Inside this issue:

Message from the Editor 2
Message from the Program Chairs 3
CINF Symposia List - 2020 Spring Meeting 5
Awards and Scholarships 7
Reports
  Report on the Council Agenda for March 25, 2020 11
  Reproducible Data Analysis and Publishing in Chemistry with R: Creating a Workshop Experience During the ACS National Meeting 14
Book Review 22
Twenty-five Years Ago in Anaheim 26
Sponsor Announcements 28
2020 CINF Officers and Functionaries 35
Contributors to Spring 2020 CIB Issue 36
Technical Program—Symposia and Presentations 37
Technical Program—With Abstracts 47
Message from the Editor

Well, here’s the spring issue of the Chemical Information Bulletin, the one that would have been published before the National Meeting in Philadelphia.

With a few minor adjustments, this is the issue we planned to publish before the meeting was cancelled, including the technical program. I did omit several pages related to the schedule, such as the meetings and social events.

Please see Program Chairs Sue Cardinal and Ye Li’s letter on the next page to find see which symposia have been rescheduled for the upcoming meetings in San Francisco, San Antonio, and Atlanta. We also have a book review from Bob Buntrock, a “Twenty-five years ago” essay from Wendy Warr, and a report about the Reproducible Data Analysis and Publishing in Chemistry with R workshop that Ye Li, Steven Wathen, and Donna Wrublewski taught last August at the San Diego meeting.

If you are interested in contributing an article or feature to a future issue, please reach out to one of the CiB editors (David Shobe for summer, me for fall, and Judith Currano for winter). We are also looking for someone going forward to edit the spring issue, so feel free to contact any of us for more information.

Stay safe, everyone.

Teri M. Vogel, Co-Editor
UC San Diego
tmvogel@ucsd.edu
Message from the Program Chairs

What happened to the Philadelphia meeting? Unfortunately, the meeting to be held on March 22 - 25 at the Pennsylvania Convention Center was canceled because of the COVID-19 worldwide pandemic. Our organizers and speakers needed to decide which of 4 options they would take with their presentations/programs.

1. Speakers/ presenters could withdraw their presentation by emailing maps@acs.org.

2. Speakers/ presenters could post their slides and/ or a video screencast of their talk online in SciMeetings (https://Scimeetings.acs.org/ACSSpring2020) and still receive credit for the presentation.

3. Organizers could resubmit the symposium for a future ACS Meeting. Fall 2020 is a tight timeline with abstracts being due on April 20th. Spring 2021 is much more doable with the Call for Papers being due at the end of June 2020.

4. Organizers could set up a virtual webinar for their symposia. To our knowledge, none choose this path.

To date, a few people have deposited their presentation into SciMeetings. Two symposia will be rescheduled for the Fall 2020 National Meeting in San Francisco:

- Current State of FAIR Chemistry Data - Moving Forward became “Making Chemistry FAIRer”
- The poster session “Pushing the Boundaries of Chemical Information & Cheminformatics” was combined with “Multidisciplinary Thematic Poster Crawl: Perspectives on Chemical Health & Safety”

Several symposia are being rescheduled to the Spring 2021 meeting in San Antonio. In fact, MPPG decided to reuse the **Macromolecular Chemistry: The Second Century** theme:

- Cultivating Good Data Practices Among Chemists
- AI-based Big Data Application in Drug Discovery
- Data Exchange and Integration among Open Chemical Information Resources
- Macromolecular Information for the Second Century: Digital Representations, Identifiers, and Data Exchange
- The Current State of FAIR Chemistry Data - Other Disciplines, Chemistry and Physical Sciences
- Scientific Visualizations and Creative Presentations
- AI Meets Cheminformatics

One symposium will be rescheduled for the Fall 2021 meeting in Atlanta

- The Current State of FAIR Chemistry Data - General FAIR Landscape, Publisher Perspectives
Would you like to speak? The Call for Papers for the San Francisco meeting is open for abstract submissions from now until April 20. See https://callforpapers.acs.org/sanfrancisco2020/CINF. Please don’t wait until the last minute. Current plans are that the San Francisco meeting will be held on August 16-20, 2020.

Do you have an idea for a great program? Symposium and organizer ideas can be sent any time via email to committee members or via this Google form at https://forms.gle/BuGwyrwx3RdBWMXGA. Thank you to all that have suggested topical symposia in the past. Would you like to organize a program or recommend someone else? The next opportunities for these programs and organizers will be for the San Antonio meeting in March 2021.

If your program is ready to be submitted to the Call for Papers for the San Antonio meeting, please use this Google Form at https://forms.gle/9o3UdFH4fJgskVHJA. Program proposals will be due at the end of June 2020.

Thank you to past organizers, speakers, and past program planners. We couldn’t do this without your help.

Would you like to join the Programming Committee? 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. Please email Ye Li or Sue Cardinal at program@acscinf.org to discuss ways to participate.

Wishing you good health and peace during this unsettling time,

Ye Li, 2020 - 2021 Program Planner  yel@mit.edu
& Sue Cardinal, 2019 - 2020 Program Planner scardinal@library.rochester.edu
CINF Symposia List — 2020 Spring Meeting

S. Cardinal, Program Chair

SUNDAY MORNING

Useful Databases & Computation Methods for Studying Receptor-Ligand Interactions
R. J. Bienstock, Organizer; R. J. Bienstock, Presiding Papers 1-6

Macromolecular Information for the Second Century: Digital Representations, Identifiers & Data Exchange
L. R. McEwen, Organizer; V. F. Scalfani, Organizer; D. Wrublewski, Organizer; V. Scalfani, Presiding; L. R. McEwen, Presiding; D. Wrublewski, Presiding Papers 7-15

SUNDAY AFTERNOON

Data Exchange & Integration Among Open Chemical Information Resources
S. Kim, Organizer; S. Kim, Presiding Papers 16-23

Makerspace Safety: An Overview & Case Studies
N. Bharti, Organizer; S. Singh, Organizer; N. Bharti, Presiding Papers 24-29

SUNDAY EVENING

Pushing the Boundaries of Chemical Information & Cheminformatics
E. Alvaro, Organizer; J. N. Currano, Organizer; Papers 30-52

MONDAY MORNING

Machine Learning in Drug Discovery
S. Sirimulla, Organizer; S. Sirimulla, Presiding Papers 53-61

Current State of FAIR Chemistry Data: General FAIR Landscape & Other Disciplines
I. Bruno, Organizer; S. J. Chalk, Organizer; Y. Li, Organizer; L. R. McEwen, Organizer; N. Ruhs, Organizer; V. F. Scalfani, Organizer; I. Bruno, Presiding; N. Ruhs, Presiding Papers 62-72

MONDAY AFTERNOON

Machine Learning in Drug Discovery
S. Sirimulla, Organizer; S. Sirimulla, Presiding Papers 73-80

Current State of FAIR Chemistry Data: Publishers Perspectives
I. Bruno, Organizer; S. J. Chalk, Organizer; Y. Li, Organizer; L. R. McEwen, Organizer; N. Ruhs, Organizer; V. F. Scalfani, Organizer; I. Bruno, Presiding; N. Ruhs, Presiding Papers 81-89

MONDAY EVENING

Sci-Mix
S. K. Cardinal, Organizer; Papers 4, 27, 32, 35, 40, 42, 43, 45, 47, 51, 52, 57, 73, 106, 114, 130, 135, 143, 148, 161
TUESDAY MORNING

**Scientific Visualizations & Creative Presentations**
S. K. Cardinal, Organizer; H. Cheng, Organizer; S. C. Hayden, Organizer; P. J. MacDougall, Organizer; P. J. MacDougall, Presiding Papers 90-97

**Current State of FAIR Chemistry Data: Chemistry and Physical Sciences**
I. Bruno, Organizer; S. J. Chalk, Organizer; Y. Li, Organizer; L. R. McEwen, Organizer; N. Ruhs, Organizer; V. F. Scalfani, Organizer; Y. Li, Presiding; N. Ruhs, Presiding Papers 98-107

TUESDAY AFTERNOON

**Scientific Visualizations & Creative Presentations**

**AI Meets Cheminformatics**
N. Bharti, Organizer; T. Qin, Organizer; T. Qin, Presiding Papers 113-118

**Current State of FAIR Chemistry Data: Moving Forward**
I. Bruno, Organizer; S. J. Chalk, Organizer; Y. Li, Organizer; L. R. McEwen, Organizer; N. Ruhs, Organizer; V. F. Scalfani, Organizer; S. J. Chalk, Presiding; I. Bruno, Presiding Papers 119-124

WEDNESDAY MORNING

**AI Meets Cheminformatics**
N. Bharti, Organizer; T. Qin, Organizer; T. Qin, Presiding Papers 125-130

**Al-Based Big Data Application in Drug Discovery**
A. Zakharov, Organizer; q. zhu, Organizer; A. Zakharov, Presiding; Q. Zhu, Presiding Papers 131-134

**Cultivating Good Data Practices Among Chemists**
Y. Li, Organizer; S. Ward, Organizer; Y. Li, Presiding; S. Ward, Presiding Papers 135-145

WEDNESDAY AFTERNOON

**Al-based Big Data Application in Drug Discovery**
A. Zakharov, Organizer; Q. Zhu, Organizer; A. Zakharov, Presiding; Q. Zhu, Presiding Papers 146-154

**Cheminformatics for Chemists**
R. E. Belford, Organizer; T. Qin, Organizer; R. E. Belford, Presiding Papers 155-165
Awards and Scholarships

Chemical Structure Association Trust

Applications Invited for CSA Trust Grants for 2020 and 2021

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 pounds sterling equivalent to the U.S. dollar amount. Grants are awarded for specific purposes, and within one year each grantee is required to submit a brief written report detailing how the grant funds were allocated. Grantees are also requested to recognize the support of the Trust in any paper or presentation that is given as a result of that support.

Who is Eligible?

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

Which Activities are Eligible?

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

Applications must include the following documentation:

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

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

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

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

**Deadline for Applications:**

Application deadline for the 2020 Grant is April 17, 2020. Successful applicants will be notified no later than May 25, 2020. Application deadline for the 2021 Grant is April 16, 2021. Successful applicants will be notified no later than May 24, 2021.

**Address for Submission of Applications:**

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

2019

Vinicius Alves, University of North Carolina Eshelman School of Pharmacy, Chapel Hill (U.S.A.), was awarded $2,572 to present his research paper entitled “Multi-Descriptor Read Across (MuDRA) as a novel computational approach for Chemical Toxicity Prediction” at the 10th International Symposium on Computational Methods in Toxicology and Pharmacology Integrating Internet Resources that was held in Ioannina, Greece, from June 23-27, 2019.

Guilian Luchini, Colorado State University, Fort Collins, CO (U.S.A.), was awarded $1,399.00 to attend the American Chemical Society meeting that was held from August 24-29 in San Diego, CA, where he presented his research in applying often-overlooked corrections to DFT frequency calculations in an automated fashion.

Roi Rutenberg, Chemistry Department, Stanford University, Stanford, CA (U.S.A.), was awarded $2,072 for travel to visit the University of Illinois, Chicago in order to model molecular dynamic (MD) simulations at the Kral group as part of his research related to retrieving information about pEtN cellulose’s chemical structure as an individual compound and as a partner in future chemical reactions.

Monika Szabo, Monash Institute of Pharmaceutical Sciences, Monash University, Victoria, Australia, was awarded $2,000.00 for travel to attend two conferences at which she presented her research on drugs for myelofibrosis. The conferences were EFMC-ASMC’19 International Symposium on Advances in Synthetic and Medicinal Chemistry, held in Athens, Greece, from September 1-5, 2019, and the 20th SCI/RSC Medicinal Chemistry Symposium, held in Cambridge, U.K., from September 8-11, 2019.

2018

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

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

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

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

**Roshan Singh**, University of Oxford, UK, was awarded a Grant to conduct research within Dr. Marcus Lundberg’s Group at Uppsala University, Sweden, as part of a collaboration that he has set up between them and Professor Edward Solomon’s Group at Stanford University, California. He conducts research within Professor John McGrady’s group at the University of Oxford. The collaboration will look to consolidate the experiments studies on heme Fe (IV)=O complexes currently being studied by Solomon’s Group with future multi-reference 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 23-23, 2017. He works on the development of novel methodologies for the in-the-field detection of novel psychoactive substances (NPS), where chemical structure and information play a crucial role.

**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 modelling including; compound database development, virtual screening, docking, homology modelling, 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 modelling 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.
Reports

Report on the Council Agenda for March 25, 2020

The Council of the American Chemical Society will meet in Philadelphia, PA on Wednesday, March 25, 2020 from 8:00 a.m. until approximately 12:00 p.m. in the Grand Ballroom Salon E-H of the Philadelphia Marriott Downtown Hotel. All ACS members are welcome to attend, although only Councilors are permitted to vote. A continental breakfast is usually available at 7:00am for all attendees. There are six items for council action and these are noted below.

Nominations and Elections

President-Elect: The Committee on Nominations & Elections (N&E) has identified four nominees for the office of 2021 ACS President-Elect. They are as follows: Frank D. Blum, Mary K. Carroll, Michael P. Doyle, and Angela K. Wilson. The four nominees will answer questions at the Town Hall meeting that will be held on Sunday, March 22nd at 4:30pm in the Grand Ballroom Salon H of the Philadelphia Marriott Downtown Hotel. On March 25, Council will select the final two candidates whose names will appear on the fall ballot.

Other Elections

The Committee on Nominations and Elections has announced the list of nominees to represent District III and District VI on the Board of Directors for the term 2021 - 2023. Nominees for District III are Benny C. Chan, Teri Quinn Gray, Lynne P. Greenblatt, and Jeffrey L. Sturchio. Nominees for District VI are Janet L. Bryant, Paul W. Jagodzinski, Sharon P. Shoemaker, and Jeanette M. Van Emon. Ballots as appropriate are now handled by Survey & Ballot Systems, the ACS national election vendor, using the procedures developed by N&E and approved by the Council. The ballots have been emailed to the voting councilors in the two districts and the results will be announced at the Council Meeting in Philadelphia and in Chemical & Engineering News. On or before October 10, 2019, ballots listing the two candidates selected by the councilors for each district will be mailed to all members of District III and District VI for the election of a director from each district.

N&E also announced the election of Directors-at-Large that will be conducted in the fall. The candidates for a 2021 - 2023 term are Dawn A. Brooks, Wayne E. Jones, Jr., Kristin M. Omberg, and Carolyn Ribes.

ACS Dues for 2021

Council will vote on the recommendation from the Committee on Budget and Finance with regard to the 2021 membership dues (an increase of $5.00, from $175 to $180). The increases to ACS dues are based upon an escalator defined in the ACS Bylaws (Bylaw XIII, Section 3-a). The dues are calculated by multiplying the base (current) rate “by a factor which is the ratio of the revised Consumer Price Index for Urban Wage Earners and Clerical Workers (Service Category) for the second year previous to the dues year to the value of the index for the third year previous to the dues year, as published by the United States Department of Labor, with the fractional dollar amounts rounded to the nearest whole dollar”.

Report on the Council Agenda for March 25, 2020

Base rate 2020: $175.00

Change in the Consumer Price Index, Urban Wage Earners, Services Category:

| December 2018 CPI-W (Services): | $321.967 |
| December 2017 CPI-W (Services): | $313.441 |
| Change in CPI-W Index:          | 2.72%    |

2021 Dues, fully escalated: $175.00 x 1.0272 = $179.76
2021 Dues, Rounded: $180.00

Petitions for Vote

Petition to Clarify Amendments to the Standing Rules (Revised Version, Bylaw XI)

The intent of the petitioners is that amendment(s) to the Standing Rules must be reviewed by the committee responsible for the Governing Documents Function to ensure consistency with the ACS Governing Documents. This was an oversight in the new ACS Governing Documents, which requires such reviews for amendments to the Constitution and bylaws and the petitioners are now calling for an urgent correction.

Standing Rule V states that the Committee on Constitution and Bylaws (C&B) is responsible for the Governing Documents Function. The duties of C&B are listed in Sec. 1-b of this standing rule and include the following: (a) review provisions of the constitution, bylaws, and standing rules and initiate such action as may seem appropriate; and (b) interpret and initiate such action as may seem appropriate to eliminate conflicts in the constitution, bylaws, or standing rules. However, the procedure in Bylaw XI for making amendments to the standing rules contains no provision for C&B’s review before action by Council. This could lead to problems and inconsistencies within the standing rules. The standing rules include many of the provisions that formerly were in the Bylaws. C&B’s review was required for Constitution and bylaw changes in the former version of the ACS Governing Documents and is still required for amendments to the Constitution (Article XIII, Sec. 2) and amendments to the Bylaws (Bylaw X, Sec. 2, a). To ensure consistency within all documents in the ACS Governing Documents, including the standing rules, it is imperative that Bylaw XI be amended and that Council approve this urgent action petition.

Petition on Benefits and Dues

The ACS has been experiencing a decline in overall membership for more than a decade. In 2018, the Society was able to halt the decline and grow membership, but it required a great deal of effort for a small amount of growth: an increase of approximately 400 net members. In part, that strenuous effort was due to the limitations in the governing documents, which do not provide ample means to quickly adopt and sustain successful market-tested dues and benefit offers that attract new members. ACS recognizes that the society will need more flexibility with our processes and procedures to improve ACS membership for the future.

To that end, the Committee on Membership Affairs (MAC) has worked for several years to examine how ACS can meet the challenge of continued and sustainable growth through market testing initiatives. MAC continues to seek new ways to reconcile and deliver a valuable membership experience to our members by revisiting dues structures and exploring changes to provided benefits. Comments and
Report on the Council Agenda for March 25, 2020

observations offered at the Council special discussion in spring 2019 included asking the society to be more nimble and tailored in its approach to membership. The ability to deliver as such, and to provide members the value they seek in a professional association, has been hindered in many ways by a lack of control in both the dues setting process used for the last 30+ years and concurrently the benefits granted to the various membership types.

The partnership between MAC, the Society Committee on Budget and Finance (B&F), Council, other stakeholders, ACS members themselves, and staff, and the increased focus on enhancing our agility, are the keys to achieving membership goals. The petitioners seek to amend the ACS Standing Rules 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. MAC, with input from 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.

Adopting separate processes and procedures would make it easier to produce nimble change in response to market tests and most importantly provide the society with flexibility in the dues setting process to ensure the value of ACS membership remains in line with the relevant benefits we are offering. Further, it would allow for changes around benefits for current membership categories and their pricing and discounts, with full Council discussion and approval, at a pace accelerated over what is currently permitted. Thus, the society can conscientiously offer a more relevant approach to how benefits and dues are considered, but appropriately maintain proper oversight of the process.

Petition on the 2021 Schedule of Membership

The proposed Schedule of Membership is based on the ACS Governing Documents, for which a lot of text in the Schedule of Membership Specifics section was taken from the standing rules. The Petition on Benefits and Dues noted above proposes to remove the text from the standing rules so that it can be included in this schedule.

Town Hall Meeting

A Town Hall meeting organized by the Committee on Nominations and Elections is scheduled for Sunday, March 22 at 4:30 p.m. in the Grand Ballroom Salon H of the Philadelphia Marriott Downtown Hotel. It will highlight a Q&A session with the candidates for 2021 President-Elect. All ACS members are encouraged to attend. It is a great way to gather first-hand information and decide for whom you might want to vote in the fall election.

Note: The Council Agenda Book can be accessed at: https://www.acs.org/content/acs/en/about/governance/councilors.html

Respectfully submitted, February 26, 2020

CINF Councillors
Bonnie Lawlor
Andrea Twiss-Brooks
Reproducible Data Analysis and Publishing in Chemistry with R: Creating a Workshop Experience During the ACS National Meeting

Summary:

Script-based data analysis and authoring tools such as R can help researchers streamline their work-flow with research data and enhance reproducibility. At the Fall 2019 American Chemical Society National Meeting, two chemistry librarians partnered with a chemistry faculty member to offer a hands-on computer programming workshop. The workshop used the R programming environment, and was structured for beginners with no previous experience with R. Three main topics were covered during the five-hour session: introduction to R, working with QSAR data, and creating a report using R Markdown. This poster will cover the workshop’s logistics, instructional design, and content, and discuss feedback received from participants. Suggestions for future improvements will also be discussed.

1. Event Goals and Target Audience

The goal of this event was to provide an introduction to the R programming environment for chemists. Using such script-based data analysis and authoring tools will allow chemists to streamline their workflow with research data and enhance reproducibility. Cultivating these good practices enables researchers to improve transparency and reproducibility of their research process. It was a hands-on learning experience for participants and also gave them resources to share with their communities.

The intended audience was attendees at the ACS National Meeting, and included chemical information professionals, students, and librarians.

2. Description of event

A. Location and time

The event was held Saturday, August 24, 2019 from 9am to 2pm, with two short breaks and a longer break for lunch. This date was chosen in hopes that many people traveling to the American Chemical Society Fall meeting would be able to attend as part of their conference travel. The workshop was held at the Omni Hotel, an ACS Conference Hotel and in close proximity to the convention center.

B. Instructors

The three facilitators were Dr. Ye Li (Massachusetts Institute of Technology), Prof. Steven Wathen (Siena Heights University), and Dr. Donna Wrublewski (California Institute of Technology).

Ye is the Chemistry & Chemical Engineering Librarian at MIT. She is a certified Carpentries instructor and has taught many workshops on R, python, Git for reproducible research. Steve is Professor of Chemistry at Siena Heights University in Adrian, Michigan. He incorporates R programming into his lab classes, and one of his lab experiments is the basis for this workshop. Donna is the Chemistry & Chemical Engineering Librarian at Caltech. She is also a certified Carpentries instructor, and has helped develop and teach “Author Carpentry” lessons that utilize R Markdown for creating reproducible reports.
Reproducible Data Analysis and Publishing in Chemistry with R: Creating a Workshop Experience During the ACS National Meeting

3. Preplanning

A. Lesson plan

The workshop was broken down into three parts:

(1) Introduction to R and the RStudio Environment (taught by Ye)

(2) Working with QSAR data (taught by Steve)

(3) Creating a reproducible report using R Markdown (taught by Donna)

The lesson plans were stored on the Github repository to allow for collaborative editing by all three of the team members. Github also integrates with RStudio. Two options were available for learners to access the lessons: (1) require participants to download R and RStudio in advance of the workshop, and provide configuration instructions for the setup as well as instructions for accessing the Github repository; or (2) using the RStudio Cloud platform, which would allow learners to arrive with just a laptop and the ability to connect to the internet. All the lesson scripts, and required packages, would be pre-loaded into the learning space, so learners would only have to connect to the RStudio Cloud website and log in. Although the RStudio Cloud service was still technically in beta, the instructors’ prior experience with the platform led to the decision to use it, as it would eliminate potential problems with troubleshooting individual participants’ laptop software setups. It should be noted that RStudio Cloud does allow for users to download course materials through the interface. Participants were also given the link to the Github repository (https://github.com/YeLibrarian/R_ACS_20190824Materials) to access the materials after the workshop.

B. Pre-Workshop Survey

A pre-workshop survey was sent to registrants as their registrations were received and processed. Fourteen responses to the survey were received. 64% of registrants had no experience with R and/or RStudio, with 36% indicating some experience. Other data analysis or programming software with which registrants indicated familiarity included Python, Matlab, and SPSS, as well as languages such as Fortran. Finally, registrants were asked what they hoped to gain from the workshop. Responses included gaining familiarity with R, how to use R for reproducible research and document preparation, and observing workshop teaching methods.

C. Cost

The main expenditures were room rental and internet, beverage service, and lunch. Providing refreshments and lunch helped to keep the participants nearby and avoid delays getting food, as there were not many establishments nearby that were open the day of the event. This helped to keep the workshop timing roughly on schedule. The instructors donated their time, and received no other compensation.

Attendees were asked to pay a $40 registration cost in order to attend the event. This was done to (1) encourage people to attend and not cancel at the last minute; and (2) help cover the facilities and food costs. The cost of the workshop was covered by a generous sponsorship of $4000 from ACS Publications, as well as the participant registrations (approximately $700).
Reproducible Data Analysis and Publishing in Chemistry with R: Creating a Workshop Experience During the ACS National Meeting

D. Marketing

Emails were sent to various listservs, including CHMINF-L, CHED, and local ACS sections. Many Chemistry Librarians helped promote the workshop to their graduate students.

4. Execution

A. Space Logistics

A total of 21 people registered for the event and 18 attended.

The hotel conference room was adequately prepared with tables and power outlets, and WiFi was also available specifically for attendees. Counter space was available outside the room for beverage and meal setup, making it more convenient and time-saving for learners. Restrooms and other private spaces were located nearby, and the room was on the first floor and readily accessible for those using mobility assistive devices.

B. Content Delivery

The schedule for the day is included in Appendix A. The workshop was taught with a Carpentries-style approach. While one instructor was teaching at the front of the room, the other two instructors made their way around the class to answer questions one-on-one. This was facilitated by using pink and blue Post-it notes that were distributed to each participant. Instructors would ask participants to put up a blue Post-it when they had completed a task, or a pink one if they had a question or needed help.
Reproducible Data Analysis and Publishing in Chemistry with R: Creating a Workshop Experience During the ACS National Meeting

The first section was taught by Ye and focused on giving learners a basic introduction to using R in the RStudio IDE (integrated development environment). It began with an overview of reproducibility in chemistry research and factors causing irreproducible results. The discussion then covered the components that make research computationally reproducible and highlighted how tools like R could be used to create automated and open workflow. The slides for the overview are available at https://rpubs.com/YeLibrarian/slides20190824. Hands-on practice started with constructing an R project in the RStudio Cloud environment using folders to separate raw data, scripts or codes used to process data, figures, and documents that are generated during the data analysis and writing process. Using the “assignment” function in RStudio Cloud environment, students were able to “fork” a file with an outline and followed Ye to practice using R to do mathematical calculations, logic comparisons, assigning values to variables, naming variables, importing tabulated data as a data frame, subsetting data frames, calculating basic statistics, plotting a calibration curve, and saving a generated plot to a figure. A set of UV-Vis absorbance calibration data from a real lab experiment was used to practice these fundamental commands.

The second section was taught by Steve and covered using R to create QSAR (quantitative structure-activity relationship) models to predict the pKa of a series of substituted benzoic acids based on the sigma values of the substituents. This example was based on a laboratory activity that Steve teaches to his sophomore undergraduate students. This activity was chosen because it is a fairly straightforward application of chemical modeling. Enough background on QSAR was discussed so that the learners could see the benefit of using R to perform the analysis. This section built on the material Ye introduced by importing a data set, calculating statistics and a linear model based on the data, and displaying basic plots to examine the data and relationships. After demonstrating one model based on all descriptors, learners created models for other scenarios with different sets of descriptors, thus allowing them to repeat the steps they were shown.

The third section was taught by Donna and walked learners through creating a lab report and a slide presentation using an “R Markdown” document. “Markdown” is a simplified plain-text-based formatting scheme that is common on sites such as Github. “R Markdown” is a version of Markdown that allows for R programming code to be embedded into the document. Markdown documents can be exported into other formats such as HTML for websites, Microsoft Word documents, and slide presentations by using the appropriate conversion program. Citations can also be included. The RStudio interface makes this fairly straightforward, and thus conveys two main advantages:

1. The author only needs to write one document, and it can be output to many different formats, as opposed to making an entirely different version of the work for each format (such as a Word document, a PowerPoint document, an HTML document, etc.). This saves a lot of time and reduces the chance of copy/paste errors.

2. Embedding R code directly into the document allows for better reproducibility and easier updating. If a change is made to the code, any resulting changes are reflected directly in the output, as opposed to having to make a new chart or graph, and pasting into the finished document. Having the code linked directly to the output allows the reader to clearly see the analytical process.

The final product for this lesson was an R Markdown document that learners could output to a Microsoft Word file and to a slide presentation file. The R Markdown documents were heavily annotated in the hope that learners could easily follow along with the lesson right in the document in
Reproducible Data Analysis and Publishing in Chemistry with R: Creating a Workshop Experience During the ACS National Meeting

which they were working.

There was more content scheduled than actual time allowed, partly due to not being strict about the timing of breaks and lunch. As such, some of the content was a bit rushed at the end and was noted by the attendees.

All content is available at: https://github.com/YeLibrarian/R_ACS_20190824Materials

C. Feedback

Workshops taught using the Carpentries format collect feedback periodically throughout the class using the aforementioned Post-it notes. Learners are asked to contribute one thing about the workshop that could be improved, and one thing that is working well for them. For this workshop, two rounds of feedback were collected - one before lunch, and one at the end of the day. Multiple rounds of feedback allow instructors to respond to concerns in real time. Suggestions for improvement included slowing the pace and/or increasing the length of the workshop, and providing more handouts. Aspects that worked well for learners included multiple instructors and helpers to answer questions, the Post-it note system for feedback during class, and the content overall, which the learners felt was helpful for those learning R. The full collected comments are included in Appendix B.

We also sent a post-workshop survey and 10 participants responded to the survey. Among them, 70% felt more confident in using R after attending the workshop; 60% would be willing to attend a 1.5-day workshop instead of a half-day session to cover more content; 90% thought the calibration curve example was effective for learning R; 56% thought the QSAR example was effective for learning R. Participants expressed their enthusiasm with the session and also expressed the interest to explore other types of chemistry data using R.

Steve and Donna reviewing feedback.
5. Conclusions and Recommendations

Overall the organizers felt the workshop was a success and accomplished the main goal of introducing the benefits of reproducible reporting in chemistry. Logistically the event went without incident. The room was comfortable, the wireless internet was functional, and food service arrived on time and was satisfactory. This event will hopefully be a viable model for future workshops at national meetings.

Should another workshop of this type be planned, some improvements for consideration are listed below.

(1) Several learners suggested a longer workshop. Alternately, content could be chosen more judiciously, and the timing of breaks could be altered.

(2) Provide print handouts for learners to follow during the lesson. Although the scripts and files used were heavily marked with comments intended to help the participants, some remarked that it was easy to get lost as content scrolled off the screen. A printout would have been easier to follow.

(3) One issue was the low color contrast on the monitor. Using RStudio, comments in scripts appear as green, which was difficult to see on the particular screen that was in the workshop room.

6. Acknowledgements

ACS Publications is gratefully acknowledged for their generous financial sponsorship of this event, without which it would not have been able to go forward. The ACS Division of Chemical Information is also deeply thanked for their commitment and support, including logistical assistance and contingency funding.

Appendix A - Schedule (approximate)

8:30 - 9:00 AM - Pre-workshop time for computer setup and troubleshooting
9:00 - 10:15 AM - Introduction to R (Ye)
10:15 - 10:30 AM - Break
10:30 - 11:15 AM - Introduction to R (Ye)
11:15 - 11:45 AM - QSAR Modeling (Steve)
Reproducible Data Analysis and Publishing in Chemistry with R: Creating a Workshop Experience During the ACS National Meeting

11:45 AM - 12:15 PM - Lunch break
12:15 - 1:00 PM - QSAR Modeling (Steve)
1:00 - 1:15 PM - Break
1:15 - 2:00 PM - Reproducible Reporting with R Markdown (Donna)

Appendix B - Post-it Note Feedback

Break 1

Needs Improvement:

- A little fast pacing near the end of the session
- A comprehensive review document with all the commands is appreciated!
- More examples would be better
- Need more time to ask questions
- It went a little too fast at the end
- Maybe include a pre-study part, ask to pre-study some simple tasks, e.g. ask to learn string, int, num, before the session, simple but useful
- It would be useful to have a more complicated, fully finished data analysis script to see what this looks like in practice
- Either list of codes you will use or printed script to refer to
- Please don’t include shortcuts as part of an introductory course
- Jumped around a lot and rushed at the end
- Paper handout with list of tasks would have been helpful as we navigate the screen
- Cloud environment crashes on memory limit!
- Only thing I wish is that it was a longer/2 day workshop. I have a lot more questions about R I would like to know but feel like there is not enough time

Working Well:

- Worked well
- Pace was good
- Good examples
- Generally the pace was fine
- R is like other languages
- It’s great that everything you touched on was already written out as comments in the workspace
- It is reassuring to have individual help available
- I love the sticky note idea! Also that there are 3 of you here helping it makes it easier to ask questions
- Great introduction to someone completely unfamiliar with R
- Blue & pink stickies
- Hands-on
- Tricks such as Tab
- Easy to start from the tutorial file
Reproducible Data Analysis and Publishing in Chemistry with R: Creating a Workshop Experience During the ACS National Meeting

- Loved the comments in the RStudio file! They helped me when I was lost
- The comments throughout the practice script really helped me keep on track
- Ye is very clear and organized in presentation
- Great! Very practical, exactly the knowledge I need to immediately apply into research data processing

Final

Needs Improvement:

- Would be helpful if y’all created a document with a transcript of major points. Good for visual/textual learners.
- Some issues viewing the screen, particularly the green text.
- Another 1-2 hours would have been helpful.
- Would have liked more about the packages
- Maybe have a document online that has a walk-through of steps to be referenced if there is an issue
- More time for Power Point formatting would have been helpful
- It’s a little bit rushed in time
- I wish it was longer!!

Working Well:

- R Markdown rocks - thank you!
- It was very helpful to see the Word document right after typing
- Nice primer
- Organization
- Exposure to different online (ideas?)
- Liked the R Markdown
- Great! Very helpful thank you
- Very informative! Thank you!!! We need more of this…
- Wow. I loved this workshop, thank you! Very informative and very much worth $40
- Appreciate the one-on-one help! Thanks!
- The afternoon session is extremely practical and useful, especially the markdown
Book Review


As befits a review of a second edition of any book, this review will emphasize differences between the first and second editions. As the author points out in the preface to the first edition (1), that edition was well received, and reviewed in more than 60 journals and magazines. The publisher encouraged Scerri to update the first edition to coincide with the International Year of the Periodic Table (IYPT) and the sesquicentennial of Mendeleev’s publication of his periodic table. Scerri received extensive communications with suggested corrections and improvements most of which were incorporated into the second edition.

In the first edition, the chapter notes were listed at the end of the book, by chapter. In the second edition, chapter notes appear at the end of each chapter, but are accumulated into a bibliography at the end of the book, alphabetized by author, a more efficient and usable method.

In the second edition, four chapters have been added to the original 10 which are more or less intact, but updated to some extent. Some of the content of the new chapters is extracted and expanded from the original 10. For example, Chapter 11, titled The Seven Last Infra-Uranium Elements to be Discovered, is a 48-page expansion of a topic covered in only a few pages in the first edition.

The Introduction covers the cursory definition of the periodic table, and the development of the concept of the elements beginning with the Greeks and alchemy and the evolution of the table and its philosophical implications.

To summarize the original 10 chapters, Chapter One describes the timeline of the discovery of the elements, the periodic law, ordering of the elements, the modern periodic table, recent changes, understanding the periodic system, and extensions such as periodic tables of the properties of compounds rather than elements. (Not discussed in the book, the cultural ubiquity of the periodic table is shown by a number of recent, whimsical “tables” such as the Periodic Table of Deserts).

Chapter Two is titled “Quantitative Relationships Among the Elements and the Origins of the Periodic Table”. Discussed are quantitative analyses, equivalent weights, and early atomic theories including the law of triads posited by a number of chemists before Mendeleev. Chapter Three covers discoverers of the periodic system, not just one but several (five or six) after the Karlsruhe conference on atomic weights. The process was an evolution rather than an isolated proclamation.

Chapter Four discusses Mendeleev in depth, including his early life and scientific career. Beginning with his “Crucial Discovery” the evolution of his periodic table is discussed. The survivability of the elements when combined to form compounds has a long history. Other scientists had attempted to make predictions of missing elements. Mendeleev was far more successful, predicting more elements and revising atomic weights where needed. The answer to the question, “Was Mendeleev a Reductionist?”, is probably “No”. He was not only concerned with chemical properties.

