CHEMICAL Information BULLETIN

ACS Chemistry for Life®

Winter 2015 — Vol. 67, No. 4

A Publication of the
Division of Chemical Information
of the ACS

ISSN: 0364–1910

Chemical Information Bulletin

A Publication of the Division of Chemical Information of the ACS

Winter 2015 — Vol. 67, No. 4

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

ISSN: 0364–1910

Message from the Chair

CINF's Boston ACS program was an outstanding success! This was one of the largest CINF programs with over 170 papers, beating the previous records of 130-140 talks in San Diego and San Francisco. Thanks to our program chair, Erin Davis, for all her hard work!

At this meeting, CINF participated in some novel programming, including organizing a panel discussion of Careers in Chemical Information and Cheminformatics, in conjunction with the ACS Graduate and Undergraduate Program Offices. Thank you to Lori Betsock, ACS Undergraduate Program Office, for her assistance with organizing this program and providing refreshments. In addition, on Wednesday afternoon CINF hosted a Wikipedia edit-a-thon, in conjunction with the ACS Committee on Public Relations and Communications - another "first" for CINF! Thank you to Keith Lindblom, ACS National Historic Chemical Landmarks Program, for his assistance with organizing this program.

The Herman Skolnik Award symposium honoring Jürgen Bajorath provided for some excellent presentations to a standing room only audience, despite some unfortunate last minute cancelations and substitutions due to illness. Veer Shanmugasundaram put together an outstanding program highlighting Jürgen's entire career in talks from mentors, colleagues, and recent students. Also a "first" this year, we hope that the Herman Skolnik Award symposium will be published as an ACS Symposium Series book. Hopefully, this will become a recurring tradition, further highlighting and honoring our Skolnik awardee.

Thanks to our luncheon speaker, Michele Derrick, Research Scientist, Boston Museum of Fine Arts, who discussed CAMEO (The Conservation and Art Materials Encyclopedia Online (CAMEO), an electronic database that disseminates technical information on terms, materials, and techniques used in the fields of art conservation and historic preservation. This is the first time in recent history that we actually ran out of luncheon tickets and sold out with record attendance!

In the area of novel programming, the "Intercollegiate Cheminformatics Education Symposium" was funded by the ACS Innovative Projects Grant program to the Division of Chemical Information in collaboration with the Division of Chemical Education (CHED). I want to thank Robert Belford, Stuart Chalk, and Leah McEwen, for their work on this proposal. The symposium is part of a larger project within CHED in which several members of CINF are participating, the Cheminformatics OLCC (OnLine Chemistry Course, http://olcc.ccce.divched.org). It is a concurrent online conference style course covering a range of chemical information and cheminformatics skills, bringing in the expertise of a range of chemical information professionals, including chemistry librarians, research and teaching faculty, and government scientists. This course was developed with NSF funds and is running as a pilot in collaboration with four chemistry departments this fall (http://olcc.ccce.divched.org/Fall2015OLCC). Presenting at the symposium is the capstone event for the students taking the course, and provides an opportunity for them to meet each other, the course faculty, and many chemical information professionals at the ACS National Meeting in San Diego, CA.

One of the issues, which consistently emerges from our discussions, is *what added value can CINF provide to its members outside of the national meeting programming*? Only a small percentage of our division members attend the national meetings, so let me highlight some of the recent activities. Thanks to Carmen Nitsche and Belinda Hurley, webinar coordinators, CINF has been able to offer many interesting topical webinars throughout the year (http://www.acscinf.org/content/webinars).

Some ACS Symposium Series books, such as the upcoming Herman Skolnik Award Symposium book, are another information resource developed by CINF outside of national meetings. The most recent idea emerged from a Boston presentation by Rajarshi Guha and Noel O'Boyle "So I have an SD file…what do I do?" It suggests that CINF should support a repository of organized and annotated links to various cheminformatics tools available on the web. We will be developing a collated index of useful cheminformatics tools for the community. I am open to other suggestions from the division membership regarding *what CINF could and should be doing for them outside of national meetings.*

I want to thank some of our CINF colleagues for their valued service to the Division over many years, who are stepping down from their current positions: Leah McEwen as secretary, and Guenter Grethe as coordinator of the CINF Scholarship for Scientific Excellence and a representative to the ACS Multidisciplinary Program Planning Group.

We are now looking for active participants to suggest or organize symposia, coordinate webinars, and help with fundraising and social events. Please email me (Rachelleb1@gmail.com) if you would like to play a valuable role volunteering within CINF. I am looking forward to meeting you virtually or in person at the spring meeting in San Diego. Feel free to introduce yourself either way!

Rachelle Bienstock, Chair, ACS Division of Chemical Information

ACS SYMPOSIUM SERIES eBooks Sponsored by the Division of Chemical Information 2010-2014

Science and the law: analytical data in support of regulation in health, food, and the environment
Town, William G; Currano, Judith N. 2014. http://pubs.acs.org/isbn/9780841229471

The future of the history of chemical information
McEwen, Leah Rae; Buntrock, Robert E. 2014. http://pubs.acs.org/isbn/9780841229457

Special issues in data management
Xiao, Norah; McEwen, Leah Rae. 2012. http://pubs.acs.org/isbn/9780841227125

Library design, search methods, and applications of fragment-based drug design
Bienstock, Rachelle J. 2011. http://pubs.acs.org/isbn/9780841224926

Special topics in intellectual property
Twiss-Brooks, Andrea. 2010. http://pubs.acs.org/isbn/9780841225947

There are 12 books in total sponsored by CINF from 1977 to 2014. The first chapter of each is free to read. ACS members may use their "Universal Member Access" benefit (for any 25 articles, including ebooks).

Letter from the Editor

Welcome to the most comprehensive post-conference issue of *Chemical Information Bulletin* (*CIB*). At almost a hundred pages, this compilation reflects on the largest ever CINF technical program (180 presentations) involving many divisional activities for education, career guidance, collaboration and outreach, social events, and award ceremonies. Furthermore, this proceedings was made possible by the exceptional voluntary efforts of the *CIB* authors (35 article submissions).

It is like no other time in our history for CINF, echoing with the ACS meeting theme, "Innovation: from Discovery to Application," through its innovative programming, peaking at a high honor with the 2015 ChemLuminary Award for the iRAMP Chemical Safety Information Project (http://www.irampp.org), jointly funded by CINF and the Division of Chemical Health & Safety. At the Boston national meeting a full-day symposium explored progress on the iRAMP project (presentation slides). Earlier in the summer a Bi-Society Symposium on Laboratory Safety Information was jointly organized by CINF and the Chemistry Division of Special Libraries Association (report in this issue of the *CIB*). Many thanks to Leah McEwen, Ye Li, and their collaborators in other divisions for moving forward their innovative ideas and making their applications recognized by ACS!

Keep reading this *CIB* to learn news of other CINF innovative programs as well as to ponder over an interesting interpretation of the meeting theme in the Editors' Corner.

This year we continue celebrating "golden" anniversaries with a feature article of the Chemical Structure Association Trust history, kindly written by Bonnie Lawlor in collaboration with trustees. You will also find a few topical continuations from previous issues such as a write-up of the symposium "Scientific Integrity: Can We Rely on the Published Scientific Literature?" by Bill Town, a review of a hot-of-the-press book "Data Management for Researchers" by Bob Buntrock (an interview with the book author Kristin Briney), an insight into the Merck Index *Online* by Mark Archibald (an interview with Maryadele O'Neil), and an overview of "Another Committee" by Jeremy Garritano. This issue continues celebrating the Herman Skolnik Award with a remarkable story of the symposium honoring Jürgen Bajorath skillfully written by Wendy Warr.

Overall, thanks to Erin Davis for "cat herding" (to quote her interview) as Program Chair 2013-15 *plus* spring 2016, and congratulations for her stepping up as CINF Chair-Elect. In this context, let me report this year's CINF election results: 2016 Chair-Elect: Erin Davis; 2016-17 Secretary: Na (Tina) Qin; 2016-18 Councilor: Bonnie Lawlor; 2016-18 Alternate Councilor: Carmen Nitsche; 2016 Councilor: Svetlana Korolev; 2016 Alternate Councilor: Jeremy Garritano. The incoming 2016-17 Program Chair is Elsa Alvaro (member profile).

In closing, I would like to thank all authors for submitting their symposium write-ups, committee reports, feature articles, and sponsor announcements. Special thanks to Bonnie Lawlor, David Shobe, and Wendy Warr. See the Boston photos at: https://www.flickr.com/photos/cinf/sets.

I hope you enjoy reading this *Bulletin* and start rolling up your sleeves for "Computers in Chemistry."

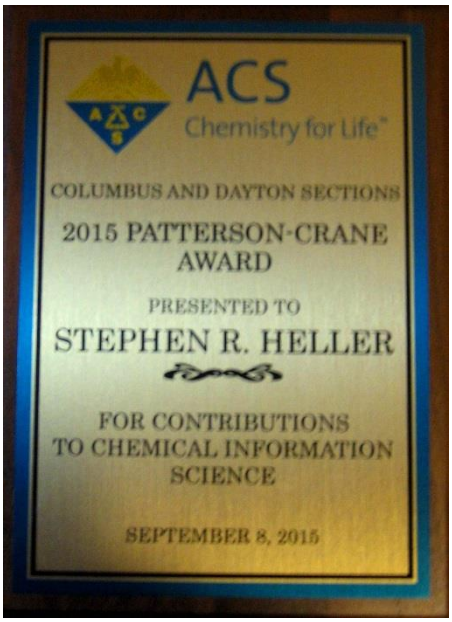
Svetlana Korolev, Editor, Chemical Information Bulletin

Awards and Scholarships

Dr. Stephen R. Heller Receives the 2015 Patterson-Crane Award

The 2015 Patterson-Crane Award ceremony was held on Tuesday evening, September 8, 2015, at the Clintonville Woman's Club in Columbus, OH to honor Dr. Stephen R. Heller for his work on the development of the IUPAC International Chemical Identifier (InChI). Dr. Heller described how and why InChI was developed and how it is being used today to find chemical information.

The Columbus and Dayton (Ohio) Sections of the American Chemical Society sponsor the Patterson-Crane Award for contributions to chemical information. It is international in scope and given in honor of two outstanding members of the sections: Austin M. Patterson (1876-1956) and E.J. Crane (1889-1966).



The biennial award consists of a $3,000 honorarium and a personalized commendation. The award is funded by a bequest of the Patterson family to the Dayton Section, by the Helen G. Crane Fund of the Columbus Foundation, and by the Patterson-Crane Award Fund of the Columbus Section.

The Austin M. Patterson Award was established in 1949 by the Dayton Section to acknowledge meritorious contributions in the field of chemical literature and especially documentation of chemistry. Dr. Patterson, the first recipient of the biennial award, was recognized for his leadership in organic chemical nomenclature and his work as editor of *Chemical Abstracts*. There were 13 additional recipients of this award, including E.J. Crane, who was editor of *Chemical Abstracts* from 1915 to1958. Subsequently, there was a desire to honor and establish an award in his memory. In February 1975 the ACS Board of Directors accepted a proposal by the Dayton and Columbus Sections for a jointly sponsored Patterson-Crane Award. Through 2010 there have been 31 recipients of the joint award. (Past recipients of the award are listed at: http://columbus.sites.acs.org/pcawardmore.htm).

Steven Rosenthal, Chair, Patterson-Crane Award Committee

2015 CINF Scholarship for Scientific Excellence Presented

The scholarship program of the Division of Chemical Information (CINF) of the American Chemical Society (ACS) is designed to reward students and postdoctoral fellows in chemical information and related sciences for scientific excellence, and to foster their involvement in CINF. Since 2005 the program has awarded scholarships at each of the ACS National Meetings, 58 scholarships in total. The awards at the 250th National Meeting in Boston were sponsored by the Royal Society of Chemistry.

Applicants presented their posters at the CINF Welcoming Reception and the Sci-Mix session, and the four winners received scholarships at the CINF Luncheon during the same meeting. Two full scholarships valued at $1,000 each were awarded to Darshan Mehta and Florian Roessler, and the third scholarship was given to a team of Ewa Gajewska and Sara Szymkuc.

The names of the recipients and the titles of their posters are (listed from left to right on the photo):

Darshan Mehta, Department of Chemical and Biochemical Engineering, The Ohio State University, Columbus, OH, USA, "*Chemical alerts and QSAR models based on dynamically-generated annotated linear structural fragments."*
Co-authors: James Rathman, Chihae Yang.

Florian Roessler, Department of Chemistry, University of Cambridge, Cambridge, UK, "*A knowledge-based approach to the parameterization of small molecule force fields based on crystal structures."*
Co-authors: Oliver Korb, Robert Glen, Peter Bond.

Sara Szymkuc, Institute of Organic Chemistry, Polish Academy of Sciences, Warsaw, Poland, "*Chess-like algorithms behind Chematica's retrosynthetic planning."*
Co-authors: Ewa Gajewska, Tomasz Klucznik, Piotr Dittwald, Michael Startek, Karol Molga. Michal Bajczyk, Bartosz Grzybowski.

Ewa Gajewska, Institute of Organic Chemistry, Polish Academy of Sciences, Warsaw, Poland, "*Retrosynthesis of complex molecules using Chematica."*
Co-authors: Sara Szymkuc, Tomasz Klucznik, Piotr Dittwald, Michael Startek, Karol Molga. Michal Bajczyk, Bartosz Grzybowski.

The next scholarships are jointly sponsored by InfoChem and Springer and will be awarded at the spring 2016 ACS National Meeting in San Diego, CA.

The coordination of the scholarships will be in the capable hands of Stuart Chalk (University of North Florida). I wish him much success for the future.

