridge, United Kingdom*

Having reached the tremendous milestone of one million organic and metal-organic crystal structures, generated by the scientific community over the last 50+ years, we explore new horizons for structural chemistry: Where next?

Many diverse insights and learnings have so far been possible with the present wealth and precision of data stored in the Cambridge Structural Database (CSD). This talk will look at where things are headed for the future. Further advances in chemical semantics will power advanced search capabilities. Better visualisation with advances in 3D, augmented reality (AR) and virtual reality (VR) will bring whole new scientific understanding. High levels of accuracy and precision of data will all be needed in order to get valid results out of the application of artificial intelligence (AI) and machine learning (ML) methods. Excellent high-quality molecular data and clear semantic understanding of it, with good metadata coverage, is a strong requirement for new and powerful algorithms and workflows to perform well. This furthermore enables us to more fully utilise the computational power the cloud flexibly provides.

The data and science strategy of the Cambridge Crystallographic Data Centre (CCDC) will have to explore how to go beyond storing structural experimental data to also storing vast amounts of physicochemical properties (in order to solve such a complex problem as solubility for example), and theoretical data coming from numerous computational approaches. On the scientific side, the insights derived from the bulk knowledge of structural chemistry are likely to be able to underpin scientific advances in a number of industry domains alongside biopharmaceutical – e.g. agrochemicals, nutraceuticals, catalysis, and functional materials, as well as nanotech. Our software suites and the plethora of functionalities of existing tools will have to evolve and be deployed such that these new areas can best be served.

For biopharmaceuticals the digital transformation trend continues: digital drug design and digital drug manufacture benefit hugely from being underpinned by the precise experimental data of one million small molecular structures. As the number of structures grows and more physical and calculated properties are associated with these, we look forward to these benefits being realised across other areas of industry and science.

## **CINF 167**

### **US EPA CompTox Chemicals Dashboard: Integrating chemistry and biology data to serve computational toxicology and environmental science**

*Antony J. Williams<sup>1</sup>, tony27587@gmail.com, Christopher Grulke<sup>2</sup>, Ann Richard<sup>1</sup>, Richard Judson<sup>1</sup>, Grace Patlewicz<sup>1</sup>, Imran Shah<sup>1</sup>, John Wambaugh<sup>1</sup>, Katie Paul-Friedman<sup>1</sup>, Jeremy Dunne<sup>1</sup>, Jeff Edwards<sup>1</sup>. (1) National*

*Center for Computational Toxicology, Environmental Protection Agency, Wake Forest, North Carolina, United States (2) National Center of Computational Toxicology, US EPA, New Hill, North Carolina, United States*

The U.S. Environmental Protection Agency (EPA) Computational Toxicology Program utilizes computational and data-driven approaches that integrate chemistry, exposure and biological data to help characterize potential risks from chemical exposure. The National Center for Computational Toxicology (NCCT) has measured, assembled and delivered an enormous quantity and diversity of data for the environmental sciences, including high-throughput *in vitro* screening data, *in vivo* and functional use data, exposure models and chemical databases with associated properties. The CompTox Chemicals Dashboard website provides access to data associated with ~900,000 chemical substances. New data are added on an ongoing basis, including the registration of new and emerging chemicals, data extracted from the literature, chemicals studied in our labs, and data of interest to specific research projects at the EPA. Hazard and exposure data have been assembled from a large number of public databases and as a result the dashboard surfaces hundreds of thousands of data points. Other data includes experimental and predicted physicochemical property data, *in vitro* bioassay data for over 4000 chemicals and 2000 assays, and millions of chemical identifiers (names and CAS Registry Numbers) to facilitate searching. Other integrated modules include an interactive read-across module, real-time physicochemical and toxicity endpoint prediction and an integrated search to PubMed. This presentation will provide an overview of the latest release of the CompTox Chemicals Dashboard and how it has developed into an integrated data hub for environmental data. *This abstract does not necessarily represent the views or policies of the U.S. Environmental Protection Agency.*

#### **CINF 168**

##### **Lessons learned in building the CompTox Chemicals Dashboard: Engineering a more sustainable web-based chemical database**