Chapter Five delves further into the development and acceptance of Mendeleev’s table, his having the most impact compared to those of the six co-developers. Prime examples of correcting atomic weights are beryllium (a long-standing controversy), uranium, tellurium and iodine. “Eka” element predictions and subsequent discoveries due to his work include gallium scandium, and germanium. He was not perfect, only about half of his predictions were valid, (i.e., the predicted elements were actually discovered), but these “failures” did not detract from the acceptance of his table and were outweighed by the successful predictions. (Scerri points out the differences in reasoning between physics and chemistry. The understanding of the former usually proceeds unambiguously from general principles whereas chemical reasoning does not. That reasoning is more inductive and large amounts of data must be considered before reaching conclusions.) Acceptance of Mendeleev’s periodic table is discussed including the joint awarding of the Davy medal to both Mendeleev and Lothar Meyer. Further aspects of “The Power of an Idea” conclude the chapter.
Chapter Six discusses the impact of research on the atomic nucleus on the development and evolution of the periodic system and is the beginning of the impact of physics on understanding the system. X-rays, radioactivity, and the discovery of the nucleus itself play a role. The discoveries come to a head with the concept of atomic number by van den Broek and experimental demonstration by Moseley. For the first time the exact number of existing elements could be determined, as well as gaps to be filled. Debates still existed, partially resolved by the discovery of isotopes.

Chapter Seven transfers focus from the nucleus to the electron, its discovery and early models of the atom. Emission spectroscopy could not be explained until Bohr explained the periodic system via arrangements of electrons in the atom, i.e. the quantum theory of the atom. This was followed by his second theory, the Aufbau Principle which was followed by additional developments including the Pauli Exclusion Principle.

In Chapter Eight, the contributions of chemists, rather than physicists, towards explaining the periodic systems are described. The contributions of Lewis, Langmuir, and Bury are discussed as well as the debates between the latter and Bohr over the prediction and discovery of element 72, hafnium.

Chapter Nine is the long and complicated story of the evolution of Bohr’s quantum theory to quantum mechanics. The contributions of Heisenberg and Schrödinger are described leading to the Hartree-Fock method. Also discussed is the assignment of electronic configurations via the Aufbau, Hund, and Pauli Exclusion principles including anomalous configurations, ab initio calculations, and the density functional approach. Chemistry has neither been completely reduced to quantum mechanics nor has it failed to do so, it is a work in progress.

Astrophysics, nucleosynthesis, and the evolution of the elements are discussed in Chapter Ten. The stability of nuclei and its relation to cosmic abundance are also covered.

The chapters new to the second edition begin with Chapter Eleven with the description of the seven last infra-uranium elements to be discovered, namely 91, protactinium; 72, hafnium; 75 rhenium; 43, technetium; 87, francium; 85, astatine; and 61, promethium. All are very interesting sagas with several blind alleys, defunct postulated elements, and rivalries. The discovery of the Oklo “natural” reactor is also described.

Chapter Twelve covers synthetic elements. Uranium was known to be the heaviest naturally occurring element. In the quest to discover transuranic elements, experiments with uranium, facilitated by the invention of the cyclotron, by bombardment with neutrons led to some possible production of transuranium elements. More significantly, these experiments led to the discovery of nuclear fission. Subsequent successful syntheses, first of neptunium, led to the discovery that such elements were part of a new series of elements, the actinides. Discovery of plutonium was followed by the discovery of elements 95-100 but nuclear stabilities were decreasing rapidly. Elements with an odd number of protons are less stable than those with an even number since spin-pairing of protons takes effect. In addition, “magic numbers” of nucleons have a stabilizing effect. The use of linear accelerators facilitated the synthesis of elements 101-106 and higher. The primary reason for alterations in the chemistry of ultra-heavy elements is relativistic effects. The increasing velocities of inner-shell electrons result in increased screening of the outermost electrons and behavior of the outermost electrons determines the chemistry. Assignment into predicted slots on the growing periodic table begin to be anomalous and recent observations of element 118 may indicate that assignment of slots on the periodic table may be impossible since the existence of discrete electron shells may disappear. As more research groups became involved, competition in discovery priority and naming occurred.

As a matter of interest, this update of the first edition in Chapter Eleven for missing elements was already done with Scerri’s book, A Tale of Seven Elements, (2) but not acknowledged here. The elements described in the second edition and their order of their presentation is the same. Chapter Eleven is 47 pages compared with the 270 pages of A Tale of Seven Elements but Chapter Eleven also covers the synthetic elements described in Chapter Twelve (35 pages) and is more current since
Chapter Thirteen discusses and illustrates the variety of forms of the periodic table. Expanding on the discussion of table forms in Chapter One, the three basic forms of short, medium-long, and long leads, of course, to differences in the number of columns. Approximately one thousand forms exist. The eight-column short form only works well for the first two rows which exhibit periodicity of eight. After two rows, the repeat number is 18 and the appearance of the lanthanide and actinide series results in further complications, but expansion to a 32-column format yields undesirable problems in display. Periodic “tables” other than those that are table-shaped are also discussed and illustrated including both two- and three-dimensional examples, produced both by “amateurs” and professionals. Scerri asks, “Is there an optimal periodic table?” An optimal table would tend to ignore utility whereas utility makes a form useful, but what is useful for one group is not that useful for others. Some proponents of alternate tables say theirs are more symmetrical and therefore more beautiful or elegant, but beauty is in the eye of the beholder (according to both Shakespeare and this reviewer). The varying philosophies also depend on the definitions of the elements as basic substances or simple substances, but the definition as “simple substances” leads to difficulties accommodating isotopes, which are not new elements but isotopic forms, which do behave somewhat differently. Also, the simple substance philosophy leads to the common placement of helium in the rare gas column rather than column two. Atomic number triads, previously discussed in Chapters Two and Thirteen, also have bearing on the form of the table.

Chapter Fourteen, “More Chemistry”, discusses trends within the periodic table other than the well-known intra-row and intracolumn relationships. Included are diagonal behavior similarities between group (n) and group (n + 10), early actinide relationships with transition metals, secondary periodicity (zigzag or alternating), knight’s move relationships, ions that imitate element superatom clusters, and first-member anomalies. The last-named phenomena apply to both physical and chemical properties, including physical state and valences, especially for first-row elements.

Conclusions

Throughout the book, Scerri discusses the controversies regarding the periodic system, as well as the “best” periodic tables, and the value of the periodic table in education. Has chemistry lately been taught as a subdivision of physics? Should it be taught more regarding the tangibles of chemical properties? The latter argument comes to a head with debates regarding the placement of helium over the other rare gases, since it is indeed one, or placing it over the Group II elements. I think most chemists would agree that the periodic system’s strength is in its predictions of the properties and reactions of the elements so helium should be in the same column as the rare gases whereas physicists, some physical chemists, quantum mechanics, and orbital filling would place helium in Group II.

This reviewer believes that the periodic table needs to be taught at different levels. To chemists of all stripes the forms that stress chemical periodicity are paramount. Fundamentals of the physics and quantum mechanics of the “construction” of the table need to be included but the complexity should be delayed until higher-level courses are taught. The periodic table is a “grabber” for all taking or encountering chemistry, especially for the first time, so unneeded complexity should be avoided in lower-level courses and in science electives for humanities students.

Scerri proposes that the various representations are a continuum between the extremes of Canham’s inorganic chemist’s table showing unusual relationships and with left-step tables. In Scerri’s opinion, the popular medium-long forms probably feature the best balance on the continuum.

Scerri also argues that the periodic table is neither a theory nor a model but an “organizing principle” and is a work in progress. It is also a prime subject of “science wars” between reductionism of all science including chemistry to physics or giving chemistry a more standalone existence. There are
also debates on whether science provides objective truth or is relativism stressing the equal validity of all forms of knowledge.

This second edition has a collected bibliography of all sources cited in end of chapter notes, books of suggested reading, a bibliography of works by Eric Scerri, and an index. Curiously, the bibliography does not include a citation to Chemical Periodicity by Sanderson (3), a compendium of periodic properties of the elements.

I believe there is enough additional and more-current information and discussion provided in this second edition to warrant acquisition even if the even edition is available.

Reviewer’s note. Not covered but important to the visually impaired is the existence of the periodic table in Braille (4). I am familiar with three other unpublished versions of the periodic table in Braille. My late wife, Gloria Kral Buntrock, was a certified Braillist, also certified in Nemeth Braille (Math and Science). She received requests from high school and college educators for three somewhat different forms for the periodic table in Braille. One was to be supplied not only with the pages in book format, but also with the pages arrayed, taped together and folded so that the blind student could experience the table in full two-dimensional format. At the time we were both employed by our consulting firm, Buntrock Associates (I was the chemistry consultant), and the Braille was supplied free of charge.

References


R. E. Buntrock
Buntrock Associates
Twenty-five Years Ago in Anaheim

Continuing our trip down memory lane, we come to spring 1995. The year began with somber news. Herman Skolnik, one of biggest names in the CINF division, passed away on December 29, 1994. Four CINF functionaries attended the memorial service on January 5, 1995 and another represented CINF at the viewing. A full obituary, by Val Metanomski, and a number of personal memories of Herman appeared in the spring 1995 issue of the CIB. Not long after Herman’s passing, Gerry Vander Stouw passed away suddenly on February 7, 1995. Gerry had worked for CAS for 30 years and had also served CINF well in many roles. I had the pleasure of working with him on more than one of the Noordwijkerhout meetings.

What of the spring 1995 ACS meeting in Anaheim, CA? At the end of my last article I said that Beilstein Information Systems promised CROSSFIRE Online for 1995, and they delivered. In Anaheim, they were showing Release 2 of CrossFire both in the Exhibition Hall and in a suite at one of the hotels. There were three gateways for accessing CrossFire: a serial connection via telephone, the Internet, or a straight network connection via TCP/IP. In addition to the demonstrations, there was a CINF talk entitled “CrossFire in Concept and Practice: The Beilstein File ... and More,” given by Sandy Lawson.

The big news at the meeting, though, was the launch of SciFinder. The CAS booth at the exhibition was dominated by SciFinder. I noted that a new feature of the booth was scheduled demonstrations on a large monitor with seating for the audience. I had a demonstration and wrote: “Explore by Research Topic has a natural language interface. If you input ‘the effect of caffeine on pregnant women’ you get a list of hits (including, interestingly, one on pregnant monkeys) with an indication of whether your main terms are closely associated with each other. SciFinder is now on the market. Companies with 20 or more chemists are being targeted, with a price of $3,000 per person per year. Users will be able to search without the taximeter running”.

In the CAS Open Meeting (which was still part of the CINF program in those days) Ken Ostrum (“Dr. O.”) told us “CAS offers information about more than 11 million documents and more than 14 million substances, but many chemists do not know how to get at it. In the race for new ideas and innovative solutions they can now have SciFinder on their desktops. The only failure of SciFinder is that you don’t need any training, so you won’t be able to go to one of Dr. O’s workshops.” Enhancements planned for SciFinder included substructure searching, reactions, and chemical catalogs. Thus, we were really seeing only the text search features at this stage.

The good Dr. O. repeated several times, to the amusemen of the audience, the advantages of CAplus as a single file offering access to both retrospective and very current information. There was even more CAS news: the inauguration of the “Web server” (cas.org). CAS, however, was not the only organization to have a home page by spring 1995. Beilstein Information Systems, for example, had http://www.beilstein.com/, a URL that now redirects you to the Beilstein Institute. ACS COMP had implemented “http://www.ssd.intel.com/ACS/acs.html.” I reported that “there are still some bugs at the moment but the service should be stable and operational by the fall 1996 meeting in Orlando, Florida”!


Other big themes in spring 1995 were combinatorial chemistry, molecular diversity, and library design. MDL, for example, launched Project Library in January 1995. Project Library was only the first in a line of products: Central Library, a client server implementation, was planned for later in 1995. MDL’s Available Chemicals Directory (ACD), a compendium of chemical products, proudly offered entries for over 200 chemical suppliers. Note that in 2020, ACD covers about 940 suppliers (according to the BIOVIA website). PSI International was showing a new product, Relational Structure Search System (RS3) Discovery, an ORACLE-based chemical information management system. I liked the look of it, and some people were keen to see a competitor to MDL’s ISIS. John Wiley had acquired the online Chemical Information System, the CIS, at about this time. Cambridge Scientific Computing (vendor of ChemDraw) changed its name to CambridgeSoft Corporation. As I said in my last article, the era of
electronic laboratory notebooks (ELNs) was beginning. Interest grew in 1995: CINF had a symposium on the topic in Anaheim.

I also gave some “personal” items of news. Bob Buntrock, having left Amoco, had formed Buntrock Associates with his wife Gloria K. Buntrock. Bonnie Lawlor, previously an Executive Vice President at ISI, resigned to seek new opportunities in electronic publishing.

If I am able to produce another “25 years article” this summer, it will cover the meeting held in Chicago, in fall 1995. That should be interesting: the Herman Skolnik Award Symposium honoring Reiner Luckenbach and Clemens Jochum is covered in detail, and the first published review of SciFinder (my own) is appended. The tale goes on...

Wendy Warr  
Wendy Warr & Associates
These were the sponsors who were committed to support CINF for the national meeting in Philadelphia.

Gold
ACS Publications

Silver
Royal Society of Chemistry

Bronze
Thieme Chemistry
C&EN

Contributor
Springer Nature
Waters Corp
Agilent Technologies
Bio-Rad Laboratories
ACS President-Elect

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.

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

Tina Qin
Chair, Fundraising Committee 2020
Email: Sponsorship@acscinf.org
Tel: 617-998-1789

The ACS CINF Division is a non-profit tax-exempt organization with taxpayer ID no. 52-6054220.
Introducing the ACS Research Data Center

For researchers in organic chemistry, ACS has developed a new tool for the sharing of primary research data. The ACS Research Data Center is a data packaging tool to assist authors in zipping their FID files, acquisition data, and processing parameters along with other appropriate FAIR metadata such as a SMILES or InChI for submission. This tool is free for any researcher to use and includes instructions for authors on how to upload their data. Find out more: https://axial.acs.org/2020/02/12/acs-research-data-center/.

Congratulations to the Recipients of the Spring 2020 ACS Publications Librarian Travel Grant

ACS Publications and the Award Committee are excited to announce the recipients of the ACS Publications Travel Grant for Librarians and Library School Students to attend the Spring 2020 ACS National Meeting & Exposition in Philadelphia: Joshua Borycz, Clair Castle, Sharon Maria S. Esposo-Betan, and Amanda McCormick. Meet the Winners: https://axial.acs.org/2020/02/26/spring-2020-acs-publications-librarian-travel-grant-winners/.

Announcing JACS Au

JACS Au (pronounced 'JACS Gold')—a sister journal to the society’s flagship Journal of the American Chemical Society (JACS) is launching in 2020. This fully open access journal will allow for the rapid dissemination of cutting-edge, high-impact research across the breadth of chemistry and all related areas intersecting with chemistry. Learn more: https://pubs.acs.org/journal/jaaucr
2020 marks the 20th anniversary of Science of Synthesis (SOS) the online synthetic methodology review compendium used by synthetic chemists worldwide. SOS, the successor to the well-known Houben-Weyl series was established in 2000 by an esteemed Editorial Board of international chemistry experts including Nobel Prize winner, Ryoji Noyori. Today under the guidance of Editor-in-Chief, Alois Fürstner, a team of eminent editors commission quality content from expert authors and ensure the selection of useful and practical methods. Chemists therefore have quick access to thorough and quality overviews on the entire range of organic synthesis topics saving them hours of searching and literature research. SOS is considered the place to begin when writing a thesis, preparing a talk, writing a paper, starting out in a new area of chemistry or preparing consultancy work.

A Place to Begin

SOS is a multi-authored reference work of synthetic methods that offers “a place to begin”. It comprises the largest collection of evaluated methods in organic synthesis worldwide. Its elegant, didactic presentation of synthetic methodology means it is an essential treatise for organic chemistry students, lecturers, and researchers. SOS is organized in a logical hierarchical system based on the target molecule to be synthesized. The critical coverage of methods is supported by information intended to help users choose the most suitable methods for their application, thus providing a strong foundation from which to develop a successful synthetic route. Within each category of product, illuminating background information such as history, nomenclature, structure, stability, reactivity, properties, safety, and environmental aspects are discussed along with a detailed selection of reliable methods. Each method and variation is accompanied by reaction schemes, tables of examples, experimental procedures, and a background discussion of mechanistic rationale, stereochemistry, scope of the reaction described and its limitations, and functional group compatibility. All published results from journals, books, and patent literature from the early 1800s until the year of publication are considered by the authors, who are among the leading experts in their field.

SoS Editorial Board

Alois Fürstner (Editor-in-Chief) Max-Planck-Institut für Kohlenforschung Germany
Erick M. Carreira ETH Zurich Switzerland
Margaret Faul AMCDN Inc. CA, USA
Shu Kobayashi University of Tokyo Japan
Guido Koch Topwax Pharma AG Switzerland
Gary A. Molander University of Pennsylvania PA, USA
Cristina Nevado University of Zurich Switzerland
Barry M. Trost Stanford University CA, USA
Shu-Li You Shanghai Institute of Organic Chemistry (SIOC) P. R. of China
20th Anniversary of Science of Synthesis: 
Your Expert Guide to Making Molecules

Evaluation of Methods by Experts

The SOS Editorial Board comprises expert chemists from academia and industry: E. M. Carreira, M. Faul, A. Fürstner (Editor-in-Chief), S. Kobayashi, G. Koch, G. Molander, C. Nevado, B. M. Trost, S.-L. You. The Editorial Board’s overarching goal is to make the suite of SOS resources the first and foremost focal point for critically evaluated information on chemical transformations for those individuals involved in the design and synthesis of organic molecules. Over 56 volume editors and 2,000 authors worldwide have contributed to the series over the last 20 years.

Online, Updated, and Topical

The online version of SOS enables text, structure, substructure, and reaction searching via a simple web-based interface with powerful functionality. Continual updating of the electronic version means that the content of SOS remains pertinent and relevant to the synthetic organic chemist’s needs. Also supplementing current content with special topics acknowledges the broad spectrum of organic chemistry today and the need for chemists to appreciate many different peripheral scientific fields in addition to the core subject area. The electronic product is designed so that it:

- Provides an exclusive overview of the synthetic chemistry literature
- Provides easy access to the best and most reliable synthetic methods in organic and organometallic chemistry
- Allows researchers to tailor their structure, text and reaction searches to accommodate their chemical information needs
- Provides personalized support for scientific queries through an Editorial Office professionally staffed by Ph.D. Chemists, and a technical support desk
- Supports the chemical community by responding to its need for relevant and value-added synthetic chemistry information

The Science of Synthesis Advisory Board comprises experts who have significant experience of chemical information systems in both industry and academia: G. Baysinger (Stanford University), L. Betschart (ETH), J. Currano (University of Pennsylvania), C. Keil (Pfizer), Y. Li (MIT), X. Li (CAS, China), and D. Wrublewski (Caltech). They regularly contribute to discussions regarding the development of the electronic product.
Women in Chemistry Award

In addition to providing an important editorial contribution for the chemistry community, the dynamic and forward-thinking SOS Editorial Board founded the first major international synthetic chemistry award for Women in Chemistry. As a result, the first award was presented to Sarah Reisman (Caltech) at the 21st ESOC in Vienna in 2019.

SOS: We Transform Synthesis!

Over 20 years some 96,000 “pages” of evaluated information have been published including over 2 million molecules, 425,000 selected reactions and no less than 54,000 experimental procedures. SOS has proven to be the perfect research companion for the organic synthetic chemist and will continue to provide the chemistry community with valuable synthetic methodology reviews in the years to come. Anyone interested in having free trial access to the online product for two weeks at their institution should contact the publisher.

Contact
Science of Synthesis/Thieme Chemistry
Thieme Group
Ruedigerstrasse 14
70469 Stuttgart
Germany

Editorial/Technical Queries:
SOS.techsupport@thieme.com

Access/Marketing:
marketing@thieme-chemistry.com

Website: sos.thieme.com
**Journal of Cheminformatics launches Educational article type**

In January, *Journal of Cheminformatics* launched a new Educational article type. Educational articles describe an aspect of cheminformatics, such as cheminformatics theory and methods, as well as tutorials explaining in detail how to perform some cheminformatics.

Educational articles can roughly be divided into two categories:

- Articles that describe in detail how common cheminformatics tasks can be performed using a particular open-source toolkit. Data, standards, and code used in this type must be fully open-source.
- Articles that describe in detail a cheminformatics algorithm. These must provide enough detail that it can be fully implemented with open-source software. Data and standards used in this type must be fully open-source.


**In Review — now available on Journal of Cheminformatics**

Our commitment to research in progress and openness inspires us to think about how to help our authors in new ways. Together with our partners at Research Square ([https://www.researchsquare.com/browse/journal/journal-of-cheminformatics](https://www.researchsquare.com/browse/journal/journal-of-cheminformatics)), we added a new option for authors when they submit to select journals, including, since January, *Journal of Cheminformatics*. This option, called In Review ([https://www.springernature.com/gp/authors/campaigns/in-review](https://www.springernature.com/gp/authors/campaigns/in-review)), gives authors insight over the progress of their manuscript through peer review. In addition, you can share your submission as a preprint — along with the editorial timeline. In Review is a unique preprint service that clearly links your manuscript to the journal reviewing it. Since offering it for *Journal of Cheminformatics*, we have seen an impressive uptake of around 50% of submissions.

**Springer Nature and OpenAIRE collaborate to further Open Science**

Since January, Springer Nature has provided OpenAIRE ([https://www.openaire.eu/](https://www.openaire.eu/)), an EU organization that facilitates openness in scholarly communication, and access to full-text articles and chapters, regardless of their open access status and access rights.

This enables OpenAIRE to extract links from articles, research data, and other scholarly outputs using text- and data-mining algorithms. Such links between research objects can then be used by both Springer Nature and OpenAIRE to provide searches and statistics through their respective portals and will be made freely available via nature.com, link.springer.com, OpenAIRE Explore ([https://explore.openaire.eu/](https://explore.openaire.eu/)), and Scholexplorer ([https://scholexplorer.openaire.eu/](https://scholexplorer.openaire.eu/)).
Through this collaboration, the partners aim to foster discoverability of datasets linked to publications and vice versa, which advances research by supporting reuse, making datasets easier to find and access.

**ioChem-BD is now a Recommended Data Repository**

Springer Nature maintains a list to help authors identify discipline-specific, community-recognized repositories to which they can submit their data. ioChem-BD (https://www.iochem-bd.org/), a repository to manage, store and publish Computational Chemistry Datasets has now been added to the list of Recommended Repositories. Spearheaded by Núra López (http://www.iciq.org/research/research_group/prof-nuria-lopez/) and her ERC Proof of Concept Grant for the project “Big Data for Catalysis” (BigData4Cat, https://cordis.europa.eu/project/id/680900), and in collaboration with ICIQ group leaders Feliu Maseras (http://www.iciq.org/research/research_group/prof-feliu-maseras/), Carles Bo (http://www.iciq.org/research/research_group/prof-carles-bo/) as well as Universitat Rovira i Virgili (http://www.urv.cat/en/), ioChem-BD was launched in 2015 and has since grown and adapted to become a useful tool for the community.

View the full list of Recommended Repositories here: https://www.springernature.com/gp/authors/research-data-policy/recommended-repositories
2020 CINF Officers and Functionaries

Chair
Jeremy Garritano
University of Virginia
jg9jh@virginia.edu

Chair-Elect
Donna Wrublewski
Caltech Library
dtwrub@caltech.edu

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

Secretary
Nick Ruhs
Florida State University
nruhs@fsu.edu

Treasurer
Stuart Chalk
University of North Florida
schalk@unf.edu

CINF Councilors
Bonnie Lawlor
chescot@aol.com

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

CINF Alternate Councilors
Rachelle Bienstock
RJB Computational Modeling LLC
rachelleb1@gmail.com

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

Archivist/Historian
Bonnie Lawlor
chescot@aol.com

Awards Committee Chair
Rajarshi Guha
Vertex Pharmaceuticals
rajarshi.guha@gmail.com

Careers Committee Chair
Neelam Bharti
Carnegie Mellon University Libraries
nbharti@andrew.cmu.edu

Communications and Publications Committee Chair
Graham Douglas (Interim)
communications@acscinf.org

Procedures Chair
Bonnie Lawlor
See Councilor

Education Committee Chair
Grace Baysinger
Stanford University
graceb@stanford.edu

Finance Committee Chair
Stuart Chalk
University of North Florida
schalk@unf.edu

Fundraising Committee Chair
Tina Qin
Harvard University
qinnamsu@gmail.com

Membership Committee Chair
Donna Wrublewski (interim)
Caltech Library
dtwrub@caltech.edu
Nominating Committee Chair
Elsa Alvaro
Northwestern University
elsa.alvaro@northwestern.edu

2019–2020 Program Committee Chair
Sue Cardinal
University of Rochester
scardinal@library.rochester.edu

2020–2021 Program Committee Chair
Ye Li
MIT
yel@mit.edu

Chemical Information Bulletin Editor, Spring
Teri Vogel and Judith Currano (interim)
See CIB Editors, Fall and Winter

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

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

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

Webmasters
Rachelle Bienstock
RJB Computational Modeling LLC
rachelleb1@gmail.com

Stuart Chalk
University of North Florida
schalk@unf.edu

Articles and Features
Wendy Warr
Donna Wrublewski

Committee Reports
Bonnie Lawlor

Book Reviewer
Robert E. Buntrock

Sponsor Information
Tina Qin
Graham Douglas

Production
Teri M. Vogel
Judith N. Currano
Wendy Warr
Bonnie Lawlor
David Shobe
Svetlana Korolev
Technical Program — 2020 Spring Meeting

SUNDAY MORNING

Pennsylvania Convention Center
Bridge Over Arch Street - Theater 7

Useful Databases & Computation Methods for Studying Receptor-Ligand Interactions

R. J. Bienstock, Organizer, Presiding

8:50 1. ChEMBL: Leading data source for machine learning applied to drug discovery. N. Bosc
9:40 3. Advances in data curation and the user interface at BindingDB. M.K. Gilson, T. Liu, L. Hwang
10:05 Intermission.
10:15 4. Robust modeling of receptor-ligand complexes in cryo-EM maps by combining glide docking and force field based refinement for cryo-EM enabled drug design of class A streptogramin antibiotics. G. van Zundert, K. Borrelli
10:40 5. High-throughput experimental and computational technologies at the National Center for Computational Toxicology. A.J. Williams, J. Wambaugh, R. Judson, K. Houck, C. Grulke, K. Paul-Friedman
11:05 6. Benchmarking indexing algorithms for in-memory molecular similarity search. m. song, C. Zhu, J. Bi

Section B

Pennsylvania Convention Center
Bridge Over Arch Street - Theater 8

Macromolecular Information for the Second Century: Digital Representations, Identifiers & Data Exchange

Cosponsored by CARB, PMSE and POLY‡

V. F. Scalfani, Organizer
L. R. McEwen, D. Wrublewski, Organizers, Presiding

V. Scalfani, Presiding

8:15 Introductory Remarks.
8:20 7. Challenges in the representation and curation of polymer data. M. Dunbar, A. Miller
9:00 9. Macromolecular data representation at the worldwide protein data bank. J.D. Westbrook, Z. Feng, V. Guranovic, C.L. Lawson, E. Peisach, C. Shao, J.Y. Young, S.K. Burley
9:20 Intermission.
9:35 10. Representation of biopharmaceutical substances in the FDA global substance registration system and in the structured product labeling. Y. Borodina, T. Peryea
9:55 11. Unknown or variable composition, complex reaction products and biological materials data management through the US EPA CompTox Chemicals Dashboard. A.J. Williams, C. Grulke, G. Patlewicz, A. Richard
10:15 Intermission.
11:05 Intermission.
11:35 15. HELM in the wild: Real-world applications in a pharmaceutical discovery environment. M. Somogyi, R. Knipsel, T. Parrott
11:55 Concluding Remarks.

SUNDAY AFTERNOON

Section A

Pennsylvania Convention Center
Bridge Over Arch Street - Theater 7

Data Exchange & Integration Among Open Chemical Information Resources

S. Kim, Organizer, Presiding

1:00 Introductory Remarks.
1:30 17. Curating ChemSpider: Challenges in chemical data management. M. Archibald
1:55 18. Open API for searching and accessing thermophysical and thermochemical properties provided by the NIST Thermodynamics Research Center. D. Riccardi, C.D. Muzny
2:20 19. Sharing and linking chemical structure data on PubChem and institutional repositories. V.F. Scalfani
2:45 Intermission.
3:00 20. Data interchange en masse via batch search functionality in the CompTox Chemicals Dashboard. A.J. Williams
4:15 23. Data exchange and integration with DrugBank, HMDB and other Canadian cheminformatics resources. D.S. Wishart
4:40 Concluding Remarks.

Section B

Pennsylvania Convention Center
Bridge Over Arch Street - Theater 8

Makerspace Safety: An Overview & Case Studies

Cosponsored by CHAS
S. Singh, Organizer
N. Bharti, Organizer, Presiding

1:30 Introductory Remarks.
1:55 25. Don't cut that! Lessons learned from a laser cutter fire. N.J. Leon, D.R. Kuespert
2:15 26. Are you ready to have a laser cutter in your makerspace? Think again! A. Lawson, S. Singh
2:35 27. Student library employees make the makerspace safe!. A. Zeidman-Karpinski, J. Murray, D. Walton
2:55 Intermission.
3:05 28. 3D printing in library makerspace: Health and safety concern. N. Bharti, S. Singh, A. Singh
3:25 29. New risk paradigm: Advancement of campus makerspaces and safe innovation. T. Durak

SUNDAY EVENING

Pennsylvania Convention Center
Exhibit Hall A

Pushing the Boundaries of Chemical Information & Cheminformatics

E. Alvaro, J. N. Currano, Organizers

5:30 - 7:30
32. Structurally annotated lists of chemicals from international frameworks for the control of chemical weapons and precursors. S. Costanzi, C. Slavick, G.D. Koblentz, R. Cupitt
34. Clustering of the phosphacin protein family and its functionally relevant groups. S. Hosler, J. Fetrow
35. Bringing order from chaos: FAIR database of sesquiterpenes organized by dihedral angles. J.M. Schrey, C. Hamann
36. Integration of heterogeneous data using the SciData Framework and JavaScript for linked data to improve the FAIRness of scientific data. S.J. Chalk, D. Johnson
37. 30-year analysis of intellectually-indexed Markush structures from patents. M. McBRIDE, E.N. Cheeseman
39. Predictions of pairs of drugs with synergistic activity against leukemia cell lines based on NCI-ALMANAC data. A.A. Lagunin, L.O. Soina
40. Deciphering chemical information of small molecules contributing to biased GPCR signaling. J. Sanchez, S. Sirimulla
43. Adapting evolutionary algorithms for autonomous machine learning in chemistry. G. Vishwakarma, M. Haghighatlari, J. Hachmann
44. Chemists’ data needs for machine learning research in academia. Y. Li
45. Predicting the glass transition behavior of polymers via integration of cheminformatics and molecular dynamics simulations. a. k, B. Rasulev, w. xia, a. alesadi
46. Insights into therapeutic fusion protein R&D from an analysis of the CAS databases. Y. Li, Y. Deng, C.Y. Liu
47. STN to the rescue: When SciFinder is not sufficient. P. Barnett
48. Reproducible data analysis and publishing in chemistry with R: Creating a workshop experience during the ACS National Meeting. Y. Li, S.P. Wathen, D. Wrublewski
49. Crystal structure data without a publication: How much more can be learnt?. S. Ward, C.A. Tovee, I. Bruno
50. Introduction of scholarly communications and the ethics of authorship into a chemical literature course. N.J. Butkovich
52. Degradable, on-patient medical record. J. Collins, K. McHugh, A. Jaklenec, R. Langer
MONDAY MORNING

Philadelphia Marriott Downtown
Liberty A

Machine Learning in Drug Discovery

Cosponsored by CINF
S. Sirimulla, Organizer, Presiding

8:00 Introductory Remarks.
8:05 53. Scaling ligand-based virtual screening to a larger purchasable chemical space. M. Alnammi, S. Liu, S. Ericksen, G.E. Ananiev, J.L. Keck, F.M. Hoffmann, S.A. Wildman, A. Gitter
8:30 54. Using machine learning methods and structural alerts for prediction of mitochondrial toxicity. G.F. Ecker, J. Hemmerich, F. Troger, B. Fuezi
8:55 55. Big errors in big data: When automated data curation misses the mark. R.D. Clark, P.R. Daga, M. Waldman
9:45 Intermission.
10:05 57. Generation of druglike molecules with generative adversarial networks (GANs). B. Ji, M. Brock, Y. Bian, S. Liu, X. He, V.H. Man, X. Xie, J. Wang
10:30 58. Synthetic feasibility and de novo molecular generation and optimization. C.W. Coley, W. Gao, K.F. Jensen
11:45 61. Disease drivers, drug discovery & drug positioning in NASH: Smart medicine outsmarts disease. P. Narayan

Section B

Philadelphia Marriott Downtown
Liberty B

Current State of FAIR Chemistry Data
General FAIR Landscape & Other Disciplines

S. J. Chalk, Y. Li, L. R. McEwen, V. F. Scalfani, Organizers
I. Bruno, N. Ruhs, Organizers, Presiding

8:00 Introductory Remarks.
8:05 62. Advancing science as a global public good. B.C. Carroll, B. Mons
8:25 63. GO FAIR in the U.S.: Advancing FAIR beyond the basics. M. Cragin, C. Kirkpatrick
8:45 64. World Data System and CoreTrustSeal: Together building trust in scientific data. R. Edmunds
9:10 65. NIST Research Data Framework and its relevance for FAIR data. R.J. Hanisch
9:30 66. NIH strategic plan for data science. S.K. Gregurick
9:50 67. Why, how and what of open science. M.G. Hicks
10:10 Intermission.
10:20 68. FAIR, FACT, and the US RCSB Protein Data Bank. J.D. Westbrook, S.K. Burley
10:40 69. RRIDs for antibodies and instruments, an idea to make chemistry more FAIR. A. Bandrowski
11:00 70. What do the FAIR Data Principles mean to the life sciences industry?. C.I. Nitsche
11:20 71. FAIR data and climate change resilience. C. Thomas, C. Riopelle
11:40 72. Open and FAIR data sharing: Geosciences benefiting from collaboration with chemistry and other sciences. S. Stall, L.R. McEwen
MONDAY AFTERNOON
Philadelphia Marriott Downtown
Liberty A

Machine Learning in Drug Discovery

S. Sirimulla, Organizer, Presiding

1:00 Introductory Remarks.
1:05 73. Proteochemometric modeling of GABA transporters. **G.F. Ecker**, S. Kickinger, A. Seiler
2:45 Intermission.
3:10 77. GNINA 1.0: Predicting protein-ligand binding affinity through cross-docking. **P. Francoeur**, D. Koes
3:35 78. Analysis of machine learning strategies to predict protein developability. **A. Golinski**, B. Hackel
4:50 Concluding Remarks.

Section B

Philadelphia Marriott Downtown
Liberty B

Current State of FAIR Chemistry Data
Publishers Perspectives

I. Bruno, S. J. Chalk, Y. Li, N. Ruhs, V. F. Scalfani, Organizers
L. R. McEwen, Organizer, Presiding
V. Scalfani, Presiding

1:30 81. Community efforts to enhance supporting information of spectral data and chemical structures. **V.F. Scalfani**, L.R. McEwen
1:50 82. Essential factors for data sharing: Publisher initiatives to help make data FAIR. **S.H. Winthrop**, M. Smyllie, S. Pauly
2:05 83. FAIR data policies at science. **J.S. Yeston**
2:20 84. Implementing FAIR data principles at American Chemical Society publications. **A.M. Hunter**
2:35 85. FAIR Chemistry: Publisher perspective. **R.J. Boucher**
2:50 Panel Discussion.
3:10 Intermission.
3:25 86. Data sharing at the Royal Society of Chemistry. **G. Jones**
3:40 87. Thieme chemistry: publisher initiatives for supporting FAIR data exchange of chemical data. **F. Shortt de Hernandez**
4:10 Panel Discussion.
MONDAY EVENING

Pennsylvania Convention Center
Exhibit Hall A

Sci-Mix

S. K. Cardinal, Organizer

8:00 - 10:00

4, 27, 32, 35, 40, 42, 43, 45, 47, 51, 52, 57, 73. See Previous Listings.