Guenter Grethe, Coordinator, CINF Scholarships for Scientific Excellence

2016 Herman Skolnik Award Announced

The American Chemical Society Division of Chemical Information is pleased to announce that Stephen Bryant and Evan Bolton have been selected to receive the 2016 Herman Skolnik Award for their work on developing, maintaining, and expanding the web-based NIH/NLM/NCBI PubChem database and related software capabilities and analytic tools to enhance the scientific discovery process. The award recognizes outstanding contributions to and achievements in the theory and practice of chemical information science and related disciplines. The prize consists of a $3,000 honorarium and a plaque. Drs. Bryant and Bolton will also be invited to present an award symposium at the fall 2016 ACS National Meeting to be held in Philadelphia, PA.

Stephen H. Bryant received his B.A. in chemistry and english from the University of Virginia, and then completed a Ph.D. in biophysics at the Johns Hopkins University School of Medicine, where he studied protein crystallography. He did postgraduate work at Johns Hopkins University Schools of Medicine and of Hygiene and Public Health, followed by a stint at Brookhaven National Laboratory working on the Protein Data Bank (PDB). Bryant then spent some time in upstate New York as a Research Scientist, Wadsworth Center for Laboratories and Research, New York State Department of Health, and Assistant Professor, Department of Biomedical Sciences, School of Public Health, State University of New York at Albany before going to the National Library of Medicine, National Institutes of Health as a Senior Investigator.

Evan Bolton received a B.S. in chemistry from Rider College in Lawrenceville, New Jersey and a Ph.D. in physical chemistry from the University of Georgia. He held positions as a computational scientist at American Cyanamid and IRL, Inc. before becoming a consultant who led the creation, use, and management of computer applications to manage scientific data. At first under contract to the National Center for Biotechnology Information (NCBI) to work on the PubChem project, Bolton joined the NCBI staff as Lead Scientist in 2004.

The awarding of the 2016 Herman Skolnik Award to Bryant and Bolton recognizes the significant contribution of the creation of the necessary computer and software systems to make PubChem information (a database of small molecules and biological activity information) easily web-accessible to biomedical researchers. Under Bryant and Bolton's leadership, the PubChem team has created a world-class resource for chemical and biological information. PubChem is the first major public database to connect cheminformatics to bioinformatics and thereby provide a unique information resource for pharmaceutical research.

From its beginning, PubChem's overriding goal has been to provide comprehensive information on the biological properties of small molecules. Since 2004, PubChem has grown to 196 million chemical substance records, encompassing 68 million unique compounds. It has been a demonstration of a public and private cooperation that has benefited the entire scientific community in collecting and integrating resources, as demonstrated by collaborations with over 250 academic and commercial organizations who have contributed records to PubChem. Biological screening results are available from over 1.1 million bioassay screens for over 3.1 million tested substances. Every day, tens of thousands of researchers from university labs, as well as pharmaceutical and biotech companies access PubChem in their drug discovery efforts.

Bryant and Bolton have provided important leadership for the PubChem project. Bryant is guided by an overarching vision of data integration, and in particular his focus on adding the third dimension to

chemical structure searches, likely guided by his early experience in structural biology and crystallography. Bolton has had a consistent eye and focus on engineering and design, and has shared his insight and expertise with others in the field, especially with the intricacies of building highly robust chemical registration systems.

Bryant and Bolton are also cited as lucid and determined advocates not only for PubChem, but also for cheminformatics and chemical sciences in general. They are valued among their colleagues, having worked with other projects such as ChEMBL (http://www.ebi.ac.uk/chembl) at EMBL-EBI, an open medicinal chemistry data resource and on the ChEBI (http://www.ebi.ac.uk/chebi) chemical ontology database.

Andrea Twiss-Brooks, Chair, CINF Awards Committee

2017 Herman Skolnik Award: Call for Nominations

The ACS Division of Chemical Information established this Award to recognize outstanding contributions to and achievements in the theory and practice of chemical information science. The Award is named in honor of the first recipient, Herman Skolnik.

By this Award, the Division of Chemical Information is committed to encouraging the continuing preparation, dissemination, and advancement of chemical information science and related disciplines through individual and team efforts. Examples of such advancement include, but are not limited to, the following:

- Design of new and unique computerized information systems;
- Preparation and dissemination of chemical information;
- Editorial innovations;
- Design of new indexing, classification, and notation systems;
- Chemical nomenclature;
- Structure-activity relationships;
- Numerical data correlation and evaluation;
- Advancement of knowledge in the field.

The Award consists of a $3,000 honorarium and a plaque. The recipient is expected to give an address at the time of the Award presentation. In recent years, an Award Symposium has been organized by the recipient.

Nominations for the Herman Skolnik Award should describe the nominee's contributions to the field of chemical information and should include supportive materials such as a biographical sketch and a list of publications and presentations. Three seconding letters are also required. Nominations and supporting material should be sent by email to awards@acscinf.org. Paper submissions will not be accepted. The deadline for nominations for the 2017 Herman Skolnik Award is June 1, 2016.

Andrea Twiss-Brooks, Chair, CINF Awards Committee

2016 Lucille M. Wert Scholarship: Call for Applications

Designed to help persons with an interest in the fields of chemistry and information to pursue graduate study in library, information, or computer science, the scholarship consists of a $1,500 honorarium. This scholarship is given annually by the Division of Chemical Information of the American Chemical Society.

The applicant must have a bachelor's degree with a major in chemistry or related disciplines (e.g., biochemistry or chemical informatics). The applicant must have been accepted (or be currently enrolled) into a graduate library, information, or computer science program in an accredited institution. Work experience in library, information or computer science is preferred.

The deadline to apply for the 2016 Lucille M. Wert Scholarship is February 1, 2016. Details on the application procedures can be found at:
http://www.acscinf.org/content/lucille-m-wert-student-scholarship.

Applications should be sent by email to: marge.matthews@outlook.com.

Marge Matthews, Coordinator, Lucille M. Wert Scholarship

2016 CINF Scholarship for Scientific Excellence: Call for Applications

The international scholarship program of the Division of Chemical Information (CINF) of the American Chemical Society (ACS) co-sponsored by InfoChem (www.infochem.de) and Springer (www.springer.com) is designed to reward students in chemical information and related sciences for scientific excellence and to foster their involvement in CINF.

Up to three scholarships valued at $1,000 each will be awarded at the 251st ACS National Meeting in San Diego, CA, March 13-17, 2016. Student applicants must be enrolled at a certified college or university; postdoctoral fellows are also invited to apply. The applicants will present a poster during the Welcoming Reception of the Division on Sunday evening at the national meeting. Additionally, they will have an option to show their posters at the Sci-Mix session on Monday night. Abstracts for the poster must be submitted through MAPS, the abstract submission system of ACS.

To apply, please inform the chair of the selection committee, Stuart Chalk, at schalk@unf.edu that you are applying for a scholarship. Submit your abstract at http://maps.acs.org using your ACS ID. If you do not have an ACS ID, follow the registration instructions. Submit your abstract in the CINF program in the session "CINF Scholarship for Scientific Excellence. Student Poster Competition." MAPS is now open and submissions are due by October 12, 2015. Additionally, please send a 2,000-word abstract describing the work to be presented to schalk@unf.edu by January 31, 2016. Any questions related to applying for one of the scholarships should be directed to the same e-mail address.

Winners will be chosen based on the content, presentation, and relevance of the poster, and their names will be announced during the Sunday reception. The content should reflect upon the student's work and describe research in the field of cheminformatics and related sciences.

Stuart Chalk, Coordinator, CINF Scholarship for Scientific Excellence

Technical Program

CINF Technical Program Highlights

The Boston 2015 National Meeting turned out to be one of CINF's biggest yet, and I daresay one of the most diverse. At 180 papers and 16 symposia, we ended up having to parallel track several sessions. This included programming a parallel session against the Herman Skolnik Award symposium, which we traditionally try to avoid. Fortunately the organizers did such an excellent job with talk quality that we drew substantial audiences for both sessions: they were standing room only (even after I stole chairs from other rooms). I can't thank all of the Boston organizers enough for all their hard work recruiting the array of speakers, drawing in cross-divisional participants and audience members. Bravo!

After an early morning symposium diving deep into the philosophy of handling data flow, this time around we brought back, on an experimental basis, CINFlash lightning talks to demonstrate workflow tools, presented in rapid succession. Feedback for the format was very positive so that the “lightning demos” is something we’ll probably see more of in the future. In the rapidly evolving cheminformatics world, keeping tabs of new tools is always a challenge. This tied in with the afternoon symposium on “Data Visualization to Guide Optimization,” which ranged from philosophical approaches on large scale data visualization (woe unto the pie chart!) to SAR tools seeking to present analysis of the chemical data itself. The symposium generated some passionate commentary from the packed crowd.

CINF was also very pleased to play a role in a careers panel breakfast, a Wikipedia chemistry edit-a-thon, and speed networking for undergraduate careers. We are focusing more and more on outreaching to other divisions as well as enhancing our role in continuing education. Chemical information is seeing increased relevancy across these fields, and I am delighted at the way the Boston National Meeting included this work.

Even “General Papers,” which often attracts the leanest crowd on Thursday, drew decent attendance well into the afternoon. This always makes me happy as it often includes students and early-stage career chemists looking to share their research with the rest of the world. This time around we had eight presentations, and some lively discussion on subjects ranging from the use of periodicals to tautomer handling in commercial screening samples. We are still in the process of collecting presentation slides from the meeting (which always takes a while), so stay tuned for their posting (if shared by speakers) on the CINF website.

CINF has been looking forward to the spring 2016 ACS National Meeting for some time: with a theme like “Computers in Chemistry” we are pretty excited about the cross-divisional opportunities. We’ve put out the call for papers covering a wide range of topics with 15 symposia, and many other co-sponsored ones. Highlights include several fine-tuned topics on large scale data management and meta-data, mining chemistry- and biology-based data across various data sources and types, and broad ethical conundrums in dealing with all these issues. We have several others off the mainstream, so be sure to check out the call for papers (link).

Thanks again for the large array of people who contributed to making the Boston program one of our biggest and best yet. See you in San Diego!

Erin Davis, Chair, CINF Program Committee

Substance Identifiers, Addressing the Challenges Presented by Chemically Modified Biologics: The Role of InChI & Related Technologies

This symposium was organized by Keith Taylor and Steve Heller and held on Sunday, August 16, 2015, at the ACS National Meeting in Boston, MA. Three papers were presented:

CINF 1. Generating canonical identifiers for glycoproteins and other chemically modified biopolymers by R. Sayle, J. May, N. O'Boyle, and E. Bolton
CINF 2. Towards addressing informatics challenges presented by antibody drug conjugates by S. C. Sukuru, T. Zhang, L. Tumey, E. Muszynska, M. Tran, and F. Loganzo
CINF 3. Representation of chemically modified proteins in the Substance Index SPL files by Y. Borodina, and G. Schadow.

These three papers dealt with three distinct phases in the development of chemically modified biologics (CMBs).

Chemically modified biologics are very important therapeutic agents; five of the top 10 drugs by sales value are chemically modified biologics. They bring unique opportunities and unique risks. It is important, as it is with all drugs, that they are identified uniquely and that associated data are made reliably accessible.

The first paper was presented by Roger Sayle (NextMove Software) and co-authored by John May, Noel O'Boyle, and Evan Bolton. In this paper Roger explored options that are currently available for identifying chemically modified biologics and discussed complementary approaches to biologics registration; one based upon expressive all-atom representations, another on tracking deltas to a reference database of protein sequences. The main characteristic of chemically modified biologics compared with conventional drug entities is their size; Humira, the top selling chronic hepatitus B (CHB) drug, consists of 1,330 amino acids and has a molecular weight of approximately 148 kDa. This alone challenges many conventional cheminformatics tools; InChI in its standard form is limited to 999 atoms and clearly cannot handle a typical CMB.

Chemical modifications can take many forms. They may be done deliberately by chemists in a laboratory, or passively during storage where methylation reactions occur. Roger gave an important example of the importance of tracking glycosylation in the biologic. Clinical trials for Erbitux (cetuximab) were conducted in California. Adverse reactions were observed in 1% of patients. Fortunately (in one respect) this percentage was high enough to require that the first dose be given in a controlled, clinical environment. Much higher numbers of reactions, and more serious ones, including anaphylactic shock, were recorded in a number of U.S. states, particularly in the South West. Many of these patients had been exposed to tick bites and developed an immune response that responded to the glycosylation pattern on Erbitux with serious consequences. This demonstrates the importance of tracking all aspects of a CMB's make-up.

Roger then discussed the existing technologies for identifying CMBs. Many, such as Hierarchical Editing Language for Macromolecules (HELM) and PDB, are based on standardized monomer dictionaries that then prove difficult to maintain. At first sight, the limited number of building blocks available in a natural biologic makes the task seem simple, but once chemical modification is included, and especially when skilled chemists get involved, the number of building blocks and sidechain modifications become limitless. Roger next raised the subject of canonicalization, the process by which many varied inputs are converted to the same unique representation.

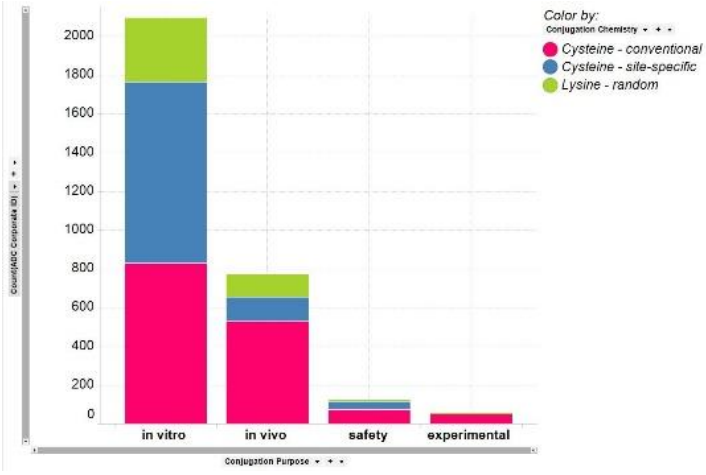
Canonicalization is important in the naming of entities. A simple example is ethanol which can be written as CH_3CH_2OH, or $HOCH_2CH_3$. A chemist would normally use the first format, but when 1,330 residues (and approximately 12,500 heavy atoms) are involved, there is scope for creativity.