**Christopher Grulke**, *grulke.chris@epa.gov*, Antony J. Williams, Amar Singh, Jeremy Dunne, Ann Richard, Jeff Edwards. *National Center of Computational Toxicology, US EPA, New Hill, North Carolina, United States*

The development of the US-EPA's CompTox Chemicals Dashboard (<https://comptox.epa.gov/dashboard>) has led to a remarkable advance in the availability of information for chemicals of environmental interest. The Dashboard is a publicly accessible website providing access to data for ~900,000 chemical substances and is accessed by thousands of users per day. It provides access to a wide array of experimental and predicted physicochemical properties, *in vitro* bioactivity and *in vivo* toxicity data, product-use information and integrated linkages to a growing list of literature, toxicology, and analytical chemistry websites. However, to ensure standards of data quality and integrity, this assembly of data has required hundreds of hours of manual curation and data checking. Whereas the application has become a key public source of content and has integrated many disparate data sources, the degree of manual intervention to piece these data together prior to a public release for ever-increasing content sources has become growth-limiting and unsustainable. Here, we will describe how data streams are currently stitched together to support the Comptox Chemical Dashboard, the inefficiencies caused by the siloed data management practices employed in the past, and a vision for how our underlying data management systems will change to facilitate the support of a sustainable, "ever-green" dashboard in the future as our software applications utilizing these data continue to expand. *This abstract does not necessarily represent the views or policies of the U.S. Environmental Protection Agency.*

#### **CINF 169**

##### **In the world of free, is there room for subscription solutions?**

**Jonathan W. Taylor**, *jtaylor@cas.org*, Mindy A. Pozenel. *CAS, Columbus, Ohio, United States*

Budget cuts and budget restrictions provide a great incentive to find free chemical information on the internet. And, there is much readily available. So why would anybody want to pay for data? Why subscribe to SciFinder<sup>®</sup>? Why use CAS products? Because the availability of information is increasing exponentially, there are so many sources with different definitions of quality and few if any sources are comprehensive. Using the right tools to efficiently access and apply scientific information is critical. CAS combines domain expertise and human-curated

data with technology and data analytics to enhance an organization's digital transformation. By integrating proprietary data with CAS scientific information, organizations can better anticipate and seize opportunities, drive groundbreaking innovation, and solve tough challenges. Literature analysis fuels the insights needed to focus the artificial intelligence efforts on such initiatives as predictive drug design. This presentation will highlight CAS cheminformatics approaches.

#### **CINF 170**

##### **Google BigQuery for analysis of scientific datasets**

**Stephen Boyer**<sup>1</sup>, *skboyer@google.com*, **Ian Wetherbee**<sup>2</sup>, *wetherbee@google.com*, **Lutz Weber**<sup>3</sup>, **Jane Frommer**<sup>1</sup>. (1) Collabra Inc, San Jose, California, United States (2) Google, Mountain View, California, United States (3) OntoChem IT Solutions, Halle, Germany

Google BigQuery provides access to a massive cloud-based relational database with the potential to significantly enhance the way the scientific community uses information. BigQuery can process billions of rows and terabytes of data in tens of seconds, and scale to petabytes of data at very low costs. Data providers can contribute database tables for public access or protect their tables with Access Control Lists (ACLs) for private or subscriber only access. Data from tables uploaded by different providers can be jointly analyzed using SQL, and kept updated by each provider so users can focus on analyzing data instead of maintaining database infrastructure. To date much of the early content in BigQuery was of a non-scientific nature; accelerated efforts to increase the scientific content are underway. While BigQuery has broad applications across multiple domains, this talk will focus on efforts to develop and populate BigQuery with datasets relevant to the life sciences - chemistry and pharmaceuticals in particular. Resources under development will be presented to demonstrate how this publicly available platform and content can be leveraged to address questions that are difficult to answer otherwise. Use cases will demonstrate cross-mapping scientific content from disparate sources with capabilities beyond what is easily done today.