TUESDAY MORNING

Pennsylvania Convention Center
Bridge Over Arch Street - Theater 7

Scientific Visualizations & Creative Presentations

Cosponsored by CPRC and PROF
Financially supported by Waters, Agilent, C&E News
P. J. MacDougall, Organizer, Presiding

8:00 Introductory Remarks.
8:05 90. Compound Interest: Graphical chemistry communication. A. Brunning
8:50 91. Tell your science story with visuals: Other thousand words. G. Zaidan
9:15 92. How C&EN creates effective data visualizations. R. Bryson
9:40 93. Using personas and infographics to make scientific data come alive to multiple audiences. M.A. Strausbaugh, L.E. Dragin, D.L. Daniels
10:05 Intermission.
10:20 94. Authenticity and creativity in chemical communication. R.M. Burks
10:45 95. Thirty years of gloop, slime, and cabbage juice: How do we utilize current communication methods to deliver effective messages to the public?. A. Serfis
11:10 96. Data storytelling: Intersecting data, stories and technology to create visual inspiration. R. Torres
11:35 97. Communicating science with little (or no) budget: Design rules and tricks for the non-artist. K. Deards

Pennsylvania Convention Center
115A

Current State of FAIR Chemistry Data
Chemistry & Physical Sciences

I. Bruno, S. J. Chalk, L. R. McEwen, V. F. Scalfani, Organizers
Y. Li, N. Ruhs, Organizers, Presiding

8:20 Introductory Remarks.
8:25 98. Supporting the FAIR challenge in chemical information. A. Sanford, A. Jacobs
8:45 99. FAIR data for the chemistry life cycle. M. Clark, T. Slater, A. de Waard
9:05 100. Associating live analytical data to synthetic chemistry experiments: Applying FAIR principles across the scientific experimentation lifecycle. A.A. Anderson, M. Boruta
9:45 102. UK physical science data-science service: FAIR resource for chemistry in UK. S.J. Coles, N. Knight
10:05 Intermission.
10:40 104. Baby steps toward FAIR chemistry data. M.E. Lafferty
11:00 105. Organic reaction data: Status, desiderata, and outlook for reaction informatics. C.W. Coley

TUESDAY AFTERNOON

Pennsylvania Convention Center
Bridge Over Arch Street - Theater 7

Scientific Visualizations & Creative Presentations

Cosponsored by CPRC‡ and PROF
Financially supported by Waters, Agilent, C&E News
S. K. Cardinal, H. Cheng, P. J. MacDougall, Organizers
S. C. Hayden, Organizer, Presiding

1:00 108. Communication is a rare event: let's make it happen. C.T. Hunt
1:50 110. CANVAS: Six pointers for formulating your message. W.T. Lambert
2:15 111. 2020 molecular visualization and picotechnology. P.J. MacDougall
2:40 112. Translational toxicology: Simultaneous data visualisation across phases. A. Cooke, E. Champness, T. Mansley, M.D. Segall

Pennsylvania Convention Center
Bridge Over Arch Street - Theater 7

AI Meets Cheminformatics

N. Bharti, T. Qin, Organizers
T. Qin, Presiding

3:30 113. Using machine learning to predict human health from biofluid-based metabolomics. E. Evans, d. sontag, E. Alm
4:50 117. Retrosynthesis with transformer and molecular grammars. S. Sosnin, M. Fedorov
5:10 118. CINFull materials: Insights on development of custom feature set data for machine learning in materials. C. Baddeley
**Pennsylvania Convention Center**  
**115A**

**Current State of FAIR Chemistry Data**  
**Moving Forward**

Y. Li, L. R. McEwen, N. Ruhs, V. F. Scalfani, *Organizers*  
I. Bruno, S. J. Chalk, *Organizers, Presiding*

1:10 119. Current state of FAIR chemical data standards: IUPAC’s role. **L.R. McEwen**

1:30 120. IUPAC efforts in the area of FAIR management of spectroscopic data. **R.M. Hanson, D. Jeannerat**


2:10 123. GO-FAIR chemistry implementation network (ChIN). **S.J. Coles, S.J. Chalk, J.G. Frey, E.L. Willighagen, N. Knight**

2:10 124. GO-FAIR ChIN: Review, panel discussion and open conversation about FAIR in chemistry. **S.J. Chalk, S.J. Coles, J.G. Frey, E.L. Willighagen**

**WEDNESDAY MORNING**

**Pennsylvania Convention Center**  
**Bridge Over Arch Street - Theater 7**

**AI Meets Cheminformatics**

N. Bharti, T. Qin, *Organizers*  
T. Qin, *Presiding*

8:00 125. ChEMU shared task: Chemical entity recognition and event extraction of chemical reactions from patents. **D. Nguyen, C. Druckenbrodt, K. Verspoor, S. Akhondi, C. Thorne, Z. Zhai**

8:20 126. Using chemical ontologies to create molecular prediction systems for any molecular property. **L. Weber, M. Irmer, C. Bobach, K. Kruse**

8:40 127. Evolutionary fuzzy rule induction processes for subgroup discovery in chemical datasets. **P. Kowalczyk**

9:00 128. Machine learning transition temperatures from 2D structure. **A. Sifain, B. Gifford, J.A. Morrill, B. Rice, S.H. Yalkowsky, B.C. Barnes**

9:20 129. Learning the physical properties of organic molecules using graph-based models. **M. Afzal, Y. An, A. Chandrasekaran, A.R. Browning, M.D. Halls**


**Chemical Information Bulletin, 2020, 72 (1)**

Section A
10:15 Introductory Remarks.
10:20 131. Practical applications of deep learning to imputation of drug discovery data. T. Whitehead, B. Irwin, M.D. Segall, G. Conduit
10:45 132. AI-driven 3D design for orphan targets from big PDB/SAR data. D. Kireev

Section B
Pennsylvania Convention Center
Bridge Over Arch Street - Theater 8

Cultivating Good Data Practices Among Chemists
Cosponsored by CHED and COMP
Y. Li, S. Ward, Organizers, Presiding

8:00 Introductory Remarks.
8:05 135. Research context and research chemistry: Information literacy, scholarly publishing, and data management in the chemistry curriculum. J.D. Borycz
8:25 136. Cartooning the data champions at the University of Cambridge. C. Castle
8:45 137. Fostering data stewardship through innovative data literacy partnerships. N. Ruhs
9:25 139. "Don't trust your future self": Wholistic data practices one conversation at a time at Cornell University. L.R. McEwen
9:45 Intermission.
9:55 140. What's in a (file) name? Introducing data management skills in an undergraduate laboratory course. J.N. Currano
10:15 141. New frontier in lab data management: Getting instrument data into an electronic notebook. J. Hockaday, R. Macneil
10:35 142. Use of electronic lab notebook with advanced cheminformatics in academia. J. Lee
10:55 143. Cultivating good chemistry data in open-source interdisciplinary repository. T. Qin
11:15 144. Cultivating good data practices: Perspective from a structural database. S. Ward, I. Bruno, M.P. Lightfoot
11:35 145. I read the paper, where is the data?. E. Bolton, B. Shoemaker, A. Gindulyte, T. Cheng, J. Zhang, P. Thiessen
11:55 Concluding Remarks.

WEDNESDAY AFTERNOON
Pennsylvania Convention Center
Bridge Over Arch Street - Theater 7

AI-based Big Data Application in Drug Discovery
Q. Zhu, Organizer
A. Zakharov, Organizer, Presiding
Q. Zhu, Presiding

1:00 146. How machine learning will enrich SAVI (Synthetically Accessible Virtual Inventory). V. Delanee, H. Patel, M.C. Nicklaus

Chemical Information Bulletin, 2020, 72 (1)
1:25 147. Accelerating hit discovery with multiobjective deep reinforcement learning. M. Popova, O. Isayev
2:40 Intermission.
2:55 150. Accurate prediction of nuisance compounds based on large libraries of HTS data. V.M. Alves, S. Capuzzi, R.C. Braga, D. Korn, K. Bowler, J.E. Hochuli, A. Yasgar, A. Simeonov, G. Bantukallu, E. Muratov, A. Zakharov, A. Tropsha
5:00 Concluding Remarks.

Pennsylvania Convention Center
Bridge Over Arch Street - Theater 8

Cheminformatics for Chemists

T. Qin, Organizer
R. E. Belford, Organizer, Presiding

1:00 155. No coding required: Processing chemical data with KNIME. M. Archibald
1:20 156. Computational notebooks for cheminformatics. P. Kowalczyk
1:40 157. Teaching introductory cheminformatics and machine learning with Mathematica. J. Schrier
2:00 158. Chemical structures in the Google era. G.M. Banik, S. Tenney, A. Koenig, A. Kreckmann
2:20 Intermission.
2:50 160. DevOps in digital chemistry: Cheminformatics and data science toolchains. T. Gressling
3:10 161. Lead-based virtual screening and prediction of EGFR inhibitors using PubChem’s database with data mining and machine learning algorithms. K. He
3:30 Intermission.
3:40 162. Unlocking data science and cheminformatics insights for chemists. F. van den Broek
4:00 163. Beyond the bench: Leveraging laboratory expertise in your cheminformatics development. M. Prissel
ChEMBL: Leading data source for machine learning applied to drug discovery

Nicolas Bosc, nbosc@ebi.ac.uk. ChEMBL group, EMBL-EBI, Hinxton, Cambridge, CB10 1SD, Wellcome Genome Campus, United Kingdom

For more than a decade, the ChEMBL database has grown to incorporate ever more data, extracted from the scientific literature or directly deposited. Initially focused on small molecules bioactivities, it has evolved to fulfill (and indeed, sometimes to exceed) the expectations of the scientific community in the challenging field of drug discovery. Now also providing drug data such as development status, mechanisms of action, indications and withdrawal reasons, ChEMBL supports more than ever pharmaceutical research. The database comprises in its latest version [1] 1.1 million assays for 1.9 million compounds and more than 8,000 protein targets, with 26,000 indications and 4700 mechanisms. All of this, freely accessible to the global community. This information can be used as a source of evidence but many scientists also find in ChEMBL the data they need to build structure activity relationship models [2]. With the excitement surrounding machine learning and AI, ChEMBL has a central role to play, addressing the need for more and more quality data. Through this presentation, I will describe several important details about the ChEMBL database and how its data are used to develop predictive models.

PubChem for structure-activity relationship (SAR) studies

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

PubChem (https://pubchem.ncbi.nlm.nih.gov) is a large chemical information resource, with more than 236 million depositor-provided substance descriptions, 96 million unique chemical structures, and 268 million bioactivity outcomes from one million assays covering around ten thousand unique protein target sequences. This presentation provides an overview of PubChem’s data, tools, and services useful for structure-activity relationship (SAR) studies. While most bioactivity data in PubChem are generated from high-throughput screening (HTS), PubChem also has a substantial amount of bioactivity data extracted from scientific articles and patent documents, thanks to contribution by other databases like ChEMBL, Guide to PHARMACOLOGY, and BindingDB. In addition, through data integration with other information resources, PubChem contains a wide range of annotations, including pharmacology, toxicology, drug target, metabolism, chemical vendors, scientific articles, patents, and many others.

PubChem supports various types of chemical structure searches, including identify search, 2-D and 3-D similarity searches, substructure and superstructure searches, molecular formula search. In addition, PubChem Target View pages provide users with easy access to the bioactivity data for a given gene/protein target. Programmatic access to PubChem’s bioactivity data enables one to build an automated workflow that exploits this data. In addition, through
PubChemRDF, users can integrate PubChem’s data into their own in-house data on a local computing machine.

CINF 3

Advances in data curation and the user interface at BindingDB

Michael K. Gilson, mgilson@ucsd.edu, Tiqing Liu, Linda Hwang. School of Pharmacy and Pharmaceutical Sci., U. C. San Diego, La Jolla, California, United States

BindingDB (www.bindingdb.org) is the first publicly accessible database of experimental protein-small molecule binding data. In recent years, we have focused our curation efforts on patent data, as these are not, to our knowledge, handled by any other similar database. In particular, although SureChEMBL’s automated curation extracts many compounds from patents, it does not provide affinity data for these compounds. In addition, it is easier to extract data from US patents than from scientific articles, for several reasons. Indeed, semi-automation of the process of curating binding data from US Patents has allowed us to process over 1,000 US Patents a year, leading to extraction of >200,000 affinity data for >100,000 compounds during the 12-month period from Nov 1, 2018 to October 31, 2019 (http://bindingdb.org/bind/ByPatent.jsp). I will discuss the curation process and the data, as well as recent enhancements to the BindingDB user interface aimed at improving the viewing format as well as query and browsing capabilities.

CINF 4

Robust modeling of receptor-ligand complexes in cryo-EM maps by combining glide docking and force field based refinement for cryo-EM enabled drug design of class A streptogramin antibiotics

Gydo van Zundert¹, gydo.vanzundert@schrodinger.com, Kenneth Borrelli². (1) Schrodinger, New York, New York, United States (2) Life Science, Schrodinger, New York, New York, United States

Building accurate atomic models of receptor-ligand complexes in cryo-EM Coulomb potential maps is often a challenging task as the resolution is often too low to create an unambiguous pose. To this end, we developed two computational tools: we implemented our advanced OPLS3e/VSGB force field and implicit solvent model as a restraint model in the popular PHENIX real space refine tool; and we created a variation of Glide Small Molecule Docking that includes a real-space cross-correlation term, to allow the identification of a set of possible ligand conformations that are consistent with the EM data. We validated our tools individually by re-refining and re-docking deposited cryo-EM based receptor-ligand complexes, showing similar and higher quality structure and poses, respectively, based on energy calculations, correlation values and MolProbity metrics. Finally, we show how these tools were used as part of a cryo-EM enabled structure-based drug discovery effort to design class A streptogramin antibiotics with improved resistance profiles.

CINF 5

High-throughput experimental and computational technologies at the National Center for Computational Toxicology
Antony J. Williams, tony27587@gmail.com, John Wambaugh, Richard Judson, Keith Houck, Christopher Grulke, Katie Paul-Friedman. Center for Computational Toxicology and Exposure, Environmental Protection Agency, Research Triangle Park, Durham, North Carolina, United States

The U.S. Environmental Protection Agency (EPA) is faced with the challenge of efficiently evaluating chemical safety often with access to only sparse toxicity data. The increasing number of chemicals found in commerce and the environment, together with the time and resource requirements for traditional toxicity testing and exposure characterization, requires new approaches to The U.S. Environmental Protection Agency (EPA) is faced with the challenge of efficiently evaluating chemical safety often with access to only sparse toxicity data. The increasing number of chemicals found in commerce and the environment, together with the time and resource requirements for traditional toxicity testing and exposure characterization, requires new approaches to be developed. In 2005, EPA embraced computational toxicology (CompTox) to deliver results and applications across a broad range of environmental health problems. This work includes 1) the Toxicity Forecaster (ToxCast) project for in vitro high-throughput screening (HTS) of environmental chemicals; 2) high-throughput toxicokinetics (HTTK) and 3) high-throughput transcriptomics (HTTr) for cost-efficient screening of thousands of chemicals. One aspect of this work has been the delivery of a number of web-based “dashboards” providing access to experimental and predicted data ensuring community access to data streams that can be of value to researchers. This presentation will provide an overview of the CompTox Chemicals Dashboard and how this freely available community resource provides access to experimental and predicted data generated within the EPA’s National Center for Computational Toxicology, and is an information hub for data aggregated from public databases. This includes specific efforts to aggregate data associated with pesticides and the generation of large volumes of in vitro bioactivity data associated with 100s of assays.

CINF 6

Benchmarking indexing algorithms for in-memory molecular similarity search
Minghu song¹, songminghu2004@yahoo.com, Chunjiang Zhu², Jinbo Bi². (1) Biomedical Engineering Department, University of Connecticut, Mansfield Center, Connecticut, United States (2) Department of Computer Science and Engineering, University of Connecticut, MANSFIELD CENTER, Connecticut, United States

Molecular similarity search is one of popular chemoinformatics tools to rapidly identify molecules that are structurally similar to a given query compound. With the increasing number of compounds available in chemical databases, the traditional exhaustive approach becomes computationally prohibited. A variety of indexing algorithms have been developed to speed up the overall search process. However, most state-of-the-art indexing algorithms, such as Hnsw and Onng, were originally invented for the multimedia database search and only implemented the metric space such as the Euclidean distances, and therefore cannot be readily applied to molecular similarity search problems which usually compute the Jaccard-Tanimoto distance metric. Furthermore, there are increasing interests in the computational biology and chemoinformatics community to develop robust benchmarking systems so that both users and developers can systematically evaluate various computational algorithms or tools for a given biomedical task in an unbiased fashion. In this study, we first implement Python or C++ codes to enable the Tanimoto similarity search capability for some recent graph or inverted indexing algorithms; Moreover, we provide the first unbiased benchmark to thoroughly evaluate the large-scale molecular similarity searching performance of these recent indexing algorithms initially introduced for multimedia data retrieval. To avoid the computational environment de-
dependency issue, two separated benchmarks are built based on currently popular container technologies, Docker and Singularity. Both benchmark systems are extendable to incorporate other new indexing algorithms or benchmarking datasets, and customize different parameter settings for a particular indexing algorithm.

Benchmarking architecture of various indexing algorithms for molecular similarity search

CINF 7

Challenges in the representation and curation of polymer data

Matthew Dunbar, mdunbar@cas.org, Adam Miller. CAS, a division of the American Chemical Society, Columbus, Ohio, United States

In the world of polymer chemistry, chemical entities become defined by more than simply their molecular structure. Defining unique polymers in a way that is useful for chemists, materials scientists and engineers requires us to flip our model for their representation – properties must be the core anchor for a definition of uniqueness. Bridging this representation to polymers published in the literature can be especially challenging: with partial or inconsistent definition, it can be a challenge to extricate enough information to disambiguate them effectively. In this talk, we will highlight these challenges and CAS’s perspective on them, and discuss the ways that newer representation modes can support the community.
Challenging polymers: From the perspective of a structural database
CINF 8

Suzanna Ward, ward@ccdc.cam.ac.uk, Ian Bruno, Matthew P. Lightfoot. Cambridge Crystallographic Data Centre, Cambridge, United Kingdom

Representing chemical data in standard ways can be challenging but representing polymers adds a whole new level of complexity. The Cambridge Structural Database (CSD) is a data repository of over one million experimental crystal structures. Each entry in the CSD goes extensive curation and every underlying data set is enhanced with descriptive metadata, as well as 2D and 3D representations to ensure that the data is as discoverable and re-usable as possible so that researchers can gain new insights from this wealth of data. Polymers now account for over 10% of the structures in the CSD but the proportion of time we spend curating this type of structure is significantly more. With so many polymers in the database we have developed standard processes in terms of chemical representation and nomenclature as well as search strategies to enable users to locate and learn from this data.

This talk will highlight our current efforts to represent polymers consistently within the CSD and how our tactics compare and translate to other data resources. We will detail some of the challenges we face, how standards have evolved over the 50 plus years since the database was first established and what we can learn from our experiences. Finally, as we enter the second century of macromolecular science we will explore future opportunities for evolving how polymers can be represented in a domain specific resource like the CSD.

CINF 9

Macromolecular data representation at the worldwide protein data bank
John D. Westbrook1,2 john.westbrook@rcsb.org, Zukang Feng1,2, Vladimir Guranovic1,2, Catherine L. Lawson2,3, Ezra Peisach1,2, Chenghua Shao1,2, Jasmine Y. Young1,2, Stephen K. Burley1,2,4. (1) Research Collaboratory for Structural Bioinformatics Protein Data Bank, Rutgers, State University of New Jersey, Piscataway, New Jersey, United States (2) Institute for Quantitative Biomedicine, Rutgers, State University of New Jersey, Piscataway, New Jersey, United States (3) EMDataResource, Rutgers, State University of New Jersey, Piscataway, New Jersey, United States (4) Research Collaboratory for Structural Bioinformatics Protein Data Bank, San Diego Supercomputer Center, University of California, San Diego, La Jolla, California, United States

The Protein Data Bank (PDB) is the global repository for experimentally-determined three-dimensional structures of biological macromolecules. The Worldwide Protein Data Bank (wwPDB) is the international collaboration that manages the PDB archive according to the FAIR Principles: Findability, Accessibility, Interoperability, and Reusability. The PDB archive holds and disseminates >157,000 structures of biological macromolecules (e.g., proteins, DNA, RNA). From its inception in 1971 as the first open-access, digital-data resource in biology, PDB data architecture has evolved significantly. Starting from 72-column punch cards, today’s PDB data model incorporates production and analysis of structure data coming from macromolecular crystallography, 3D electron microscopy, and NMR spectroscopy. This transformation has paralleled advances in science, structural biology, and information technology, culminating in adoption of the PDBx/mmCIF format (http://mmcif.wwpdb.org) as the master format for the PDB.

Development of new PDB data specifications has been guided by input from community standards bodies, the outcomes of targeted workshops, and recommendations developed by working groups of community experts. The scope of content has been further shaped by
increasing community appreciation of the need for robust assessment of data quality and reproducibility. Much of this community engagement has been orchestrated by the wwPDB partnership, which convenes methods-specific validation task force meetings. Of particular importance the evolution of PDB data architecture, the wwPDB works closely with a standing PDBx/mmCIF Working Group (http://www.wwpdb.org/task/mmcif).

This talk will emphasize the important role of community engagement in developing the PDB data architecture, and how this community-focused approach continues to be applied to the challenging problem of extending the data architecture to support broader confederation of data resources in support of structural biology.

CINF 10

Representation of biopharmaceutical substances in the FDA global substance registration system and in the structured product labeling

Yulia Borodina, yulia.borodina@fda.hhs.gov, Tyler Peryea, Tyler.Peryea@fda.hhs.gov.
US FDA, Silver Spring, Maryland, United States

The FDA Global Substance Registration System (G-SRS) is a database application used by the FDA to register, maintain, index and analyze substance data based on the IDMP ISO 11238 data model. FDA experts use this system to maintain detailed information on substances found in FDA-regulated drugs, biologics, supplements, cosmetics and foods and to establish a UNique Ingredient Identifier (UNII) for every substance considered unique and unambiguous. To maintain the complex array of substances relevant to FDA’s mission, the FDA G-SRS system must support detailed structural information on a wide variety of diverse materials including small molecules, polymers, monoclonal antibodies, as well as plant and animal-derived materials. The FDA G-SRS is built on the freely distributable G-SRS software written in a collaboration between FDA and the National Center for Advancing Translational Science (NCATS).

Structured Product Labeling (SPL) is a document markup standard approved by Health Level Seven (HL7) and adopted by FDA as a mechanism for exchanging product and facility information. In the past few years the standard has been gradually extended to cover a growing number of product-related topics. One of the challenges has been to adapt the standard for describing substances used as ingredients in products. For small molecule ingredients inclusion in the SPL file of the structure representation by a MOL file and the IUPAC chemical identifier InChI was typically sufficient to both convey the structure and uniquely identify it. However, for more complex materials, such as biopharmaceuticals, more significant adjustments were necessary. The main challenges have been representing the polydispersity of materials and ensuring that the information is conveyed in an unambiguous and reproducible way.

We describe how the structural information on biopharmaceutical substances is represented in the GSRs and SPL, how we approach uniqueness check and similarity analysis, and how we can convey information provided by other sources such as International Nonproprietary Names (INN), US Adopted Names (USAN), UniProt, and Chemical Abstract Service (CAS).

CINF 11

Unknown or variable composition, complex reaction products and biological materials data management through the US EPA CompTox Chemicals Dashboard
Antony J. Williams, tony27587@gmail.com, Christopher Grulke, Grace Patlewicz, Ann Richard. Center for Computational Toxicology and Exposure, Environmental Protection Agency, Research Triangle Park, Durham, North Carolina, United States

Chemical substances of unknown or variable composition, complex reaction products, and biological materials (UVCBs) are a common form of chemical substance in EPA related chemical inventories. Tens of thousands of these substances can be represented using CAS registry numbers, descriptive text and, in some cases, a chemical structure representation. Markush representations have long been used to represent polymers, polyorganics, and other groups of chemicals sharing a general structural frame, but with variable chain lengths, repeating units, indeterminate locations of substituents on rings, etc. Using state-of-the-art Markush structure drawing and cheminformatics capabilities, unambiguous structural classes for a number of chemical types have been defined. These structure-based classes are being used to enumerate chemical members of a particular class, for example surfactants and polychlorinated biphenyls, as well as to create hierarchies (i.e., subclasses) for effective communication and structure-activity relationship (SAR) modeling activities. These approaches have specifically been applied to lists of per- and polyfluoro alkyl substances (PFAS chemicals) and although chemical names for commonly occurring PFAS have long been used to convey PFAS chemical class membership (e.g., fluorotelomers, perfluoroalkylsulphonic acids, etc.). This and other expert means of classification have drawbacks when individual PFAS span multiple possible classes and as PFAS lists extend into partially halogenated, ether, ester, aromatic, and multifunctional PFAS. This presentation will provide an overview of how the data management for UVCB chemicals is performed in our chemical registration system and how the data are delivered to the community via the freely accessible web-based CompTox Chemicals Dashboard.

CINF 12

BigSMILES: Digitalization scheme for macromolecules

Tzyy-Shyang Lin, ts.lin.92@gmail.com, Bradley D. Olsen. Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

Structurally based identifiers such as SMILES, molfile and InChI have been shown to be exceptionally useful in modern chemical information and cheminformatics research by enabling the digitalization of small molecules. However, since most existing identifiers are designed under the premise that each chemical species corresponds to one well-defined chemical structure, their applicability to polymers, which are intrinsically stochastic molecules that are usually ensembles of molecules with a distribution of chemical structures, are markedly limited.

To provide support over polymers, a new line notation, BigSMILES, that is built on top of the popular line notation SMILES, is proposed. In BigSMILES, ensembles of polymeric fragments are represented by “stochastic objects” that encodes the chemical structures by specifying the structures of the constituent repeating units and the permissible set of connectivity patterns between the repeating units. This allows efficient and compact representation of the ensemble of possible molecular structures for a polymer based on its fundamental polymerization chemistry. Through a few simple extensions to the SMILES syntax, the new BigSMILES system can easily encode a variety of polymer chemistries, including but not limited to linear polymers formed from chain and step polymerization, including random or block copolymers, branched polymers, polymer networks and ring polymers.
Aside from the line notation, an accompanying standard data format is also developed to store any additional quantifications on the distribution of the ensemble prescribed by the line notation. This information could be used to reconstruct the ensemble and infer properties necessary for informatics studies. With the two components, the BigSMILES digitalization scheme offers a multi-purpose system that can serve as both compact identifiers and comprehensive standard polymer data sharing scheme that could be useful to both the polymer and the cheminformatics communities.

CINF 13

Adapting cheminformatics identifiers for metal-organic frameworks

**Benjamin Bucior¹**, ben.bucior@gmail.com, **Andrew S. Rosen⁴**, **Maciej Haranczyk²**, **Zhen-peng Yao³**, **Michael Ziebe⁴**, **Omar K. Farha⁵**, **Joseph T. Hupp⁵**, **J I. Siepmann⁶**, **Alan Aspuru-Guzik⁷,³**, **Randall Snurr¹**. (1) Chemical and Biological Engineering, Northwestern University, Evanston, Illinois, United States (2) IMDEA Materials Institute, Madrid, Spain (3) Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts, United States (4) Chemistry, UC-Berkeley, Berkeley, California, United States (5) Chemistry, Northwestern University, Evanston, Illinois, United States (6) Chemistry, University of Minnesota, Minneapolis, Minnesota, United States (7) Chemistry and Department of Computer Science, University of Toronto, Toronto, Ontario, Canada

Metal-organic frameworks (MOFs) are a class of nanoporous materials, sometimes referred to as coordination polymers. They have been explored for many applications including gas separations and catalysis. MOFs are synthesized in a building-block approach from inorganic “nodes” and organic “linkers,” which assemble into various topologies, providing an enormous design space with many thousands of MOFs synthesized to date. However, MOFs have conventionally been identified with an ad hoc nomenclature. As such, multiple names can refer to the same MOF, and there is no guarantee that a given name is unique (e.g. Cu-BTC, HKUST-1, and MOF-199 describe the same material, while MOF-48 has been used to describe two unrelated MOF structures a decade apart). This makes it difficult to carry out large-scale data mining efforts, easily search for specific MOF structures, and link MOF datasets together. In this work, our team has developed two MOF identifier formats: MOFid provides detailed information about the building blocks using a SMILES-based representation, and MOFkey provides a compact barcode derived from InChIKeys. Our method deconstructs MOFs into their basic building blocks and topology, which improves interoperability with existing chemical databases while also avoiding certain inconsistencies when directly applying cheminformatics representations to periodic crystal structures. Applying MOFid and MOFkey has enabled us to rapidly search within MOF structural databases, find duplicate MOFs in common between databases, and discover new insights. Through the process of designing MOFid and MOFkey, we provide a perspective on opportunities for the community to facilitate data reuse, improve searchability, and rapidly apply cheminformatics analysis to complex crystalline materials.

CINF 14

Current progress in HELM developments and data integration

**David Deng¹**, dengw2@gmail.com, **Tony Yuan¹**, **Jinbo Lee²**, **Rajeev Hotchandani¹**. (1) Scilligence, Cambridge, Massachusetts, United States (2) Scilligence Corporation, Cambridge, Massachusetts, United States
Biologic representation and registration has always been a challenge due to the variety of custom notations. Since its creation by Pfizer scientists in 2008 and introduction by the Pistoia Alliance in 2013, HELM (Hierarchical Editing Language for Macromolecules) notation has been increasingly adopted as an industry standard for biologic sequence representation and registration. Biologic representation and registration has always been a challenge due to the variety of custom notations. Since its creation by Pfizer scientists in 2008 and introduction by the Pistoia Alliance in 2013, HELM (Hierarchical Editing Language for Macromolecules) notation has been increasingly adopted as an industry standard for biologic sequence representation and registration. With its ability and flexibility to represent various types of complex macromolecules, scientists can now save biologic sequences in an exchangeable format rather than that of their own organization. To help expand HELM adoption, Pistoia Alliance released an open source HELM web editor and various informatics vendors have added HELM notation support.

As a Pistoia Alliance member, Scilligence has long been assisting HELM 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 15

HELM in the wild: Real-world applications in a pharmaceutical discovery environment

Mark Somogyi, Roland Knipsel, Tim Parrott, tim.parrott@rocketmail.com. ChemAxon, San Diego, California, United States

Hierarchical Editing Language for Macromolecules (HELM) is used to represent, store, and search a variety of biological macromolecules including proteins, siRNAs, ADCs, etc. In this presentation we will discuss how scientists at several partner organizations are using HELM tools, how we have integrated these tools with other discovery tools, and some of the challenges encountered in those implementations.

CINF 16

Chemical information integration and data exchange in PubChem

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

Many open access databases exist for chemical information. These databases may focus on one topic to another. For example, while the International Chemical Safety Card (ICSC: https://www.ilo.org/dyn/icsc/showcard.home), OSHA Occupational Chemical Database (https://www.osha.gov/chemicaldata/), and NIOSH Pocket Guide to Chemical Hazards (NPG: https://www.cdc.gov/niosh/npg/) are mainly focusing on the occupational hazardous substance safety information, MassBank of North America (MoNA: http://mona.fiehnlab.ucdavis.edu/), the SpectraBase at Bio-Rad (http://www.bio-rad.com/specetroscopy), and NIST Mass Spectrometry Data Center (https://chemdata.nist.gov/) cover spectral data. When searching chemical information, scientists sometimes need to browse from one dataset to another in order to find the right data. PubChem (https://pubchem.ncbi.nlm.nih.gov/), as an integrated open chemical database, aggregated chemical information from many open and partial open databases to let scientist shop chemical
information at one central location. Moreover, PubChem provided data comparison, possibly, showing data under use conditions or other factors, and also allowing scientists to track to the data provenances. In this presentation, we will discuss the opportunities and challenges in the data integration and exchange while handling hundreds of datasets.

CINF 17

Curating ChemSpider: Challenges in chemical data management

Mark Archibald, archibaldm@rsc.org. Royal Society of Chemistry, Cambridge, United Kingdom

This talk will discuss the challenges of managing large-scale chemical datasets and how the Royal Society of Chemistry approaches the curation of new and existing ChemSpider data.

CINF 18

Open API for searching and accessing thermophysical and thermochemical properties provided by the NIST Thermodynamics Research Center

Demian Riccardi, demian.riccardi@nist.gov, Chris D. Muzny. Thermodynamics Research Center, NIST, Boulder, Colorado, United States

The NIST Thermodynamics Research Center (TRC) provides critically evaluated thermophysical and thermochemical property data complete with provenance and uncertainty. The collection/evaluation is an intensive process that is limited to published experimental data. We will present a new, open RESTful API for searching and retrieving data that is currently available in the ThermoML archive. Major goals of the API include: interoperability with other open sources of chemical information; enabling the discovery of data that may be used for validation, comparison, and the design of new experimental and theoretical investigations. Further, we will discuss methods for users to post structured information that TRC may use for its own discovery of reliable sources of thermodynamics data.

CINF 19

Sharing and linking chemical structure data on Pub Chem and institutional repositories

Vincent F. Scalfani, vincent.scalfani@gmail.com. University Libraries, University of Alabama, Tuscaloosa, Alabama, United States

Chemical structures in University theses and dissertations are often never published in the chemical literature. As a result, these structures are challenging to discover, as they are not indexed by secondary abstract and indexing operations. Many university libraries have started to share theses and dissertations openly on their digital institutional repository websites. While open sharing of full text theses and dissertations has certainly improved access to university chemical information via keyword searching, specific chemical structure and search discovery is still largely absent for university theses and dissertations. This presentation will provide an update on how the University of Alabama Libraries is extracting chemical structures from theses and dissertations and sharing them openly on PubChem. The PubChem substance records are then linked to the dissertation/thesis record at the Institutional Repository and Catalog, greatly increasing the discovery of chemical structure data from university
research. We will discuss our workflows, challenges and opportunities with our approach, in particular considering aspects of linking data, data quality, provenance, and copyright.

**CINF 20**

**Data interchange en masse via batch search functionality in the CompTox Chemicals Dashboard**

Antony J. Williams, tony27587@gmail.com. Center for Computational Toxicology and Exposure, Environmental Protection Agency, Research Triangle Park, Durham, North Carolina, United States

There are a myriad of public free access chemistry databases now accessible to the community to source various types of data to support, for example, drug discovery, environmental chemistry and materials research. These databases can be generalized content and contain tens of millions of chemicals (e.g. PubChem and ChemSpider) or be smaller “boutique” databases focused on specific data collections (e.g. SuperDrug and DrugBank). In general many of these sites offer flexible search capabilities including mass, formula, structure, substructure and structure similarity searches as well as providing downloads of their data for inclusion into other databases. The US EPA CompTox Chemicals Dashboard is a web-based application providing access to ~880,000 chemical substances and diverse data types including physicochemical property, toxicity and bioactivity data. While the application supports the users with the expected single chemical search (based on CASRN, chemical name, InChI Key etc) one of the most powerful pieces of functionality is the batch search that allows a user to search of thousands of chemicals at a time. Batch searching using each of, or a combination of chemical names, CAS RNs, InChIs, masses or formulae as inputs facilitates downloads of data *en masse* into either Excel spreadsheet, comma or tab-separated value files, or into an SDF file containing thousands of chemicals. This presentation will provide an overview of the batch search capability in the dashboard and the rich data streams that are made accessible by this functionality and available for exchange with other systems. It will also review the Open data that has been made available for download from the site so that it can be reused and repurposed in other systems.