In Roger's opinion, a chemical identifier should be independent of the input representation or file format, and there should be equivalence between small molecules, peptide and proteins, which are best determined by a single identifier, preferably the existing standard InChI. Currently, InChI has an atom limit, but Roger was able to use a modified version of the algorithm without the atom limit. He provided data from the peptide UTP10_KLULA. It has a sequence of 1,774 amino acids, with 28,509 atoms. An InChI and an InChI Key could be generated in 73.2 seconds. Roger then converted the sequence to a SMILES string and using OEChem's SMILES Canonicalization Time; the canonical SMILES was generated in 0.4 seconds. This demonstrates that there is much scope to improve the performance of the InChI algorithm.

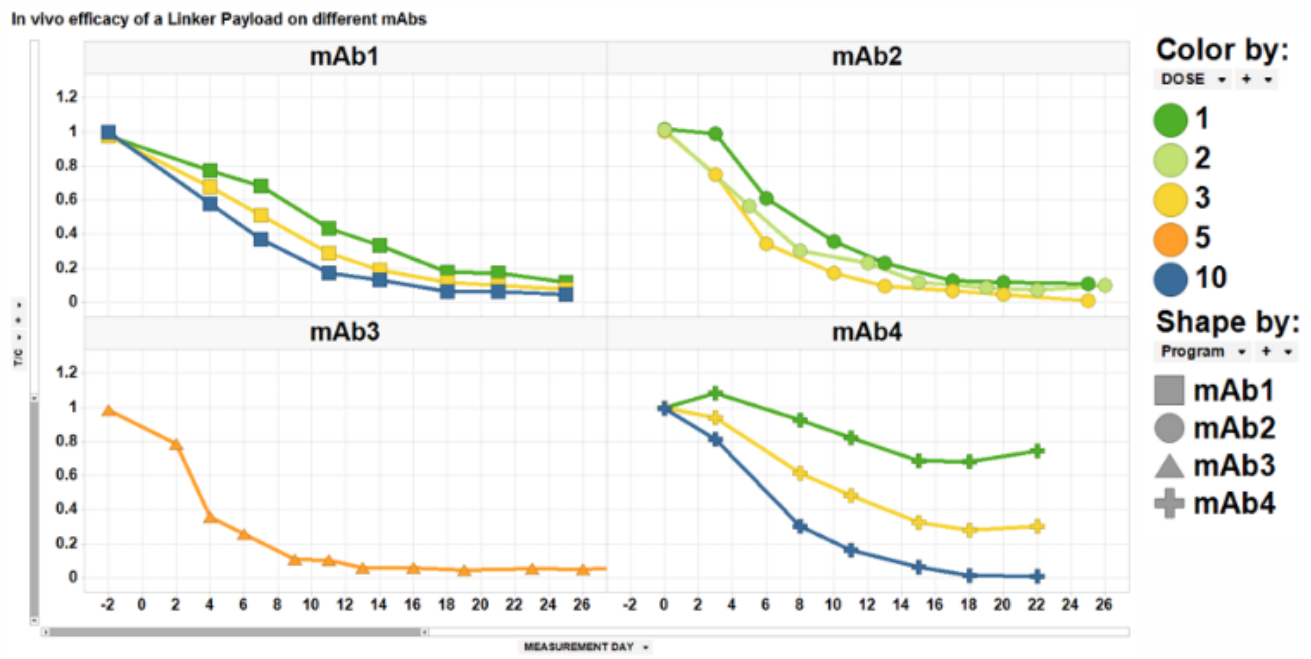
Roger then moved on to discussing a new database structure based on a Directed Acyclic Graph (DAG) for characterization and searching of biologics. This approach enabled the building of a DAG for all 540,546 protein sequences in uniprot_sprot, which contains over 192 million amino acids. This data structure allows close analogues to be identified much faster than using NCBI blastp. For example, all 540,546 sequences can be queried against this database (i.e., all-against-all) in about 9 minutes 30 seconds on a single core on a laptop and the sequence from PDB 1CRN (crambin 46AA) is canonically named as [L25I]P01542 in 0.002 seconds.

Roger concluded with the statement that "InChI for large molecules" can be achieved, and remain compatible with small molecule InChI identifiers, through the evolution of ever better canonicalization algorithms. In addition, he directed a jab at journal reviewers who claim that the run-time of canonicalization algorithms is a non-issue, and not an area ripe for improvement; these reviewers are very mistaken.

The second paper was presented by Chetan Sukaru (Pfizer), and co-authored by Tianhong Zhang, Lawrence Tumey, Elwira Muszynska, Megan Tran, and Frank Loganzo. The team has developed a novel *in silico* tool called Antibody Conjugate Tracker (ACT). ACT is designed to characterize each Antibody Drug Conjugates (ADC) efficiently, and its molecular components, namely the antibody, linker-payload and payload. The ACT provides a unique *in silico* environment with structured metadata that enables comprehensive data analytics on ADCs. Based on their experiences with the ACT, the authors proposed novel descriptors to parse and analyze Antibody Drug Conjugates data that could improve our understanding and accelerate the discovery of potential therapeutic ADCs. ADCs at Pfizer are given corporate IDs based on the parent antibody, the linker technology, and the drug payload.



Using the ACT data, it is trivial to visualize the distribution of ADCs quickly for *select* conjugation chemistries and purposes.



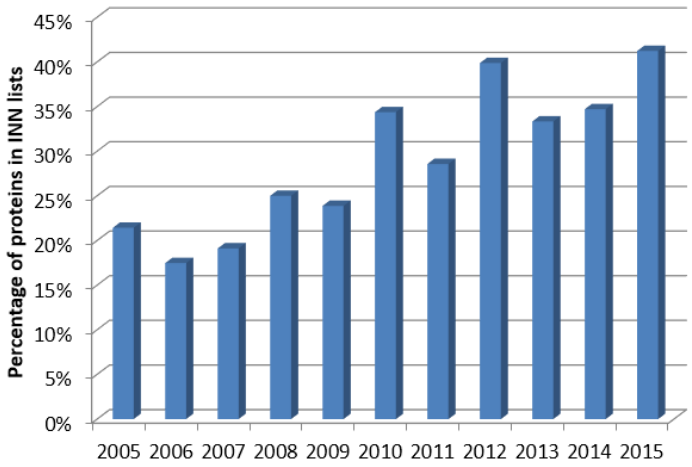
Structured metadata in ACT enables *in vivo* data comparison of a given linker-payload across different antibodies (program/antigen specific).

In the concluding remarks, Chetan stated that despite the rising interest in ADCs, there is still a gap in the informatics infrastructure/technology to support their discovery and development. With the ACT, Chetan and his colleagues have attempted to bridge the gap and address some of the informatics challenges presented by ADCs. The deliverable is that the structured metadata incorporated in the Antibody Conjugate Tracker not only keep track of all the ADCs, but also enhance *in silico* data analytics and visualization. Finally, a similar approach with standardized descriptors could help develop substance identifiers for other chemically modified biologics, too.

The final paper was presented by Yulia Borodina (FDA), co-authored by Gunther Schado, who described the ongoing work and challenges involved in incorporating CMBs into the FDA's Substance Index Structured Product Labeling (SPL) Files. Currently the FDA's Substance Registration System (SRS) contains information on 98,000 substances. The following entities are represented: small molecules, polymers, biopolymers, plant parts, tissue parts, vaccines, etc. The information (chemical structures, names, protein and nucleic acid sequence, taxonomic information) is highly curated. Each substance has a Unique Ingredient Identifier (UNII).

The SRS contains over 1,500 proteins (2% of all registered substances). Some are considered to be confidential, but over 1,100 are in the public domain and are targeted for public release.

Yulia provided an informative bar chart that showed the increasing role of proteins in marketed drugs.



The challenges that need to be addressed are: reliable electronic exchange of protein information, and the unique identification of protein substances. The following information has to be captured for each CMB:

- Amino acid sequences of chains
- Covalent connections between/within chains
- Modifications of natural amino acids
- If amino acid is modified by a synthetic polymer, structure and characteristics of that polymer
- Sites and type of glycosylation
- Structure of glycan
- The frequency of modification (which may be an average)
- Co-factor enzyme interactions.

Dinutuximab and dinutuximab *beta* differ only by their glycan composition, as do erythropoietin (EPO) *alpha*, *beta*, *delta* and *omega*.

The information will be published in the Substance Index SPL files as an XML document that is Health Level 7 compliant. It has been adopted by the FDA as a mechanism for exchanging product information electronically and it has also been adopted by ISO 11238 as an exchange standard for medicinal substances. The information is available through FDA Online Label Repository and DailyMed. It is up to date (new information or changes are added daily), and it is free.

Yulia then described the XML markup that is being used for CMBs. Currently, about 800 files have been generated for proteins that do not have polymeric modifications. After review, these files will be posted on the SPL website: http://www.fda.gov/ForIndustry/DataStandards/StructuredProductLabeling/ucm377913.htm

An SPL Implementation Guide with Validation Procedures is under update and will be made available at: http://www.fda.gov/ForIndustry/DataStandards/StructuredProductLabeling/default.htm. Further work is underway to update the SPL model to handle synthetic polymers and protein-polymeric conjugates.

Conclusion

The three papers covered many of the challenges of handling CMBs from the fundamental cheminformatics, through managing them in the laboratory, to finally registering and publishing the information with the FDA. The session was well attended, especially considering that it started on Sunday at 8:30am. The audience was somewhat variable, but we had an average of 20-25 attendees.

Keith Taylor, Symposium Organizer

Careers in Chemical Information and Cheminformatics Panel Discussion & Brunch

While enjoying a delicious breakfast off the beaten path in the Boston Convention and Exhibition Center on August 16th, we met professionals who pursue careers in cheminformatics and related fields. Lori Betsock from the ACS Undergraduate Programs Office shared a few words about the networking events for undergraduate students and resources on the College to Career website (http://www.acs.org/content/acs/en/careers/college-to-career.html). The moderator, Rachelle Bienstock, introduced the panelists, who were: Sean Ekins, Kevin Theisen, Christopher Lipinski, Carmen Nitsche, Rajarshi Guha, Erja Kajosalo, and Thomas Marman. They told us their career stories and gave some advice. Two key themes ran through all of their talks: 1) collaborate and network, and 2) choose work that you enjoy and for which have a passion.

Sean Ekins led off, describing himself as a "Serial Collaborator." He has worked for over twenty years in cheminformatics on rare disease drug discovery, ADME/Tox and transport modeling, developing mobile apps for chemistry, and spreading the word. His history includes spending time on QSAR, PharmaForce and the "rule of five." He uses computational approaches, working with chemical structures, "fishing" through chemical libraries, and always looking for collaborators with data. Sean is especially interested in neglected and rare diseases such as Ebola, tuberculosis, Sanfilippo Syndrome, and Chagas disease. He runs his own company, consults for others (mainly writing grants), and collaborates with pharmaceutical companies in order to scale his work for more diseases.

Sean started out with an applied biology degree from Nottingham Polytechnic (UK) and then went into pharmacology at Arberdeen University. He had a postdoc at Eli Lilly, then worked at Pfizer and then went back to Eli Lilly. Next he worked for a startup company, Concurrent Pharmaceuticals, and after that started writing grants for small businesses at GeneGo. Over time, he observed a trend of working with smaller and smaller companies: they move quickly and have lots of ideas. His expertise is in collaborating, not programming or chemistry. Sean advises us *to learn how to collaborate and find good collaborators*. *Don't be afraid to connect with people*. *You may get ignored, but many would like to team up* (in one case he shared a room at a conference and met a collaborator). *Also important, learn how to publish and talk, as a way to give back and share openly*. *In the future, the datasets will grow and we'll need new algorithms, data visualization, and mining approaches*. (Sean's slides)

Kevin Theisen is a software developer and the founder of iChem Labs (http://www.ichemlabs.com/), a successful, small software company. His specialty is visualization tools and graphics for communicating, interpreting, and interacting with chemical information. His company has developed multiple versions of the ChemDoodle software with evolving graphics, mobile software, and 3D. Recently, they released the BioTuple software for bioinformatics. Kevin started as an undergraduate student interested in NMR simulation. His education, a B.A. at Rutgers and an M.S. at Berkeley, was useful for honing skills. He discovered that it was fine to give up a Ph.D. and pursue other goals. Some of his success was related to passion and timing (the 2007 Apple smart phone with HTML 5 came out at a fortunate time for him). His parents were programmers who made him take classes. He found a way to apply that learning in a way that appealed to him. Kevin's advice is *to give means, opportunity, encouragement, and investment into people who have passion. Opinions matter even if people discourage you*.

Chris Lipinski, the author of the "rule of five," has a rich history working for thirty-two years at Pfizer. He continues to consult and do science in retirement, publish articles, give presentations at conferences, and travel a lot (though less now than before). His undergraduate degree is from San Francisco State College and his Ph.D. in physical organic chemistry from Berkeley. Chris wanted to

work with pharmaceuticals. He worked at Caltech doing total synthesis. His background in physical organic chemistry caused him to think differently from other colleagues. When exploring inactive and active compounds, he asked what the structure was that caused this activity level. He focused on ADME (absorption, distribution, metabolism, and excretion). Chris left medicinal chemistry at the top: he had a brand new lab with robotics, computers, electronic systems, and materials science. His advice is *to look at what you enjoy doing and to find out what makes you happy. Also learn how to keep track of literature* (Chris had PDF files everywhere and ChemWorx helped). *You need to nurture your career and build your resume now. Networking is key. Know more about people who work outside of your group! Recruiters appreciate this*.

Carmen Nitsche works in pharmaceutical consulting. She was not officially educated and trained to do this work, so how did she get there? Her degrees in chemistry are a B.S. from the University of Minnesota and an M.S. from the University of California at Berkeley. After graduating, she began working in labs at Atlantic Richfield and then at Los Alamos National Laboratory. Her path changed after she met a librarian and learned Dialog. She enjoyed the new way of searching *Chemical Abstracts*. Carmen joined Nalco Chemical Company (Illinois) working in a library and information services group. This became an eye opening career path. She had been doing information research by hand and then CAS came out with SciFinder, an end-user search tool. Next, she moved into sales, and then into business development at MDL, Symyx, and Accelrys, where she became Vice President, gaining experience with intercompany partnerships, mergers, and acquisitions.