#### **CINF 171**

##### **Google BigQuery for analysis of scientific datasets: Interactive exploration and analysis of the data using KNIME analytics platform**

**Gregory Landrum**<sup>1</sup>, *greg.landrum@gmail.com*, **Martyna Pawletta**<sup>2</sup>, **Jeanette Prinz**<sup>2</sup>. (1) KNIME AG, Basel, Switzerland (2) KNIME GmbH, Berlin, Germany

The availability of scientific datasets in Google BigQuery (announced elsewhere in this session) opens new possibilities for the high-performance exploration and analysis of public life-sciences data. Here we present a number of practical examples of how one can effectively work with these datasets using the open-source KNIME Analytics Platform. KNIME is a general purpose tool, and it has particularly excellent support for chemical and biological data. KNIME is broadly used in the life sciences for interactive data processing, analysis, visualization, and machine learning. In this presentation, we will cover use cases relevant for chemical, biological, and pharmaceutical research. We focus especially on linking different datasets together to answer some relevant and interesting questions in the field of life sciences. We will show how the new data sources available in BigQuery are complementary to standard open data sources in our field and highlight the utility of scientific content in Google BigQuery. The workflows presented will all be made freely available so that they can be used and expanded by others.

#### **CINF 172**

##### **ChEMBL, SureChEMBL and UniChem: Web-based chemistry databases for drug discovery and chemical research**

**Ricardo Arcila**, *arcila@ebi.ac.uk*. ChEMBL Group, EBI - EMBL, Cambridge, United Kingdom

ChEMBL is a large, open-access drug discovery database that aims to capture Medicinal Chemistry data and

knowledge across the pharmaceutical research and development process. Information about small molecules and their biological activity is extracted from the full text articles of several core Medicinal Chemistry journals and integrated with data on approved drugs and clinical development candidates, such as mechanism of action and therapeutic indications.

**SureChEMBL** is a publicly available large-scale resource containing compounds extracted from the full text, images and attachments of patent documents. The data are extracted from the patent literature according to an automated text and image-mining pipeline on a daily basis. SureChEMBL provides access to a previously unavailable, open and timely set of annotated compound-patent associations, complemented with sophisticated combined structure and keyword-based search capabilities against the compound repository and patent document corpus.

**UniChem** is a low-maintenance, fast and freely available compound identifier mapping service. Its purpose is to optimize the efficiency with which structure-based hyperlinks may be built and maintained between chemistry-based resources. Primarily, this service has been designed to maintain cross references between EBI chemistry resources. These include primary chemistry resources, i.e., ChEMBL and ChEBI, and other resources where the main focus is not small molecules, but which may nevertheless contain some small molecule information e.g.: Gene Expression Atlas, PDBe. Currently we have cross-references to more than 158 million compounds in 39 different databases such as PubChem and ZINC.

The screenshot displays the ChEMBL Compounds search results for 'Dopamine'. The interface includes a search bar at the top with the query 'Dopamine'. Below the search bar, there are navigation options like 'Table', 'Cards', 'Graph', and 'Heatmap'. The main content area shows a list of 79 compounds, with the first 24 visible. Each compound entry includes its ChEMBL ID, name, maximum phase, full text count, and alignment score. A sidebar on the left provides filters for Type, Max Phase, #ROS Violations, Molecular Weight, and AllogP, each with a corresponding bar chart showing the distribution of compounds across the filter categories.

## CINF 173

### CIRCA: Your cheminformatics assistant

**Thomas D. Griffin**, [tdg@us.ibm.com](mailto:tdg@us.ibm.com), **Eric W. Louie**, **Stephen Boyer**, **Laura Anderson**, **Kristin Schmidt**, **Daniel P. Sanders**. IBM Research - Almaden, San Jose, California, United States

The CIRCA project (Chemical Information Resources for Cognitive Analytics) is an informatics cloud solution that applies computer curation, automated natural language processing (NLP) and machine learning techniques

to authoritative scientific content, with molecular information at the center. CIRCA's technical data corpus includes 22M patents, 28M MEDLINE abstracts, 2M full text articles and other public databases. Each document is annotated to extract chemicals, reactions, polymers, inorganic materials, drugs, species, gene sequences, substances and effects. We will detail some of the challenges in computer curation of this large data corpus, and the technical approaches we have developed to address them. Data is indexed as structured and full text fields, as well as for chemical structure similarity. We will show how this database can be a foundation for computational discovery through cognitive applications. We will introduce new features, including search capabilities from mobile devices.