**CINF 21**

**Integration of US EPA webtools via web-services**

Todd Martin¹, martin.todd@epa.gov, Antony Williams¹, Valery Tkachenko². (1) Center for Computational Toxicology and Exposure, US Environmental Protection Agency, Research Triangle Park, North Carolina, United States (2) Science Data Experts LLC, Rockville, Maryland, United States

WebTEST (Web-services Toxicity Estimation Software Tool) was developed within the US EPA’s Chemicals Dashboard (https://comptox.epa.gov/dashboard) to provide real time estimates of toxicity values (e.g. acute mammalian toxicity and Ames mutagenicity) and physical properties (e.g. melting point and water solubility). WebTEST estimates toxicities/properties using several QSAR (quantitative structure activity relationship) methodologies. QSAR models correlate toxicity with molecular structure using molecular descriptors as independent variables. A web interface was created to allow users to make predictions for single chemicals (by drawing them or by searching in the database of ~880,000 chemicals underlying the Dashboard). A batch mode has also been developed to enable prediction for multiple
chemicals. Users can also access WebTEST predictions via web-services (such as using a simple URL). Recently EPA has released the Chemical Transformation Simulator (CTS, https://qed.epacdx.net/cts/), a web-based tool for predicting environmental and biological transformation pathways and physicochemical properties of organic chemicals. WebTEST’s web-services chemicals. Users can also access WebTEST predictions via web-services (such as using a simple URL). Recently EPA has released the Chemical Transformation Simulator (CTS, https://qed.epacdx.net/cts/), a web-based tool for predicting environmental and biological transformation pathways and physicochemical properties of organic chemicals. WebTEST’s web-services are being utilized by EPA’s CTS (Chemical Transformation Simulator) to provide additional estimates of physicochemical properties. Finally, the CTS web-service is utilized within WebTEST to predict environmental transformation products and simultaneously estimate their toxicities.

Interface for single chemical predictions on the EPA Chemicals Dashboard

**CINF 22**

**Exchange and integration of crystallographic data and knowledge**

*Ian Bruno*, bruno@ccdc.cam.ac.uk, Suzanna Ward. Cambridge Crystallographic Data Centre, Cambridge, United Kingdom

For a number of years now, the Cambridge Crystallographic Data Centre (CCDC) has provided free online access to what is now over one million crystal structures in the Cambridge Structural Database (CSD). Many of these structures have a Digital Object Identifier associated with which enables them to be linked to from a range of other resources. To many of these
structures we can also assign standard InChIs that can be used in combination with the DOI as the basis for integration of small molecule crystal structure data with both chemical and biological information resources.

The structures in the CSD have been determined experimentally. This means that they are prone to experimental artefacts that if not treated appropriately can result in mis-assignment of InChIs. This can be avoided if we ensure that we have a validated chemical representation of the molecules in the crystal structure and additional metadata that can be used to verify correspondences between resources. Automated and manual curation of data and metadata is key to being able to reliably integrate data with other resources. The aggregation of crystal structure data in the CSD can be further used as a foundation for exchange and integration of knowledge derived from experimental results with other data resources and workflows. Knowledge about molecular geometry, for example, can be used to validate 3D data being deposited in other resources such as the Protein Data Bank (PDB) and the results of this validation included in reports that accompany the data.

This presentation will offer an overview of work that the CCDC has undertaken to provide integrations between the CSD and chemistry information resources such as PubChem, ChemSpider, Drugbank, the Pesticides Properties DataBase and others. It will discuss the challenges associated with ensuring that the data exchanged is sufficiently reliable and touch on considerations around respecting restrictions required to sustain repository funding models. It will also describe how derived knowledge can be integrated with curation workflows in other systems to help ensure the quality of data resources in biology as well as chemistry.

CINF 23

Data exchange and integration with DrugBank, HMDB and other Canadian cheminformatics resources

David S. Wishart, dwishart@ualberta.ca. Biological Sciences, University of Alberta, Edmonton, Alberta, Canada

For the past 15 years my group at the University of Alberta has been developing and maintaining a large number of open access chemical information resources for applications in drug discovery, metabolomics and exposome research. Some of our better known databases include DrugBank and the Human Metabolome Database (or HMDB). In this presentation I will briefly describe the history behind these resources and the development of several other widely used cheminformatics databases (such as YMDB, FooDB, PathBank, ECMDB, ClassyFire, BioTransformerDB, etc.) by members of my laboratory. In particular, I will describe several aspects of the collection, curation, presentation and quality control of these open access data resources. I will also describe the efforts we have made to improve the ability to link and exchange data between our databases and other chemical information databases. These include the development of NMR exchange data formats, the creation of spectral hash tags (SPLASH), the development of structural and functional ontologies, the creation of tools for structure regularization, the development of automated chemical annotation tools and the automated linking to other databases. Many of these tools are also freely available and it is my hope that these can help in making other chemical information resources more consistent, more integrated and much more FAIR (findable, accessible, interoperable and reusable).
CINF 24

Makerspace in academic institutions: Operation, safety and challenges

Shailendra Singh¹, shailen2@andrew.cmu.edu, Jesse Decker². (1) Environmental Health and safety, Carnegie Mellon University, Pittsburgh, Pennsylvania, United States (2) College of Engineering, University of Wisconsin-Madison, Madison, Wisconsin, United States

UW-Madison made a large commitment to undergraduate makerspaces in 2001 with the design of their Engineering Centers Building TEAM Lab. It included dedicated spaces for wood, metal and welding work along with an expansive design and fabrication area. Beyond just providing space, the College of Engineering funded a significant workforce of professionals to educate students on the equipment and materials that were available. In 2018, The College of Engineering added Makerspace for 3-D printing, soldering, laser cutting and materials work; again, adding more professional staff to oversee the spaces. These professional staff are assisted by large numbers of undergraduate employees that allows for extended work hours. While initially trainings were given in person for the woodwork, metalwork and welding equipment, the popularity of the TEAM Lab caused the college to go to an online safety training system with only select higher-hazard equipment continuing to require in-person training.

Learn the steps UW-Madison’s College of Engineering took to ensure that their makerspaces were both cutting edge and safe for their student users.

CINF 25

Don't cut that! Lessons learned from a laser cutter fire

Nathaniel J. Leon, Daniel R. Kuespert, dankuespert@me.com. G.W.C. Whiting School of Engineering, The Johns Hopkins University, Columbia, Maryland, United States

We recount an August 2019 fire in a laser cutter in a university makerspace, resulting from an attempt to cut closely-spaced lines in foam-core board. Student emergency response to this incident was suboptimal, and the resulting investigation revealed at least one other similar fire that went unreported. Lessons learned from the incident relate to establishment of technical design standards that respect the limits of safe machine operation, recognition that students using makerspaces are generally novices on the use of equipment, and changes to training to incorporate adherence to technical standards and improvements in emergency response & reporting training.

CINF 26

Are you ready to have a laser cutter in your makerspace? Think again!

Andrew Lawson, alawson@andrew.cmu.edu, Shailendra Singh. Environmental Health and safety, Carnegie Mellon University, Pittsburgh, Pennsylvania, United States

The objectives of this session will be to describe the strategies and lessons learned that EH&S and their partners have developed to reduce the risks associated with laser cutter usage. This session will cover some of the challenges EH&S has worked to overcome since the first laser cutter was purchased at Carnegie Mellon University in 2007 and still encounters today. We will be presenting two case studies that demonstrate the unintended or unknown consequences of purchasing a laser cutter. This session will be very useful for faculty and staff who support teaching, research, and maker spaces in an academic institution.
CINF 27

Student library employees make the makerspace safe!

Annie Zeidman-Karpinski, annie@uoregon.edu, Jimmy Murray, Dean Walton. University of Oregon, Eugene, Oregon, United States

Our presentation will cover the following:

- An overview of the DeArmond Makerspace, specifically highlighting usage data in the context of our large public campus.
- Our path through the campus administration approval process which included discussions with campus Environmental Health & Safety as well as our General Counsel’s office.
- A discussion of our current safety program including information about how our materials have evolved over the last three years to meet the ongoing needs of our users and specifically our use of the Red Safety Sheet Notebook.
- Finally, we will present a case study about how we involve our student employees with staffing the space. Our students have taken the lead on safety trainings and workshops, which in turn, means that we have developed a rich student employment program that engages our employees and contributes to their success.

CINF 28

3D printing in library makerspace: Health and safety concern

Neelam Bharti¹, nbharti@andrew.cmu.edu, Shailendra Singh², Aastha Singh³. (1) Mellon Institute Library, Carnegie Mellon University, Pittsburgh, Pennsylvania, United States (2) Environmental Health and safety, Carnegie Mellon University, Pittsburgh, Pennsylvania, United States (3) Obama Academy, Pittsburgh, Pennsylvania, United States

3D printing is a form of technology that is steadily gaining popularity in library makerspaces, homes, and workspaces around the globe. Several types of 3D printing technologies and materials are available, but most common 3D printers use plastic filament for printing objects. Several research articles report that industrial-sized 3D printers emit harmful ultrafine particles (UFPs) while printing with plastic filament, but there is no information on UFP emissions from portable 3D printers. We will discuss the results of our UFP emission study using portable printers and how we can operate these printers safely.

CINF 29

New risk paradigm: Advancement of campus makerspaces and safe innovation

Tolga Durak, tdurak@mit.edu. Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

This session will provide an overview of the evolution of makerspaces in higher education and at MIT. Over the recent years, the communal makerspaces have prioritized fostering unrestricted making via a community effect. In this ecosystem, students are often the community members who serve as stewards of the space/resources and educate/govern users in safe making practices. Thus, creating a new challenge of balancing the design, daily oversight and
the operational parameters to enable maximum utilization of such innovative spaces with management of risk and broader perspective on regulatory compliance. During the session, the panelists will share insight and case studies that highlight the evolution of communal and other types of makerspaces at MIT. The discussion will include information on utilization of smart design with embedded safety features and the structure of safety ecosystem framework that promotes and enables safety culture as well as reasonable access.

CINF 30

Batch searching chemicals data in the US-EPA CompTox Chemicals Dashboard

Antony J. Williams, tony27587@gmail.com, Christopher Grulke, Jeremy Dunne, Jeff Edwards. Center for Computational Toxicology and Exposure, Environmental Protection Agency, Research Triangle Park, Durham, North Carolina, United States

While many publicly accessible databases provide access to rich data streams including experimental and predicted physicochemical property data, toxicity data and chemical synonyms and identifiers, intuitive batch searches allowing querying using thousands of various identifiers (e.g. Names, InChls, CAS RNs etc.) are lacking. The US EPA CompTox Chemicals Dashboard is a web-based application providing access to ~880,000 chemical substances and associated property, toxicity and bioactivity data. One of the most powerful pieces of functionality is the batch search that allows a user to search of thousands of chemicals at a time, using each of, or a combination of chemical names, CAS RNs, InChls, masses or formulae. These inputs can be used to harvest data into either Excel, comma or tab-separated value files, or into an SDF file containing thousands of chemicals. This poster will provide an overview of the batch search capability in the dashboard and the rich data streams that are made accessible by this functionality.

CINF 31

Detecting chemical reactions in patents

Hiyori Yoshikawa¹, Karin Verspoor¹, Timothy Baldwin¹, Dat Nguyen¹, Zenan Zhai¹, Saber Akhondi², Camilo Thorne³, c.thorne.1@elsevier.com, Christian Druckenbrodt³. (1) University of Melbourne, Melbourne, Victoria, Australia (2) Elsevier, Amsterdam, Netherlands (3) Elsevier, Frankfurt, Germany

Extracting chemical reactions from patents is a crucial task for chemists working on chemical exploration. In this presentation we introduce the novel task of detecting the textual spans that describe or refer to chemical reactions within patents. We formulate this task as a paragraph-level sequence tagging problem, where the system is required to return a sequence of paragraphs which contain a description of a reaction. To address this new task, we construct a silver-standard dataset from an existing proprietary database of chemical reactions manually extracted from patents. We introduce several baseline methods for the task and evaluate them over our dataset. Through error analysis, we discuss what makes the task complex and challenging, and suggest possible directions for future research.
Structurally annotated lists of chemicals from international frameworks for the control of chemical weapons and precursors

Stefano Costanzi¹, costanzi@american.edu, Charlotte Slavick¹, cs1229@student.american.edu, Gregory D. Koblentz², Richard Cupitt³. (1) Department of Chemistry, American University, Washington, District of Columbia, United States (2) Schar School of Policy and Government, George Mason University, Arlington, Virginia, United States (3) Stimson Center, Washington, District of Columbia, United States

International frameworks for the nonproliferation of chemical weapons (CW) and intergovernmental organizations have assembled lists of chemical warfare agents and precursors for their synthesis that are subject to controls and verification regimes. Some of these lists are compilations of individual chemicals. Other ones comprise individual chemicals as well as families of chemicals identified by a common scaffold with variable substituents.¹,²

To support these multilateral efforts to stem CW proliferation, we are putting together a website (https://costanziresearch.com/cw-control-lists/) in which the lists of chemical warfare agents and precursors from the Chemical Weapons Convention (CWC), the Australia Group (AG), the Wassenaar Arrangement and the World Customs Organization (WCO) are collected into manually curated tables.

First, in our tables, all entries are annotated with chemical structures (exact structures for individual chemicals and Markush structures for families of chemicals). This is important because chemicals are better described by structures than names. A single letter difference in a chemical name can account for an important difference in the chemical structure, marking the watershed between chemicals that are controlled and those that are not. Hence, documents annotated with structures will make it easier for chemists and scientific advisors to communicate these differences to policy makers.

Our tables also are annotated with information that highlights the overlaps within the various lists, noting for each entry of each list whether that chemical is covered by one or more additional lists, either as an individual chemical or as a member of a family of chemicals. Importantly, to further highlight overlaps and differences between the lists, we also provide a synoptic table in which all the lists are provided side-by-side on one page.

Finally, to facilitate the work of chemists working in CW nonproliferation, we are annotating our tables with simplified molecular-input line-entry system (SMILES) notations, and downloadable 2D coordinates

CINF 33

Cys. sqlite: Structured-information approach to the comprehensive analysis of cysteine disulfide bonds in the protein databank

Demian Riccardi², demian.riccardi@nist.gov, Theodore L. Fobe¹, Andrei F. kazakov². (1) Chemical and Biological Engineering, University of Colorado Boulder, Boulder, Colorado, United States (2) Thermodynamics Research Center, NIST, Boulder, Colorado, United States
Cysteine is a multifaceted amino acid that is central to the structure and function of many proteins. A disulfide bond formed between two cysteines restrains protein conformations through the strong covalent bond and torsions about the bond that prefer, energetically, ±90°. In this study, we transform over 35,000 Protein Databank files (PDBx/mmCIFs) into a single file, the SQLite database (Cys.sqlite). The database schema is designed to accommodate the structural information on both oxidized and reduced cysteines and to retain essential protein metadata to establish informational and biological provenance. Cys.sqlite contains over 27,000 sequences, 95,000 peptide chains, and 500,000 cysteines (700,000 structural conformers); there are over 300,000 cysteine disulfide bond conformations from structures solved with all available experimental methods. The structural information is analyzed with respect to sequence identity cutoff, the experimental method, and energetics of the disulfide.

CINF 34

Clustering of the phosducin protein family and its functionally relevant groups

Sarah Hosler1, sehosler44@gmail.com, Jacquelyn Fetrow2. (1) Biology, Albright College, Andover, New Jersey, United States (2) Chemistry and Biochemistry, Albright College, Reading, Pennsylvania, United States

The goal is to functionally classify proteins using their active sites. This method will lead to a detailed understanding of the molecular functional mechanisms. This research can be further used in phosducin research and drug development. The phosducin proteins are mainly found in the retina and bind to G-proteins. Phosducins are also found in other parts of the body, but its other functions are not determined. We classified the phosducin family using auto multi-iterative search sequencing technique (MISST), a set of computer scripts that incorporates Active Site Profiling (ASP), and Deacon Active Site Profiler 3 (DASP3). The program searches protein databases to cluster proteins with similar active site profiles. We have classified phosducins into two functional families. We are also classifying the phosducins on whether they have the binding site. This analysis reveals interesting functional details of this protein family.

CINF 35

Bringing order from chaos: FAIR database of sesquiterpenes organized by dihedral angles

Julie M. Schrey, julie.schrey001@albright.edu, Christian Hamann. Chemistry & Biochemistry, Albright College, Reading, Pennsylvania, United States

Farnesyl diphosphate is a biomolecule transformed via a series of carbocation intermediates into the vast number of sesquiterpenes found in nature. Sesquiterpenes are fifteen-carbon secondary metabolites that serve a variety of functions. For example, they aid plants in the protection against predators and in the attraction of pollinators. In addition, they are widely used in the flavors and fragrances industry: some common examples are santalene, which is the characteristic sandalwood scent, and zingiberene, the characteristic flavor of ginger. Scientists estimate there may be 1,000-5,000 sesquiterpene molecules. Each could have 3-5 carbocation intermediates and multiple conformers of these, so in the study of sesquiterpene biosynthesis there could be more than 100,000 isomers or conformers to track! This incredible structural diversity results in a major data organization issue; currently there is no...
universal way to organize this overwhelming dataset. Further complicating the issue, the ma-
majority of datasets regarding sesquiterpene biosynthesis exist within supporting information
files or personal research files in a variety of file formats. In the growing movement to make
chemical data more FAIR (findable, accessible, interoperable, and reusable), this chaos calls
for the formation of a FAIR database of sesquiterpenes. The novel approach described herein
uses dihedral angle analysis to uniquely characterize both carbocation pathway intermediates
and their resulting natural products. A systematic numbering system, related to but independ-
ent of connectivity, has been applied to all data disinterred from the supporting information
produced by several research groups. The combination of unique characterization and sys-
tematic numbering can facilitate communication amongst different research groups. A more
accurate and more useful scientific literature results, as different groups can more easily vet
and contextualize new results. Further utility is realized as the database may be mined for
new perspectives on the biosynthesis of sesquiterpenes. In summary, the chaos of a vast
sesquiterpene database can be brought to order using dihedral angle analysis and a system-
atic numbering system to develop a searchable database with predictive power, significantly
enhancing FAIRness of this dataset.

CINF 36

Integration of heterogeneous data using the SciData Framework and JavaScript for
linked data to improve the FAIRness of scientific data

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

The heterogeneity of scientific data and its distribution across different general and domain
repositories, databases, and at various application programmin interfaces (APIs) makes it dif-
ficult to integrate. This is a significant hurdle to both big and small data applications in the
chemical sciences. Achieving interoperability of data is one of the tenets of Findbale, Accessi-
ble, Interoperable, and Reuseable (FAIR) principles.

The SciData Framework is an digital representation of the scientific data model focusing on
the aggregation of data and metadata around three primary sections - methodology, system,
and dataset. Generically, scientific experiments are carried out to study a specific system, us-
ing a specific methodology, to collect one or more piece of experimental data. The SciData
Framework is structured in such a way that it can be used to organize data from any discipline
at any level of detail.

In this presentation the use of the SciData framework to organize data/metadata in the Ja-
vaScript for Linked Data (JSON-LD) is applied to various heterogeneous datasets such that
the JSON-LD files can then be converted to Resource Description Framework (RDF) triples,
ingested into a graph database, and the subsequently searched using SPARQL and SHACL
queries. The conversion process, as will interesting correlations in the data, will be discussed.
It is proposed that data encoded in the SciData Framework is FAIR for both humans (the
JSON-LD file) and machines (RDF) and could be the foundation of the next generation digital research notebooks
CINF 37

30-year analysis of intellectually-indexed Markush structures from patents

MATTHEW MCBRIDE, MMCBRIDE@CAS.ORG, Elaine N. Cheeseman. Science IP, Chemical Abstracts Service, Newtown Square, Pennsylvania, United States

Markush structures are representations of chemical structure used to indicate a group of related chemical compounds commonly located in patent claims and descriptions, and are key resources in intellectual property searching, such as patentability, Freedom to Operate, and validity searches. MARPAT, CAS’s source of Markush structures from patents and patent applications, recently achieved two milestones – it is now 30 years old and contains more than 500,000 patent records. An analysis of the Markush disclosures in MARPAT over time, the changes in complexity and diversity, and commercial domains covered will be presented.

CINF 38

IUPAC SMILES+ project: Developing a validation suite

Vincent F. Scalfani¹, vincent.scalfani@gmail.com, Christopher Grulke², Evan Bolton³, Gregory Landrum⁴, Helen Cooke⁵, Issaku Yamada⁶, John J. Irwin⁷, Jose L. Medina-Franco⁸, Miguel Quirós⁹, Oliver Koehler¹⁰, Susan Richardson¹¹, Francis Atkinson¹². (1) University Libraries, University of Alabama, Tuscaloosa, Alabama, United States (2) National Center of Computational Toxicology, US EPA, New Hill, North Carolina, United States (3) National Center for Biotechnology Information, Bethesda, Maryland, United States (4) T5 Informatics GmbH, Basel, Switzerland (5) Nantwich Museum, Cheshire, United Kingdom (6) The Noguchi Institute, Tokyo, Japan (7) Pharmaceutical Chemistry, University of California San Francisco, San Rafael, California, United States (8) Universidad Nacional Autonoma de Mexico, 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) Cambridge Crystallographic Data Centre, Cambridge, United Kingdom

The IUPAC SMILES+ project seeks to create up-to-date documentation of the SMILES specification. One component of the project is to develop a validation suite, useful for cheminformatics toolkit testing. The SMILES validation suite would serve to help identify toolkit SMILES reading differences, and demonstrate the standard interpretation of SMILES. This poster presentation will highlight our efforts with creating a validation suite using InChI. We will also discuss our efforts to use a custom JSON validation format, which may be useful to present a lossless standard interpretation of SMILES.

CINF 39

Predictions of pairs of drugs with synergistic activity against leukemia cell lines based on NCI-ALMANAC data

Alexey A. Lagunin²,¹, alexey.lagunin@ibmc.msk.ru, Liubov O. Soina². (1) Department of Bioinformatics, Institute of Biomedical Chemistry, Moscow, Russian Federation (2) Department of Bioinformatics, Pirogov Russian National Research Medical University, Moscow, Russian Federation
Drug resistance of tumor cells is one of the main problems in oncology. A combination of several drugs is used to overcome tumor resistance. There are a huge number of possible combinations of drugs. It is not possible to test all combinations of pairs of drugs experimentally to identify synergy. We created classification SAR models for prediction of synergistic pair drug combinations against 6 leukemia cell lines based on the data from NCI-ALMANAC database (https://dtp.cancer.gov/ncialmanac/). NCI-ALMANAC database contains the results of experimental testing of 5232 drug pairs in cell lines of NCI-60 panel. The modified version of PASS (substructural PoSMNA (Pairs of Substances Multilevel Neighbourhoods of Atoms) descriptors and naïve Bayes like algorithm) and GUSAR (electro-topological QNA (Quantitative Neighbourhoods of Atoms) descriptors and self-consistent regression) software were used for the creation of SAR models. The accuracy of SAR models was verified using the 5-fold cross-validation procedure. For the best models, AUC and balanced accuracy were more than 0.85 and more than 0.78, respectively. The SAR models were used to screen potentially synergistic antitumor pairs of FDA approved drugs acting against 6 leukemia cell lines. The pairs of drugs with potential synergistic antitumor effect against all 6 leukemia cell lines were found.

CINF 40

Deciphering chemical information of small molecules contributing to biased GPCR signaling

Jason Sanchez, jsanchez6@miners.utep.edu, Suman Sirimulla, ssirimulla@utep.edu.
School of Pharmacy, The University of Texas at El Paso, El Paso, Texas, United States

G-Protein Coupled Receptors (GPCRs) are able to modulate cellular responses by means of different signaling cascades. These cascades involve other downstream players such as G-proteins, β-Arrestin (β-Arr), and Mitogen-Activated Protein Kinases (MAPK). A ligand is termed biased when it preferentially activates one particular GPCR signaling cascade over another. In this project, we explored the recently available public database, BiasDB, that complied information about ligand-protein combinations resulting in signaling bias. This database contains 618 cases of signaling bias describing specific G-protein (Gs, Gi, Go, Gq), β-Arr and MAPK bias. Unique pharmacophores that contribute to the biased signaling and the machine learning models trained on this data to characterize ligands as specific biased ligands will be presented.

CINF 41

Undefined stereochemistry in ChemSpider: Application of machine learning

Mark Archibald, archibaldm@rsc.org. Royal Society of Chemistry, Cambridge, United Kingdom

Structures submitted to ChemSpider that contain undefined stereochemistry range from 'almost certainly correct' to 'almost certainly wrong'. A set of rules and specified exceptions can be used to try to identify structures to keep or discard, but this approach requires human review to check for false positives. A machine-learning approach to identifying which structures to keep was developed, eliminating the need for human intervention.
CINF 42

Generating 3D transition state structures with deep learning

Lagnajit Pattanaik\textsuperscript{1}, lagnajit@mit.edu, John Ingraham\textsuperscript{2}, Colin Grambow\textsuperscript{1}, Tommi Jaakkola\textsuperscript{2}, Regina Barzilay\textsuperscript{2}, Klavs F. Jensen\textsuperscript{1}, William H. Green\textsuperscript{1}. (1) Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States (2) Computer Science & Artificial Intelligence Lab, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

The generation of complex chemical mechanisms requires exploration of large molecular spaces. For example, the Reaction Mechanism Generator (RMG)--an open-source software developed at MIT and successfully applied to modeling systems such as combustion and pyrolysis--typically scans 10,000 to 1,000,000 species during a single run. A crucial task in this exploration involves accurate calculation of reaction kinetics, for which correct transition state structures are required. Popular methods for optimizing transition state structures, such as the Berny algorithm, require good initial guesses for successful convergence. However, obtaining these initial guesses can be a major challenge. Interpolation methods such as linear and quadratic synchronous transit have high failure rates, while computational methods such as nudged elastic band, growing string method, and freezing string method require significant time for convergence. Expert-guided and data-driven methods such as KinBot and AutoTST overcome these challenges, but they are restricted to explicit reaction families without the ability to extrapolate to new chemistries. Here, we present a machine learning algorithm to generate 3D transition state structures. The method is conditioned on 3D geometries of the reactants and products and uses a distance matrix representation as input to a message passing neural network. The final output of the network is the xyz coordinates of the transition state geometry. The method is trained on an internal set of isomerization transition states created by successful application of the growing string method. On a held-out test set, our algorithm generates correct geometries for 76\% of all reactions and 91\% of reactions falling within RMG reaction families (i.e. well-known gas-phase reaction types), where each success is verified by an intrinsic reaction coordinate calculation correctly matching the IRC reactant and product to the ground-truth reactant and product. An additional evaluation shows that the method translates to new chemistries.

CINF 43

Adapting evolutionary algorithms for autonomous machine learning in chemistry

Gaurav Vishwakarma, gvishwak@buffalo.edu, Mojtaba Haghighati, Johannes Hachmann. Dept of Chemical and Biological Engineering, University at Buffalo, SUNY, Buffalo, New York, United States

Over the past few years, there has been a rising trend in the use of data-driven approaches for the discovery and design of new chemistry. These strategies encompass virtual high-throughput exploration of chemical space and analysis of the generated data using data mining techniques, and our group has developed a cyberinfrastructure for accomplishing these tasks. However, in order to accelerate this process, it is critical to identify and optimize several steps in our research paradigm and we therefore employ a Genetic Algorithm (GA) for generating high quality solutions to search and optimization problems.
Rational model selection and hyperparameter optimization are important concerns for the efficient and successful use of machine learning but have so far largely remained unexplored by this community. To address these issues, we use GA to automatically perform these tasks and present results for training an artificial neural network to predict refractive indices of one hundred thousand small organic molecules. Further, we benchmark its performance against other algorithms popular in the data science community and the results show that our implementation outperforms these methods both in terms of time and accuracy. The effectiveness of our implementation is further demonstrated via a scenario involving multi-objective optimization for model selection.

Next, we use the GA-optimized machine learning model for small organic molecules to ‘reverse engineer’ new molecules that will have a high refractive index. Since chemical space is infinite and its exhaustive exploration is impractical for any real application, it is desirable that this exploration proceeds in a direction biased towards a set of target properties. In principle, a fitness function for GA can be designed such that it penalizes sub-structures that suppress the desired property in a molecule. However, the biggest challenge in using GA for high-throughput screening is the encoding of molecules into a GA chromosome. We modify our implementation of GA to not only handle this by using a node-network of python objects for every molecule, but also keep track of the connections between molecules which can later be used for creating structural descriptors in a machine learning model. Using GA, we propose new molecules with high refractive index and validate the predicted values using Density Functional Theory.

CINF 44

Chemists’ data needs for machine learning research in academia

Ye Li, yel@mit.edu. MIT Libraries, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

Applied data science and machine learning (ML) are making differences in the chemistry domain including molecular design of drugs and materials as well as prediction and optimization of synthesis routes and production processes. To train effective ML models for these emerging areas, a large amount and a variety of well-curated experimental datasets are essential. Among such datasets needed in chemistry, the structured and curated datasets are often-times locked down in proprietary indexing databases, corporate in-house databases, or reference works, either in print or digitized; while the unstructured data scattered in the published research articles. In this presentation, we will first analyze the recent publications reporting ML research in chemistry and examine some ML research projects at MIT to highlight the types and extent of chemistry datasets needed for ML. Then, we will discuss the availability of these datasets and texts as well as the economic, legal and technical challenges in obtaining access to these datasets for data and text mining purposes. To tackle these challenges, we will share some strategies on exploring and developing collaborations among researchers, librarians, publishers, and database providers.

CINF 45

Predicting the glass transition behavior of polymers via integration of cheminformatics and molecular dynamics simulations
Anas k¹, anas.karuth@gmail.com, Bakhtiyor Rasulev², wenjie xia³, amirhadi alesadi³. (1) coating and polymeric materials, North Dakota State University, Fargo, North Dakota, United States (2) Department of Coatings and Polymeric Materials, North Dakota State University, Fargo, North Dakota, United States (3) civil and environmental engineering, North Dakota State University, Fargo, North Dakota, United States

The glass transition temperature (Tg) is one of the most important thermophysical properties of polymers. The substantial change in polymer dynamics at glass transition temperature causes a major change in physical properties, including mechanical modulus, density, specific heat, damping characteristics, dielectric properties of a polymer. The cheminformatics-based approach based on machine learning algorithms is often applied to predict the quantitative relationships between key molecular descriptors and properties of investigated polymers. In this work, we develop an innovative modeling framework by integrating cheminformatics and coarse-grained molecular dynamics (MD) to predict the glass transition temperature of a diverse set of hundred polymers. This synergistic approach provides valuable insights into the roles of key molecular parameters (i.e., cohesive interactions, chain stiffness, and branching) in influencing the glass transition temperature of polymers.

CINF 46

Insights into therapeutic fusion protein R&D from an analysis of the CAS databases

Yingzhu Li, yingzhuli@cas.org, Yi Deng, Cynthia Y. Liu. CAS, Columbus, Ohio, United States

Therapeutic fusion proteins have emerged as a promising class of biologics. Since the first therapeutic fusion protein, etanercept (Enbrel), was approved by the U.S. FDA in 1998, total global sales have been growing at a rate of ~6 percent annually, with the market expected to reach an estimated $24 billion by 2025. In collaboration with the National Science Library of the Chinese Academy of Science, CAS conducted multifaceted analyses of 30 years’ worth of curated therapeutic fusion protein information in the CAS reference and substance collections. Our analyses not only revealed an interesting global R&D landscape and patent flow pattern, but also covered fusion protein classifications based on half-life extending and activity components, target analysis, mechanism of action and disease indications, among others. Our big data analysis showcased a wealth of biologics information in CAS databases and suggested promising alternative applications of fusion proteins in vaccines, as well as gene and cell therapies.

CINF 47

STN to the rescue: When SciFinder is not sufficient

Philip Barnett¹,², pb405@caa.columbia.edu. (1) Science/Engineering Library, City College of New York (CUNY), New York, New York, United States (2) Ph.D. Program in Chemistry, The Graduate Center of the City University of New York, New York, New York, United States

The launch of SciFinder by Chemical Abstracts Service enabled chemists and chemistry students to retrieve references from the Chemical Abstracts database with little or no training. However, non-experienced searchers need to be taught when SciFinder is sufficient and reliable and when searching the more powerful command driven version of Chemical Abstracts.
on STN is necessary. Some searches such as a substructure search, limited by subheading, refined by document type, then analyzed by company-organization work as well on SciFinder as on STN. But other searches are either difficult on SciFinder or yield inaccurate retrieval. Complete searching often requires beginning with a hedge, all the possible ways a topic can be described. A typical hedge, easily entered on STN, often contains more terms then SciFinder can handle in a single search, and the search needs to be executed in several steps, combining these sets, while dealing with the limits on the number of references that can saved in an answer set. Focused substance searches in STN, like CASRN(s)molecular orbital?, executed in two slightly different ways in SciFinder show that truncation is not effective. The searches molecular orbital of substance name, and molecular orbitals of substance name, and choosing the "as entered" option yield different non-overlapping results, only some of which are in the STN search. The other options "containing the two concepts" or "the two concepts were present anywhere in the record" contain many more references than the STN search. A Substance Identifier search refined by Research Topic molecular orbital, on STN is necessary. Some searches such as a substructure search, limited by subheading, refined by document type, then analyzed by company-organization work as well on SciFinder as on STN. But other searches are either difficult on SciFinder or yield inaccurate retrieval. Complete searching often requires beginning with a hedge, all the possible ways a topic can be described. A typical hedge, easily entered on STN, often contains more terms then SciFinder can handle in a single search, and the search needs to be executed in several steps, combining these sets, while dealing with the limits on the number of references that can saved in an answer set. Focused substance searches in STN, like CASRN(s)molecular orbital?, executed in two slightly different ways in SciFinder show that truncation is not effective. The searches molecular orbital of substance name, and molecular orbitals of substance name, and choosing the "as entered" option yield different non-overlapping results, only some of which are in the STN search. The other options "containing the two concepts" or "the two concepts were present anywhere in the record" contain many more references than the STN search. A Substance Identifier search refined by Research Topic molecular orbital, even Categorized, yields a cumbersome number of references. A simple two term Reference search using "and" instead of "in" gives one fewer choice in articles containing both search terms. Some searches are amenable when limited to title words only. STN has this feature. But even selecting the "as entered" option in SciFinder only retrieves some of the articles compared to the title word limited STN search.

CINF 48

Reproducible data analysis and publishing in chemistry with R: Creating a workshop experience during the ACS National Meeting

Ye Li², yel@mit.edu, Steven P. Wathen³, Donna T Wrublewski¹, dtwrub@caltech.edu. (1) Caltech Library, California Institute of Technology, Pasadena, California, United States (2) MIT Libraries, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States (3) Chemistry, Siena Heights University, Adrian, Michigan, United States

Script-based data analysis and authoring tools such as R can help researchers streamline their workflow with research data and enhance reproducibility. At the Fall 2019 American Chemical Society National Meeting, two chemistry librarians partnered with a chemistry faculty member to offer a hands-on computer programming workshop. The workshop used the R programming environment, and was structured for beginners with no previous experience with R. Three main topics were covered during the five-hour session: introduction to R, working with QSAR data, and creating a report using R Markdown. This poster will cover the
workshop’s logistics, instructional design, and content, and discuss feedback received from participants. Suggestions for future improvements will also be discussed.

CINF 49

Crystal structure data without a publication: How much more can be learnt?
Suzanna Ward, ward@ccdc.cam.ac.uk, Clare A. Tovee, Ian Bruno. Cambridge Crystallographic Data Centre, Cambridge, United Kingdom

Storing chemical information in a drawer or on a hard-drive doesn’t benefit the researcher or the research community. With a recent increased focus on data and a resurgence of machine learning to utilise this data, there has also become a greater imperative to share chemical data and information reliably and efficiently.