Most recently, Carmen started a consultancy for many organizations at Pistoia Alliance, a not-for-profit members' organization committed to lowering the barriers to innovation in life sciences R&D, by improving the interoperability of R&D business processes through precompetitive collaboration (http://www.pistoiaalliance.org). Her observation: *Basic work doesn't need to be duplicated, especially regarding neglected diseases. In the online world, there is an intense need for information management, text-mining, and big data analysis. It is important to network and stay in touch with what is going on. Actively look for mentors who have an interest in you and offer a reality check*.

Rajarshi Guha works at the National Institutes of Health (NIH) in the National Center for Advancing Translational Science (NCATS) where they support high-throughput screening platforms and methodology to enable better screening. They construct libraries, and predict biological activities and physical properties. This highly collaborative group switches from small molecules to pathways to data mining images to building web apps. They write project proposals and do screening with biologists and chemists.

Rajarshi has a Ph.D. in chemistry from Pennsylvania State University and did a postdoc at Indiana University with David Wild. He ultimately filled a job opening in the screening center of NIH. Rajarshi appreciates an advisor who enables independence and understands his multiple interests. He can experiment with software, hardware, open source code, blogging, activities with the ACS divisions (through CINF he meets everyone in cheminformatics), and writing a short paper. He learns quickly on the job, uses his strong programming skills, and grasps data mining. There are many opportunities available at the NIH that are not available at "traditional" pharmaceutical companies. His advice: *Learn to code, work with others, and learn science. Focus on how to solve the problem, not how to write code. Current challenges are data size issues, complexity issues, and how to automate. If you are interested in working with him, NCATS has internships.* See https://ncats.nih.gov/ncgc/work.

Erja Kajosalo has worked in three countries and has had three careers. She is currently the Chemical Information Librarian at Massachusetts Institute of Technology (MIT) where she does reference work,

instruction, and collection management. Erja decides what to buy for the library, reviews annual subscriptions, and justifies funding for large databases like SciFinder and Reaxys. Her education started in Finland with a master's degree in chemistry. She fell in love and moved to Canada where her master's degree was not accepted, so she became a laboratory technician. After discovering a passion for computers, Erja became a programmer for almost five years, but this job wasn't a good match. At the University of Alberta she studied information and learned that combining science with librarianship was very useful. This led to her current position at MIT in Boston where she was hired to work with the chemistry department. Now she follows cutting edge chemistry closely as a way to keep in touch. She can apply past learnings by creating web applications and selecting a library management system.

Libraries continue to be user-centered, but the information format and services are changing to online. Librarians don't see users as often. The historic materials are important, leading libraries to digitize and store information for the long term. While books are still useful, they are no longer getting as much use, so library spaces are changing into visualization labs and collaborative places to learn about geographic information systems (GIS) and bioinformatics. Her advice: *If you want to get into this field, a science background is helpful. Whatever you do, when you make choices, look beyond what you have done before. Because of constant change, it is important to learn new things continually and do new projects. Keep up with cutting edge chemistry by following where faculty are researching, what journals are important, and what new journals are coming out. Faculty research interests change over time and it is important to look beyond your liaison departments.*

Thomas Marman works in the Pfizer patent department on the Global Legal Information Science Team. Patents are used to protect the company's interests and can exclude others from making or selling similar products for twenty years. After the expiration, anyone can make the products. Thomas likes collecting background information for patent attorneys by searching chemistry and biology databases to find anything that has been published prior to filing of a patent. In addition to finding written materials, he also looks for what people have said.

Thomas started out with a Ph.D. from Texas A&M University, specializing in organic chemistry. His organic synthesis postdoc was at ICSN-CNRS in Gif-sur-Yvette, France. Then he worked at a small company called ANGUS Chemical on process development. This company was purchased by Dow Chemical where he worked in the business and technical services group as a technical specialist. Next, Thomas transferred to the Pfizer patent department where he learned more about patents and patent law. His observation: *Generally patent law is learned on the job. Law is different from science because it is very fuzzy. You look at a problem from many different perspectives. Strong searching expertise helps find hidden information. A strong technical background helps a person understand patent claims and communicate with others at various expertise levels in engineering, medicine, or materials. Know people's strengths in order to support your projects.*

Susan Cardinal, Panel Organizer

Lisa Balbes, CINF Careers Chair 2006-08, and author of "Nontraditional Careers for Chemists: New Formulas in Chemistry" (2006), is the winner of the Howard & Sally Peters Award given by the ACS Division of Chemistry & the Law in recognizing achievements addressing nontraditional careers for chemists. *Chemical & Engineering News,* September 7, 2015.

Wikipedia and Chemistry: Collaborations in Science and Education

When this symposium was proposed, we were unsure what response to expect. As it turned out, we were delighted with the quality and the variety of presentations. The topics ranged from descriptions of Wikipedia itself, to use of the site for cheminformatics, through to use in teaching information literacy. The themes chosen were collaborations and education, which both fit naturally with the mission of Wikipedia. The session was cross-listed by the Division of Chemical Education.

Elsa Alvaro (Northwestern University) began by examining Wikipedia from the librarian's perspective; how is the chemistry content organized and categorized, how has the quality and size grown, who is writing it, and how is it used and cited? Wikipedia is beginning to be used and cited in the chemical literature. Using data extracted from the Scopus database, Elsa found that the number of research articles citing Wikipedia entries in chemistry journals had grown to over 300 articles per year by 2014. The number of English Wikipedia pages under the chemistry category and subcategories has gone up to over 20,000. Something listed under the "chemistry" top-level category may end up being quite unrelated to our field, making it hard to use categories for such classifications. The Joker (comics) page is a good example of that. Chemistry pages also overlap with other disciplines such as physics, biology, and earth sciences. Over 678,000 editors (of whom about 150,000 are registered) contributed to these Wikipedia chemistry pages, while 8% of the registered editors account for 80% of the revisions.

Martin Walker (SUNY Potsdam) described how collaborations (WikiProjects) work within Wikipedia to coordinate editing and standards. He then went on to give a history of collaborations among the English language Wikipedia chemists and other chemistry resources such as ChemSpider and CAS. Martin also described new initiatives such as WIKIDATA for more stable data validation and informal collaboration with IUPAC for validating definitions. Future collaborations among chemists and Wikipedians rely on mutual trust and common goals in making Wikipedia "a richer and more reliable source for chemical information."

Guido Herrmann (Thieme) continued the theme of collaboration by showing the mutual benefits from a relationship between the German language Wikipedia chemists and the well-respected RÖMPP German language encyclopedia published by Thieme. RÖMPP is able to benefit from the mainstream exposure in Wikipedia, while Wikipedia is able to receive expert guidance and use of the valuable RÖMPP collection. The concise and expert-curated content in RÖMPP, and the more detailed and comprehensive articles in Wikipedia complement each other and are interlinked, which allow users to choose the preferred format. The collaboration between the two editorial teams was mutually beneficial.

Jian Zhang (PubChem) showed how PubChem has been working with the Wikipedia and Wikidata communities to improve links between the sites. PubChem Compound Summary pages now all include Wikipedia links. Although much has been done, some errors remain, and PubChem plans to collaborate further to ensure that links from Wikipedia lead to the correct records in PubChem. PubChem's data provenance information can add value to Wikipedia. In addition, those data and information held by the two sites in common could be used in validation; or the discrepancies in details could be annotated for further evaluation. The PubChem team is planning to annotate the information from Wikipedia API further, to correct and validate PubChem compound identifiers (CIDs) in Wikipedia, and also to get new chemical and drug records from PubChem into Wikipedia.

Roger Sayle (NextMove Software) demonstrated the value of data found within Wikipedia. As might be expected, Wikipedia can be used to provide glossary terms; its value comes from the large number of synonyms found in redirects, which often include “street” names and vernacular terms which may be absent from more formal resources. This is particularly seen in medicine, and drug terms in particular, where connections are only possible through the use of Wikipedia. A collaboration between NextMove and AstraZeneca demonstrated this advantage and also used the linking in Chembox to PubChem CID for SMILES retrieval. Another example used the cross references to the International Classification of Diseases (ICD) in the Diseasebox. More examples include providing multilingual support, named reaction, and parts of speech. In all, the use of boxes, categories, templates, and redirects provided by Wikipedia and Wikitionary supplements the traditional lexicons and ontologies in cheminformatics research.

The second half of the symposium emphasized the educational value of Wikipedia. Adam Hyland (Wiki Education Foundation) described the work of his organization in working with educators to produce viable classroom projects, where Wikipedia articles are edited and even created by students. Students benefit from seeing their work having a real-world application, and they quickly appreciate the importance of copyright and the citation of reliable sources. They also learn to work together and see the value of constructive criticism in improving their writing. Adam further introduced the support Wiki Ed could provide for instructors and students who use Wikipedia editing in their classes.

Ye Li (University of Michigan) then showed specific examples of such classroom teaching, which has long been part of the UM chemistry curriculum. The interaction of students and instructors with the Wikipedia community turns out to be vital, and can be very positive for students when handled well. The editing of Wikipedia trains students in a wide range of information literacy skills, and they align very well with the standards and guidelines set out by the Association of College and Research Libraries (ACRL). Meanwhile, there are also significant gaps between the controlled learning environment and the open and diverse Wikipedia community, such as semester-based cycle versus long-term commitment, grading needs versus bit-by-bit collaborative editing style, and academic value versus neutral point of view. To avoid frustrations caused by these gaps, frequent communication, careful students’ training plans, transparency of class project progress, and working with the Wiki Ed Foundation are crucial.

Keith Lindblom (ACS Office of Public Affairs) described how his office has worked with Wikipedians to engage chemists more in editing Wikipedia articles and contributing to this important resource. This fits well with ACS’s mission to promote chemistry among the general public, who frequently use Wikipedia to learn about scientists and their ideas. The effort is a part of the Chemistry Ambassadors program. ACS has organized several edit-a-thons and was organizing one during the ACS Boston conference. Speaking and writing simply are the keys to reaching a broader audience as chemistry ambassadors.

Antony Williams (U.S. Environmental Protection Agency) showed the value of the MediaWiki platform. He began by showing the rich data within Wikipedia, and the WikiProjects that support the work. He went on to explain how the collaborative environment of a wiki is able to produce a variety of other sites such as VIPEr (educational materials for inorganic chemistry), the University of California, Davis ChemWiki (virtual textbook) and the CINF wikibook (cheminformatics education). Open data within Wikipedia can be harvested to produce sites such as ScientistsDB, SciMobileApps, and SciDBs. The LearnChemistry wiki shows the flexibility of MediaWiki in education. Chemicalprobes.org (for probing biological activity) and Adverse Outcome Pathway (AOP) wiki (for adverse effects) help find connections between chemicals and diseases. Antony emphasized that MediaWiki platforms as a

proven concept have demonstrated their value in leveraging scientists' contributions and enhancing collaborations.

Chiara Ceci of the Royal Society of Chemistry (RSC) described the work of Andy Mabbett, the RSC's "Wikipedian in Residence" in connecting British chemists with the wiki communities. Wikipedia edit-a-thons have been organized at RSC sites, and Wikipedia editors have been provided with free access to RSC journals. Like ACS, RSC sees the value of Wikipedia in promoting chemistry among the general public, and Mabbett was able to help with this in a variety of venues. One of RSC's surveys showed that 48% of people find Wikipedia as a trustworthy source of information on chemicals and chemistry in everyday life (with 27% finding untrustworthy and 25% not knowing). Chiara also demonstrated the multimedia contributions, including images and voices, and international collaborations, through their effort. Much of their work can be found at Wikipedia: GLAM/Royal Society of Chemistry.

The session ended with a lively panel discussion, with questions on many different topics, ranging from enriching Wikipedia itself to copyright issues and instruction needs. A good time was had by all!

Three days later, a group of conference attendees and Wikipedians gathered for an edit-a-thon, a coordinated effort to create and edit Wikipedia articles. The session began with a tutorial by John Sadowski (John's ACS Member Spotlight) on editing Wikipedia articles, including an explanation of the main site policies. About thirty attendees (and a few remote participants) then pitched in and worked on biographical articles on "notable chemists," and more than a dozen of these were started or expanded, aided by a collection of useful books provided by ACS for the occasion.

Martin Walker and Ye Li, Symposium Organizers

C&EN
CHEMICAL & ENGINEERING NEWS

Working with Wikipedia
Volume 93 Issue 36 | pp. 36-37
Issue Date: September 14, 2015
http://cen.acs.org/articles/93/i36/Working-Wikipedia.html

Retrosynthesis, Synthesis Planning, Reaction Prediction:

When Will Computers Meet the Needs of the Synthetic Chemist?

On a hot and humid, yet sunny, Monday in Boston we were treated to a *tour de force* of the current thoughts on how the machine may aid (or even eclipse) the synthetic chemist. This full-day symposium covered the full spectrum of past, current and hopefully future masters of the art. A wide variety of papers was given, each taking a slightly different view on the topic.

Juergen Swienty-Busch (Elsevier) spoke about the next steps of your synthesis. He described the problem related to searching for substances, and for reactions, and the issues with atom mapping. He described Reaxys' approach to reaction similarity, and reaction classification. He went on to describe how the Reaxys taxonomy was developed and how this is applied to Ask Reaxys. Finally, he pulled all of this together to describe how Reaxys uses all of these underpinning technologies in order to solve the synthetic chemist's problems.

Peter Johnson (University of Leeds) described some new advances from Leeds, in conjunction with the Chem21 IMI project, aimed at developing new methods for addressing key bottlenecks in synthetic processes. He is particularly involved with work package 5: assessing greenness of chemistry. He discussed the work that has gone into the creation and development of the Chem21 Reaction database. Synthetic chemists are working on data entry, appropriate atom mapping, and, crucially, entry of all important synthetic data. The system proceeds to provide a green score for the reaction. Peter ended by describing work for a reaction database looking at biocatalysed reactions.