#### CINF 174

##### **Transforming quality chemical data handbooks into web-based chemistry databases**

**John Rumble**<sup>1</sup>, [john.rumble@randrdata.com](mailto:john.rumble@randrdata.com), **Fiona Macdonald**<sup>2</sup>. (1) R&R Data Services, Gaithersburg, Maryland, United States (2) Taylor and Francis, Boca Raton, Florida, United States

Traditional printed chemistry handbooks were of two primary types: general comprehensive handbooks compiled by multiple authors under the leadership of a senior editor and topic-specific handbooks edited by one or two experts focused on a small chemistry sub-discipline. The success of these handbooks depended on several factors, including the completeness of coverage, currency of data, and degree of quality assessments. The most successful handbooks were in printed by many years, in some cases many decades and went through numerous updates, topic expansion, and author and editor changes. The longevity also presented challenges such as consistency as new information and data were added, harmonization of terminology, nomenclature, and data representation as new knowledge was developed, and categorization of fundamental property data in terms of new applications, e.g., how to group data relevant to diverse topics such as chemical safety, environment chemistry, and water chemistry. The digitization of chemistry data is now over four decades old and web-based chemistry databases more than 20 years old. The size and diversity of web-based chemistry databases is virtually impossible to determine, but the functionality of traditional handbook, formerly printed but now web-based, remains critically important to modern chemistry. This includes coverage, currency, and quality. And of these three, high quality continues to be of utmost importance. In this talk, we will explore how the *CRC Handbook of Chemistry and Physics*, now in its 100<sup>th</sup> Edition, has addressed quality issues in the digital age. Challenges include defining shared properties across hundreds of data collections authored by different experts, harmonizing data values with major data centers, coordinating data evaluation criteria across fields of different scientific maturity, and maintaining data quality as experimental procedures become automated.

#### CINF 175

##### **Does bigger mean better in the world of chemistry databases?**

**Antony J. Williams**<sup>1</sup>, [tony27587@gmail.com](mailto:tony27587@gmail.com), **Chris Southan**<sup>2</sup>. (1) National Center for Computational Toxicology, Environmental Protection Agency, Wake Forest, North Carolina, United States (2) TW2Informatics, Gothenburg, Sweden

The internet has changed the way we access chemistry data as well as providing access to data that can quickly proliferate and becomes referenceable. Web access to chemical structures and their integration with biological data has become massively enabling with numbers for UniChem, PubChem and ChemSpider reaching 157, 97 and 71 million respectively (at the time of writing). A range of specialist databases small enough to be curated have stand-alone utility and synergies when integrated into the larger collections. These include DrugBank, BindingDB, ChEBI, and many others. Databases of any size have inherent quality challenges but at large scale various forms of "noise" accumulate to problematic levels. The unfortunate consequence is that "bigger gets worse". This is particularly associated with large uncurated submissions from vendors and automated document extractions (even though these are high-value). Virtual enumerations and circularity between overlapping sources add to the problem. As a result of some of the noise in the larger databases the value becomes highly dependent on the specific applications. An example includes using the databases to support non-targeted analysis. This presentation covers examples of these noise and quality issues and suggests at least some options to ameliorate the problem. *This abstract does not necessarily represent the views or policies of the U.S. Environmental Protection Agency.*

**Evolution of the public chemistry databases: Past and the future**

**Valery Tkachenko**, *tkachenko.valery@gmail.com*, Rick Zakharov. Science Data Software, Rockville, Maryland, United States

Over the last few years we have seen a tremendous growth in various chemical databases. As a result we have now a variety of scientific resources, combined into a broad network and indexed through the directories like BioSharing and re3data. Such network, while growing quickly, is still in early days of adopting semantic web standards and does not yet support deep data indexing and discoverability, leave alone that mechanisms of intellectual properties protection are as simple as making data public or private at best. The lack of standards and well defined models to describe a scientific information structure even further inhibits free information flow which is essential for scientific discovery.

In this talk we will share our experience spanning through decades of building chemical databases like PubChem, ChemSpider, OpenPHACTS and National Database Services and will outline fundamental problems associated with chemical databases as such as well as data quality and approaches for the modern architecture of the large-scale chemical databases.