The Cambridge Structural Database (CSD) is a data repository of over one million experimental crystal structures. Over the last few years chemists have been encouraged to share their crystallographic data through the CSD without the need for an accompanying journal article. These structures, called CSD Communications, undergo the same curation and validation processes as a structure associated with a traditional paper and receive a full data citation.

This poster will highlight recent strategies to engage the community and increase the volume of data shared through CSD Communications. We will also demonstrate key trends in chemical data and compare the quality of data shared in this way with data published in a scientific journal. The poster will additionally explore some examples of the potential benefits of publishing data as a CSD Communication, both to the author and to the wider research community. Can we learn more with more data and if so, how can we further increase incentives to encourage more data sharing?

CINF 50

Introduction of scholarly communications and the ethics of authorship into a chemical literature course

Nancy J. Butkovich, njb2@psu.edu. University Libraries, Pennsylvania State University, State College, Pennsylvania, United States

As undergraduate students become more involved in research, some knowledge of the scholarly communication and ethics regarding authorship and publishing become more important. This poster discusses a three part module sequence in a chemical literature course which covers topics such as copyright, peer review, open access, predatory publishing, author disambiguation, human subjects research, and the ethics of authorship. A wide range of resources - such as the ACS Ethical Guidelines to Publication of Chemical Research, Institutional Review Board documents for the university, ORCiD, Retraction Watch, and SHERPA/RoMEO - are used in the sequence. Selected course materials will also be displayed
InChI OER integration with the LibreText

Robert E. Belford1, rebelford@ualr.edu, Vincent F. Scafani2, vincent.scafani@gmail.com, Tina Qin3, qinna.msu@gmail.com, Jordi Cuadros4, Ehren C. Bucholtz5, Nathan Brown6, Andrew P. Cornelli7, Tanya Gupta8, Martin A. Walker9, Steven P. Wathen10, Delmar S. Larsen11. (1) Univ of Arkansas at Little Rock, Little Rock, Arkansas, United States (2) University Libraries, University of Alabama, Tuscaloosa, Alabama, United States (3) Cabot Library, Harvard University, Cambridge, Massachusetts, United States (4) Departament Estadistica Aplicada, IQS, Barcelona, Spain (5) Basic Sciences, St. Louis College of Pharmacy, St. Louis, Missouri, United States (6) The Institute of Cancer Research, Sutton, United Kingdom (7) Chemistry/Biology, University of Arkansas at Little Rock, Little Rock, Arkansas, United States (8) Chemistry, South Dakota State University, Brookings, South Dakota, United States (9) SUNY Potsdam, Potsdam, New York, United States (10) Chemistry, Siena Heights University, Adrian, Michigan, United States (11) Department of Chemistry, Univ California Davis, Davis, California, United States

The InChI OER (Open Education Resource) is a project developed by the InChI Trust and supported by IUPAC to provide online resources that can facilitate the adoption and adaption of InChI into the workflow of practicing scientists (https://www.inchi-trust.org/oer/). One of the strategies is to create educational textbook content involving applications of InChI for courses within the undergraduate chemistry curriculum. LibreText (https://chem.libretexts.org/) is a multidisciplinary OER HyperLibrary for collaborative generation and hosting of textbooks that evolved out of the ChemWiki at UC-Davis.

This poster will focus on InChI OER content created for an intercollegiate course in cheminformatics that was hosted on LibreText and taught on five campuses during the Fall of 2019 (https://chem.libretexts.org/link?143689). This presentation will provide information on both the use of the InChI OER and LibreText, and how through this synergistic project, material in the InChI OER can now easily be integrated throughout the HyperLibrary (and beyond), into core textbooks of the chemistry undergraduate curriculum like organic and biochemistry.

Degradable, on-patient medical record

Joe Collins1, joecol@mit.edu, Kevin McHugh2, Ana Jaklenec3, Robert Langer1. (1) MIT, Cambridge, Massachusetts, United States (2) Rice, Houston, Texas, United States

Accurate medical recordkeeping is a significant challenge in low-resource settings where centralized databases do not exist. Here, we present an approach to encode medical history on a patient in a way that is invisible to the naked eye yet detectable when exposed to near-infrared light. The specific medical information is encoded via the spatial distribution of biocompatible, near-infrared quantum dots (NIR QDs). The NIR QDs, encapsulated in biodegradable polymer microparticles, can be embedded in dissolvable microneedles and administered in select patterns to the dermis of patients. This invisible mark can be detected through specially adapted smart phones and the specific pattern can be read via pattern recognition software. In this way, medical information (e.g. vaccination statues, date of administration) can be recorded without the need for paper records or a centralised database. This new way of data storage may find particular use during disease outbreak or mass vaccination.
campaigns and may open new avenues for decentralized data storage and biosensing that could help shift the way medical care is provided and recorded in both the developing and developed world.

CINF 53

Scaling ligand-based virtual screening to a larger purchasable chemical space

Moayad Alnammi¹, alnammi@wisc.edu, Shengchao Liu⁴, Spencer Ericksen², Gene E. Ananiev³, James L. Keck⁴, F M. Hoffmann³, Scott A. Wildman⁴, Anthony Gitter⁴. (1) Computer Sciences, UW-Madison, Madison, Wisconsin, United States (2) Small Molecule Screening Facility, University of Wisconsin-Madison, Madison, Wisconsin, United States (3) UW Carbone Cancer Center, University of Wisconsin - Madison, Madison, Wisconsin, United States (4) University of Wisconsin-Madison, Madison, Wisconsin, United States

For the task of small molecule bioactivity prediction, we consider the practical efficacy of ligand-based machine learning methods in a realistic scenario with a large space of purchasable compounds and budget constraints. When faced with a prospective pool of millions or tens of millions of compounds, what virtual screening strategy should be employed by an academic lab on a constrained budget? We highlight our recent efforts to prioritize and test compounds in this setting using both one round and iterative strategies for compound screening. Given a fixed budget, the one round strategy exhausts the entire budget on compounds to be screened in a single iteration. In contrast, the iterative strategy distributes its budget across a number of batch screening cycles where each iteration provides feedback for subsequent compound prioritization.

To demonstrate the one round strategy, we showcase our successful virtual screening effort on the PriA-SSB protein-protein interaction. Our pipeline involves training ligand-based machine learning models on data from 427,000 previously assayed compounds and predicting on a prospective pool of 8.1 million purchasable compounds. The pipeline first uses cross-validation to identify the best performing supervised learning model, in this case a random forest. Subsequently, this model is then trained and scores the 8.1 million compound library according to predicted activity. A budget and additional filters for cost, compound availability, and delivery time are then imposed to select the most promising compounds. These compounds are then purchased from vendors and screened locally. For comparison, we use a chemical structure similarity baseline (ECFP4 tanimoto) that is given an equivalent budget. Our screening results demonstrate effective retrieval of chemically diverse hits from the large purchasable space. The random forest achieves a 48% hit rate with a 701 compound budget, outperforming the similarity baseline that achieves a 36% hit rate with a 705 compound budget.

CINF 54

Using machine learning methods and structural alerts for prediction of mitochondrial toxicity

Gerhard F. Ecker, gerhard.f.ecker@univie.ac.at, Jennifer Hemmerich, Florentina Trogger, Barbara Fuezi. Department of Pharmaceutical Chemistry, University of Vienna, Wien, Austria
Over the last few years more and more organ and idiosyncratic toxicities were linked to mitochondrial toxicity. Despite the increasing evidence, in vitro testing for such toxicities is still under development. Therefore, in silico approaches could be very beneficial to indicate hazards early in the drug development pipeline. In this paper we present the largest so far published dataset on mitochondrial toxicity by combining multiple endpoints. A thorough data analysis shows that molecules causing mitochondrial toxicity can be distinguished by physicochemical properties. Training a set of machine learning models as well as a deep learning model showed that the deep learning model as well as a gradient boosting model had a good performance on the training and test set. Furthermore, the combination of machine learning and deep learning with structural alerts derived from the training set proved to be a powerful tool to predict hazards and derive mechanistic insight.

CINF 55

Big errors in big data: When automated data curation misses the mark

Robert D. Clark\textsuperscript{1}, bob@simulations-plus.com, Pankaj R. Daga\textsuperscript{2}, Marvin Waldman\textsuperscript{1}. (1) Simulations Plus, Inc., Lancaster, California, United States (2) Simulations-plus Inc, Lancaster, California, United States

Cheminformatics modeling has always drawn heavily on compilations as a source of data. The most desirable ones were assembled by experts in a particular field, typically as literature reviews. Compilations assembled by cheminformatics groups specifically for model building purposes tended to be less carefully curated but were attractive because they were typically large - sometimes because they were compilations of compilations. The advent of sophisticated data mining tools has made it much easier to automate compilation, and the need for big data to drive deep learning has brought us into the age of Big Data. The sheer size of the data sets involved has driven development of automated curation techniques, particularly ones focused on standardizing structures. Unfortunately, the tools involved sometimes introduce systematic errors or over-standardize data; this talk will explore examples of the kinds of errors that can be inadvertently introduced in the process.

CINF 56

LigandNet: Compilation of ligand-based models for predicting ligand bioactivities

Daniel Castaneda-Mogollon\textsuperscript{1}, Dewan Shresta\textsuperscript{1}, Md Mahmudulla Hassan\textsuperscript{1}, Tudor I. Oprea\textsuperscript{2}, Suman Sirimulla\textsuperscript{1}, ssirimulla@utep.edu. (1) School of Pharmacy, The University of Texas at El Paso, El Paso, Texas, United States (2) University of New Mexico, Albuquerque, New Mexico, United States

LigandNet is a Machine Learning toolbox that combines different models into an open source platform so it can be used to predict if a ligand may bind to a specific protein or not. The models are trained on Pharos, a NIH public database that presents the data from the Target Central Resource Database. Machine learning models for about 800 proteins were developed and the models were trained on TACC resources, more specifically, Stampede2 and Lonestar5 clusters. Different machine learning approaches such as Random Forest, Support Vector Machine and Neural Networks were employed for training the models.
The development of modern machine learning technologies has highly increased the scale of computer-aided drug design in drug discovery and development. In this study we proposed a successful application of advanced deep learning technology, namely generative adversarial network (GANs), to create models for the design of druglike molecules with properties similar to known drugs. The combination of two deep neural networks, discriminator and generator, composes the GANs architectures. In the process of training, two neural networks contest with each other and were trained simultaneously. The discriminator was trained to be able to accurately discriminate between “real” and “fake” compounds derived from the generator; and the generator was trained to generate “fake” compounds to “fool” the discriminative network. The purpose of the generator is to generate compounds which will not be identified as “fake” ones by discriminator, while the purpose of the discriminator is to identify compounds coming from the generative network as “fake” ones. As for the training sets, which include about 8000 drug molecules and 10,000 screening compounds from ZINC database, four types of molecular fingerprints including FP2, FP3, FP4 and MACCS were introduced to describe the molecules with diverse structural characteristics. In addition to fingerprints, chemical SMILES descriptor was also employed to shape features of molecules. Unlike the fingerprints, this descriptor will be utilized in a recurrent neural network (RNN) to facilitate the design of druglike molecules directly. We expect the creative chemistry enabled by the advanced deep learning technique has a great application in modern drug development and discovery.
Synthetic feasibility and de novo molecular generation and optimization

**Connor W. Coley***, ccoley@mit.edu, Wenhao Gao†, Klavs F. Jensen†. (1) Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States (2) Broad Institute of MIT and Harvard, Cambridge, Massachusetts, United States

There is substantial interest in de novo molecular generation and optimization techniques as a way to propose new molecules during early-stage drug discovery. Deep generative models and other inverse design techniques allow for the multi-objective optimization of surrogate models (e.g., for activity, druglikeness) without relying on brute-force screening of a virtual chemical library. However, their utility is limited by an ignorance of synthesizability. This talk will describe an evaluation of several state-of-the-art methods in terms of their abilities to generate synthesizable molecules as judged by a data-driven chemical synthesis planning program. I will summarize emerging methods that explicitly integrate synthetic feasibility into their generative processes, which is an important step toward overcoming this limitation.

Molecular grammars: Accurate and universal way to construct molecules in silico

**Sergey Sosnin**, sergey.sosnin@skoltech.ru, Maxim Fedorov. Skolkovo Institute of Science and Technology, Moscow, Russian Federation

Retrieving compounds that satisfy desired chemical properties is a so-called inverse-QSAR problem. Recently, several approaches for the direct generation of structures of organic compounds by deep recurrent neural networks (LSTM, GRU) have been proposed. However, these approaches were initially designed for solving natural language processing problems, and have limitations applied to the generation of chemical structures. The general approach is the use of SMILES notation as a linear representation of organic structures, although it may lead to syntactically and chemically invalid compounds. Furthermore, there is a philosophical problem: organic chemists think more in terms of functional groups rather than atoms and bonds. To address these issues, we propose a molecular grammar approach: a molecule is constructed by sequential application of molecular grammar rules, whereby a grammar rule can be regarded as a structural fragment with “sticky bonds” and some additional criteria which guarantee connectivity between rules. In this approach, a generated molecule is structurally correct by design. We define a binary operation (+) for the concatenation of two rules into a larger one so that one can produce complicated building blocks from simple ones. We experimentally demonstrate that one can construct a sub-optimal grammar for a particular task. We investigated the applicability of molecular grammars to several inverse-QSAR problems and showed that this approach could effectively generate compounds with desired properties; however, the main advantage of the method is its universality.
Antimicrobial peptides (AMPs) are promising candidates in the fight against multi drug-resistant pathogens owing to AMPs’ broad range of activities and low toxicity. Nonetheless, identification of AMPs through wet-lab experiments is still expensive and time consuming. In particular, short-length (≤ 30 AA) AMPs are promising drug agents which have enhanced antimicrobial activities, higher stability, lower toxicity to human cells, and low production cost. However, existing AMP prediction methods achieved only 60 to 70% accuracy for short length AMPs; the problem is due to the largely mixed sequences of different lengths in the training dataset. To provide a solution to short AMP prediction, we have developed Deep-AmPEP30, which is a deep learning method based on an optimal feature set of reduced amino acids composition and convolutional neural network. Deep-AmPEP30 yields an improved performance over existing machine learning-based methods. To show usage of Deep-AmPEP30 in the field of drug discovery, we have screened all open reading frames from the genome sequence of *Candida glabrata* for potential AMPs. *Candida glabrata* is a gut commensal fungus expected to interact with and/or inhibit other microbes in the gut. Selected high-scoring peptides were subjected to experimental validation for antimicrobial activities and a 20-AA peptide P3 (FWELWKFLKSLWSIFPRRRRP) that showed strong anti-bacteria activity against *Bacillus subtilis* and *Vibrio parahaemolyticus* were identified. This peptide has a potency comparable to that of ampicillin in the bacterial inhibition assay.
Comparison of our prediction models with existing methods for short AMP prediction using the benchmark dataset. All values were multiplied by 100.

<table>
<thead>
<tr>
<th>Method</th>
<th>ACC</th>
<th>AUC-ROC</th>
<th>AUC-PR</th>
<th>Kappa</th>
<th>Sn</th>
<th>Sp</th>
<th>MCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>iAMP-2L</td>
<td>65.43</td>
<td>-</td>
<td>-</td>
<td>31.85</td>
<td>82.98</td>
<td>47.87</td>
<td>32.95</td>
</tr>
<tr>
<td>iAMPpred</td>
<td>70.74</td>
<td>-</td>
<td>-</td>
<td>41.49</td>
<td>80.85</td>
<td>60.64</td>
<td>42.36</td>
</tr>
<tr>
<td>AmPEP</td>
<td>68.09</td>
<td>75.14</td>
<td>68.63</td>
<td>36.17</td>
<td>93.62</td>
<td>42.55</td>
<td>42.07</td>
</tr>
<tr>
<td>AMP Scanner DNN</td>
<td>73.40</td>
<td>80.66</td>
<td>77.78</td>
<td>46.81</td>
<td>80.85</td>
<td>65.96</td>
<td>47.34</td>
</tr>
<tr>
<td>RF-AmPEP30</td>
<td>76.60</td>
<td>84.76</td>
<td>85.82</td>
<td>53.19</td>
<td>78.72</td>
<td>74.47</td>
<td>53.24</td>
</tr>
<tr>
<td>Deep-AmPEP30</td>
<td>77.13</td>
<td>85.31</td>
<td>85.36</td>
<td>54.26</td>
<td>76.60</td>
<td>77.66</td>
<td>54.26</td>
</tr>
</tbody>
</table>

Anti-bacteria effect of three top-ranked predicted AMPs from screening the C. glabrata genome against three different bacteria species. Treatment showing an inhibitory effect against the assayed bacteria is highlighted by a red (strong) or pink (mild) box.

CINF 61

Disease drivers, drug discovery & drug positioning in NASH: Smart medicine out-smarts disease

Prakash Narayan, pnarayan@angion.com. Preclinical Research, Angion Biomedica Corp., Uniondale, New York, United States
Non-alcoholic steatohepatitis (NASH) remains a challenge for the gastroenterologist as results from clinical trials have been dismal. There is now growing appreciation that hard endpoints such as steatosis, inflammation and scarring are not necessarily driven by a single mechanism across animal models and in patients but rather via disease-driving networks comprising genes and gene clusters. Identification of cross-species i.e. animal-human disease-driving networks and development of drugs that engage and neutralize these networks might translate to clinical success in at least a subset of NASH patients sharing that network. Furthermore, identification of potential non-responders can spare that subset of any adverse drug effects. Presented herein is an exercise that identifies a common and early disease-driving network in NASH and applies machine learning to repurpose existing drugs or identify novel chemical entities that target that network.

Advancing science as a global public good

Bonnie C. Carroll\textsuperscript{2,1}, bcarroll@iiaweb.com, Barend Mons\textsuperscript{3,4,1}. (1) Committee on Data (CODATA), International Science Council (ISC), Paris, France (2) Information International Associates, Oak Ridge, Tennessee, United States (3) Leiden University Medical Center (LUMC), Leiden, Netherlands (4) International Support and Coordination Office, GO FAIR, Leiden, Netherlands

CODATA is the Committee on Data of the International Science Council (ICS). CODATA exists to promote global collaboration to improve the availability and usability of data for all areas of research. CODATA supports the principle that data produced by research and susceptible to be used for research should be as open as possible and as closed as necessary. CODATA works also to advance the interoperability and the usability of such data: research data should be intelligently open or FAIR. By promoting the policy, technological and cultural changes that are essential to make research data more widely available and more usable, CODATA works in three main goal areas: 1) promoting principles, policies and practices for
Open Data and Open Science; 2) advancing the frontiers of data science; and 3) building capacity for Open Science by improving data skills and the functions of national science systems needed to support open data.

We will discuss the technical activities and accomplishments of CODATA in the context of the International Science Council's 2019-21 Action Plan of "Advancing Science as a Global Public Good", in its own programs, and in its cooperation with other data organizations such as the Research Data Alliance, World Data System, Scientific Unions, and GO FAIR. Some of the projects that will be discussed include data science schools; the development of regional platforms such as the African Open Science Platform; development of policy through papers and policies for data sharing; organization of workshops to advance interdisciplinary data sharing and opened, FAIR data; the establishment of an International Data Week, and projects that are addressing "Tackling Complexity: Data-Driven Interdisciplinarity" to make data work for cross-domain challenges. Finally, we will end on the point of: Unions, Alliances, Societies, Networks – what role can/do/should they play?

CODATA: Committee on Data International Science Council

CINF 63

GO FAIR in the U.S.: Advancing FAIR beyond the basics

Melissa Cragin, mcragin@sdsc.edu, Christine Kirkpatrick. SDSC, UCSD, San Diego, California, United States

Based at the San Diego Supercomputer Center, the US GO FAIR office is one of several GO FAIR Coordination and Implementation offices joining an international effort to increase the availability of machine-readable data and metadata for use in research and development. Production of research data that are FAIR (Findable, Accessible, Interoperable, and Reusable) – including improved access to semantically enhanced data, utilizing existing standards – can facilitate scientific discovery and enhance the societal and economic benefit of research and development (R&D) funding.
Creating and nurturing an internet of FAIR data will require a combination of more data stewards, better tools, and increased FAIR data stewardship skills across research computing staff, libraries, research staff (PIs, postdocs, graduate students), university leaders and funders, and (software) developers. To address this, the US GO FAIR office is leading the development of a national pool of experts who will be available for training activities and data fairification events. Leveraging the curriculum developed by GO FAIR leaders in the EU, an initial GO FAIR Data Stewardship training session was held in May, 2019 at SDSC. In February, 2020, a “Train-the-Trainer” (T3) workshop funded by the NSF will be held at Georgia Tech. The workshop will bring together practitioners and consultants from academia, non-profits, and industry, to participate in advanced lessons in the use of semantic tools, and consider best practices in “how to teach it.” Collaboration with the U.S. regional Big Data Innovation Hubs extends the US GO FAIR reach to additional cross-sector research communities, and provides opportunities for joint efforts to address scientific grand challenges.

The curriculum and outcomes of the T3 workshop will be shared, as well as plans for future FAIR-related activities led by SDSC.

CINF 64

World Data System and CoreTrustSeal: Together building trust in scientific data

Rorie Edmunds\textsuperscript{1,2}, rorie.edmunds@icsu-wds.org. (1) World Data System of the International Science Council, Tokyo, Japan (2) 2. CoreTrustSeal Standards and Certification Board, Tokyo, Japan

The World Data System (WDS) is a Research Programme established by the International Science Council (formerly ICSU) to help realize the vision of universal and equitable access to scientific data and information. To this end, WDS promotes long-term data stewardship that will deliver quality-assured scientific data, data services, products, and information; and builds ‘communities of excellence’ by helping organizations become persistent, robust, and trustworthy components of a data infrastructure from which a knowledge system that is both interoperable and distributed can be based.

This presentation will give an overview of current status of the World Data System and its endeavours. In particular, it will focus on the CoreTrustSeal standard for certifying the trustworthiness of data repositories established in September 2017 by WDS and the Data Seal of Approval (DSA) under the umbrella of the Research Data Alliance (RDA). CoreTrustSeal sets overarching criteria that embrace the prior standards of WDS and DSA, as well as other certification standards and data principles. It works with candidate repositories to meet 16 Requirements that will ensure they are persistent trustworthy sources of data and data services. Since its formation, around 70 repositories have been CoreTrustSeal certified, a similar number are being processed, and demand is increasing.

Although at the repository rather than dataset level, the CoreTrustSeal certification has been recognized as a reasonable proxy for FAIR. Moreover, for data to stay authentic and FAIR over time requires contextual information and Trustworthy Data Repositories that will actively support and maintain the FAIRness of these data. To answer the question, ‘Who can we trust to enable FAIR?’, the CoreTrustSeal Board has worked with the data repository community to develop TRUST (Transparency, Responsibility, User Community, and Sustainability, and Technology) Principles that try to capture the essential elements needed for concise and measurable approaches to trustworthiness. The presentation will briefly introduce the TRUST
Principles, as well as a potential extension to the CoreTrustSeal certification that has been developed by a joint WDS–RDA Working Group to examine data fitness for use.

CINF 65

NIST Research Data Framework and its relevance for FAIR data

Robert J. Hanisch, robert.hanisch@nist.gov. Director, Office of Data and Informatics, National Institute of Standards and Technology, Material Measurement Laboratory, Gaithersburg, Maryland, United States

The US National Institute of Standards and Technology (NIST) is considering the development of a Research Data Framework, patterned on the well-known NIST Cybersecurity Framework (https://www.nist.gov/cyberframework) in which community consensus was established around “standards, guidelines, and best practices to manage cybersecurity-related risk.” The Research Data Framework (RDaF) seeks to identify interests, obligations, costs, benefits, and risks surrounding the generation, analysis, curation, preservation, distribution, and re-use of research data. This is a broad and dynamic environment, and it is already clear that few people or organizations fully comprehend the complexity of the research data ecosystem. Many stakeholders share interests in research data (both produced through research or used by researchers in unanticipated contexts), including government agencies, funders (both public and private), data centers, repositories, universities and university associations, libraries, professional societies and associations, tool and infrastructure providers, academic publishers, policy organizations, advocacy groups, national and international data organizations, topical data initiatives, and, of course, individual researchers in all fields of science and humanities. A number of stakeholders have multiple roles in this rich but complex space. Open science and open data are trending and important topics, intersecting with key issues such as reproducibility, replicability, and reliability of research findings. The FAIR data guidelines are an important element of the research data ecosystem, and I will describe the RDaF in general as well as specific actions NIST is taking toward increasing the FAIRness of its research data.

CINF 66

NIH strategic plan for data science

Susan K. Gregurick, susan.gregurick@nih.gov. Office of Data Science Strategy, National Institutes of Health, Bethesda, Maryland, United States

"Storing, managing, standardizing and publishing the vast amounts of data produced by biomedical research is a critical mission for the National Institutes of Health (NIH). In support of this effort, NIH released its first Strategic Plan for Data Science in June 2018 that provides a roadmap for modernizing the NIH-funded biomedical data science ecosystem. NIH is actively implementing its strategy and will continue to seek community input during the implementation phase."
Why, how and what of open science

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

Making data FAIR is a common subject for symposia and discussions. But how do we make FAIR data useful? Moving from an analogue to digital world in research is by no means straightforward. When it comes to sharing research data most scientists have an idea of how they would go about doing it. Pre-requisites are workflow tools in smart laboratories, such as ELNs and LIMS capturing analytics data directly and feeding into domain specific repositories. However, the questions of why open science is needed and what do we envisage being able to accomplish through it have not yet been addressed in sufficient detail. To avoid building a bigger and bigger haystack of data, answering the why helps us work out the what.

FAIR, FACT, and the US RCSB Protein Data Bank

John D. Westbrook1,2, john.westbrook@rcsb.org, Stephen K. Burley1,2,3, (1) Research Collaboratory for Structural Bioinformatics Protein Data Bank, Rutgers, State University of New Jersey, Piscataway, New Jersey, United States (2) Institute for Quantitative Biomedicine, Rutgers, State University of New Jersey, Piscataway, New Jersey, United States (3) Rutgers Cancer Institute of New Jersey, Rutgers, State University of New Jersey, Piscataway, New Jersey, United States

The Protein Data Bank (PDB) was established in 1971 as the first open-access, digital data resource in all of biology. Current archival holdings encompass ~160,000 experimentally determined, atomic-level structures of proteins, DNA, RNA, and their complexes with approved drugs or other small molecules. Since 2003, the PDB has been managed according to the FAIR Principles (Findability-Accessibility-Interoperability-Reusability) by the Worldwide Protein Data Bank (wwPDB, wwpdb.org) partnership, including the US Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB), the Protein Data Bank in Europe (PDBe), Protein Data Bank Japan (PDBj), and BioMagResBank (BMRB).

OneDep, the wwPDB single global deposition-validation-biocuration system, ensures that all three wwPDB regional data centers (US RCSB PDB, PDBe, and PDBj) provide the same data processing experience to PDB Data Depositors around the world. The system supports deposition of atomic coordinates plus experimental data and related metadata for structures coming from macromolecular crystallography, nuclear magnetic resonance spectroscopy, and 3D electron microscopy. To provide for continuous improvement of data quality and future extensibility, the system depends on the PDBx/mmCIF data dictionary and data management framework throughout the deposition-validation-biocuration process.

Like its wwPDB partners, the US RCSB PDB adheres to and promotes the FACT Principles (Fairness-Accuracy-Confidentiality-Transparency) when delivering PDB data through its two web portals (RCSB.org and PDB101.RCSB.org). On an annual basis, RCSB.org serves millions of scientific researchers and their trainees in every sovereign nation around the world by furnishing continuously updated PDB data integrated with ~40 external data resources. In parallel, PDB101.RCSB.org delivers introductory materials from the PDB archive relating to basic biology, biomedicine, and bioenergy to hundreds of thousands of educators and their
students worldwide every year.

This talk will exemplify how the US RCSB PDB operates according to the FAIR and FACT Principles

CINF 69

RRIDs for antibodies and instruments, an idea to make chemistry more FAIR

Anita Bandrowski1,2, abandrowski@ncmir.ucsd.edu. (1) Neuroscience, ucsd, La Jolla, California, United States (2) SciCrunch Inc, San Diego, California, United States

The RRID, research resource identifier, initiative is a system for identifying the following in the methods sections of journal articles: antibodies, cell lines, plasmids, organisms, and tools, which may include software and hardware used to conduct the experiments. Similar to the CAS system for chemicals, the RRID is a persistent unique identifier, provided by an authoritative database for a particular resource type. An example is the Antibody Registry, which provides RRIDs for antibody reagents, while Cellosaurus provides RRIDs for cell lines. The RRID system is built on top of these core authorities and allows authors to search for all RRIDs in a single portal (rrid.site), pointed to by multiple publishers including NatureResearch, AACR, Wiley and Sons, and PLoS. Making chemistry journal articles more FAIR should include properly labeling all resources including antibodies and statistical or physical tools. Statistical tools, especially those that are created by individual laboratories may be quite useful yet are not often subjected to thorough quality control, resulting in bugs are only uncovered years later as was recently demonstrated by Neupane et al, 2019 (https://pubs.acs.org/doi/10.1021/acs.orglett.9b03216). These researchers uncovered that chemical shift value calculations for NMR experiments were effected by operating system used. In the field of Neuroimaging, a similar discovery was made roughly a decade ago, where one of the most popular MRI programs was observed to change the answer for a measurement of cortical thickness depending on the operating system that it was running on. In neuroimaging, the package was updated by source developers and is a known bug tested for when code is released. The problem, however, remains in the literature because the bug potentially effected hundreds of studies that can't be identified because authors failed to add sufficiently granular information about software into their manuscripts. We have found that the use of cell lines that are misidentified or contaminated at the repository where they are purchased, may effect as much as 16% of the articles that use cell lines. Use of the RRID system allows anyone to reveal the articles that may be affected by a particular problem resource with nothing more than the RRID and google scholar. We have also shown that the use of the RRID system by authors before publication substantially reduces the use of problematic reagents in the scientific literature.

CINF 70

What do the FAIR Data Principles mean to the life sciences industry?

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

While the FAIR Data Principles initially sprung up from a desire to foster findability, accessibility, interoperability and reuse of scholarly data, industry has begun to recognize that this
approach is relevant to their own commercial challenges around data management, business partnerships and pre-competitive collaboration. In this paper we will survey the emerging industrial FAIR landscape, with particular emphasis on the efforts at the Pistoia Alliance, a precompetitive collaboration community serving life sciences R&D.

CINF 71

FAIR data and climate change resilience

Camille Thomas¹, cthomas5@fsu.edu, Cameron Riopelle², criopelle@miami.edu. (1) Florida State University, Tallahassee, Florida, United States (2) U of Miami, Miami, Florida, United States

Researchers are beginning to take to heart the FAIR Data Principles: in the ideal setting, data should be findable, accessible, interoperable, and reusable. However, just because data in the short-term sense meet (or even exceed) FAIR standards, we seek to spark discussion on the long-term concerns of managing scientific data (and more broadly, the products and by-products of the research process) in our era of climate change and flux. In the past year, we have studied academic and white papers on data management, with particular attention to issues of avoiding climate change-related data loss, and we will present our findings. Broadly, we introduce the state of the field of data management, current best practices in data resilience, and trends which will make these issues and protocols increasingly relevant.

CINF 72

Open and FAIR data sharing: Geosciences benefiting from collaboration with chemistry and other sciences

Shelley Stall¹, sstall@agu.org, Leah R. McEwen². (1) American Geophysical Union, Washington, District of Columbia, United States (2) Clark Library, Cornell University, Ithaca, New York, United States

Research transparency and rigor is supported by well-documented data that is as open as possible. Making data products easy to discover and evaluate as a separate research product from a paper brings new ways for other researchers and potential collaborators to learn of our work. This is especially important for cross-domain and transdisciplinary efforts. Within the Earth, space, and environmental sciences, journals are implementing common author guidelines for data and code that support the published research, and requiring they be deposited in a community-accepted repository that supports the FAIR Principles (where FAIR stands for Findable, Accessible, Interoperable, and Reusable). Data and code are cited in the reference section of the paper along with availability statements that clarify access to these research products. Similar efforts implementing the FAIR Principles are underway by our colleagues at the International Union of Pure and Applied Chemistry (IUPAC) that makes possible a valuable opportunity to collaborate on a common message and share ideas.

More work is needed across all domains to ensure the culture change needed to implement open and FAIR data sharing is sustainable and adopted across the international scientific community. Encouraging data and code sharing across all domains, credit for well-documented digital products, and adequate funding for the repositories and infrastructure needed to support the digital elements of our research is essential.
CINF 73

Proteochemometric modeling of GABA transporters

Gerhard F. Ecker, gerhard.f.ecker@univie.ac.at, Stefanie Kickinger, Anna Seiler. Department of Pharmaceutical Chemistry, University of Vienna, Wien, Austria

Proteochemometric modeling (PCM) combines ligand information as well as target information in order to predict an output variable of interest (e.g. activity of a compound). The big advantage of PCM compared to conventional Quantitative Structure-Activity Relationship (QSAR) modeling is, that by creating a single model one can not only predict the affinity of a diverse set of compounds to a diverse set of targets, but also extrapolate the specific ligand-protein interactions that might be relevant for binding. Moreover, by combining different datasets also targets with sparse data-availability can be included and modelled. Since PCM has already been successfully applied to SLC and ABC transporters, we wanted to investigate whether PCM is a potential new method to predict the activity of neurotransmitter transporter such as the four GABA transporters (GAT1-3, BGT1).

First, we compiled a dataset by collecting bioactivities for all four GABA transporters from the ChEMBL25 database and PubChem. After rigorous curation of the data, we obtained a dataset of 317 unique compounds that were tested at least in one of the four transporters in a [3H]GABA uptake assay (85 compounds were tested in all transporters). For model building, 25 physicochemical RDKit descriptors as well as Z3-scales and Z5-scales descriptors for the orthosteric binding pocket of the GABA transporters were calculated. Subsequently, regression models by applying machine learning techniques such as partial-least-square(PLS) and random forest were derived. Finally, we compared the performance of the PCM models to conventional QSAR models. Overall, we observed that PCM models showed similar performance as the QSAR models for an external test set. Remarkably, by analyzing the importance of the descriptors used in the PCM models, we identified residues as relevant for binding that already had been confirmed in mutational studies. Thus, PCM is a new approach that not only helps to identify new active molecules on the GABA transporters, but also is capable to shed light on the possible binding modes of these compounds.

CINF 74

Deep convolutional neural networks for prediction and biomolecular modeling of cancer driver mutations and design of selective inhibitors

Gennady M. Verkhivker1,2,3, verkhivk@chapman.edu, Steven Agajanian2, Oluyemi Odeyemi2. (1) Biomedical and Pharmaceutical Sciences, Chapman University School of Pharmacy, Irvine, California, United States (2) Graduate Program in Computational and Data Sciences, Schmid College of Science and Technology, Chapman University, Orange, California, United States (3) Department of Pharmacology, University of California San Diego, San Diego, California, United States

Development of machine learning solutions for prediction of functional and clinical significance of cancer driver genes and mutations are paramount in modern biomedical research and have gained a significant momentum in a recent decade. Given rapid proliferation of machine learning tools to address biological problems, there are several fundamental questions arising in the context of classification of cancer driver mutations. Will deep learning make all other models obsolete? Can deep learning models achieve robust recognition of cancer driver
mutations based solely on nucleotide information? What is the role of functional and structural predictors derived from biophysical analysis? In this work, we have explored and integrated different machine learning approaches, including deep learning and convolutional neural networks for prediction and classification of cancer driver mutations by using and consolidating information across large datasets of functional and clinical mutations. The classifiers were benchmarked against their tree-based alternatives in order to evaluate the performance on a relative scale. We then integrated DNA-based scores generated by convolutional neural networks with various categories of conservational, evolutionary and functional features into a generalized random forest classifier. The results of this study have demonstrated that convolutional neural networks can learn high level features from genomic information that are complementary to the ensemble-based predictors often employed for classification of cancer mutations. By combining deep learning-generated scores with only two main ensemble-based functional features, we can achieve a superior performance of various machine learning classifiers. Our findings have also suggested that synergy of nucleotide-based deep learning scores and integrated metrics derived from protein sequence conservation scores can allow for robust classification of cancer driver mutations with a limited number of highly informative features. Machine learning predictions are leveraged in molecular simulations, protein stability and network-based analysis of cancer mutations in the protein kinase genes to obtain insights about molecular signatures of driver mutations and enhance the interpretability of cancer-specific classification models. We discuss how these approaches can be used in the design of selective and allosteric inhibitors of cancer-associated protein kinases and oncogenic mutants.