Marc Nicklaus (National Institutes of Health) began by asking the question that defines this work: "What can I make reliably and cheaply?" After answering that question, you can search for those compounds that should be good for testing. To answer the first question Marc suggests we need good rules, and predictions, and available and inexpensive materials. Marc's team needed to create forward synthetic routes, and they have examined some 1,500 transforms from LHASA, and looked at the robustness of the transforms (yield, reliability, thermodynamics, etc.), resulting in (so far) 13 transforms. They have also gathered building block information from Sigma Aldrich, looking for materials with high availability, low costs, etc. The initial work has resulted in a significant number of new compounds (with a low overlap with PubChem). Marc expects that building this work up to cover all 1,500 transforms from LHASA, and all Sigma Aldrich's 3 million building blocks, will result in an enormous number of new, potentially interesting compounds.

Roger Sayle (NextMove Software) has worked with a number of large pharmaceutical companies, and gave us some insights into the gold mine of information contained in these companies' electronic lab notebooks (ELN). A significant number of reactions fail, and by and large, these reactions never make it into the published (journal or patent) literature. Roger gave the example of a 50 by 50 library made by GSK (GlaxoSmithKline): 566 compounds were not synthesized, and assay results were reported for only 1,706, so, despite the use of reliable, predictable chemistry, a number of compounds were not made, and did not deliver results. Roger went on to describe an example of computational forensic chemistry whereby his team had examined over 11 million patents and extracted approximately 2 million reactions. He noted that the sixth most common reaction was the transformation of NO_2 to NH_2, but that the introduction of NO_2 was very rare. Roger concluded that NO_2 is delivered into most compounds via a building block. In 1990, Suzuki coupling appeared in almost no patents, but by 2013 it accounted for 7% of reactions.

Jonathan Goodman (University of Cambridge) delivered his usual amusing and animated insight into synthesis questions. While aiming for a world in which computers and machines could happily replace the bench chemist, he concluded that such a world is a long way off. He stressed that (fortunately) there are lots of complex pieces and parts required that make up a chemist. He described the various data that the chemist might take from systems to help in the decision process for synthesis design (literature analysis, spectral analysis, model design, property prediction etc.). His group has recently published an *in silico* inspired synthesis of Dolabriferol (2012). He concluded by discussing the future needs of a synthetic chemist, and how we might go about designing synthesis systems and machines. He noted that there are many obstacles: the current start of the art is uncertain, each step is uncertain (purification, selectivity, etc.), even reliable reactions can fail, and, ultimately, the question of whether a reaction is intended for discovery, process, or "just" publication. Jonathan felt that a successful retrosynthesis tool would need many things to determine success, including data on performance; use in industry and publishing; experimental verification of new examples; and discussion on the generality of reactions.

Brian Masek (Certara Inc.) introduced a problem in which a simple analysis of compound space for a set of 80 generic reactions, and a database with 1,000 reactants per reaction class, and schemes of 5 steps generates a space of at least 3×10^{27}, so the problem becomes one of what is interesting and what can be made. Brian then described the process Certara has developed to perform *de novo* design, and then perform the retrosynthetic analysis. The system is "biased" towards pharmaceutical chemistry, and hence typically involves short(er) routes. Certara has focused on high probability, generic reactions. Brian examined a set of predicted analogues of Abilify and their predicted syntheses. The best predicted routes and several published routes were given in a blind test to practicing synthetic chemists. The published routes scored better (8.2 versus 7.5 out of 10) but the results give rise to hope.

John Figueras (retired) described the SynTree application he has developed. This works on a simple MacBook and can be downloaded from his website (http://www.jfigueras.com/COMPUTER%20Progs/Chemistry.html). The program is intuitive and seemingly gives the chemist complete control, by enabling selection of precursors at each branch of the tree. The system comprises various modules which together enable the synthetic analysis. There are two main modules. One, handling transforms, currently contains 240 transforms, and a variety of different atom types to enable appropriate mapping and definition. The tools are basic operations that add or remove atoms, change the order of bonds, and alter the hydrogen count at an atom. The second module is IPLists, a list of interference groups, that is, those groups that should not be allowed in a particular reaction. The program was demonstrated to show its ease of use.

Orr Ravitz (John Wiley & Sons) gave a good overview of the field. He said that systems should be productive, efficient, and creative, and help identify opportunities. He outlined a number of the key terms and nomenclature around retrosynthesis. He reviewed a number of initiatives that came before Automated Reasoning in Chemistry (ARChem), and then proceeded to describe the ARChem system in detail, including the generation and curation of rules, and the use of differing reaction databases to provide background and credibility to each prediction. He ended by announcing the launch of a new service from John Wiley & Sons which is underpinned by the ARChem software.

Valentina Eigner-Pitto (InfoChem GmbH) gave an overview of the development of various tools from InfoChem. She thanked the previous speaker who had covered a number of the key definitions already. She described some of the work with which InfoChem has been involved, especially with the process chemistry department at AstraZeneca. The main tools she described were IC_{SYNTH} and IC_{FRP}

and the SPRESI database. IC_{SYNTH} is a retrosynthetic planning tool, while IC_{FRP} is a forward reaction planner. The advantage of InfoChem's approach is the automatic generation of transform libraries from any reaction database, using IC_{MAP} and CLASSIFY. SPRESI data have been an advantage over time and have helped to develop and optimize InfoChem's algorithms using big quantities of data.

Bartosz Grzybowski (Ulsan National Institute of Science and Technology) discussed the evolution of his Chematica system. He takes the analogy of the expert chess system with six different pieces, and on average 10 rules to define how the pieces move. The rules for organic synthesis are many, many times more complex. Chematica currently has defined in excess of 20,000 rules. Using similar logic from chess computers, the system does not follow a single path, but rather checks back to see if it follows an earlier "less satisfactory" path, so it may get to a superior position that necessarily always follows the best path at each stage. Bartosz followed this up with a video demonstration of the Chematica software.

Alexandre Varnek (University of Strasbourg) started by explaining why chemical reactions are difficult objects (many species; two types, namely reactants and products; multi-steps; dependence upon reaction conditions, etc.). He described his team's work on the condensed graph of a reaction (CGR). CGR may be used in a variety of ways including reaction searching, reaction data curation, reaction classification, analysis and visualization of reaction data, predictive models for reaction conditions, and models for kinetic and thermodynamic properties. Alexandre proceeded to show examples for each. Atom-atom mapping and quality of the underlying data were flagged as bottlenecks in the CGR generation process. Alexandre then described approaches to structure-reactive modelling, with many of the data coming from PhD and habilitation theses. He finished by showing the web page (http://infochim.u-strasbg.fr/webserv/VSEngine.html) where a number of his tools may be accessed.

Timur Madzhidov (Kazan Federal University) has been studying protecting group chemistry. He referred to Greene's *Protective Groups in Organic Synthesis*, noting that the reactivity charts result from manual analysis of relatively small datasets. He proceeded to describe how he and his colleagues have analyzed approximately 142,000 reactions from Reaxys using the condensed graph of reaction approach. Their initial analysis showed some disagreements with Greene's standard text. Timur finished by describing a prototype expert system to provide synthetic chemists with detailed recommendations of experimental conditions, in order to achieve the desired transformation.

Lee-peng Wang (University of California, Davis) described his *ab initio* nano-reactor. He described the problem associated with trying to describe events which occur infrequently on the time scales used during the calculations. In an effort to force the system, his algorithms squeeze the system hence forcing the temperature up, increasing the likelihood of reactions occurring. He corrects the pathway information with periodic minimizations. He finally described the application of the nudged elastic band method.

Acknowledgment

Many thanks to my co-chair Wendy Warr (Wendy Warr & Associates) for her great assistance in the preparation of this fine symposium, encouraging submissions, and working with me through the pains of new MAPS abstract submission system.

David Evans, Symposium Organizer

Conflict of Interest: David Evans is an employee of Reed Elsevier Properties SA, a member of the RELX Group.
All comments herein are David's own and do not necessarily reflect the views of the RELX Group.

Enabling Machines to "Read" the Chemical Literature:

Techniques, Case Studies and Opportunities

This symposium covered many themes: text-mining for chemicals, genes and proteins; relating chemical entities to ontologies; extraction of chemical properties (especially from tables), and their association with compounds; interpreting structures in the CHEMKIN format; and chemical image to structure conversion. Broadly, the symposium was organized such as to flow from text-mining from patents, to general text-mining, and finally to mining chemical structure information from images.

Obdulia Rabal (University of Navarra) gave a talk entitled "CHEMDNER-Patents: automatic recognition of chemical and biological entities in patents." She described the upcoming CHEMDNER-Patents challenge, which consists of three tasks for participants' systems to attempt: chemical entity recognition, chemical passage detection, and gene and protein recognition. For the challenge a corpus of 21,000 patent abstracts was assembled from various patent offices (WIPO, EPO, USPTO, CIPO, DPMA, SIPO); these abstracts were then manually annotated for chemicals, genes and proteins. The entities in the corpus are further classified into seven classes for chemicals and eight for genes and proteins, indicating the type of mention, for example, FAMILY for a family of compounds. The results will be available in the proceedings of the upcoming BioCreative V workshop. The corpus and more information are available from: http://www.biocreative.org.

George Papadatos (European Bioinformatics Institute, EBI) gave a talk entitled "SureChEMBL: an open patent chemistry resource." He gave an overview of SureChEMBL's functionality. Through a collaboration with Open PHACTS, SciBite is now providing biochemical annotations, for example, for genes, proteins, and disease. Currently these annotations are available on-demand, but it is hoped that they can be integrated into their database and made available via the Open PHACTS API later this year. Currently, about 80,000 novel compounds are being added to SureChEMBL each month. George finished by giving an overview of EBI's work, using the RDKit, to analyze the chemical space of a patent to identify the key compounds. More details on this are available here. SureChEMBL is available from www.surechembl.org.

Christopher Southan (University of Edinburgh) gave a talk entitled "Deuterogate: causes and consequences of automated extraction of patent-specified virtual deuterated drugs feeding into PubChem." Chris talked about the issues raised by the large number of deuterated compounds being added to PubChem, which are primarily being extracted from the USPTO patent Complex Work Units (CWUs). ChemDraw files are associated with each image in a patent. Many of these are simply deuterated versions of existing drugs and it is highly unlikely that the vast majority of these have been synthesized, meaning that they present a growing source of virtual compounds. While there was a surge in deuterated compounds disclosure five years ago, the number of compounds being disclosed in recent years has declined, while still remaining significant.

Lutz Weber (OntoChem) gave a talk entitled "Evaluating U.S. patent full-text documents with chemical ontologies." Lutz described OntoChem's UIMA-based OCMiner pipeline for document annotation. The system supports a wide variety of entity types, for example, chemistry, proteins, anatomy, species, diseases, and cell lines. Lutz presented benchmarks on the ChEBI patent set, showing high precision (96%) on long names, but he cautioned about the use of the corpus in general, as for shorter entities, 65% of the system's false positives were omissions in the corpus rather than true false positives. Lutz presented advances in OCMiner's formula detector. He then described some of the uses of chemical ontologies, both structure- and usage-based, for example, knowing that a compound is an anti-infective implies that it is an antibacterial or an antiviral. Looking specifically at the challenges of

structure-based ontology classification, he presented a comparison between OntoChem's SODIAC system and the similar software Classyfire (from the University of Alberta), highlighting some of the challenges in harmonizing ontology representation. Finally Lutz presented some use cases for chemical ontologies, for example, homonym resolution, document classification, and anaphora resolution.

Valery Tkachenko (Royal Society of Chemistry, RSC) gave a talk entitled "Text-mining to produce large chemistry datasets for community access." He covered the RSC's recent collaboration with NextMove Software on text-mining melting points and NMR from the U.S. patent literature. Over 200,000 melting points were extracted and used to build a model to predict melting points. This model's errors were comparable to observed experimental error in the patent data. The predictions were also more precise than a previous model built using smaller, more curated data sets. Valery demonstrated trends in 1H-NMR frequency over time and the number of NMR spectra extracted. Finally, he described the RSC's upcoming experimental data repository, which will have specific support for various different experimental properties, for example, chemical reactions, measured properties (melting points, log*P*, etc.), and spectra.

Richard West (Northeastern University) gave a talk entitled "Identifying chemical species in combustion models." He talked about the challenges of interpreting the CHEMKIN format, which is prevalent in combustion research. The format is fixed width and relies on nicknames to identify the chemical species involved. Unfortunately these nicknames are frequently ambiguous, and different nicknames may be used in different models to describe the same species. Richard has worked on a system for deducing the intended structure of the chemical species from the nickname. This is done by solving the constraints on what the compound can be, based on the reactions it takes part in (he used the analogy of solving a Sudoku puzzle). By using this technique Richard has been able to identify cases where the same species has erroneously been included in a model twice, and examples where the same species has vastly different predicted combustion energies in different models.

Tong-Ying (Tony) Wu (Linguamatics) gave a talk entitled "Text mining the chemical literature to find chemicals in context." Tony talked about using Linguamatics I2E to extract data from tables. This included association of compounds numbers with compounds, and resolving compounds referenced by compound number in tables. More tricky cases (e.g., resolving the meaning of "+++" when used as a measure of activity) are also supported. Linguamatics have identified patterns allowing the high precision identification of chemical compounds and are planning to use them to construct annotated corpora from English and Chinese text.

Daniel Lowe (NextMove Software) gave a talk entitled "Unlocking chemical information from tables and legacy articles." He talked about how he uses grammars to describe chemical properties, such as melting points, and NMR spectra. These grammars are then used to generate a state machine to recognize the property efficiently, and a multi-state machine parser to parse the properties efficiently. He talked about some of the challenges of table extraction from U.S. patents. For example, the XML provided by the USPTO describes the appearance of the tables rather than the semantics, so a multi-line row is presented as multiple rows requiring heuristics to give the intended semantics. Finally, Daniel talked about promising work on extracting melting points and NMR from post-2,000 Royal Society of Chemistry journal articles and the difficulties of extracting data from the older articles, for example, headings and paragraphs need to be perceived from PDF, with OCR errors (especially in important compounds and important symbols like the degree symbol).