CINF 75

Multitask prediction: How to evaluate the performance of prediction and screening methods?

Ekaterina Sosnina¹,⁵, Ekaterina.sosnina@skoltech.ru, Dmitry I. Osolodkin²,³,⁴, Maxim Fedorov¹. (1) Skolkovo Institute of Science and Technology, Moscow, Russian Federation (2) Department of Chemistry, Lomonosov Moscow State University, Moscow, Russian Federation (3) Chumakov Federal Scientific Center for Research and Development of Immune and Biological Products RAS, Moscow, Russian Federation (4) I M Sechenov First Moscow State Medical University, Moscow, Russian Federation (5) Institute of Physiologically Active Compounds RAS, Chernogolovka, Russian Federation

Over the past decade, virtual screening methodology is changed due to increasing the amount of experimental data and introduction the multitask prediction techniques. Previously used evaluation metrics became unreliable for the multitarget prediction due to inconsistency of prediction efficacy for different targets within one model. The evaluation metrics which are suitable for the multitask prediction are used in the recommender systems and can be adapted for drug discovery challenges.

In the present work, we assessed the prediction and screening quality (the ability to create the list of 10/50/100/500 most prospective compounds) by evaluation metrics, which were both used in single-target prediction and borrowed from the recommender systems. The prediction efficacy was evaluated by MAE, MSE, RMSE in case of regression models, and ROC AUC curve, Breese score, precision/recall, and precision-recall curve in case of classification models. The screening efficacy was assessed by MAP@k, MAR@k, coverage in the context of binary (active or inactive compounds), and multi-class (high-, middle-, low-, inactive classes) classification. Prediction and screening were performed by models based on two main
recommender system techniques: collaborative filtering (Surprise implementation) and content-based filtering (Sparse Group Inductive Matrix Completion (SGIMC) method). Besides, for a screening task, the selection based on ‘popularity’ (frequency of recognition) and a random selection was assessed and compared with model-based results. We used ViralCHEMBL dataset, containing 650K interactions on activity or inactivity of 250K compounds against 185 viral species; an All-Assay-Max2 dataset, containing 0.5M compounds, more than 4K assays, and 1.4M of their dose-response activity interactions; manually collected dataset from ChEMBL v.25, containing 16K interactions of more than 4K compounds toward 5 types of muscarinic acetylcholine receptors.

It was shown that multi-target datasets should be characterized using a Long Tail plot, as well as serendipity and diversity metrics. It allows for selecting optimal prediction algorithms and evaluation metrics. We demonstrate that MAP@k and MAR@k metrics are mandatory for screening evaluation: even the high ROC AUC score (more than 0.9) does not guarantee satisfying results for every target. For sparsed, imbalanced data with a ‘long tail’ individual AP@k and AR@k are preferable for screening quality assessment.

CINF 76

Comparison of ensemble learning methodologies in drug discovery
Pankaj R. Daga¹, pankaj@simulations-plus.com, Michael Lawless¹, Marvin Waldman², Robert D. Clark². (1) Simulations-plus Inc, Lancaster, California, United States (2) Simulations Plus, Inc., San Diego, California, United States

Building predictive models for various ADMET end-points e.g., solubility, intrinsic clearance, or hERG toxicity, requires the inclusion of diverse chemotypes to ensure their applicability to a wide range of chemical structures. Building an ensemble of models tends to yield better results when there is significant diversity amongst the models. In this presentation, the performance of different ensemble methods will be compared. We used different ensemble algorithms, e.g, Random Forest, and XGBoost as well as our proprietary Artificial Neural Network and Support Vector Machine ensemble methods to build models from the same training compounds and descriptor pools. These methods were used to build regression and classification models of different end-points, pertaining to drug discovery. The strengths and limitations of these methods will be discussed along with their performance.

CINF 77

GNINA 1.0: Predicting protein-ligand binding affinity through cross-docking
Paul Francoeur, PAF46@pitt.edu, David Koes. Computational and Systems Biology, University of Pittsburgh, Pittsburgh, Pennsylvania, United States

Currently, machine learning approaches have had significant success in predicting protein-ligand binding affinities. However, recent papers have shown that there is significant bias in some commonly used benchmarking datasets, resulting in overly optimistic results of the models. Additionally, training data generation is typically done through redocking, which is not analogous to the use-case in drug discovery as one is trying to fit either similar molecules to known binders or fit known binders into new targets. Lastly, there is a lack of available, high quality, and labelled data to train our models on. In order to alleviate these concerns we
present an extension of GNINA, a grid-based convolutional neural network, that is trained on a new set of cross-docked poses in order to generate a more reliable model. Cross-Docking provides both a method of data augmentation and a way to generate contexts for training data that is more similar to the use-case of drug discovery, by grouping the PDB by binding pockets as defined by Pocketome and docking every ligand with every receptor in a given pocket.

We show that the performance of the new data is similar to training on clustered cross-validated splits of the PDBbind database, can generalize to the PDBbind models, and that models trained on redocking with the PDBbind data fail to generalize to cross-docked poses. We also explore the extent that these structure-based models actually use protein-ligand interactions to make their predictions versus learning ligand-only cheminformatic models.

CINF 78

Analysis of machine learning strategies to predict protein developability

Alexander Golinski, golin010@umn.edu, Benjamin Hackel. Chemical Engineering, University of Minnesota, Minneapolis, Minnesota, United States

Jain et al. studied biologics entering clinical trials and identified a connection between late phase failures and poor developability (stability, solubility, and producibility). Predicting protein developability based upon amino acid sequence remains a challenge due to a large difference between theoretical mutational diversity \(10^{24}\) for our model system and the throughput of traditional assays \(10^2/\text{week}\). To bridge this gap, we propose the utilization of 3 classes of high-throughput \(10^5-10^8/\text{week}\) developability assays to empower deep data collection to inform and evaluate sequence models of developability.

We utilized the small protein scaffold GP2 with two diversified loops as a proof of concept for predicting developability from loop sequence. \(10^5\) mutants were experimentally scored in the high-throughput assays representative of protein stability and expression. Our proposed model utilizes two stages of prediction: first to use amino acid sequence to predict high-throughput assay performance, and second to combine sequence and assay information to predict developability features such as recombinant yield and thermostability. The ability to connect amino acid sequence to performance is complicated by varying sizes of input, a categorical interpretation of amino acids causing a sparse representation, and biological noise in experimentation. Thus, we will describe how adjustments to the model impact performance with noisy data. This includes comparing recurrent and convolutional neural networks which treat the loop sequences as an entity rather than unique interactions between residues. We also discuss using embedding layers rather than one-hot encodings to reduce sparseness by placing chemically similar residues closer together. Finally, we discuss our attempts of handing experimental noise by choosing observation requirements and discretizing inputs.

At present, our model was able to reduce continued experimentation of poorly developable mutants by lowering the false positive rate of producible proteins by 5%. Further analysis of assay results confirms an increase in median stability of a disulfide stabilized loop, and 2-9% increase in Spearman’s rho when utilizing an embedding and recurrent neural network over traditional modeling strategies. Continued progress will identify a more developable subset of mutants for future candidate discovery and expand the assays to assess alternative protein scaffolds, enzymes, and anti-microbial peptides.
Challenges and progress in combining docking programs with deep neural networks

**Joseph A. Morrone, jamorron@us.ibm.com, Jeffrey K. Weber, Wendy D. Cornell. IBM T.J. Watson Research Center, Yorktown Heights, New York, United States**

The combination of deep learning and computer-aided, structure-based drug design is an area of great promise, although recent work has put forth that issues such as dataset bias can negatively impact model development. We present a simple, modular graph-based convolutional neural network that takes the output of a docking procedure as input to generate models for activity and binding mode prediction. We show, in agreement with recent literature, that dataset bias drives many of the promising results on virtual screening that have previously been reported. However, we also show that our neural network is capable of learning from protein structural information when, as in the case of binding mode prediction, an unbiased dataset is constructed. We develop a deep learning model for binding mode prediction that uses docking ranking as input in combination with docking structures that outperforms the baseline docking program in a variety of tests, including on cross-docking datasets that mimic real-world docking use cases.

Schematic of a modular, graph-based neural network architecture that takes the three-dimensional structure of a given protein-ligand complex as input.
CINF 80

Virtual screening with convolutional neural networks

Jocelyn Sunseri, jss97@pitt.edu, David Koes. Computational and Systems Biology, University of Pittsburgh, Pittsburgh, Pennsylvania, United States

Virtual screening is a fundamental task in computational drug discovery, and significant research has been dedicated to improving it. Developing novel methods that result in consistently improved enrichment would reduce drug development costs and could enhance available treatment for many human diseases. To that end, we have developed novel scoring functions based on three-dimensional grid based convolutional neural networks. These networks are intended for general-purpose modeling tasks, but here we specifically investigate their performance at virtual screening. We compare networks resulting from different training strategies and datasets as well as several model architectures. We also compare multiple methods for performing the screening itself, including a method based on comparing generated grids to the reference grids associated with screening compounds. In addition to raw performance, especially assessments of generalization, we also investigate properties such as the score distributions of actives and inactives with a given scoring function. Finally, we describe using our library for generating and transforming molecular grids, libmolgrid, for training and deploying a model for virtual screening. libmolgrid is written in C++/CUDA with Python bindings, and is intended to expedite and simplify the process of training grid-based molecular models.

CINF 81

Community efforts to enhance supporting information of spectral data and chemical structures

Vincent F. Scalfani1, vincent.scalfani@gmail.com, Leah R. McEwen2. (1) University Libraries, University of Alabama, Tuscaloosa, Alabama, United States (2) Clark Library, Cornell University, Ithaca, New York, United States

In March 2019, the NSF Office of Advanced Cyberinfrastructure sponsored a community workshop to help advance FAIR publishing guidelines for machine-readable spectral data and chemical structures. The workshop brought together key stakeholders in the chemical community including librarians, researchers, database managers, software providers, and publishers. One of the major outcomes from the workshop was a structured template approach to organizing and packaging machine-readable spectral data and chemical structures alongside publications. This presentation will update the community on our efforts with developing workflows and pilots associated with this enhanced machine-readable spectra/structure supporting information.

CINF 82

Essential factors for data sharing: Publisher initiatives to help make data FAIR

Springer Nature’s 2019 white paper set out five key challenging areas for data sharing based on the responses of more than 11,000 researchers. Here we will provide a quick overview of the activities Springer Nature is undertaking to meet some of these, focusing on clear policies, credit and practical help.

CINF 83

FAIR data policies at science

Jake S. Yeston, jyeston@aaas.org. Science Magazine, Washington, District of Columbia, United States

Science has been working with communities in multiple disciplines to encourage best practices in making experimental data findable, accessible, interoperable, and reusable (FAIR) after publication. This talk will focus on policies and comparative challenges in chemistry specifically, including efforts to broaden the use of repositories beyond those that archive crystallographic data.

CINF 84

Implementing FAIR data principles at American Chemical Society publications

Angela M. Hunter, a_hunter@acs.org. American Chemical Society, Washington, District of Columbia, United States

With increasing awareness on the importance of FAIR data principles, American Chemical Society (ACS) Publications is working to incorporate these standards within the journals. While the Supporting Information in all ACS journals is freely available and deposited to figshare, authors often choose to only include static data images which hinders discoverability and interaction with the data. In order to evaluate author willingness to support FAIR data principles, a pilot program has been initiated at select ACS journals to encourage authors to submit primary NMR data as ‘Supporting Information for Publication’. This presentation will focus on the initiatives to improve data policies and benefits for authors, reviewers, and readers.

CINF 85

FAIR Chemistry: Publisher perspective

Raymond J. Boucher, rboucher@wiley.com. John Wiley and Sons Ltd, Chichester, United Kingdom

As a Publisher, Wiley are committed to supporting the aspirations of the chemistry community through openness, accessibility and transparency of research and data. We support the fundamental aims to implement FAIR (Findable, Accessible, Interoperable, and Re-usable) data principles to protect the long-term integrity of research by making data, methodologies and reporting standards openly available and to facilitate rapid and effective research discovery.
and business innovation based on reuse of research data. We are working with various bodies to provide solutions and an infrastructure that are necessary to undertake research according to the FAIR data principles and to comply with funder requests to share data as part of the publication process. This presentation will outline how Wiley intends to support open research by:

- Helping to define minimum information (MI) standards for data and machine-readable metadata, as well as open data standards that will enable more efficient connection and linking of publications.
- Promoting the use of FAIR Data principles and raise awareness amongst researchers and professionals and to highlight developments that provide demonstrable value to the community.
- Encouraging authors of articles published in Wiley/Wiley-VCH journals to share their research data including, but not limited to: raw data, processed data, software, algorithms, protocols, methods, materials.
- Developing processes to integrate research data into the publication process to replace the model of supplemental material with the model of depositing open and FAIR research data in public research data repositories and citing DOIs and stable identifiers of those datasets in articles.

Wiley’s user community consists of authors, editors and researchers. Our aim is to support the development of an infrastructure that enables chemists to develop good practices with their research data from creation to preservation.

**CINF 86**

**Data sharing at the Royal Society of Chemistry**

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

This talk will cover policies, workflows and exemplar projects for chemical data sharing at the Royal Society of Chemistry, as well provide updates on collaborative efforts with researchers and publishers to set common guidelines and widen researcher participation for FAIR-er supplementary information.

**CINF 87**

**Thieme chemistry: publisher initiatives for supporting FAIR data exchange of chemical data**

**Fiona Shortt de Hernandez, fiona.shortt@thieme.de. Thieme Chemistry, Thieme Group, Stuttgart, Germany**

Thieme was the first publisher to make primary chemistry data available in its online journals worldwide – thus allowing scientists to share scientific data gathered from experimental measurements. As a publisher Thieme now continues to work toward finding solutions for chemists and enabling FAIR data exchange where possible. New initiatives include text and data mining (TDM) collaborations with academia and also partnerships, for example, the PubChem collaboration enabled by technology partner, InfoChem, which resulted in the addition of nearly 700,000 chemical substance records, nearly 700,000 scientific article descriptions, and
over 1.2 million links between chemicals and articles to the PubChem database.

CINF 88
Aligning data practices for the 21st Century

Thomas B. Faust, thomas.faust@hindawi.com, Catriona MacCallum, catrina.maccallum@hindawi.com. Hindawi Limited, London, United Kingdom

Hindawi is committed to Open Science, working with the academic community and driving forward open data initiatives. Our aim is to foster an open access approach to all research outputs (including data, code and articles), while ensuring that appropriate attribution is given to those who invest time, thought and funds into their creation. Although many initiatives have been led by open access publishers, much of the publishing agenda around data sharing has been primarily driven by the biomedical community. Hindawi’s own publishing portfolio includes many physical and technical sciences journals, as well as biomedicine and the natural sciences, and data sharing practices across these disciplines are still being developed. In 2018, we required all authors, regardless of discipline to provide a Data Availability Statement alongside their papers and are gradually increasing the strength of this policy to mandate the deposition of relevant data in publicly available repositories. But just making the data underlying a paper publicly available is insufficient if others are to reuse that information reliably. We therefore also endorse the FAIR Guiding Principles, where the focus is on stewardship, reuse and discovery rather than openness per se. There are significant challenges to all those involved in funding, creating, curating, disseminating and discovering data, regardless of discipline; by the time you are ready to submit to a journal it is generally too late. Here we discuss differences among research disciplines and communities of practice and how some of the barriers to sharing and reusing data - and the associated metadata - can be overcome so that the researchers generating the data can benefit as much as those who seek to mine and re-use it. The latest initiatives implemented by Hindawi, or under discussion with our academic community of Editors, will be presented.

CINF 89
Chemical data in the Google era

Gregory M. Banik¹, gregory_banik@bio-rad.com, Leah R. McEwen², Stuart J. Chalk³, Sara Tenney⁴, Amanda Koenig⁴. (1) Bio Rad Laboratory, Informatics Division, Philadelphia, Pennsylvania, United States (2) Clark Library, Cornell University, Ithaca, New York, United States (3) Department of Chemistry, University of North Florida, Jacksonville, Florida, United States (4) Publications Division, American Chemical Society, Washington DC, District of Columbia, United States

The ACS Guide to Scholarly Communication is being expanded with new topics beyond the scientific journal, including trends and best practices for exchanging research data in the digital environment. Most of the ACS Guide to Scholarly Communication focuses on best practices for communicating scholarly information for human consumption. In the newest edition of The Guide, a new part has been added on best practices for communicating scholarly information for machine consumption. Machines are becoming an increasingly larger and more important part of how scientists are recognized and rewarded for their research contributions. Those scientists who engage in data sharing and learn the best practices to make the data
associated with their published works more machine-friendly will have a distinct competitive advantage relative to their peers.

The new part, entitled “Data in the Google Era”, consists of three chapters:

**Data Sharing** – An overview of the current landscape of sharing digital chemical data as well as practical guidance for best practices in preparing data and information in digital form that can be transmitted or processed

**Digital Chemical Data** – An overview of the essential types of digital chemical data for authors, reviewers, and publishers.

**Chemical Structures in the Google Era** – An overview of subtle, critical machine interpretation issues problems that authors, reviewers, and publishers will face when sharing chemical structure information in digital formats.

Highlights from each of these chapters will be given in this talk.

**CINF 90**

**Compound Interest: Graphical chemistry communication**

*Andy Brunning, ndbrning@gmail.com. Compound Interest, Ely, United Kingdom*

Posters and infographics are an effective way of communicating chemistry, making chemical information accessible and helping the public to engage with the subject. This talk will discuss tools, methods and strategies for sharing chemistry research in graphical form, as well as highlighting some of the challenges and how they can be overcome. It will also discuss the benefits and pitfalls of using infographics to communicate about chemistry via social media.
CINF 91

Tell your science story with visuals: Other thousand words

George Zaidan, G_Zaidan@acs.org. American Chemical Society, Washington, District of Columbia, United States

Whether you’re teaching middle-schoolers or writing a JACS paper, visuals can help tell your science story. A great visual can be powerful, effective, and memorable; bad ones can be confusing, distracting, and undercut your argument. Executive Producer of ACS Reactions George Zaidan will lead discussion on several case studies of excellent (and not-so-excellent) visual storytelling. By the end of the session, you’ll be able to recognize (and avoid) the most common pitfalls and follow a few basic rules to make your visuals stand out.

CINF 92

How C&EN creates effective data visualizations

Robert Bryson, r_bryson@acs.org. Publications, The American Chemical Society, Alexandria, Virginia, United States

Science journalists are often tasked with the challenge of how to translate a large amount of data or a tough-to-understand scientific concept into a meaningful and effective visualization. Learn how C&EN sifts through the data and the science to figure out the story we want to tell and how we want to tell it—and how those decisions change for digital and print presentations. C&EN Creative Director Rob Bryson will share some work C&EN has done and some insight into the process of building data and scientific visualizations.

CINF 93

Using personas and infographics to make scientific data come alive to multiple audiences

Molly A. Strausbaugh, mstrausbaugh@cas.org, Lisa E. Dragin, Devin L. Daniels. CAS, Columbus, Ohio, United States

Communication is an art, so why not use art to communicate? In the development of chemical information solutions, diverse groups of stakeholders must be aligned, including those in software development and product management. From developers to leadership, decision makers need a thorough understanding of the scientific problems users face. Communications that illustrate how a chemical information solution solves a real-world challenge faced by users, customers and leaders, allows the scientific solution to come alive. At CAS, we leverage user personas, including photos, generalized quotes, key themes and workflows, to make potential users real to our software development teams. Illustrations of the entire user ecosystem highlight interactions with others involved in their processes and infographics summarize key points. Together, these visual aids enable high-level alignment to the most critical information, while providing entry points for those interested in a deeper dive. Scientific information is valuable in its own right, but increases in value when it is communicated in a way that is easier to comprehend. Art helps increase comprehension and allows more people to
leverage the information in their day-to-day lives.
CINF 94

Authenticity and creativity in chemical communication

Raychelle M. Burks, rburks@stedwards.edu. Chemistry, St. Edward’s University, Austin, Texas, United States

GIFs, memes, infographics, animations, red-hot RTs, viral videos, top TikToks, innovative Instagrams… We have many ways to engage in science communication, but what are the best practices for our communities, our messages, and ourselves? To help us tailor our approach, relevant research and practitioner experiences will be shared.

CINF 95

Thirty years of gloop, slime, and cabbage juice: How do we utilize current communication methods to deliver effective messages to the public?

Alexa Serfis, alexa.serfis@slu.edu. Chemistry, Saint Louis University, St. Louis, Missouri, United States

Public outreach activities are a great way to bring attention to the exciting field of chemistry. The ways in which our messages can be communicated to the public need to be identified, then utilized in a manner that will reach our target audience. The messages must also be understood. The last thirty years has brought many changes in communication modes. Electronic modes have replaced print materials in many communities. The ability to reach large numbers of people with just a few clicks of the mouse presents opportunities, and challenges, for communicating science to the public. Identifying the target audience and channels to reach them is key. Collaborating with organizations outside of ACS is a great way to reach the public. Web presence and social media avenues can be leveraged to engage our peers as collaborators in outreach efforts. This presentation will cover successful communication practices for reaching the public, and establishing collaborations both internal and external to ACS.

CINF 96

Data storytelling: Intersecting data, stories and technology to create visual inspiration

Rosa Torres, rosa@dataimprint.org. Data Solutions, Vereti Concepts, Media, Pennsylvania, United States

Data today is more abundant than ever and because of this making data available at large has become a bigger challenge. There are many ways to represent information, the challenges lie in making the data meaningful and consumable to broader audiences. Data storytelling through data visualization is becoming a communication medium in its own right, able to inform and engage people. It is essential in making data more accessible. Data storytelling starts with a story. We will look at the main constructs of storytelling, such as building narratives that answer "What is happening?", identification of conflicts and obstacles that answer "What is the issue?" and the transformation or resolution to the issue that inform the purpose of the story. We will also discuss and show the process of transforming data into visual communications.
CINF 97

Communicating science with little (or no) budget: Design rules and tricks for the non-artist

Kiyomi Deards, kdeards2@unl.edu. Research and Instructional Services, University of Nebraska-Lincoln, Lincoln, Nebraska, United States

This presentation is for the self-proclaimed non-artist scientist who wants to communicate science effectively but has little (or no) budget to hire professionals to create and edit images (artwork, tables, graphs), websites, presentation slides, and publications. For this scientist, learning basic easy-to-apply design rules and tricks can facilitate the preparation of scientific material. The speaker has experience designing formal and informal presentations, creating videos and podcasts, working with graphic designers, and designing websites. The speaker will provide tips and suggestions based on her own experiences, collaborations, and acting as a consultant for informal science communication projects. Moreover, strategies for using advanced Google Slides and PowerPoint options, graphic design, and universal design principles (including accessibility) will be shared. Finally, finding and using open access images and other creative ways to maximize your budget will conclude this talk.

CINF 98

Supporting the FAIR challenge in chemical information

Adam Sanford, asanford@cas.org, Andrea Jacobs. CAS, a division of the American Chemical Society, Columbus, Ohio, United States

As FAIR data principles have taken hold in the academic and corporate research communities, funders and publishers have increased their demands for improved data management and accessibility in their research workflows. Supporting the scientific community for 113 years, CAS has held a front seat to many such transformations. The challenges and benefits of the core elements of FAIR are critical to understand as we consider the true impact of this movement. Several solutions to address the core principles, from findability enabled by persistent identifiers to accessibility through trusted and secure repositories and more, will be explored.

CINF 99

FAIR data for the chemistry life cycle

Matthew Clark³, mclark@drmatthewclark.com, Ted Slater², t.slater@elsevier.com, Anita de Waard¹, a.dewaard@elsevier.com. (1) Elsevier, Jericho, Vermont, United States (2) Elsevier, Suttons Bay, Michigan, United States (3) Elsevier, Philadelphia, Pennsylvania, United States

The concept of FAIR data has been developed to help exchange and use data for scientific discovery. The most widespread FAIR data resource to date has been the document object identifier, or DOI, which greatly simplifies finding the right document. We will discuss Elsevier FAIR efforts around chemistry to support the entire lifecycle from conception, acquiring reagents, and characterizing the results with NMR and other instruments. In particular, we will
describe two examples of these efforts: 1) The connection of multiple resources which include vendors, chemical reactions, and biological data on Elsevier Entellect, 2) Enabling the creation of ‘Research Objects’ connecting datasets, workflows, and analytical tools in the Mendeley Data Research Object Composer Tool.

CINF 100

**Associating live analytical data to synthetic chemistry experiments: Applying FAIR principles across the scientific experimentation lifecycle**

Andrew A. Anderson, Michael Boruta, michael.boruta@acdlabs.com. ACD/Labs, Toronto, Ontario, Canada

Chemists have at their disposal a wealth of software applications which can effectively describe experiment material utilization, procedure descriptions, observational commentaries, and result summaries—all consistent with FAIR practices.

Traditional approaches to aggregating detailed summaries of synthetic chemistry experiments, however, requires the abstraction of reaction-characterizing analytical data. Rich analytical datasets are reduced to tabular numerical data fields and static graphical depictions.

This presentation will describe efforts to use newer digital data processing, interpretation, and storage functionalities which simultaneously allow for automated result analysis, while preserving full-fidelity datasets, in accordance with FAIR principles and practices. Select examples of HT synthesis experiments, coupled to automated analysis of chromatographic and spectral characterization, will be presented.

CINF 101

**Is FAIR crystallographic data FAIR enough?**

Ian Bruno, bruno@ccdc.cam.ac.uk, Suzanna Ward. Cambridge Crystallographic Data Centre, Cambridge, United Kingdom

The crystallographic community is often heralded for its foresight and pioneering work to develop sound data management and dissemination practices. As early as the 1940s, crystallography journals were publishing both derived and processed data. Since the 1960s, data have been collected, compiled and made available by crystallographic databases and repositories spanning chemistry and biology. In the early 1990s, the community developed semantic representation formats that anticipated opportunities offered by the digital revolution and laid the foundation for one of the first data journals. Today there is a rich tapestry of services underpinned by the adoption of standards that position the sharing of crystallographic data favourably with respect to compliance with the FAIR data principles. But is FAIR crystallographic data FAIR enough or is there more that could and should be done? And is just abiding by the FAIR Data Principles enough to fully realise the benefits of reliably and sustainably sharing research data? This presentation will explore these questions with reference to crystallography and with an eye to informing FAIR activities taking place across the domain of chemistry.
CINF 102
UK physical science data-science service: FAIR resource for chemistry in UK

Simon J. Coles, s.j.coles@soton.ac.uk, Nicola Knight, N.Knight@soton.ac.uk. University of Southampton, Hampshire, United Kingdom

The UK Physical Sciences Data-science Service (PSDS) recently came into being as the successor to the UK National Chemical Database Service (NCDS). The NCDS acted as a portal to a range of, mainly proprietary, databases that had their access secured at a national scale – the primary reason for the NCDS was therefore to provide access to expensive databases that otherwise were unaffordable at the individual or institutional scale. The PSDS has now assumed this role and in the summer of 2019 a new portal to these databases was developed. The PSDS was however also funded to be more ambitious regarding access to chemical (and related sciences) data and the FAIR principles are at its core.

The mission of the PSDS is to be a driving force and enabler for data-driven physical sciences research. Over the coming years the PSDS aims to become an aggregator of chemical data – as opposed to providing access to individual databases. In this talk we discuss our plans to create a FAIRer platform containing both open and proprietary data. Making data more Findable, searchable and Accessible through new interfaces, improving interoperability through data structure and indexes, and supporting and driving Reusability by allowing users to output an aggregated dataset with associated provenance information. As part of the PSDS service, we will also produce a FAIR data resources catalogue. Underpinning this will be a taxonomy to enable machine understanding, grouping and accessing of physical sciences data resources in different ways.

CINF 103
Towards improving the research data ecosystem

Evan Bolton, bolton@ncbi.nlm.nih.gov, Tiejun Cheng, Jian Zhang, Paul Thiessen, Asta Gindulyte, Ben Shoemaker, Bo Yu. National Center for Biotechnology Information, National Institutes of Health, Bethesda, Maryland, United States

Chemistry information is critical for nearly every type of scientific research. PubChem integrates 100s of data sources. It is a massive system with web pages for compounds but also many other entities types, such as substances, bioassays, proteins, genes, pathways, data sources, patents, and publications. PubChem works with many communities to integrate their content but also to improve data standards, best practices, machine readability, FAIR-ness, and the general flow of information. This talk will give an overview in ways PubChem is helping to improve the research data ecosystem.

CINF 104
Baby steps toward FAIR chemistry data

Margaret E. Lafferty, mlaffert@umn.edu. University of Minnesota, Minneapolis, Minnesota, United States
For the past three years, the University of Minnesota Libraries have been working with a multi-disciplinary, multi-institutional research center to develop and refine a process for researchers to make data associated with articles available at the time of publication via an institutional repository. After an extensive pilot period, the project has expanded to include all publications authored by center researchers. A goal from the start of the project has been to advance data sharing in chemistry. This talk will discuss the ways in which the project has been successful at doing that, where there is room for improvement, and what factors have helped and hindered progress towards FAIR data for the research center.

CINF 105

Organic reaction data: Status, desiderata, and outlook for reaction informatics

Connor W. Coley1,2, ccoley@mit.edu. (1) Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States (2) Broad Institute of MIT and Harvard, Cambridge, Massachusetts, United States

This talk will describe my experience with mining organic reaction data for various reaction informatics applications. I will describe the current availability and quality of reaction data from the perspective of a cheminformatics researcher, and the applications that these data are compatible with. I will attempt to identify the primary desiderata for how reaction data should be acquired, stored, and shared, and how this would enable many additional applications for which the current data will not suffice.

CINF 106

Development and application of a Python library for the semantic representation of scientific data

Dylan Johnson, N01448636@unf.edu, Stuart J. Chalk. Department of Chemistry, University of North Florida, Jacksonville, Florida, United States

Before the advent of computers and digital storage, scientific data was represented via human-readable, text and image-based articles, much as it is today. Wide adoption of computers and digital storage revolutionized the accessibility of scientific information for humans, but the revolution stopped short of making the information accessible to computers. Though distribution of computer-accessible versions of scientific data should have coincided with the adoption of digital data publication, a widely adopted system for storing and distributing computer-accessible scientific data is still absent.

A data model and ontology for semantic representation of scientific data, SciData, has been developed as a means of storing and distributing data for computer accessibility and machine actionability. Wide adoption of SciData would permit aggregation and reconciliation of data from heterogenous datasets by computers to elucidate knowledge that may not otherwise be derived.

The largest hurdle to adoption of a computer-accessible data model, such as SciData, is finding a way to seamlessly incorporate generation of SciData documents into the standard scientific data publication process. The rapid growth and adoption of the Python programming
is required at this stage, the ultimate goal is to make the python library accessible to all scientists so that SciData documents can be produced along with any publication of data. This talk will therefore describe our efforts to develop and implement the SciData python library as well as how other scientists can adopt the python library for production of SciData documents from any type of scientific data.

CINF 107

Toward a globally unique persistent resource identifier for chemical substances

Stuart J. Chalk\textsuperscript{1}, schalk@unf.edu, Simon J. Coles\textsuperscript{2}, Jeremy G. Frey\textsuperscript{2}, Egon L. Willighagen\textsuperscript{3}. (1) Department of Chemistry, University of North Florida, Jacksonville, Florida, United States (2) University of Southampton, Southampton, United Kingdom (3) Biology, Maastricht University, Maastricht, Netherlands

As the promise of Findable, Accessible, Interoperable and Reusable (FAIR) data drives data science, the need for Globally Unique Persistent Resource Identifiers (GUPRIs) on all 'things' that are part of the scientific enterprise is imperative. Given GUPRIs for every type of thing, we can then assign to them labels in different languages, from different domains, and at different levels of education, and subsequently start connecting the things together.

For the chemical sciences, there are a number of important 'things' that we need to identify on the landscape of FAIR data, tools, and services. A fundamental part of this 'Internet of FAIR data and services' is assignment of a GUPRI for each distinct chemical substance, whether it be an element, ion, compound, complex, or other type of matter.

This talk will discuss the evaluation of existing identifiers for chemical substances relative to the criteria for a GUPRI, the subsequent conclusions drawn from the analysis, and the a nominal recommendation for a GUPRI for chemical substances.

CINF 108

Communication is a rare event: let’s make it happen

Catherine T. Hunt, catherinehunt55@gmail.com. Rohm and Haas/Dow (retired), Ambler, Pennsylvania, United States

A wonderful mentor and totally engaging speaker once said to me, “Communication is a rare event”. I was speechless… and good thing. They continued, “Effective communication must be 100%-100%; 50-50 will just not do.” In short, before, during, and after your presentation or discussion you must ensure that your audience is 100% with you. Over the years, this philosophy has served me very well.

Start, by taking control: a) make your slides easy to read (24 pt. font MINIMUM) and impactful (people remember pictures and stories, not words), b) have your own pointer/slide advancer and use it (mine is a Logitech R800 with a timer), and c) arrive early to check-out the room and make it your own. Once you get up, actively engage your audience. Be creative, share your passion, and, as they say in Missouri, “Show them, don’t just tell them”. Check-in with your audience during your talk. Talk with them, not at them. Then, finish strong with a
compelling conclusion summarized on an eye-catching and memorable slide complete with website and contact information. And, one last thing, always leave time for questions. If they don’t ask you one, here is your chance to ask them what they heard, or what they plan to do next.

We’ll walk through how to build a slide deck that is interesting, fast moving, and hits all the bullet points that we need to convey. See you in Philly for a rare event! @KatieChemist

CINF 109

Gateway to science

Frankie K. Wood-Black\textsuperscript{1,2}, fwoodblack90@gmail.com. (1) Phys. Sci., Eng. and PTEC, Northern Oklahoma College, Tonkawa, Oklahoma, United States (2) Sophic Pursuits, Inc., Tonkawa, Oklahoma, United States

How do we as scientists change the world? You might think that is the discovery of a miracle cure, or figuring out the mechanism to isolate a critical element, or the invention of a material that can stop a bullet, but I think it is something more fundamental. It is the art of painting a picture with words. If you can’t get people excited about science, the next miracle cure won’t receive the funding necessary to proceed through development. If you can’t inspire the community, there won’t be the drive to pursue optimization of the processes. The art of communication is crucial to the overall process of discovery and research. In the words of James Murdoch, “Through basic science literacy, people can understand the policy choices we need to be making. Scientists are not necessarily the greatest communicators, but science and communication is one of the fundamentals we need to address.”

CINF 110

CANVAS: Six pointers for formulating your message

William T. Lambert, willyT1392@gmail.com. Discovery R&D, Corteva Agriscience, Westfield, Indiana, United States

Our world is overflowing with data. In order to make informed decisions about the many challenges that lie before us, it is critical that we extract and understand the information embedded in these data. As scientists, we stand at the crossroads of data generation and the dissemination of information, and thus we have a vital role to play in this process. In this talk, I will describe six general principles (encoded by the acronym CANVAS) for creating effective presentations of scientific data that will help your audience absorb your message. My goal is to inspire you to create thoughtful visuals that will enhance the sharing of information and foster sound decision-making.
Most molecules are nanoscale objects, and thus too small to take any kind of “picture” of them. However, this is not the main challenge of molecular visualization, whether in the classroom or in a chemistry publication. After all, artists and/or computer graphics have been
fantastically successful in illustrating abstract or otherwise non-photogenic objects. Rather, it is the quantum mechanical nature of molecules that provides the greatest challenge. There aren’t any material “surfaces” to render, and virtually all molecular properties have a probabilistic nature which can be hard to capture in an image.

The concept of the “charge cloud” is introduced early in the chemistry curriculum, and is one that is readily grasped. Visualizations based directly and entirely on this entity can make many concepts of chemical reactivity more accessible to students. They can also be very effective in elucidating newly discovered molecular properties. We will present such a molecular visualization tool that reveals subatomic features of molecules that assist in explaining chemical concepts and/or novel molecular properties. This tool visualizes topological features of the Laplacian of the electronic charge density, which are measurable, reproducible, and often very transferable between molecules. The same methodology that is applied to small inorganic or organic molecules, can also applied to large biomolecules. Instead of a variety of visual textures for molecular models in the different subdisciplines of chemistry, which can disorient students, there can be greater consistency across the curriculum.

The subatomic features that are revealed in the molecular visualizations measure from 5 to 50 picometers, and their systematic investigation can help advance picotechnology. Molecular visualization research can both tell your chemistry story, and be a key element in advancing it.

Model of a diiron nitrogen reduction catalyst with picoscale subatomic features visualized with volume rendering of the Laplacian of the charge density.
CINF 112

Translational toxicology: Simultaneous data visualisation across phases

Aishling Cooke¹, aishling@optibrium.com, Edmund Champness¹, ed.champness@optibrium.com, Tamsin Mansley², mansleyte@gmail.com, Matthew D. Segall³, matthew.d.segall@gmail.com. (1) Optibrium Ltd, Cambridge, United Kingdom (2) Optibrium, Ltd, Groton, Massachusetts, United States (3) R&D, Optibrium Limited, Cambridgeshire, United Kingdom

Drug discovery and development is a knowledge-intensive process that can benefit from analysis of different data types, produced across different stages of the drug development process. These integrative approaches can facilitate better decision making throughout the drug discovery and development pipeline, improving lead optimisation and enhancing risk assessment. A pre-requisite for this is the availability of relevant, high quality data sets, yet to take advantage of such information it is important to apply powerful and flexible methods for data query, read-across, analysis and visualisation.

In this presentation we will introduce the Enhancing TRANsional SAFety Assessment through Integrative Knowledge Management (eTRANSAFE) project, in which an integrative data infrastructure is being constructed to employ innovative computational methods and tools that aim to drastically improve the feasibility and reliability of translational safety assessment during the drug development process. As part of this infrastructure, there is a significant focus on approaches to investigate and understand translational data using data visualisation. A key challenge is using visualisation to simultaneously facilitate high-level appraisal of the data and analysis of individual safety studies across the pre-clinical, clinical and post-marketing phases. We will outline approaches taken to understand the requirements of such a system and highlight the many challenges faced when bringing multi-phase data together. We will also describe and exemplify an interactive data visualisation system for investigation of translational data, with a range of data visualisations needed to elucidate the relationships that exist within the data.

CINF 113

Using machine learning to predict human health from biofluid-based metabolomics

Ethan Evans, eevans@mit.edu, david sontag, Eric Alm. mit, Cambridge, Massachusetts, United States

Personalizing healthcare requires accurate predictions of an individual’s health state. One route to doing so involves the use of minimally invasive biofluid collection, coupled with untargeted metabolomics that provides a wealth of data concerning the person’s state. To date, much focus has been on targeted studies and on developing a handful of biomarkers from those metabolites that can be chemically identified and biologically reasoned about. Here, we wish to evaluate the potential of the complete output of untargeted metabolomics data for developing health state profiles using machine learning techniques. Where possible, we reprocess and analyze 37 published metabolomics studies performed across a range of human health states. In general, we find strong predictive performance using simple and interpretable models. We demonstrate that internal to a study, there may exist significant mutual information between various metabolites and the observed health state, across different chromatographic techniques. Additionally, we find that features frequently not analyzed due to falling
above classic, but arbitrary, statistical significance values (p > 0.05 for instance) are generally capable of strong predictive performance suggesting a wealth of additional information in un-targeted metabolomics data. These results suggest a data-driven approach for advancing and further personalizing healthcare decisions.

CINF 114

Novel neural network model for table classification in chemical patents

Zenan Zhai1, Karin Verspoor1, Trevor Cohn1, Saber Akhondi2, Christian Druckenbrodt3, Camilo Thorne3, c.thorne.1@elsevier.com, Dat Nguyen1. (1) University of Melbourne, Melbourne, Victoria, Australia (2) Elsevier, Amsterdam, Netherlands (3) Elsevier, Frankfurt, Germany

Chemical patent is the most common form for disclosing novel compounds and reactions, hence is an important resource for research in both industry and academia. Key data in patents is usually presented in tables. However, both the number and the size of tables can be very large in patent documents. Categorizing tables based on their content type can help save efforts for finding tables containing information that is highly related to new inventions in patents. Previous works mainly focus on classifying tables based on their structure, ignoring semantic information. Motivated by the emerging need for table semantic analysis in the chemical patent domain, we propose a new task and a new dataset ChemTables for categorizing tables in chemical patents based on the semantics of their contents. State-of-the-art neural network models proposed for interpreting table semantics such as Table-BERT and TabNet ignore either the structural information of tables or the sequential information between table cells. In this presentation, we describe a novel neural model which leverages both sequential and structural information in tables for table classification. The experimental results on our new dataset ChemTables show that our model outperforms conventional machine learning models Naive Bayes and SVM, and the two neural models Table-BERT and TabNet.

CINF 115

Hybrid-type dynamic optimization of reaction conditions for organic synthesis

Youngchun Kwon1,2, kown10@gmail.com, Jin Woo Park1, Dongseon Lee1, Taesin Ha1, Byungkwon Choi1, Junwon Jang1, Joonke Cho1, Youn-Suk Choi1. (1) Samsung advanced institute of technology, Suwon, Korea (the Republic of) (2) Computer Science and Engineering, Seoul National University, Seoul, Korea (the Republic of)

A lot of effort has been made over the past few years to develop various deep learning technologies in the field of materials screening and design. By the way, materials synthesis is another major barrier in materials development as it is still carried out laboriously by human researchers. Therefore, several groups are trying to automate the synthetic process by combining novel robots and artificial intelligences (Als). And one of the essential AI technologies is the optimization of various reaction conditions, such as reagents, concentrations, temperature, and time, to obtain maximum yield with minimal resources. Traditionally, a number of active learning algorithms such as Bayesian optimization (BO) and random forest have been applied to solve this multi-armed bandit problem by adopting the strategies of exploitation and exploration. However, they relied only on real-time data generated from their experimental setups to train and update the active learning models. Generally, it takes a long time to accumulate enough data and eventually to reach the optimal condition. To overcome the issue, a
A hybrid-type dynamic optimization (HDO) algorithm that utilizes both previous literature and current experimental data is proposed. Unlike other optimization techniques, HDO algorithm uses two acquisition functions and recommends the most favorable experimental condition by comparing the relative advantages of exploitation and exploration. The first acquisition function is based on the deep neural network (DNN) to predict the reaction yield trained with abundant literature data. It has characteristics that emphasize the exploitation by selecting conditions with high yield but low uncertainty. The other acquisition function is based on conventional BO multiplied by user-defined weight and trained using current experimental results. The DNN extracts valuable information from rich existing data and applies it to current optimization problem when the reliability is guaranteed. And BO complements the weaknesses of the DNN model when previous data lack relevant knowledge and current experimental data can provide more powerful information. In the evaluation test, HDO showed optimization speeds 2.7 times and 1.3 times faster than random search and BO, respectively. We expect the proposed methodology to be a useful tool not only in autonomous synthesis systems but also in manual synthesis experiments by chemists.

CINF 116

Machine learning methods for hansen solubility parameters and their application in predicting solvent-polymer interactions

Aditya Sonpal\textsuperscript{1,2}, adityaso@buffalo.edu, Johannes Hachmann\textsuperscript{2}, Mojtaba Haghighatlari\textsuperscript{2}. (1) Chemical and Biological Engineering, University at Buffalo, Buffalo, New York, United States (2) Dept of Chemical and Biological Engineering, University at Buffalo, SUNY, Buffalo, New York, United States

Hansen solubility parameters (HSP) are widely recognized as a viable method to numerically estimate the interactions between solvents and other chemical compounds. A data-driven approach to predict the HSP is attractive as conventional molecular modeling of solvent systems is complex, time-consuming, and expensive. In the presented work, we employ a variety of supervised learning techniques such as simple linear regression methods (e.g., elastic nets), kernel-based methods (e.g., support vector machines), ensemble-based methods (e.g., random forest regressor), and artificial neural networks to predict the HSPs of solvents. In addition, we perform a hyper-parameter optimization of these models using genetic algorithms. In order to build physics into our data-derived prediction models, we explore the utility of \textit{first-principles} molecular descriptors of our solvent systems from electronic structure calculations as features. Here, we present a comparison of results from different algorithms and different combinations of features, and perform a cost-benefit analysis of using expensive features in these models. We then apply this model to understand the behavior of different polymer-solvent systems. The results indicate that simple linear regression models can be used to give accurate predictions for HSPs. The accuracy of these predictions is comparable to, or an improvement on the best existing models reported in the literature. The inclusion of features from electronic structure calculations only marginally improves the accuracy of our models. All models were generated using our group’s machine learning and informatics program suit ChemML, along with a module to predict solvent-polymer interactions using the concept of the 3-dimensional Hansen space. The simple linear regression models readily yield insights into those molecular features that have a predominant impact on the HSP.
Retrosynthesis with transformer and molecular grammars

Sergey Sosnin, sergey.sosnin@skoltech.ru, Maxim Fedorov. Skolkovo Institute of Science and Technology, Moscow, Russian Federation

Nowadays, computational planning of organic synthesis is of great interest, primarily due to the progress in deep learning. However, current approaches that are based on the generation of reactants and products as SMILES strings have some limitations: SMILES can be syntactically invalid, or a molecule itself can be invalid in terms of valency rules or broken cycles. To overcome these limitations, we used “molecular grammars” - templates for building organic compounds as a sequential application of grammar rules. One can regard these rules as molecular fragments with “sticky bonds” by which two rules can be matched together to construct a larger molecular fragment. These rules can be extracted from any chemical dataset by a deterministic algorithm. We used Transformer to generate either possible products of a reaction if reactants are known (direct mode) or for the generation of possible reactant structures for a compound of interest (reverse mode). The main advantages of our approach are (i) the model extracts all information from data; there is no need for explicit rules (ii) the model can operate with an arbitrary number of reagents and products (iii) the model is robust to noisy training data (iv) generated molecules are valid in terms of valency, cycles, and bonds. We compared the effectiveness of our approach with the IBM RXN model on the reaction dataset obtained from patents (USPTO). Our molecular grammar model outperformed the baseline IBM RXN model and achieved accuracy of top-1 direct prediction of 90% (vs. 89% for IBM RXN). In reverse mode, our model gained 77% for the top 5 predictions. We believe that molecular Transformer with molecular grammars can be an efficient tool for retrosynthetic analysis.

CINF 118

CINFull materials: Insights on development of custom feature set data for machine learning in materials

Christopher Baddeley, cbaddeley@cas.org. CAS, a division of the American Chemical Society, Columbus, Ohio, United States

We have seen the use of Machine Learning (ML) increase significantly across multiple scientific domains over the past decade. Now we have seen increased investiture in the materials space. While public datasets that support pharmaceutical research are widespread this is not the case for materials. More often than not we have seen Materials companies focus their ML exploration into specific needs that underpin their business model where little public data is available. This is due to a highly faceted nature of the materials industry where data is not readily transferable between areas. Here we explore some of the insights and learnings that CAS has discovered in development of custom datasets for materials ML.
CINF 119

Current state of FAIR chemical data standards: IUPAC's role

Leah R. McEwen¹,², lrm1@cornell.edu. (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

Without data standards, scientific communication is greatly diminished in the long run. The International Union of Pure and Applied Chemistry (IUPAC), in cooperation with many partners, has played a critical role in promulgating authoritative chemical data standards for over 100 years. To facilitate accurate exchange of chemical data in the digital environment, it is necessary to develop interoperable representations that are readable by both humans and machines according to the FAIR Data Principles (Findable, Accessible, Interoperable, Reusable).

The IUPAC Committee on Publications and Cheminformatics Data Standards (CPCDS) is charged with promoting "the compatibility of the electronic transmission, storage, and management of digital content through the development of standards...for human and machine-readable chemical information." The committee is actively pursuing this remit through community engagement around use cases for consistent machine-readable data representation, and review of existing IUPAC assets and dissemination procedures relative to the FAIR criteria. As a result, we have initiated several new projects to expand machine representation of chemical data classes and enhance machine accessibility to IUPAC data outputs. This presentation will outline the roadmap for IUPAC's ongoing commitment and strategic directions for supporting FAIR and accurate exchange of chemical data worldwide.

CINF 120

IUPAC efforts in the area of FAIR management of spectroscopic data

Robert M. Hanson¹, hansonr@stolaf.edu, Damien Jeannerat². (1) Chemistry, St Olaf College, Northfield, Minnesota, United States (2) University of Geneva, Geneva, Switzerland

In this presentation we will summarize recent progress sponsored by IUPAC in relation to developing standards for the FAIR management of spectroscopic data. Discussions to date have had support and input from a wide variety of stakeholders in this area, particularly in the area of NMR spectroscopy. The goal of the project is to establish a set of clear guidelines for the validation, storage, and retrieval of spectroscopic data and analyses that will revolutionize the way we work with raw and analyzed data throughout the publication process and beyond. All interested parties, from graduate students to publishers to database managers to those interested in capitalizing on novel ways of using digital data are encouraged to come find out how they might get involved and contribute to this project.
Towards machine accessible period table data

Evan Bolton¹, bolton@ncbi.nlm.nih.gov, Asta Gindulyte¹, Jian Zhang¹, Paul Thiessen¹, Leah R. McEwen², Juris Meija³. ¹National Center for Biotechnology Information, National Institutes of Health, Bethesda, Maryland, United States ²Clark Library, Cornell University, Ithaca, New York, United States ³National Research Council Canada, Ottawa, Ontario, Canada

Chemists are very familiar with the iconic periodic table. We learn about it at an early stage. One can find these in print in any chemistry text book. It may be amazing to learn that the numbers behind the elements, atomic weights and naturel isotopic abundance, are not set in stone. The IUPAC commission responsible for these values regular update their values, especially as the instrumentation improves. Strangely, in this day and age of computers, it is not yet possible to access all information about the elements in a way a computer can readily understand. This talk will highlight some of the difficulties encountered when attempting to create the PubChem periodic table (https://pubchem.ncbi.nlm.nih.gov/periodic-table/) and highlight steps within IUPAC to improve machine access to and best practices when using the most current atomic weights and isotopic abundances.

Digital representation of units of measure: CODATA task group

Robert J. Hanisch⁴, robert.hanisch@nist.gov, Stuart J. Chalk¹, Jeremy G. Frey², Simon Cox³. ¹Department of Chemistry, University of North Florida, Jacksonville, Florida, United States ²University of Southampton, Southampton, United Kingdom ³Commonwealth Scientific and Industrial Research Organisation (CSIRO), Clayton South Victoria, Victoria, Australia ⁴Director, Office of Data and Informatics, National Institute of Standards and Technology, Material Measurement Laboratory, Gaithersburg, Maryland, United States

The current state of the digital representation of units of measure (DRUM) across domains is a significant problem relative to the interoperability of data and it needs to be addressed immediately. Across the scientific disciplines there is a wide variety of knowledge about, focus
on, and care with the recording of a unit of measure with each piece of experimental, calculated, modeled or derived data. Much information is available for annotation of units for humans, however there is no authoritative source for how to represent and store units of measures (in any units’ system) in digital systems. This is a fundamental problem for data science currently and a major problem for the future integration of large, heterogeneous datasets both within and across disciplines. It is the most important single issue for the development of general or domain repositories, for the ideas behind Big Data and Open Data and the implementation of systems that support Findable, Accessible, Interoperable, and Reusable (FAIR) data.

Every measurement made where a numerical value is generated requires the recording of and association with a unit of measure. In the current research environment, where the paradigm is shifting to the digital publication of research data in openly accessible formats, researchers annotate a unit of measure by adding a string of characters to a numeric value in a computer system (database, spreadsheet, text file etc.). While the researcher may well report the unit in a common unit system (e.g. the SI) there are no guidelines about the string of characters to use in a computer system. As a result, there is a significant problem in normalization of units and this is a significant barrier to the interoperability of data.

The intention of the CODATA Task Group is to publish a recommended specification for FAIR compliant, machine actionable encoding of units of measure, supported by a Units of Measure Interoperability Service (UMIS) that supports translation among different conventions for units representation. Owing to resource limitations the initial outputs are likely to be constrained to a subset of the most common units of measure.

CINF 123

GO-FAIR chemistry implementation network (ChIN)

Simon J. Coles\textsuperscript{2}, s.j.coles@soton.ac.uk, Stuart J. Chalk\textsuperscript{1}, schalk@unf.edu, Jeremy G. Frey\textsuperscript{2}, Egon L. Willighagen\textsuperscript{3}, Nicola Knight\textsuperscript{2}, N.Knight@soton.ac.uk. (1) Department of Chemistry, University of North Florida, Jacksonville, Florida, United States (2) University of Southampton, Southampton, United Kingdom (3) Bio, Maastricht University, Maastricht, Netherlands

The Chemistry Implementation Network (ChIN) comes out of the GO FAIR project, where there are a range of networks covering both technical/political and discipline engagement aspects of FAIR. The principles and objectives of the ChIN have been endorsed by IUPAC and the two will work together, along with Research Data Alliance led initiatives, to embed FAIR principles into chemistry practice.

CINF 124

GO-FAIR ChIN: Review, panel discussion and open conversation about FAIR in chemistry

Stuart J. Chalk\textsuperscript{1}, Simon J. Coles\textsuperscript{2}, s.j.coles@soton.ac.uk, Jeremy G. Frey\textsuperscript{2}, Egon L. Willighagen\textsuperscript{3}. (1) Department of Chemistry, University of North Florida, Jacksonville, Florida, United States (2) University of Southampton, Southampton, United Kingdom (3) Biology, Maastricht University, Maastricht, Netherlands
In this session we will review the community perspectives on FAIR in chemistry from the symposium presenters and then conduct in a mediated workshop style with the aim of exploring current and potential practice / projects on two themes “Maybe not FAIR but FAIRer Chemistry” and “Your role in FAIR” i.e. in what ways can we move forward – and what can individuals do to make a difference. Both a panel discussion and open community discussion will be conducted with the goal of developing a current status of FAIR in chemistry as well as identification of action items to move FAIR forward in chemistry.

**CINF 125**

**ChEMU shared task: Chemical entity recognition and event extraction of chemical reactions from patents**

Dat Nguyen¹, Christian Druckenbrodt³, Karin Verspoor¹, Saber Akhondi², Camilo Thorne³, c.thorne.1@elsevier.com, Zenan Zhai¹. (1) University of Melbourne, Melbourne, Victoria, Australia (2) Elsevier, Amsterdam, Netherlands (3) Elsevier, Frankfurt, Germany

We introduce a new chemical information extraction shared task, named ChEMU, part of the 11th Conference and Labs of the Evaluation Forum (CLEF-2020). ChEMU proposes two key information extraction tasks over chemical reactions from patents. Task 1 --- named entity recognition (NER) --- is to identify specific types of chemical compounds, i.e. to assign the label of a chemical compound according to the role for which the chemical compound plays within a chemical reaction. Task 2 --- event extraction over chemical reactions (EE) --- involves on the other hand event trigger detection and argument recognition. We will publicly release reaction-specific gold standards --- that we describe in this presentation --- derived from patent literature and annotated by chemists (for NER and EE, resp.) in early 2020. Thereafter academic or industrial teams working in the field will be encouraged to participate in a shared evaluation campaign, by developing and contributing NER and EE models. The models will be then evaluated (by measuring their recall, precision and F1 scores), compared and jointly discussed at CLEF-2020.

**CINF 126**

**Using chemical ontologies to create molecular prediction systems for any molecular property**

Lutz Weber, lutz.weber@ontochem.com, Matthias Irmer, Claudia Bobach, Konstantin Kruse. OntoChem GmbH, Halle/Saale, Germany

We have created structure based chemical ontologies that are used to classify chemical compounds automatically. These classifications can be used with success in semantic search engines to find all representatives of a chemical class. In the present paper we would like to demonstrate use cases when utilizing these chemical classes as features in typical machine learning approaches.

Thus, we have used the co-occurrence of chemical compounds with biological and physico-chemical properties in scientific articles to train models that predict properties of novel compounds that did not occur in those training sets. One example is the prediction of hepatotoxicity as well as bioavailability. In principle, one can use any property that is found in the textual
vicinity of compounds to build such predictive models. Criteria will be presented that allow to judge the quality and predictive power of such models.

CINF 127

Evolutionary fuzzy rule induction processes for subgroup discovery in chemical datasets

**Paul Kowalczyk**, paul.kowalczyk@solvay.com. Research & Innovation, Solvay, Bristol, Pennsylvania, United States

Subgroup discovery (SD) tries to find relationships between different properties of a dataset with respect to one target variable. SD is an AI tool that ‘stands’ between classification and description. It has a target value, as do classification techniques, but, its objective is to describe a dataset, rather than predict new data. A researcher would use subgroup discovery to find simple, general, and interesting patterns with respect to a target variable of interest. The target variable may be binary, categorical, or continuous.

Evolutionary fuzzy systems are a union of evolutionary algorithms and fuzzy logic. Evolutionary algorithms are stochastic algorithms that borrow from natural evolution to solve complex optimization problems. This family of algorithms is capable of efficiently searching large, complex datasets. Fuzzy logic is an extension of traditional set theory, able to model and deal with imprecise data.

We will demonstrate how a researcher might use evolutionary fuzzy rule induction processes for subgroup discovery in chemical datasets. Use cases will be presented for identifying structure alerts, describing compound activity, and describing compound selectivity.

The input required is a representation for each molecule in the dataset (e.g., SMILES string), and one, or more, target variables. The output is a set of rules, presented both as text and graphically, and quality measures for the rules (e.g., true positive rate, false positive rate).

The work will be presented as a collection of R and Jupyter notebooks. All datasets and notebooks will be made available.

CINF 128

Machine learning transition temperatures from 2D structure

Andrew Sifain\(^3\), sifain@usc.edu, Brendan Gifford\(^1\), Jason A. Morrill\(^4\), Betsy Rice\(^2\), Samuel H. Yalkowsky\(^5\), Brian C. Barnes\(^2\). (1) Los Alamos National Laboratory, Los Alamos, New Mexico, United States (2) U.S. Army Research Laboratory, Aberdeen Proving Ground, Maryland, United States (3) US Army Research Laboratory, Claymont, Delaware, United States (4) Department of Chemistry Box 1059, William Jewell College, Liberty, Missouri, United States (5) College of Pharmacy, University of Arizona, Tucson, Arizona, United States

Phase transitions such as melting and boiling characterize a material's interaction with the environment. Further, properties such as transition temperature can influence whether a material is suitable for a given application. Examples of phase-driven applications include drug de-
livery, melt-casting explosives, or energy harvesting using phase-change materials. However, a well-known difficulty of materials discovery is that chemical synthesis is costly and time-consuming. Therefore, identifying potential candidates may proceed a long and arduous synthesis task. A priori knowledge of transition temperatures would limit the chemical space of candidate compounds, thus expediting discovery.

Predicting physicochemical properties has long been the work of linear regression-based quantitative structure-property relationships. These models have had success but are often limited by their inability to find nonlinear mappings from easy-to-derive descriptors. Recently, there has been a strong interest in applying the tools of data science to chemistry. Nonlinear machine learning (ML) algorithms have made their presence from the atomic scale to the continuum. This work builds on these modern-day predictive tools for transition temperatures. Specifically, we propose a set of features made up of group-constitutive and geometrical descriptors, previously shown to map to enthalpy and entropy—two thermodynamic quantities that drive phase transitions. A notable advantage of the method is that descriptors are derived purely from a compound's SMILES string. Thus, besides relatively simple structural characteristics such as connectivity and hybridization, there are no high-level or numerically-intensive calculations necessary. While it is generally a challenge to train models to limited experimental data, we find that a concise set of domain-specific descriptors, combined with robust and efficient nonlinear ML algorithms, provide an appealing framework for predicting transition temperatures in a diverse set of compounds.

CINF 129
Learning the physical properties of organic molecules using graph-based models

Mohammad Atif Faiz Afzal, m27@buffalo.edu, Yuling An, Anand Chandrasekaran, Andrea R. Browning, Mathew D. Halls. Materials Science, Schrödinger Inc., Portland, Oregon, United States

Physical properties of materials, such as melting point, glass transition temperature, and solubility parameters, are ubiquitous in many different materials applications. Developing fast and reliable computational approaches for these properties has been a challenging avenue for computational researchers for a long time. Molecular dynamics (MD) has shown potential for most of such properties, however, they are computationally intensive, and performing simulations for a large library of candidates is not a viable option. The recent development in advanced machine learning techniques, especially graph convolutional neural networks, has shown tremendous promise in the accurate predictions of material properties. However, most of the work in the past has focused on electronic properties. In this work, we apply graph convolutional neural networks for predicting the physical properties of organic molecules, with a focus on melting point and solubility parameters. We demonstrate that graph-based models are highly efficient, and in some cases better than physics-based models, in predicting the physical properties of materials. We also present the learning curve analysis of the models and evaluate their performance with the data size. Additionally, we use the best models from this work to predict the properties of materials from existing databases whose properties are unknown.
CINF 130

Extracting an empirical intermetallic hydride design principle from limited data via interpretable machine learning

Matthew Witman\(^1\), mwitman1@gmail.com, Sanliang Ling\(^2\), David Grant\(^2\), Gavin Walker\(^2\), Sanpan Agarwal\(^1\), Vitalie Stavila\(^1\), Mark Allendorf\(^2\). (1) Sandia National Laboratories, Livermore, California, United States (2) University of Nottingham, Nottingham, United Kingdom

An open question in the metal hydride community is whether there are simple, physics-based design rules that dictate the thermodynamic properties of these materials across the variety of structures and chemistry they can exhibit. While black box machine learning-based algorithms can predict these properties with some success, it is difficult to extract the physical basis on which these predictions are made and therefore to \textit{a priori} design novel materials exhibiting a desired property value. In this work we demonstrate how feature importance, as identified by a gradient boosting tree regressor, uncovers the strong dependence of the metal hydride equilibrium $H_2$ pressure on a volume-based descriptor that can be computed from just the elemental composition of the intermetallic alloy. Elucidation of this simple structure-property relationship is valid across a range of compositions, metal substitutions, and structural classes exhibited by intermetallic hydrides. This permits rational targeting of novel intermetallics for high-pressure hydrogen storage (low-stability hydrides) by their descriptor values, and we predict a known intermetallic to form a low-stability hydride (as confirmed by density functional theory calculations) that has not yet been experimentally investigated.

CINF 131

Practical applications of deep learning to imputation of drug discovery data

Thomas Whitehead\(^2\), Benedict Irwin\(^1\), Matthew D. Segall\(^1\), matthew.d.segall@gmail.com, Gareth Conduit\(^2,3\). (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 neural network method and its application to impute compound bioactivities and properties [1]. Unlike conventional machine learning approaches, this method can train directly on sparse, noisy bioactivity data, typical of those available in drug discovery. In combination with molecular descriptors, this enables it to learn directly from correlations between activities measured in different assays as well as structure-activity relationships. Furthermore, the model provides a robust estimate of the confidence in each prediction, enabling attention to be focused on only the most accurate results. We present a case study that demonstrates that the neural network method outperforms traditional quantitative structure-activity relationship models and discuss how these results can be used to fill in missing data, predict compound activity profiles and identify new active compounds.
CINF 132
Al-driven 3D design for orphan targets from big PDB/SAR data

Dmitri Kireev, dmitri.kireev@unc.edu. University of North Carolina, Chapel Hill, North Carolina, United States

Small molecule probes are instrumental in the discovery and validation of novel therapeutic targets. However, high-affinity ligands are unavailable for a vast majority of potentially druggable proteins. Many of those orphan (i.e., ligandless) proteins are involved in cellular signaling, protein-protein interactions or transcription machinery and have critical roles in disease. Conventional computational strategies, such as QSAR or structure-based design, are often suboptimal for such hard-to-target proteins. Here, we present FRASE-bot, a new artificial intelligence (AI)-based tool to assemble ligands for orphan proteins from fragments directly in their putative binding sites. FRASE-bot exploits the previously developed concept of FRAGments in Structural Environments (FRASE). We successfully used FRASE-based design to develop potent in vivo antitumor agents. However, the published FRASE-based approach has a critical limitation: it only allows to exploit PDB/SAR information within a given protein family, hence precluding the ligand discovery for the majority of novel targets of interest belonging to understudied families. In contrast, FRASE-bot allows exploiting the full body of 3D structural and SAR data to assemble a ligand in the binding site of any orphan protein. It makes use of convolutional neural networks (ConvNet) to (i) locate potential FRASE sites in the protein of interest, (ii) calculate fitness scores for all potential FRASEs, and (iii) assess contributions of all respective fragments to the binding affinity in the context of a ligand. Technical details of the method, as well as preliminary benchmark studies will be discussed.

CINF 133
Predicting synergism of anti-HIV drug pair combinations

Alexey A. Lagunin1,2, alexey.lagunin@ibmc.msk.ru, Sergey M. Ivanov1,2, Roman Y. Shobik2, Alexandra S. Khromova2, Anton S. Smirnov2, Nadezhda Y. Bizyukova1,2, Dmitry A. Filimonov1, Olga A. Tarasova1, Vladimir Poroikov1. (1) Department of Bioinformatics, Institute of Biomedical Chemistry, Moscow, Russian Federation (2) Department of Bioinformatics, Pirogov Russian National Research Medical University, Moscow, Russian Federation

Drug combinations are commonly used for the treatment of infectious diseases. Such therapy has several advantages in comparison with monotherapy, including higher efficacy and reduction of drug resistance and side effects. The number of pairwise combinations for approved drugs is enormous. Therefore, there is a need to develop a rational approach for the selection of the most prospective drug pairs for experimental testing and clinical trials. Based on the earlier published experimental data of multiplex screening of interacting compounds against HIV available for approximately 500 000 combinations of 1000 FDA approved drugs, we have created SAR models predicting synergism of anti-HIV drug pair combinations. The modified versions of PASS and GUSAR software which are based on substructural PoSMNA (Pairs of Substances Multilevel Neighbourhoods of Atoms) and electro-topological QNA (Quantitative Neighbourhoods of Atoms) descriptors, naïve Bayes like algorithm and self-consistent regression, respectively, were used to build the SAR models. The accuracy of prediction, calculated by the leave-one-out and 20-fold cross-validation procedures for the best SAR models, was more 0.85. The created models were used for selection of possible synergistic pairwise anti-HIV combinations for new FDA approved drugs.
Target-based vs inverse target-based drug discovery: Machine learning assisted balance between in-house compounds and commercial screening compounds

Mihyun Kim, mhmjmj@hanmail.net, Surendra Kumar, Cheongyun Jang. Pharmacy, Gachon University, Incheon city, Korea, Republic of, Korea (the Republic of)

Identification of a novel target and validation of the target (for proof of concept) is a major concern and the pivotal issue in target-based drug discovery. Starting HTS screening (virtual or real) with a chosen target, we can expect to find hit compounds against the target under the assumption that our current chemical library is enough sample to represent the population of chemical space. Unfortunately, when observing very low hit rates, limited efficacy of hits, or non-specific binding, we can suspect the target is not druggable for small molecules to give up the target. Otherwise, we can conduct various studies to overcome low-confidence on the target. However, despite advances in biological technology, insufficient characterization on target structures, their function, and their biological cascade, or the uncertain correlation of targets between species show the drawback of target-based drug discovery - this is one reason of productivity crisis in drug R&D. In addition, the quality or efficiency of compound libraries cannot be fully considered during the process.

In contrast to the approach, synthetic chemists have developed compounds and their synthetic methods based on their synthetic efficiency together with available phenotype assays. If machine learning helps to refine the process, current chemistry-initiated drug discovery, we expect the refined process will be inverse target-based drug discovery to compensate current target-based drug discovery. In my talk, I introduce chemistry-oriented synthesis (ChOS) of unprecedented drug scaffolds and their target profiling as an inverse target-based drug discovery. In addition, I suggest how to develop in-house compounds through ‘target (T)-ring (R)’ bidirectional screening. I hope that my studies can elicit the compensatory concept: from target to compounds as the initial query of drug discovery.

Chemistry-oriented Synthesis (ChOS)
CINF 135

Research context and research chemistry: Information literacy, scholarly publishing, and data management in the chemistry curriculum

Joshua D. Borycz, joshua.d.borycz@vanderbilt.edu. Library, Vanderbilt University, Nashville, Tennessee, United States

Exponential growth in the number of chemistry publications produced each year has made sorting through research topics increasingly difficult. This growth does have the advantage of exposing some of the flaws with the publication process that future chemists might be able to improve. Issues such as lack of reproducibility, overemphasis on citation metrics, data falsification, and predatory journals will have to be addressed by the next generation of scientists. To that end a new type of chemical literature course in the undergraduate and graduate curriculum at Vanderbilt University is being introduced. This course will be divided into three sections. The first half will focus on information seeking and literacy by covering contentious topics within chemistry (e.g., fracking, nuclear energy, thalidomide, and research related to the supposed vaccine/autism link) and teaching students about the databases and techniques that can help them identify trustworthy information. The second portion will focus on the history, present, and future of scholarly publishing. Students will learn about current practices by hearing from a journal editor and an acquisitions librarian that frequently negotiates with publishers. The final portion will focus on data management as a potential solution to some of the issues with publishing. Students will learn how to manage, find, store, and share data related to their research. Other curricula for scientific courses will be compared to this course by distributing a survey to faculty and staff at colleges and universities. Student learning will be assessed by using a pre/post test to determine their confidence and comfort with finding, describing, publishing, and sharing research articles and data.

CINF 136

Cartooning the data champions at the University of Cambridge

Clair Castle, cmc32@cam.ac.uk. Chemistry, University of Cambridge, Cambridge, United Kingdom

Data Champions (DCs) are volunteers who advise members of the research community at the University of Cambridge on proper handling of research data. In this, they promote good research data management (RDM) and support FAIR (Findable, Accessible, Interoperable, and Re-usable) research principles. DCs can be researchers (including graduate students), data managers, IT professionals, librarians, or data scientists. The programme is therefore an innovative way of cultivating good data practices across the University, including at the Department of Chemistry where there is effectively a DC ‘hub’. Ultimately the DCs aim to influence a change from the existing, more closed research culture at the University to an open one. One example of trying to influence this change has been through an innovative and fun project to create DC cartoons for use as an advocacy tool for good practice in RDM.

DCs undertook a stakeholder analysis exercise to try and work out: a) why RDM is of value to different stakeholders; b) their possible objections to RDM; and c) what responses a DC could formulate to these objections. The idea behind this was that if a DC was stuck in a lift, or sat next to someone at a college dinner or a meeting for example, and is having a conversation with them about RDM, and that person raised an objection to it, this could be rebutted with a
suitable response prepared in advance. This presentation discusses the project timeline, and the challenges involved in creating a range of postcards featuring cartoon scenarios where DCs have an opportunity to advocate good practice in RDM.