Igor Filippov (VIF Innovations) gave a talk entitled “Chemical structure identification and retrieval with OSRA.” OSRA is a tool allowing the conversion of chemical structure diagrams to computer readable structure formats. Igor discussed the segmentation procedure used to detect which parts of an image contain a chemical diagram. The precision of this process has been significantly improved in the most recent version (v2.01) with minimal loss of recall. OSRA, notably, also supports the recognition of chemical reactions from diagrams. Future improvements to OSRA will be driven by feedback from the user community.

Bryn Reinstadler (IBM Almaden) gave a talk entitled “P-OSRA: translating polymer images to text using extensions of open source software.” Bryn discussed the need for building databases of polymer structures. IBM is tackling this in multiple ways, with one group working on source and structure-based polymer chemical name to structure conversion, while Bryn’s work covers the conversion of polymers represented as images. P-OSRA is an extension of OSRA, allowing the repeat brackets that are frequently found around repeat units to be precisely recognized. The system works by removing the brackets then passing the bracket-less structure to OSRA. After recognition the position of the bracket is used to infer which atoms it refers to, and an extension of SMILES is used to capture this information. Multiple repeat brackets as found in, for example, block polymers, are also supported.

Aniko Valko (Keymodule) gave a talk entitled “Practical case studies of the application of CLiDE for the efficient extraction of chemical structures from documents.” CLiDE is a tool for extracting structures from chemical structure diagrams presented as images. This can either be done semi-automatically, in which case results are manually checked and can be edited (CLiDE Standard or Professional), or, alternatively, fully automatically for bulk extraction (CLiDE Batch). Aniko presented four case studies of CLiDE’s use: a WO patent, a Japanese patent, a European patent and a journal article. These highlighted some of the more challenging cases that CLiDE supports, for example, identification and extraction of structures from tables. A common cause of issues was the source images being of poor quality, for example, with unclean lines or even lines of missing pixels in the images! CLiDE uses a filter to ignore structures that appear to have been generated from non-structure diagram sources, for example, graphs, and text. This feature was shown repeatedly to be highly discriminatory at filtering out bad structures often removing 90% of the “garbage” structures. While R-group labels are supported, more complex Markush features are not yet supported, for example, positional variation, frequency variation, and attachment point indicators.

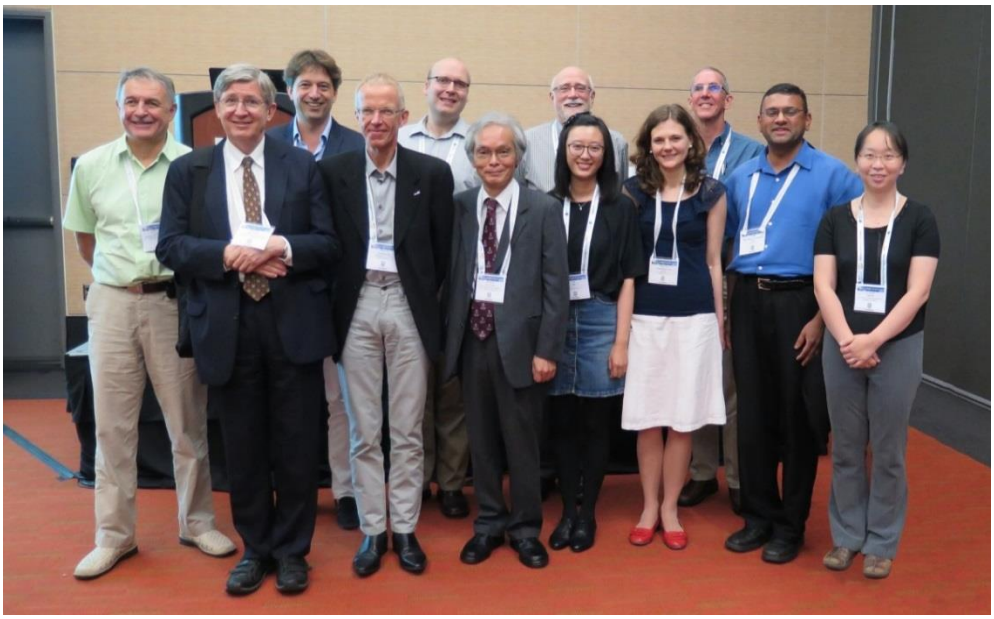
Daniel Lowe, Symposium Organizer

Herman Skolnik Award Symposium 2015 Honoring Jürgen Bajorath

Introduction

Veerabahu (Veer) Shanmugasundaram of Pfizer, who chaired the symposium, gave a brief introduction highlighting Jürgen's achievements. (A lengthier tribute has appeared at http://bulletin.acscinf.org/node/655.) Jürgen obtained his diploma (M.S.) and Ph.D. degrees (under Wolfram Saenger) in biochemistry from the Free University of Berlin. He then did postdoctoral studies with Arnie Hagler at Biosym in San Diego, focusing on DFT calculations of enzyme-inhibitor complexes. At Bristol-Myers Squibb he worked on protein modeling and structure-based design projects and developed his interests in bioinformatics and cheminformatics research. During his tenure at New Chemical Entities, he firmly established himself as a thought leader in cheminformatics. After sixteen years in the United States, he returned to Germany where he is currently Professor and Chair of Life Sciences Informatics at the University of Bonn. Jürgen is a leader in the development and application of cheminformatics and computational solutions to research problems in medicinal chemistry, chemical biology, and life sciences. He has done pioneer work in compound-centric data visualization and analysis in chemistry and is widely recognized for his seminal and prolific research work in several areas that are of interest to industry. His research interests include large-scale graphical SAR analysis, navigating high-dimensional space, multi-target modeling, machine learning, and virtual screening.

The award symposium was divided into four sections. The first three speakers were "people Jürgen has looked up to:" Tony Hopfinger, Gerry Maggiora, and Peter Willett, all of them former Herman Skolnik Award winners. (Arnie Hagler was also to have been in this group, but he was unable to attend.) The next speakers were Jürgen's colleagues and peers: Alexandre Varnek, Kimito Funatsu, Gisbert Schneider, Pat Walters, and Veerabahu Shanmugasundaram. They were followed by some of Jürgen's present and past students: Ye Hu, Eugen Lounkine, and Anne Mai Wassermann. Finally, Jürgen himself gave the award address.



Front row, L to R: Alexandre Varnek, Peter Willett, Jürgen Bajorath, Kimito Funatsu, Ye Hu, Anne Mai Wassermann, Veerabahu Shanmugasundaram, Jane Tseng
Back row, L to R: Gisbert Schneider, Eugen Lounkine, Gerry Maggiora, Pat Walters

Receptor-independent ligand activity models and receptor-dependent activity models

Jane Tseng presented the first talk on behalf of Tony Hopfinger of the University of New Mexico, who was unable to attend. In developing predictive methods to construct ligand-receptor binding models, most often to estimate IC_{50} values in the format of QSAR models, contributions from the receptor have been neglected. In the beginning, when protein-ligand structures were not available, the original goal of 4D-QSAR analysis[1] was to develop a methodology to complement Comparative Molecular Field Analysis (CoMFA).[2] In CoMFA, descriptors are calculated as grid point interactions between a probe atom and the target molecules and only one conformation of each compound is considered, not a conformational ensemble profile, as in the 4D-QSAR method. A new use of 4D-QSAR is to permit the parsing of information content arising from receptor-independent (RI) ligand activity models, as opposed to receptor-dependent (RD) models. To what extent is an RI ligand activity model (i.e., classic QSAR) of value in drug design applications?

4D-QSAR includes the conformational flexibility and the freedom of alignment by ensemble averaging in the conventional 3D descriptors found in traditional 3D-QSAR methods. Thus, the "fourth dimension" of the method is ensemble sampling of the spatial features of the members of a training set. In this approach, the descriptors are the occupancy frequencies of the different atom types in the cubic grid cells during the molecular dynamics simulation time, according to each trial alignment, corresponding to an ensemble averaging of conformational behavior.[3,4] The grid cell occupancy descriptors (GCODs) are generated for a number of different atom types (e.g., nonpolar, hydrogen bond acceptor, aromatic), called interaction pharmacophore elements (IPEs). The variable selection is made using a genetic function algorithm (GFA).[5] Multiple good QSAR models can be generated in the GFA step and the best model has to be established.

The 4D-QSAR methodology can be used in a receptor-dependent (RD) mode when the geometry of the receptor is available. In the RD-QSAR analysis, models are derived from the 3D structure of the multiple ligand-receptor complex conformations. This approach provides an explicit simulation of the induced-fit process, using the structure of the ligand-receptor complex, where both ligand and receptor are allowed to be completely flexible by the use of molecular dynamics simulation. RD-QSAR is used to gather binding interaction energies, as descriptors, from the interaction between the analogue molecules and the receptor.[6] The RD-4D-QSAR approach[7] employs a novel receptor-pruning technique to permit effective processing of ligands with the lining of the binding site wrapped about them. Data reduction, QSAR model construction, and identification of possible pharmacophore sites are achieved by a three-step statistical analysis consisting of genetic algorithm optimization followed by backward elimination, multidimensional regression and ending with another genetic algorithm optimization.[8]

The paradigm of 4D-QSAR analysis does appear to afford identical and comparative model development capabilities for both RI and RD studies. Both numeric and actual spatial pharmacophore subtractions of RI- and RD-QSAR models developed from training sets in which receptor information is available can be performed and a general assessment of lost design information in an RI study can be made.

Jane presented results for six ligand-receptor systems for which both an RI- and an RD-4D-QSAR analysis model had been constructed. She presented tentative conclusions from comparisons of the RI- and RD-4D-QSAR models and their pharmacophore sites (GCODs). The RD models are about the same "quality" (in terms of r^2 *v*alues) as the RI models, but usually have fewer GCOD terms. The RD models usually contain one or more ligand-receptor based GCODs, but receptor-only based

GCODs are not common in the RD models. Eye-ball selected clusters of “common” GCODs account for about 50% to 80% of the variance explained by the RI and the RD models.

Tony speculates that for ligand-receptor pharmacophore-based QSAR models, 20% to 40% of the targeted information in an RD-QSAR model is different from that of its corresponding RI-QSAR model. There are no discernible differences in atom-types or GCOD occupancy values, but RD GCODS are found near receptor walls whereas RI GCODs are found in “open” receptor space which is often most occupied by ligand atoms. This type of comparison of RI- and RD-QSAR models is only possible for datasets where explicit ligand-to-receptor binding occurs, and the identical pharmacophore generating methodology must be used on both the RI dataset and the corresponding RD dataset. The major finding of a 20 to 40% difference in a RI-4D-QSAR model and its corresponding RD-4D-QSAR model may be at odds with Dick Cramer’s recent success[9] in correctly predicting 12 ligands from an RI ligand-receptor model.

Non-specificity of drug-target interactions

Gerry Maggiora of the Translational Genomics Research Institute, Tucson, gave this talk. System complexity, non-specificity, and biological reductionism are issues confronting drug discovery. Complex systems, such as biosystems, weather systems, and traffic systems, have numerous interacting component parts and unpredictable behavior. They are non-computable and have emergent properties. Emergent properties arise out of more fundamental entities and yet are irreducible with respect to them.

Biological systems are structurally and functionally complex. The central dogma of molecular biology, DNA makes RNA makes protein, is an overly simplistic concept.[10] An organism’s phenotype is influenced by genomic and epigenetic phenomena, the latter being linked to a variety of biological sensors that are able to sense their environments and influence the functions the system can carry out. The discovery of microRNA revealed that part of “junk DNA” is actually transcribed by the machinery in cells into bits of RNA that are fundamental controllers of life.

Biological systems have a hierarchical structure: population, organism, organs, tissues, cells, organelles, molecules. The reductionist approach seeks to decompose biological systems into their constituent parts in an effort to understand the biology induced by these parts. Moving up the biological hierarchy, function is reintroduced and the size and complexity of the systems tend to increase.[11]

The notion of specificity in biological systems has a long history. Notions of specificity and reductionism led to the single-target hypothesis which is still a major model in drug discovery. Adverse drug reactions and repurposed drugs imply a greater lack of specificity in biosystems than is generally assumed. The emerging field of polypharmacology addresses the interaction of drugs with multiple targets. A published analysis of the drug-target network[12] suggests a need to update the single drug-single target paradigm. Many analyses of drug-target interactions have been reported.[13-20] There are also many drug-target databases.[21-27]

Data quality is an issue for these databases: the data are obtained by different methodologies and different experimental protocols in different laboratories, and drug-target interactions are predicted. The data are inconsistent within and among databases. The data may not be complete, that is, it may not relate all selected compounds to all selected targets,[15] and drug-target space (all compounds

against all targets) may not be completely covered. So, how promiscuous are drugs and how much biology is affected by the introduction of a single drug?

In the drug discovery landscape, the organismal level can be related to the molecular level by highly complex empirical models through to simpler mechanistic models; a mechanistic model relating a molecule to an organism is less physiologically relevant than an empirical model. Empirical models relate to phenotypic screening, mechanistic models to target-oriented screening.

Target-oriented drug discovery requires well validated targets. Sufficient details of the full mechanism of action are generally lacking, but, on the plus side, target-oriented screening is generally amenable to a high-throughput format. Screening hits are typically more limited than those obtained in phenotypic screens; it is unlikely that "inactive" regions of chemical space will be considered further; SAR typically neglects interactions with other targets; and follow-on phenotypic screens are required to assess biological efficacy.

Gerry gave as an example the use of imatinib in chronic myeloid leukemia. The target is Bcr-Abl kinase, a constitutively active product of the BCR/ABL fusion gene. Imatinib binds to the ATP binding region of Bcr-Abl kinase. The STITCH 4.0 database, however, shows that imatinib interacts with at least 10 different proteins, accounting perhaps for 24 or more adverse drug reactions observed for imatinib. Gerry also showed a network of imatinib interactions with the 52 proteins that interact in any fashion with imatinib. The Drug2Gene database records 41 proteins associated with imatinib binding. BCR/ABL behavior is complex[28,29] and drug resistance, both *de novo* and secondary, is observed. It is a scientific aphorism that in an experiment it is difficult to find what you are *not* looking for.

Lessons can be learned from metabolic engineering: numerous metabolic engineering studies show that metabolic networks cannot be regulated by perturbing a single network component. Manipulating single genes or gene products does not affect phenotype; or the genes' influence on phenotype does not arise in a simple, obvious fashion. Introduction of any xenobiotic into a biosystem affects multiple, and in many cases diverse targets, so there is a significant degree of "mechanistic uncertainty" in target-oriented drug discovery.

Phenotypic screening[30] has thus re-emerged. Phenotypic methods, which rely much less on mechanistic details, can provide a more robust platform. They employ a phenomenological (empirical) approach that is function-based rather than mechanism-based. They are hypothesis-driven, and similar to statistically based systems models. Phenotype-based approaches are closer to "intrinsic biology" with increased likelihood of finding viable leads. They typically generate a greater number of diverse hits than target-based screens. Functional responses are ideally, but not always, related to disease states. Phenotype-based approaches are particularly useful in cases where the biology is not clearly understood. Phenotypic screens are inherently multi-target screens but are target- and mechanism-agnostic. Promiscuity may be a virtue in phenotypic screens. The use of high-throughput formats is limited, but improvements are on the way. Determining the mechanism of action may be an issue, and incorrect target determination can cause significant problems. Gerry made two final observations: if your only tool is a hammer, all problems begin to look like nails; and the bigger the hammer, the easier it is to pound the nails.

Molecular similarity approaches in cheminformatics

Peter Willett of the University of Sheffield outlined the early history of molecular similarity, and presented a bibliometric analysis. As Rouvray[31] noted "Similarity is ubiquitous in scope,

interdisciplinary in nature, and seemingly boundless in its ramification". Mendeleev's 1869 discovery of the Periodic Table is often cited as the first example of similarity concepts in chemistry, but there are many other historical examples.[32] Computational measures of similarity are of great importance for cheminformatics, as a result of the "similar property principle", which states that structurally similar molecules have similar properties. There are many exceptions to the principle but it is still a useful rule-of-thumb. It is generally ascribed to a book by Johnson and Maggiora,[33] but Johnson and Maggiora had earlier ascribed it to a 1980 work by Wilkins and Randic.[34] The principle was in fact widely understood, even if not expressed in explicit form, much earlier than that, all the way back to 1868.[35] Analogous similarity relationships in geography[36] and social networks[37] have been referenced in recent publications in cheminformatics, the latter in studies of chemical space networks by Jürgen Bajorath. The cluster hypothesis underlying document clustering[38] spurred Peter's own studies of chemical clustering, given the analogies between cheminformatics and information retrieval.

The similarity principle provides not only a rationale for using similarity techniques in cheminformatics but also a way of validating them, for example in comparison of measures for similarity searching where benchmark datasets of actives and inactives are used to evaluate the relative effectiveness of different measures on the extent to which nearest neighbors of known actives are also active. There are analogous validation approaches in clustering and diversity applications, for example, all the molecules in a given cluster should have broadly similar properties.

The earliest example of clustering chemical databases was work[39] at ICI Pharmaceuticals Division in which fragment-based similarities were used to cluster around a known active if there were at least some number of nearest neighbors above a similarity threshold. Common structural features in such clusters were identified. Adamson and Bush[40,41] were the first to use 2D substructure searching features in a comparison of the effectiveness of similarity measures for single-linkage clustering. Fingerprint-based measures are still the most common 40 years later.

At Sheffield University extensive comparative studies of a wide range of similarity measures and clustering methods were carried out by Willett and Winterman,[42-44] using the Adamson-Bush evaluation procedures. Fragment occurrences were found to be slightly better than incidences. The Tanimoto coefficient was found to be the most effective coefficient of those tested, and it is still the standard for similarity applications in cheminformatics. Ward's hierarchic agglomerative method[45] has since proved to be the preferred clustering method, but the non-hierarchic, nearest-neighbor method of Jarvis and Patrick[46] was for years a cost-effective alternative, given the algorithmic complexity of Ward's method.

The use of the similar property principle for ligand-based virtual screening was initially studied in the mid-1980s at Lederle Laboratories,[47] the Upjohn Company, and Pfizer in the United Kingdom together with Sheffield University.[44,48] The use of substructure-searching fragments and simple association coefficients is effective and efficient in operation, and is a simple enhancement of existing database software; there was therefore a rapid take-up, and 30 years later this is still a standard approach to virtual screening. Many other 2D and 3D approaches are now available, but they are still less widely used. Perhaps the main enhancement since the initial work is the use of data fusion methods, as first studied at Merck,[49-51] and at Sheffield.[52-55]

Developments in combinatorial chemistry and high-throughput screening in the early 1990s spurred interest in the selection of diverse sets of compounds, [56,57] but work on compound selection had been undertaken several years previously at Upjohn, and at Pfizer in the United Kingdom together with Sheffield University, based directly on the similarity measures that had been developed previously for

clustering and similarity searching. Methods included cluster-based selection,[43] and dissimilarity-based selection to optimize a diversity index.[48,58] The latter, using the Kennard-Stone algorithm, is now widely implemented as MaxMin.[59]

Peter concluded his talk with a bibliometric analysis of the literature of molecular similarity, as reflected in the Web of Science database. He found 86,663 citations to 2,980 articles on molecular similarity, with an *h*-index of 114 and a mean of 29.1 citations per article. The distribution of author contributions is highly skewed: Jürgen Bajorath is the most prolific author, with 95 of the 2,980 articles (Peter himself is close behind with 88), but there are 6,579 singleton contributions. As regards organizations, Sheffield has published largest number of articles (111), but there are 1,014 singletons amongst the 1,767 distinct organizations. Ten organizations, including five private-sector ones (AstraZeneca, GlaxoSmithKline, Merck, Novartis, and Pfizer) have 50 or more articles. Thirty of the 2,980 articles have 250 or more citations; the top five are by Allen *et al.*,[60], with 1400 citations, Klebe *et al.*,[61] Willett *et al.*,[62] Tropsha *et al.*,[63] and Bemis and Murcko.[64]

The citations appeared in 3,977 distinct publications, with the most frequent being *J. Chem. Inf. Model.* (2724), *J. Med. Chem.* (2075), *Bioorg. Med. Chem.* (1019), *Bioorg. Med. Chem. Lett.* (986), *J. Comput.-Aided Mol. Design* (734), *Eur. J. Med. Chem.* (695), *Mol. Inf.* (645), *J. Mol. Graphics Modell.* (461), *PLoS One* (429), and *J. Am. Chem. Soc.* (372). The citing journals come from 202 distinct Web of Science subject categories: the methodological tools developed by the molecular similarity community are thus clearly of very broad applicability. Jürgen Bajorath has 11,037 citations to his 452 articles (120 of them in *J. Chem. Inf. Model*.) with an *h*-index of 48 and a mean of 24.4 citations per article. Of these, 16 have 100 citations or more, the top five[65-69] illustrating Jürgen's contributions to multiple fields of chemistry and the life sciences.

Generative topographic mapping

Alexandre (Sasha) Varnek, of the University of Strasbourg, France, described this tool for chemical space analysis. There are many ways of visualizing chemical space. In descriptor-based chemical space, where a D-dimensional vector represents each molecule, two popular approaches are used: similarity network graphs, and dimensionality reduction techniques which transfer the objects from the D-dimensional chemical space into a latent space of 2 or 3 dimensions.

Principal component analysis (PCA) and self-organizing Kohonen maps (SOM) are commonly used for exploration of large chemical spaces but both have drawbacks. PCA processes nonlinear data poorly. SOM is a nonlinear method and due to its topology-preserving character, it provides more information-rich plots than PCA, but it suffers from its purely empirical nature and it lacks solid statistical foundations.

Generative Topographic Mapping (GTM)[70,71] is a probabilistic extension of SOM. GTM relates the latent space with a 2D "rubber sheet" (or manifold*)* injected into the high-dimensional data space. The visualization plot is obtained by projecting the data points onto the manifold and then letting the rubber sheet relax to its original form. GTM generates a data probability distribution in both initial and latent data spaces. GTM can thus be used not only to visualize the data, but also for structure-property modeling tasks.[72]

Sasha showed a probability density distribution in the latent space. Projection of an object on a GTM is described by the probability distribution ("responsibilities") over the lattice nodes. Using GTM, one can, for each molecule, evaluate the probability of finding it in a point on the grid. There are two

possibilities: one can use the responsibilities as molecular descriptors which can be used for predictions, or one can prepare an “activity landscape” to make predictions.

In the course of this project, Sasha’s team has developed several utilities named ISIDA[73-75]/GTM (where ISIDA stands for *In Silico* Design and Data Analysis descriptors). They allow QSAR models to be created by GTM, and optimized and visualized, and the activity can be mapped. Chemical space maps can be used as a virtual screening tool. Sasha showed a GTM activity landscape of the stability of Lu^{3+} complexes with organic molecules;[76] strong and weak binders were clearly differentiated.

An activity landscape can be used directly to predict activities of test compounds using the distribution of responsibilities. In particular, in each node the product of activity landscape value for the training set and responsibility of the given test compounds is calculated by summation over all nodes of the map. It has been shown[76] that the performance of GTM-based regression models is similar to that obtained with four popular machine-learning methods (random forest, k-NN, M5P regression tree and PLS) and ISIDA fragment descriptors. By comparing GTM activity landscapes built both on predicted and experimental activities, one may visually assess the model’s performance and identify the areas in the chemical space corresponding to reliable predictions.

Sasha reported some work on a GTM-based model’s applicability domain.[77] The Biopharmaceutics Drug Disposition Classification System (BDDCS), based on solubility and degree of metabolism, is used by agencies such as FDA for granting biowaivers. Sasha and his co-workers have described the modeling in two-dimensional latent space for the four classes of the BDDCS using VolSurf descriptors. Three new definitions of the applicability domain (AD) of models were suggested: one class-independent AD which considers the GTM likelihood, and two class-dependent ADs considering either the predominant class in a given node of the map or informational entropy. The class entropy AD was found to be the most efficient for the BDDCS modeling. The predominant class AD can be directly visualized on GTM maps, which helps the interpretation of the model.

Sasha’s team has also studied a database of more than 2 million compounds containing 37 subsets coming from catalogs of 36 chemical suppliers, and the NCI database. The researchers focused both on the parameters able to characterize the whole dataset, and on the analysis of individual libraries, to see how they covered the chemical space, to what extent they overlap, and which library has compounds possessing a particular activity profile. GTM incremental learning[78] is a solution for such large datasets.

Sasha showed a GTM of the entire database built on MOE descriptors. Each data point represented a molecule and the data were colored according to molecular weight. The left hand side of the map was populated by light molecules and the right-hand one by heavier molecules. Instead of using each data point, you can use a data density distribution function represented by the ensemble of cumulated responsibilities. The density maps can also be built for the individual libraries. You can color the same GTM map by different properties or activities to visualize different property landscapes. Superposition of different activity landscapes helps you to select areas populated by compounds with particular activity profiles. The data coverage can be measured by normalized Shannon’s entropy calculated directly from the responsibilities. Surprisingly, the small ASINEX library covers the entire latent space more uniformly than the large Enamine library.

Sasha concluded with a few details of Stargate GTM (S-GTM), in which one GTM connects activity space and descriptor space. S-GTM can be used to predict a pharmacological profile and to discover

structures corresponding to a given pharmacological profile. The method has been applied to a set of eight GPCR activities.

Development of a knowledge-generating platform from drug discovery through to production

Kimito Funatsu of the University of Tokyo described a knowledge-unifying platform driven by big data. While massive amounts of quantitative data have accumulated across the pipeline of drug discovery, all the way from a candidate's initial discovery up to its production process, knowledge of and data analysis for each of the discovery and production processes has remained isolated. The big data in Kimito's project consist of a large virtual library containing chemical structures of drug candidates, interaction data between many proteins and many drug candidates, and plant operating data and product quality data. The objectives are: automated generation of a huge virtual library, discovery of new drugs, and acquisition of synthetic routes from the library; construction of a mathematical model derived from many proteins versus many compounds together with other biological information, and extraction of a guide for drug discovery; and knowledge extraction for process monitoring and control, plus development of the automated construction of a soft sensor model and a model maintenance system for process monitoring.

Prof. Okuno's group at Kyoto University is working on ligand-target information. Problems in the threefold, chemical-target-phenotype model of drug discovery include a shortage of experimental compound-protein interaction data, and compound-phenotype association data; and a lack of information on direct associations between target protein and phenotype. From mathematical models (logistic regression, PLS and SVM), predictions can help to fill the gaps. In previous work for predicting compound-protein interactions using information about chemical structures and protein sequences, the researchers used SVM, which trains up to 250,000 interactions but it is hard to learn larger scale data because of memory and computation time limits. They are trying to apply deep learning to train millions of interactions. In the prediction of protein-phenotype associations, they have compared the performance of PLS and SVM with that of logistic regression. They have demonstrated useful accuracy and high speed, but the number of proteins with positive weights is limited. In future work they aim for large-scale prediction of associations for all possible combinations between compounds and phenotypes, and they plan interaction prediction using deep learning, learning from a bigger dataset of interactions, and interpreting a deep belief network derived from big data.

Dr. Taiji's group at RIKEN is working on a very large scale virtual library (billions of compounds) with a synthetic route for all compounds, for assessment of synthetic feasibility. The massive generation of chemical structures using transformational rules involves rewriting of the transform-oriented synthesis planning (TOSP) generator[79] to allow parallel-processing, and validation of the transform and fragment data.