CINF 137

Fostering data stewardship through innovative data literacy partnerships

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

The scientific research community is currently experiencing an evolution in the amount of research data being produced and how one interacts with that data. Because of this trend, proper management and stewardship of data is crucial. In order to meet these needs, academic libraries have been increasingly developing training opportunities focused on fostering good data practices. However, the breadth of trainings offered is often limited by individual librarian and/or functional specialist capacity. One solution to addressing this issue, and thus sustaining data information literacy programs at any academic library, is identifying and executing collaborations with campus partners. For example, in 2014, Florida State University (FSU) Libraries partnered with the FSU Office of Research and Office of Proposal Development when launching a new initiative for research data management support. Building upon this successful model, FSU STEM Libraries recently launched a new STEM data literacy workshop series in collaboration with campus partners. The overarching goal of the workshops was to promote good data practice among scientists at Florida State University by providing trainings on various data analysis and visualization tools. This presentation will cover the initial development of the workshop series and its evolution into its current form. A brief discussion of ongoing and future efforts to promote data stewardship to chemists and the broader scientific community will also be provided.

CINF 138

Data in the disciplines: Partnership to create a chemical data workshop for small colleges

**Parvaneh Abbaspour**, parvaneh@lclark.edu, **Ye Li**, yel@mit.edu, **Robin Ford**, Eli Gandour-Rood, John Repplinger. (1) Watzek Library, Lewis & Clark College, Portland, Oregon, United States (2) MIT Libraries, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States (3) Collins Memorial Library, University of Puget Sound, Tacoma, Washington, United States (4) Hatfield Library, Willamette University, Salem, Oregon, United States (5) Reed Library, Reed College, Portland, Oregon, United States

To equip chemists in small colleges with data management and sharing skills, librarians from five institutions (Lewis & Clark College, Reed College, University of Puget Sound, Whitman College, and Willamette University) in the Pacific Northwest partnered with an external chemistry librarian to plan a chemical data curation workshop for faculty and students. The 1.5 day workshop took place on June 6th and 7th, 2019 and included a full day of instruction and a half-day of collaborative creation of a module for faculty attendees to integrate into their own undergraduate curriculum. This presentation will introduce how we designed, delivered and as-
essed the workshop. We will also reflect on the feedback we received from workshop participants and share our thoughts on improving the workshop for broader adoption.

CINF 139

"Don't trust your future self": Wholistic data practices one conversation at a time at Cornell University

Leah R. McEwen, lrm1@cornell.edu. Clark Library, Cornell University, Ithaca, New York, United States

Quality data are timeless and depend on detailed and enduring description and clear presentation to support findings, further study and downstream application. However, in the course of research, scientists often focus more on interpretation of the experiment at hand and less on managing individual data values and files. Can you unambiguously interpret your data from your notes last year? How do you know if the data you are analyzing are accurate and complete? What happens if data are lost due to technical failures, misfiling, or the original collector moves on?

Robust data management and notetaking practices are highly critical, highly contextual and highly personal for successful research. Cornell University has taken a responsive approach to research data support through a network of campus consultants with diverse expertise and backgrounds that address data management throughout the research life cycle. This talk will highlight some of the programs and services offered and lessons learned by this team that have intersected with the needs of chemical researchers managing their data.

CINF 140

What’s in a (file) name? Introducing data management skills in an undergraduate laboratory course

Judith N. Currano, currano@pobox.upenn.edu. Chemistry Library, University of Pennsylvania, Jenkintown, Pennsylvania, United States

Recent years have seen an increase in the number of funders requiring researchers to deposit copies of their research data in open repositories. This requires researchers to think carefully about their data and how useable it is to individuals who have not been involved in its collection. Even when they are not planning to make their data openly accessible, individual researchers need to make sure that they are understandable to and useable by collaborators, successors, and even themselves in three-to-five years' time. Learning about personal data management practices in a laboratory course can help students to develop habits that can ease the transition to independent research and publication of data in an era of openness. This presentation describes a curriculum used in a physical chemistry laboratory class that teaches students basic data management principles, including the FAIR data principles, things to keep in mind when collecting the data themselves, file naming conventions, sound data organization practices, criteria for choosing file formats, and data annotation rationales and practices. Students received two assignments asking them to prepare data sets taken in class for sharing and for long-term preservation.
New frontier in lab data management: Getting instrument data into an electronic notebook

Joanne Hockaday¹, joanne.hockaday@martel-instruments.com, Rory Macneil², rmacneil@researchspace.com. (1) Martel Instruments, Durham, United Kingdom (2) Research Space, Edinburgh, United Kingdom

The volume of data generated by lab instruments is increasing exponentially. The pressure to capture and accurately record this data, and associate it with writeups of experiments which make use of the data, is also increasing in order to ensure compliance with funder, regulatory and institutional data management requirements, and to maximize productivity.

This presentation explores an initiative designed to turn the problems created by these pressures into a solution that enables better data management.

The first part of the solution enables automated aggregation of data from multiple instruments, tacking the barrier that has historically existed because each instrument produced data in a proprietary format, making aggregation of data difficult and often impossible.

The second part of the solution enables convenient recording of experiments and sharing of experimental data in a flexible and secure platform – a modern, affordable electronic lab notebook.

The third part of the solution enables automatic entry of the aggregated instrument data into the electronic lab notebook in an organized fashion, into preconfigured folders.

The fourth part of the solution enables researchers to search and select particular, relevant, subsets of instrument data and associate them with the writeup of the relevant experiment(s) in the electronic notebook.

The principles and solution described above will be illustrated with one or more actual case studies to demonstrate how they help improve researchers’ efficiency with data workflows.

Use of electronic lab notebook with advanced cheminformatics in academia

Jinbo Lee, jlee@scilligence.com. Scilligence Corporation, Cambridge, Massachusetts, United States

Electronic Lab Notebook (ELN) has been widely used in pharma/biotech industry. However, adoption of ELN in academia has been lagging, especially ELN with advanced cheminformatics. The differences between ELN with advanced cheminformatics with 21CFR Part 11 compliance and basic system for file archiving purpose will be discussed. With case studies as examples, the speaker will talk about some of benefits of ELN use in academia environments including data standardization, advanced cheminformatics and bioinformatics, collaboration across research groups or organizations, knowledge transfer, and integration with other databases such as Reaxys and SciFinder.
CINF 143

Cultivating good chemistry data in open-source interdisciplinary repository

Tina Qin, qinna.msu@gmail.com. Harvard Library, Harvard University, Cambridge, Massachusetts, United States

Dataverse is a Harvard University-housed free open-source application of interdisciplinary repository to share, preserve, cite, explore, and analyze research data. Dataverse adopt FAIR Data Principles and have been broadly used in social sciences, physics, and humanities, but rarely by chemistry researchers. The presentation will explore introducing chemistry data into Dataverse, including a survey of the landscape of chemistry data needs, delivering tailored trainings of the repository application, collaborations of subject librarians and repository managers/programmers and case studies of different type of chemistry datasets. Aim of this project is to develop the best practices of service design, roles and responsibilities, workflows, and early metrics of adoption of chemistry data in an open-source interdisciplinary repository. The presentation will also share unique challenges of chemistry data associated with FAIR Principles and their solutions/suggestions, for instance, structural data visualization, Metadata schemas, InChI/SMILE implementation, data security and intellectual property issues with chemistry data.

CINF 144

Cultivating good data practices: Perspective from a structural database

Suzanna Ward¹, ward@ccdc.cam.ac.uk, Ian Bruno², Matthew P. Lightfoot². (1) CCDC, Cambridge, United Kingdom (2) Cambridge Crystallographic Data Centre, Cambridge, United Kingdom

The Cambridge Structural Database (CSD) was founded on a passionately held belief that collective use of data would lead to the discovery of new knowledge which transcends the results of individual experiments. Deposition and curation processes as well as the cultivation of good data sharing practices in the community have enabled that vision to come to fruition. The database is now used extensively in education as well as in industry particularly as part of drug development and materials design.

As a proponent of the FAIR data principles, the CCDC has developed a workflow for datasets deposited at the CCDC which seeks to promote these guiding principles by enabling depositors to provide information which renders the datasets more Findable, Accessible, Interoperable and Reusable. A related aspect of our recent work in this area has been to help educate others on the value of good data practices. As a non-profit organisation the CCDC has always had a keen interest in developing material to help others to use structural data to teach both chemical concepts and crystallography. More recently we have realised that we also need to use our position as a domain specific repository to engage students and researchers worldwide on best practices when it comes to data sharing too.

This presentation will highlight some of our recent activities in this area including the generation of new deposition guidelines and preservation policies through to our involvement in schools and workshops worldwide. We will share what we have learnt from our experiences and look at what more could be done to increase engagement from an early stage in the career of a researcher.
CINF 145

I read the paper, where is the data?

Evan Bolton, bolton@ncbi.nlm.nih.gov, Ben Shoemaker, Asta Gindulyte, Tiejun Cheng, Jian Zhang, Paul Thiessen. National Center for Biotechnology Information, Bethesda, Maryland, United States

Scientists, institutes, businesses, and governments generate great quantities of chemistry-associated data and publish many papers every year. Chemistry-associated data is heavily accessed by many millions of researchers and students monthly. Yet, if you ask most researchers where the data is used to create their published paper and whether they allow other researchers to access this content, it may give them pause. This talk will discuss the research data ecosystem and highlight ways scientists can share information to enhance reusability. In addition, using PubChem Upload as an example, case studies will be provided.

CINF 146

How machine learning will enrich SAVI (Synthetically Accessible Virtual Inventory)

Victorien Delannee¹, victorien.delannee@nih.gov, Hitesh Patel¹, Marc C. Nicklaus². (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

SAVI (Synthetically Accessible Virtual Inventory) is a project of computationally generating a very large database of reliably and inexpensively synthesizable screening sample structures that have desirable properties for the drug development process. While most of the current similar approaches are based on machine learning methods learning from scratch, SAVI is designed as an expert system combining a generic knowledge of reactions and reasoning rules created by human experts, encoded in the LHASA project's transform description in CHMTRN and PATRAN. Since 2014, SAVI has grown to reach more than 2 billion compounds and has led to multiple insights and related projects to improve SAVI. The presentation will address how machine learning will enrich SAVI and illustrate it through different related projects.

CINF 147

Accelerating hit discovery with multiobjective deep reinforcement learning

Mariya Popova, Olexandr Isayev, olexandr@olexandrisayev.com. Department of Chemistry, Carnegie Mellon University, Pittsburgh, Pennsylvania, United States

Designing new molecules with a set of predefined properties is a core problem in modern drug discovery and development. There is a growing need for de-novo design methods that would address this problem. We present two complementary approaches: ReLeaSE, a sequence SMILES based and MolecularRNN, the graph recurrent generative model for molecular structures. Our model generates diverse realistic molecular graphs after likelihood pre-training on a big database of molecules. Further, the model is tuned with a multiobjective policy gradient algorithm, provided a critic that estimates the reward for the multiple properties of
We show a significant distribution shift to the desired range for biological activity, physical, and ADMET properties outperforming state-of-the-art works. Finally, we show experimental validation of our computational hits on the example of the design of selective and duf-kinase inhibitors.

CINF 148

Machine learning vs. AI methods for modeling cardiotoxicity in the ‘big data’ era

Vishal Siramshetty, siramshettyv2@nih.gov, Dac-Trung Nguyen, Noel T. Southall, Natalia J. Martinez, Anton Simeonov, Alexey V. Zakharov. Division of Preclinical Innovation, NCATS (National Insitstues of Health), Rockville, Maryland, United States

Cardiotoxicity has been a leading concern behind the recall of several marketed drugs and failure of clinical candidates. Inhibition of the potassium ion channel whose alpha subunit is encoded by human Ether-a-go-go-Related Gene (hERG) leads to prolonged QT interval in the cardiac action potential. Fatal cases of cardiac arrhythmia have been reported due to drug-induced hERG channel inhibition. Several computational techniques, including machine learning (ML), have been employed to develop QSAR models for hERG. However, models based on artificial intelligence (AI) methods have been rarely reported. To this end, we performed a comprehensive comparison of binary classification models developed using the state-of-the-art ML and deep learning methods employing molecular fingerprints, whole molecular descriptors, SMILES and latent descriptors. Latent descriptors are numerical descriptors derived from the latent space of the autoencoder and adversarial autoencoder models built using all small molecules available at ChEMBL and PubChem databases. The training data (~6000 compounds) for the QSAR model was extracted from ChEMBL while the validation data (~1800 compounds) was generated internally using FluxOR\textsuperscript{TM} K\textsuperscript{+} channel assay. The latter set was pooled from an in-house compound library which was included while building the encoder models. We noticed that the performance of deep learning models was on par with the state-of-the-art ML method (Random Forests). Interestingly, the models based on latent descriptors performed well on validation data. Further, we compared our models with the models based on graph convolutional networks (GCN) that use molecular graphs as features. All our models outperformed the GCN models on validation data. Overall, this study highlights the application of AI methods to chemistry big data, readily available from publicly-accessible compound databases, for generation of novel descriptors that are promising for QSAR modeling. However, at least in case of cardiotoxicity, we demonstrate that such models do not add a significant advantage over ML models based on conventional descriptors. We believe a large benchmark study on diverse drug discovery data sets will validate our findings.

CINF 149

Large-scale rrediction of acute toxicity using multitask deep learning

Sankalp Jain\textsuperscript{1}, sankalp.jain87@gmail.com, Vishal Siramshetty\textsuperscript{1}, Vinicius M. Alves\textsuperscript{2}, Dac-Trung Nguyen\textsuperscript{1}, Noel Southall\textsuperscript{1}, Eugene Muratov\textsuperscript{3}, Mark C. Nicklaus\textsuperscript{4}, Anton Simeonov\textsuperscript{1}, Alexander Tropsha\textsuperscript{2}, Alexey Zakharov\textsuperscript{1}. (1) National Center for Advancing Translational Sciences, Rockville, Maryland, United States (2) Univ of North Carolina, Chapel Hill, North Carolina, United States (3) Medicinal Chem Natural Products, University of North Carolina, Chapel Hill,
Owing to its direct relationship to biological interactions, prediction of acute toxicity by organic compounds is one of the most challenging tasks in medicinal chemistry and pharmacology. Furthermore, the depiction of a compound for multiple toxicity endpoints, different organisms and multiple routes of administration add additional layers of complexity. Thus, there is a strong need and interest in the development of computational techniques that could aid reliable prediction of these toxicity endpoints. A large amount of data relevant for modeling of these toxicity endpoints has been accumulated and available in commercial and open source databases. In this study, we used the acute toxicity data from the RTECS compendium, accessible via the ChemIDplus database, to build a multi-task deep neural network (DNN) to predict a total of 59 toxicity endpoints spanning different species and routes of administration. We employed different molecular features (fingerprints, physicochemical descriptors, molecular graphs) for model development. In comparison to Random Forests, single-task DNN and graph convolutional neural network, the multi-task model demonstrated superior improvement in performance. In particular the smaller tasks (those with less than 100 compounds altogether in training and test sets) were significantly better predicted by the multi-task DNN model. In the light of a recent study that addressed 29 acute toxicity endpoints, we report that our model is not only able to predict higher number of endpoints but also demonstrated a superior predictive power. Furthermore, we analyzed the datasets to identify chemotypes specific to different acute toxicity endpoints.

CINF 150

Accurate prediction of nuisance compounds based on large libraries of HTS data

Vinicius M. Alves\(^1,2\), vinicius.m.alves@gmail.com, Stephen Capuzzi\(^1\), stephen_capuzzi@vrtx.com, Rodolpho C. Braga\(^3\), Daniel Korn\(^1\), Kyle Bowler\(^1\), Joshua E. Hochuli\(^1\), Adam Yasgar\(^4\), Anton Simeonov\(^4\), Ganesha Bantukallu\(^4\), Eugene Muratov\(^1\), Alexey Zakharov\(^2\), alexey.zakharov.ibmc@gmail.com, Alexander Tropsha\(^1\), alex_tropsha@unc.edu. (1) Laboratory for Molecular Modeling, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States (2) UNC Catalyst for Rare Diseases, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States (3) InsilicAll, Sao Paulo, SP, Brazil (4) NIH/NHGRI/NCGC, Rockville, Maryland, United States (5) National Center for Advancing Translational Sciences, Rockville, Maryland, United States

High-throughput screening (HTS) approaches have enabled rapid routine testing of millions of compounds towards the identification of novel “hit” molecules for therapeutic targets. However, many nuisance compounds, commonly labeled as PAINS, or Pan-Assay INterference compounds, can interfere with the assay detection technology, leading to a false positive response and thus are of great interest and importance to the drug discovery community. In this study, we have generated, integrated, and curated large libraries of HTS data for the following most common interference assays: colloidal aggregation of beta-lactamase and cruzain, FireFly and Nano fluorescence imaging, thiol reactivity, and redox activity. We have developed several computational Quantitative Structure-Interference Relationship (QSIR) models of nuisance behavior-based on HTS data mainly designed for this study. These QSIR models, developed with several hundred thousand compounds tested under various experimental conditions, were developed using multiple machine learning algorithms, including both classic methods, such as Random Forest and Gradient Boosting Trees, as well as modern approaches, such as Multi-Descriptor Read Across and Deep Learning. All the algorithms exhibited
similar behavior. We succeed in developing predictive models for all the assays, with correct classification rate (CCR) ranging between 60 and 80%. These models can be used in consensus to make an accurate and experimentally relevant prediction of potential nuisance behavior. In an effort to reduce wasted time and resource during pre-clinical drug discovery, we propose that these QSIR models be used to flag potential nuisance compounds instead of currently accepted and limited structural alert-based approaches. These models can be applied both prospectively during library design/annotation and retrospectively during HTS triage. We have developed an application tool for prediction of aggregation behavior named SCAM (Small, Colloidally-Aggregating Molecules) Detector, which we will present a live demonstration.

CINF 151

Protein-ligand binding predictions with graph neural networks

Brandon Anderson, branderson@gmail.com, Bastiaan Bergman, Jared Thompson, Stefan Schrödl. Atomwise, San Francisco, California, United States

Virtual high-throughput screening of protein-ligand binding affinity is a valuable tool in computational drug design. Here, we present recent results in the prediction of protein-ligand binding affinity using graph convolutional neural networks. Our network uses input features only upon atomic identities and pairwise relative positions. At each level of the network, we update representation for both atoms and the edges between them, analogous to message passing networks. We apply our network using a variety of metrics against open source screening benchmark datasets, and demonstrate an improvement over other physics-based neural network techniques for virtual screening.

CINF 152

Application of deep learning in purchasing compound libraries for non-target based drug discovery

Milad Salem, miladsalem@knights.ucf.edu, Arash Keshavarzi Arshad, arashka@knights.ucf.edu, Jiann Shiu Yuan. (1) Electrical and computer Engineering, University of Central Florida, Orlando, Florida, United States (2) Burnett School of Biomedical Sciences, University of Central Florida, Orlando, Florida, United States

Compound screening is a vital and impactful stage of non-target based early drug discovery. Achieving a high hit rate demands intelligent selection of the input libraries. If the chosen scaffolds have high potency, low toxicity, and high diversity, the drug discovery process has a higher chance of success. Virtual screening has proven to increase the efficiency of the early stage via predicting the activity of compounds in-silico. With the introduction of deep learning models to virtual screening, the accuracy of the models have often improved yet the interpretability of the predictions have decreased. This fact presents a dilemma in using deep learning models for compound selection. The higher accuracy increases the reliability of the selection, yet the lack of interpretability hinders the experts’ decision making. Therefore, being able to rely on the deep learning model’s prediction and interpret the output are crucial in employing the trained model in compound selection. In this work, a practical guideline is presented for using network’s predictions in order to select libraries for screening. This guideline manages the output of the model, the confidence of the prediction, and interpretation of the results. A
pragmatic training pipeline is presented alongside the selection guideline which includes transfer learning and external validation to demonstrate how training can be improved to attain better selections. Overall, the presented approach results in the selection of diverse compounds with high probability of activity and low chance of toxicity, enabling the use of deep learning models in drug repurposing or early stages of non-target based drug discovery.

CINF 153

Detection of novel TLR and STING ligands among one billion structures from the Synthetically Accessible Virtual Inventory (SAVI)

Dmitry Druzhilovskiy¹, explorermf@yandex.ru, Dmitry A. Filimonov¹, Pavel Pogodin¹, Leonid Stolbov¹, Polina Savosina¹, Nadya Tarasova², Marc C. Nicklaus³, Vladimir Poroikov¹. (1) Department for Bioinformatics, Institute of Biomedical Chemistry, Moscow, Russian Federation (2) Center for Cancer Research, National Cancer Institute, Frederick, Maryland, United States (3) NCI-Frederick Bldg 376 RM 207, Natl Inst Health NCI Ft Detrick, Frederick, Maryland, United States

Recent progress in generating large virtual chemical libraries has added billions of diverse chemical structures that significantly extend opportunities for the discovery of new pharmaceutical agents. Despite the achievements of highly active antiretroviral therapy treatment of HIV infection and its comorbidities, discovery of safer and more efficacious medicines is still required. Agonists of toll-like receptors (TLR) and stimulators of IFN genes (STING) enhance innate and adaptive immunity, including inducing antitumor immune responses. We have thus tried to identify such novel molecules with the desirable biological activities among about one billion structures included in the SAVI library. To identify the potential hits, we applied in silico screening based on the modified version of computer program PASS (Prediction of Activity Spectra for Substances) jointly with the similarity estimates based on MNA (Multilevel Neighborhoods of Atoms) and QNA (Quantitative Neighborhoods of Atoms) descriptors. After the application of this combined procedure to the analysis of the whole SAVI library, about four thousand molecules were selected for further analysis by molecular docking using the ICM-Pro program. The application of the docking procedure for cleaning of the PASS training sets led to a significant increase in the accuracy of structure-activity relationships, which improves the results of prediction. Different strategies of virtual screening in large databases will be discussed, and the results of the analysis of the SAVI library will be presented.

CINF 154

Multitask prediction using techniques of recommender systems

Ekaterina Sosnina¹,Ekaterina.sosnina@skoltech.ru, Dmitry I. Osolodkin²,³, Maxim Fedorov¹. (1) Skolkovo Institute of Science and Technology, Moscow, Russian Federation (2) Department of Chemistry, Lomonosov Moscow State University, Moscow, Russian Federation (3) Chumakov Federal Scientific Center for Research and Development of Immune and Biological Products RAS, Moscow, Russian Federation (4) Institute of Physiologically Active Compounds RAS, Moscow, Russian Federation (5) I M Sechenov First Moscow State Medical University, Moscow, Russian Federation
Chemical compounds act promiscuously in biochemical processes by interacting with multiple biological targets. On the one side, it makes drug development challenging because of the necessity of understanding the whole drug-candidate interaction profile; on the other side, it allows us to realize drug repurposing strategy and use essential information about complex compound-target interactions. These interactions can be processed by multitask modeling to perform more precise predictions than in the case of ‘classical’ single-target prediction. Recommender system (RS) is one of the multitask prediction approaches, which was becoming popular in chemoinformatics recently.

We performed a comparative study of several RS techniques for assessment of the multitask classification and regression predictions. We predicted binary antiviral activity (active/inactive) based on 650K data points for around 250K compounds against 185 viral species, extracted from ViralCHEMBL v0.1 database, as well as an activity for the manually collected dataset from ChEMBL v.25, containing 16K interactions of more than 4K compounds toward 5 types of muscarinic acetylcholine receptors. We applied three RS approaches: collaborative filtering, content-based filtering, and hybrid approach. We used methods from Surprise Python package as an example of collaborative filtering technique, and utilized SGIMC method, as content-based filtering realization. We implemented a hybrid method as a consensus of both approaches. Algorithms were applied, and their performance was compared in three scenarios: i) interaction prediction for pairs of known compounds and targets, ii) prediction an interaction profile for new compounds, and iii) new targets. Prediction evaluation was performed by ROC AUC score, MAP@k, MAR@k, and coverage.

The investigation revealed some advantages and limitations of RS approaches in drug development. We demonstrated that collaborative filtering methods are easy to apply but require a high level of data completion, and the sparsity of data restricts the possibility of their efficient application. The crucial disadvantages of the approach are the inability to interpret the results and provide a prediction for the new elements which were not used in the training set. The content-based methods are more advanced and allow to predict interaction for the new elements but highly dependant on the proper compounds and viruses characterization.

CINF 155

**No coding required: Processing chemical data with KNIME**

*Mark Archibald, archibaldm@rsc.org. Royal Society of Chemistry, Cambridge, United Kingdom*

The need to learn programming skills can present a significant barrier to a non-expert wishing to explore cheminformatics. One alternative is KNIME, a free data processing platform that requires no programming expertise. This talk will introduce how to use KNIME to process chemical data, illustrated with examples of how we process data submitted to ChemSpider.

CINF 156

**Computational notebooks for cheminformatics**

*Paul Kowalczyk, paul.kowalczyk@solvay.com. Research & Innovation, Solvay, Bristol, Pennsylvania, United States*
We present a collection of Jupyter and R notebooks that may be used to introduce cheminformatics in a classroom / (computational) laboratory setting. These computational notebooks weave together text, code, and results into an interactive narrative. The notebooks touch on all aspects of workflows associated with the mining / analyzing / visualizing / reporting of cheminformatics datasets, e.g., data ingestion, curation of molecular structures (i.e., structure standardization), building training sets and test sets, machine learning, model evaluation, and presentation graphics. Studies of both organic and inorganic datasets will be presented. Examples of cheminformatics notebooks presented will include building quantitative structure activity relationships, generating structure alerts for toxicological endpoints, and predicting band gap energies. Notebooks lend themselves to the practice of reproducible research: one may re-run the analyses; run the analyses with new datasets; and modify the code for other purposes. All datasets and notebooks presented will be made available.

CINF 157

Teaching introductory cheminformatics and machine learning with Mathematica

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

Increasingly, professional chemists are required to generate and query chemical data, and then use those data to make inferences. Yet most undergraduate chemistry curricula do not include coursework in computer programming, data management, chemical informatics (“cheminformatics”), or machine learning for structure-activity modeling.

In this talk, I will describe a project-based chemical and materials informatics course for undergraduate students with no prior programming background using Mathematica. The new Molecule[] functionality in Mathematica 12 comes with built-in capabilities for parsing IUPAC, SMILES and InChI representations, generating popular representations (e.g., graphs, Coulomb matrix, radial density functions), and computing common 2D and 3D descriptors. Built-in connectivity to PubChem and ChemSpider allows for searching chemical structure, and rich Import capabilities allow for accessing spectroscopic data files. Finally, the machine learning superfunctions can be used to develop data handling skills and QSAR models. In addition to providing some brief demonstrations, and an overview of the curriculum, I will showcase undergraduate student projects from the course.

CINF 158

Chemical structures in the Google era

Gregory M. Banik¹, gregory_banik@bio-rad.com, Sara Tenney², Amanda Koenig², Alison Kreckmann². (1) Bio Rad Laboratory, Informatics Division, Philadelphia, Pennsylvania, United States (2) Publications Division, American Chemical Society, Washington, District of Columbia, United States

As chemists, we communicate the details of our research in both oral and written form. Oral communication is ephemeral—a point can be made, but it is impermanent. Lasting scholarly communication of chemistry is in written forms such as journal articles and presentations, a large portion of which will be words and numbers represented as alphanumeric text.
Machines—search engines like Google, web sites like ACS Publications, etc.—have little problem storing text and making it searchable. This is unfortunately not the case with chemical structures, which machines can have difficulty understanding.

This talk focuses on best practices for making chemical structures suitable for machine interpretation as well as human interpretation—problems of which all authors, reviewers, and publishers will face when sharing chemical structure information in digital formats in the Google era.

CINF 159

Preparing machine-learning-ready datasets using the ESCALATE ontology

Ian M. Pendleton¹, ipendlet@umich.edu, Alex J. Norquist², Joshua Schrier¹. (1) Chemistry, Haverford College, Haverford, Pennsylvania, United States (2) Haverford College, Bryn Mawr, Pennsylvania, United States

Machine learning is increasingly being called upon to aid in new materials discovery. Machine learning provide a means to map variations in experimental composition, reaction conditions, and physicochemical descriptors to possible reaction outcomes. Building successful machine learning models requires the input data be structured such that humans can accurately input relevant information and computers can operate on the collected data. Toward this end, we put forward the ESCALATE platform and associated ontology for building flexible data sets for machine learning. The ESCALATE system has been applied to the development of two distinct materials research problems: high-throughput metal halide perovskite crystal discovery and the synthetic accessibility of polar racemates. The first project, high-throughput data collection for perovskite discovery, required development of data-pipelines which integrated data storage methods, hardware-software interfaces, and simple experimental user interfaces. Since implementation, ESCALATE has enabled the discovery of 24 previously unreported perovskite systems with retrograde solubility, over 7,500 experimental runs, and real-time collaborative design-of-experiment campaigns with data scientists. The success of the predictive models developed from the perovskite dataset have guided the synthesis of high quality single crystals in previously unexplored systems. In the second project, the ESCALTE ontology was used to construct and analyze polar racemate experiments. These data enabled the development of robust visualizations, interpretable decision tree models, and unbiased assessments of human intuition through interpolative and extrapolative modeling (including gradient boosted trees, k-nearest neighbors, and support vector classifier models). Perovskite and polar racemate materials discovery will be used to underscore the flexibility of the ESCALATE approach and enable a deeper discussion into the key aspects of organizing and managing experimental data at various scales. This work provides an overview of the tools necessary for further development of real-time data capture in materials chemistry as well as outlining a path toward dataset construction and dissemination.

CINF 160

DevOps in digital chemistry: Cheminformatics and data science toolchains

Thorsten Gressling, thorsten.gressling@gressling.com. JupyterChem, Berlin, Germany
From Gaussian to Gromacs, from RDKit to quantum computing - the history of cheminformatics, which in some cases goes back more than forty years, has produced a wide range of techniques, implementations, and processes. They are still in use today. With upcoming new technologies, we can assemble these varieties more logically.

From the perspective of the digital researcher, there is a demand for building pipelines through a wide variety of subject matters like QM, MM, QSAR, and AI (data science). Another dimension is the classic IT perspective, which addresses processes like the cloud paradigm, versioning, and security, as well as the handling of high-performance computing. So the domain of digital research is a result of merging the process dimensions of chemistry and IT: Model building, Testing, Releasing, and Monitoring. We will present an ontology containing hundreds of objects in the intersection of Chemistry and IT. On that, a cluster analysis is performed to find relevant structures. With these structures, we can model processes. Then, for example, different ecosystems can be compared. Finally, it is fundamental in the construction of an artificial intelligence-based stack performing autochemistry.
CINF 161

Lead-based virtual screening and prediction of EGFR inhibitors using PubChem’s database with data mining and machine learning algorithms

Kedan He, nina.kh23@gmail.com. Physical Sciences, Eastern Connecticut State University, Willimantic, Connecticut, United States

The epidermal growth factor receptor (EGFR) is a transmembrane glycoprotein that constitutes one of four members of the erbB family of tyrosine kinase receptors. EGFR is over-expressed in a variety of tumor cell lines and has been associated with poor prognosis and decreased survival rates. The most common pharmacology approach to inhibiting EGFR has been to develop small-molecule inhibitors that bind to the intracellular portion of the receptor to prevent tyrosine kinase phosphorylation and subsequent activation of signal transduction pathways. The present study aimed to identify high-affinity compounds targeting EGFR with safer pharmacological profiles through lead-based virtual screening approaches combined with machine learning classification models and molecular docking. A collection of EGFR inhibitors were served as the query reference molecules for identification of structurally similar database compounds by Tanimoto-based similarity searching against the PubChem database. PubChem BioAssay data is used as a training and test set for developing classification models that distinguish active and inactive compounds against EGFR. The algorithms used include Random Forest, Supported Vector Machine, and Multilayer Perceptron. The highest-ranked database compounds were subjected to the classification model for predicting their bioactivities and the compounds passed both filters were successively docked upon EGFR for calculating the binding affinities. The final candidates were also further subjected to Lipinski and Verber’s filter to confirm both lead-like and drug-like properties.

CINF 162

Unlocking data science and cheminformatics insights for chemists

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

Many believe that Artificial Intelligence has the potential to revolutionise not only life sciences and healthcare, but also chemistry and chemical synthesis, which is why there has recently been a great increase in applying Machine Learning and Big Data-handling techniques to medicinal chemistry and chemical synthesis. Efficiently and reproducibly training new models and learning from the results as well as the underpinning data requires a combination of different types of information from various sources.

This talk will give an overview of how the data and modelling platform that was initially developed with data scientists and cheminformaticians in mind, could also be made into a valuable source of insights for other chemists by including the right tools and guidance to unlock power usually reserved for data scientists and those with strong programming skills.
Beyond the bench: Leveraging laboratory expertise in your cheminformatics development

Matthew Prissel, mprissel@cas.org. CAS, a division of the American Chemical Society, Columbus, Ohio, United States

The field of cheminformatics bridges the disciplines of technology, data science, and chemistry and traditionally requires expertise in each domain. The well-trained cheminformatician is able to bring significant value to R&D projects by leveraging both empirical and in silico-derived chemical information to generate insights. As off-the-shelf tools and easy-to-use data sets grow in availability, chemists are increasingly able to improve results at the bench through informatics approaches without the need of specialized technical training. This talk will focus on how chemists can capitalize on their laboratory experience and depth of knowledge of science to succeed with cheminformatics.

Fall 2019 Cheminformatics OLCC: Intercollegiate course within the LibreText HyperLibrary

Robert E. Belford¹, rebelford@ualr.edu, Ehren C. Bucholtz², Sunghwan Kim³, Jordi Cuadros⁴, Dean H. Johnston⁵, Kristen D. Fulfer⁶, Andrew P. Cornell⁷. ¹Univ of Arkansas at Little Rock, Little Rock, Arkansas, United States ²Basic Sciences, St. Louis College of Pharmacy, St. Louis, Missouri, United States ³National Library of Medicine, National Institutes of Health, Rockville, Maryland, United States ⁴Departament Estadística Aplicada, IQS, Barcelona, Spain ⁵Department of Chemistry, Otterbein University, Westerville, Ohio, United States ⁶Chemistry, Louisiana State University, Baton Rouge, Louisiana, United States ⁷Chemistry/Biology, University of Arkansas at Little Rock, Little Rock, Arkansas, United States

OLCCs (OnLine College Courses) have been organized by the ACS CHED CCCE (Committee on Computers in Chemical Education) since 1996 and allow schools to offer classes in topics they normally could not offer. These are often collaborations involving industries, chemical societies and government agencies in addition to academic institutions, and have covered topics like environmental and industrial chemistry, pharmacology, cheminformatics, and chemical health and safety. In the Fall of 2019 the third OLCC on Cheminformatics was held and offered in 5 schools in the US and Spain. This was a collaborative project with NIH PubChem and all the content is available in the public domain through the LibreText OER platform https://chem.libretexts.org/link?143689.

Part of the goal of the course was to introduce students to programming while fulfilling the cheminformatic assignments and two tracks were developed, one in Python and the other in R. Students had the option of running the software on a local host or using a Jupyter Hub on LibreText. Participants interacted with each other through the hypothes.is annotation system and this presentation will go over both the class activities and the experience of running an intercollegiate course within the LibreText HyperLibrary.
Cheminformatics education using PubChem

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

PubChem (https://pubchem.ncbi.nlm.nih.gov) is a public database containing a large corpus of freely available chemical information, collected from more than 690+ data sources including university labs, government agencies, pharmaceutical companies, chemical vendors, publishers and a number of chemical biology resources. It draws millions of unique users per month and many PubChem users are undergraduate or graduate students at academic institutions. Therefore, PubChem has a great potential as an online resource for chemical education. However, many teachers as well as students merely view PubChem as a free online reference work that can be used in traditional chemistry courses, without recognizing new educational opportunities that PubChem provides for chemistry majors in the age of big data. This presentation discusses important aspects of PubChem as a cheminformatics education resource, including data quality and accuracy, data provenance and governance, structure standardization, terminologies, and more. It showcases various PubChem tools and services for search, analysis, and download, that can be useful in the chemistry classroom. In addition, we will discuss PubChem’s education and outreach efforts for the chemical education community.