Kimito's part of the joint project concerns a soft sensor for monitoring and controlling a chemical plant. In chemical plants, efficient and stable production is required, keeping the quality of chemicals high. Operators have to monitor the operating condition of the plants and control process variables. NIR spectra, temperature, and pressure are easy to measure online. Concentration and density are difficult to measure online[80] and are predicted in this project, using a statistical model, from the NIR spectra, temperature, and pressure input to the sensor. Until recently, application of soft sensors online has not been possible because of low predictive accuracy and complex maintenance of the sensor.

Problems in soft sensor analysis include data reliability and selection; outlier detection and noise treatment; deciding on an appropriate regression method; overfitting; nonlinearity among process variables; variable selection; dynamics in the modeling process; model interpretation; model validation; applicability domain and predictive accuracy; model degradation; model maintenance; and detection and diagnosis of abnormal data. The predictive ability of soft sensors depends on the quality of database, but the amount of data in such a database is limited, so database monitoring is essential for highly predictive soft sensors. Data measured in plants are not fully exploited in process control. Soft sensors express relationships between process variables, so an efficient control method using a soft sensor model is required.

Since the predictive performance of adaptive models depends on databases, Kimito's group has proposed a database monitoring index (DMI),[81] to monitor the database and a database monitoring method using the DMI. The DMI proposed is based on similarity between two data. The more similar two data are, the smaller DMI is. New data are stored when the minimum DMI value exceeds a threshold. Through the analysis of simulation data and real industrial data, the researchers have confirmed that databases can be appropriately managed and the predictive accuracy of adaptive soft sensor models increased by using the proposed method.

The three research groups aim to establish a platform which allows them to unify knowledge about different processes, and to advance research into improved and optimized systems that view pharmaceutical development from a comprehensive, correlated, and high-level perspective.

Enabling drug discovery by computational molecular design

Gisbert Schneider, of ETH, Zürich, Switzerland, gave a talk on *de novo* drug design and target prediction. The computer-based design of drug candidates is a complementary approach to high-throughput screening; *de novo* design[82] supports drug discovery projects by generating novel pharmaceutically active agents with desired properties in a cost- and time-efficient manner. An example is the identification of novel cannabinoid-1 receptor inverse agonists for the treatment of obesity.[83] A recent publication[84] reviews software for *de novo* drug design with a special emphasis on fragment-based techniques that generate druglike, synthetically accessible compounds.

The software Design of Genuine Structures (DOGS)[85,86] features a ligand-based strategy for automated *in silico* assembly of potentially novel bioactive compounds. The construction procedure explicitly considers compound synthesizability, based on a compilation of 25,144 available synthetic building blocks and 58 established reaction principles, with 25 regioselective variants. This enables the software to suggest a synthesis route for each designed compound. The quality of the designed compounds is assessed by a graph kernel method[87,88] measuring their similarity to known bioactive ligands in terms of structural and pharmacophoric features. Virtual intermediates are compared with a template. The scoring method does not just rely on substructure similarity, and the pharmacophore comparison is very permissive compared with graph similarity. The origin of the idea was patent beating.

Combinatorial *de novo* design can also be coupled with microfluidic synthesis and analytics.[89-91] Gisbert's team has recently reported[92] a multi-objective *de novo* design study driven by synthetic tractability and aimed at the prioritization of computer-generated 5-HT_{2B} receptor ligands with accurately predicted target-binding affinities. Gaussian process models were built for 974 proteins annotated in ChEMBL, and the team designed and synthesized structurally novel, selective, nanomolar, and ligand-efficient 5-HT_{2B} modulators. The results suggest that amalgamation of

computational activity prediction and molecular design with microfluidics-assisted synthesis enables the swift generation of small molecules with the desired polypharmacology. In another example, Fasudil, a not very active, but ligand efficient Rho-kinase inhibitor was used as a template in DOGS to design a fragment-like candidate that was made and tested, and could be grown into Azosemide, approved for treatment of hypertension in Japan.

Gisbert's team has also developed an approach to target prediction. Several computational tools for predicting macromolecular targets of new chemical entities were publicly available, but none of these methods was explicitly designed to predict target engagement by *de novo* designed molecules, so the researchers devised self-organizing map-based prediction of drug equivalence relationships (SPiDER),[93] that merges the concepts of self-organizing maps, consensus scoring, and statistical analysis to identify targets for both known drugs and computer-generated molecular scaffolds. Some 15,000 drugs and druglike compounds are used as the basis for clustering and 11 targets per compound are predicted on average. The approach results in confident predictions.

The targets of natural products are largely unknown, which hampers rational drug design and optimization. Gisbert's team has developed and validated a computational method for the discovery of such targets. The technique does not require three-dimensional target models and may be applied to structurally complex natural products. The algorithm dissects the natural products into fragments and infers potential targets by comparing the fragments to drugs with known targets. Kohonen self-organizing maps and chemically advanced template search (CATS) topological pharmacophores[94] are used.[95,96]

Of the 210,213 structures in the *Dictionary of Natural Products*, 31% are fragment-like and 69% have large structures. Gisbert's system confidently predicted targets for 36% of the fragment-like products and 22% of the large ones. Sparteine is a deadly class 1a Na+ channel blocker with high ligand efficiency. Gisbert predicted that it interacted with the kappa opioid receptor. The fragment-like, synthetically tractable structures goitrin, isomacroin and graveolinine were input to SPiDER for target inference. Five out the six confidently predicted targets were correct, unreported targets, and the molecules were profoundly dissimilar to the most similar reference compound. Graveolinine shows dual target engagement (5-HT_{2B} and COX2) and could lead to polypharmacological tool compounds for example, for migraine.[97] In a prospective validation, it has been shown that fragments of the potent antitumor agent archazolid A contain relevant information regarding its polypharmacology. Biochemical and biophysical evaluation confirmed the predictions.[96] These results obtained with SPiDER corroborate the practical applicability of the approach to natural product "de-orphaning". DOGS and SPiDER lead from complex natural products to synthesizable new chemical entities.

Integrating public data sources into the drug discovery workflow

Pat Walters of Vertex Pharmaceuticals discussed two examples of work carried out at his company. The first was high-throughput screening (HTS) data analysis. Bench scientists want to be able to find hints of SAR, that is, to identify scaffold classes and related classes, and visualize activity distributions. They want to find out what is known about the activity of the compound class from both internal data and the literature, and about the properties and pharmacokinetics of the class. They want additional information, if any, from the literature about patents, properties, and synthesis. Pat listed the guiding principles for a system that meets these requirements. The first is to keep things simple: analysis tools should be intuitive, and molecules should be organized in a "medicinal chemistry driven" fashion. The second is to make the results visually compelling with a data

dashboard, and one click access to details. Above all, the system must enable answers to critical questions.

The workflow involves partitioning actives into scaffold classes; profiling each scaffold class according to activity distribution, emerging SAR, selectivity, properties and ADME, and literature information; and prioritizing scaffolds for further exploration in “analogue by catalog”, and exploratory chemistry.

Pat ran through an example, with screenshots, of identifying three ring scaffolds, keeping the most frequently occurring scaffolds, and displaying them in the HTS Viewer, with or without molecule details. Related scaffolds are identified by scaffold similarity and arranged by similarity in the HTS Viewer. Activity on the HTS target is viewed by way of boxplot distributions in which there is an adjustable activity cutoff; boxplots provide an easy comparison of related scaffolds. To evaluate selectivity, activity against other targets is studied. Users can perform a general or target class specific analysis. Selective series are identified from plots of number of active assays against numbers of assays tested. Users can drill down to activity details and compare activity and selectivity for related scaffolds by displaying boxplots alongside the activity scatterplots. “Thermometer plots” show distributions of properties related to ADME. These plots can be displayed in yet another column alongside the scaffolds, boxplots and activities.

Scientists want to answer many literature questions. What biological activities are known for this compound class? Have related compounds been in clinical trials? Is this compound class mentioned in patents? Is the class well characterized (by physical properties)? Has the synthesis of the class been reported? There are many external sources of biological activity, physical properties, synthesis, drug data, and patents. While these databases provide a wealth of information, the data are often not in a format that is easily accessible to the bench scientist. In addition, scientists may be unaware of these resources, or may not know how to access and integrate the data. While it is tempting simply to integrate large amounts of public data into in-house systems, software developers must be careful to inform, without overwhelming, the target audience.

Vertex applications, with substructure search included, link to SciFinder via the SciFinder API, and use an internal database for Reaxys, ChEMBL, and Thomson Reuters Integrity. Pat showed screenshots of the HTS Viewer links to Reaxys, ChEMBL, and Integrity data. In each case links allow the user to jump directly to the underlying data. Vertex and CAS collaborated to provide a direct link to SciFinder using the SciFinder API. The Vertex system also addresses numerous other considerations such as identification of potential false positives and negatives; compound purity; “efficiency” of hits; filtering out undesirable compounds from assays; replicates and statistics; and hit follow-up.

Pat’s second example of work carried out at Vertex concerned patent informatics. IPedia is a platform for information sharing. Vertex used to have an in-house system for capturing data from the patent offices and chemical structures were entered manually. The release of SureChEMBL has changed all that, but unfortunately SureChEMBL’s automated process extracts *all* structures including reagents, solvents and the like.

Can SureChEMBL be used to find interesting structures? Pat’s team took 30 drug patents (based on work done at AstraZeneca),[98] and eliminated three which were not in ChEMBL. They looked if the drug structure was in the SureChEMBL curated set, and tried to develop heuristics to identify the key compounds. The number of structures per SureChEMBL patent was a minimum of 11, and a maximum of 916, with a median of 161. The drug structure was found in 19 of the 27 patents. Structures were classified as interesting if they were 0.8 similar to the drug, by Tanimoto coefficient

and MDL keys, and as boring if they were less than 0.8 similar. There were 1598 interesting structures and 4357 boring ones. Descriptors for structures were frequency of occurrence in SureChEMBL, location in the patent, number of heavy atoms, molecular weight, and number of neighbors at Tanimoto 0.8. The team built a simple recursive partitioning model using the ctree method in the "party" package in R 3.0.2. Simple heuristics proved to be very effective (accuracy 0.91 and kappa 0.77). Neighbor counts identified interesting structures: interesting compounds have more neighbors. This example again illustrates how public data can be used to advantage in drug discovery projects.

"Close-in" analogue prioritization using SAR matrices

Veer Shanmugasundaram described work done at Pfizer in collaboration with Jürgen Bajorath. Jürgen has published papers on heterogeneous SAR,[99] activity cliffs versus selectivity cliffs,[100] SAR monitoring using activity landscapes,[101] and molecular mechanism based network-like similarity graphs.[102] Visualizations include network-like similarity graphs, SAR matrices, ligand-target differentiation maps, and bipartite matched molecular series graphs. Veer's talk related to SAR matrices,[103] which are designed to highlight different SAR patterns in large compound data sets. They provide chemically intuitive organization of analogue series, and easy identification of activity cliffs, providing immediate suggestions for compound design.

The SAR matrix data structure organizes compound data sets according to structurally analogous matching molecular series in a format reminiscent of conventional R-group tables. An intrinsic feature of SAR matrices is that they contain many virtual compounds that represent unexplored combinations of core structures and substituents extracted from compound datasets on the basis of the matched molecular pair formalism. These virtual compounds are candidates for further exploration but are difficult, if not impossible to prioritize on the basis of visual inspection of multiple SAR matrices.

Pfizer therefore worked with Jürgen to develop a compound neighborhood concept as an extension of the SAR matrix data structure that makes it possible to identify preferred virtual compounds for further analysis. On the basis of well-defined compound neighborhoods, the potency of virtual compounds can be predicted by considering individual contributions of core structures and substituents from neighbors. SAR-rich matrices are prioritized based on SAR patterns, property variance, and size and dimension of matrices and confidence values can be included in the matrix visualization. In extensive benchmark studies, virtual compounds have been prioritized in different datasets on the basis of multiple neighborhoods yielding accurate potency predictions.[104]

A retrospective analysis was carried out using six large sets of different G-protein coupled receptor antagonists extracted from ChEMBL for which K_i values were available. Matrix-pattern-based, matrix pattern based weighted by similarity, and analysis of variance models were compared with Jürgen's nearest neighbor analysis. The prediction accuracy (r^2) was best for analysis of variance models (between 0.7 and 0.84). Depending on the algorithmic fragmentation scheme, single-cut matrices (i.e., one exocyclic bond in a compound is systematically deleted to yield key and value fragments), dual-cut (two exocyclic bonds are simultaneously deleted), and triple-cut matrices (three exocyclic bonds are deleted) are separately generated. Veer showed boxplots of the mean error and type of matrices on the ChEMBL datasets and scatterplots of the distribution of absolute error and neighborhood similarity. The SAR matrices can be adapted for visualization in Spotfire DXP. This environment offers a structure-data viewer, filters, dynamic interactive visualizations, and a direct connection to the Pfizer database.

In summary, a visual examination of SAR using an adaptation of SAR Matrices in DXP provides a way to view, mine and interrogate single, double and triple-cut matrices dynamically, and study SAR trends quickly. Several methods that prioritize virtual compounds “to fill” close-in analogue space ranging from nearest neighbor methods, and similarity weighting to ANOVA analysis all appear to perform equally well. Predictions based on single-cut matrices are as valuable as those with more complex double- and triple-cut matrices.

The second part of Veer’s talk concerned series progression. The problem was to see if Pfizer could develop some diagnostic methods to evaluate if they were adding SAR information as chemical series progressed, and to determine whether more “close-in” analogues should be made, or whether new lead series should be identified. The strategy was a chronological analysis of “SAR information content” using SAR matrices, and using that to distinguish productive and unproductive series.

SAR matrices with a minimum of two compounds per series and a minimum of two series per matrix were used. SAR matrices were classified as old if a matrix in the previous year had the same cores and real compounds, expanded if a matrix in the previous year was a subset (had a subset of cores and real compounds), and new if no matrix in the previous year was a subset. Veer showed a number of plots of series progression, and of raw discontinuity score against average potency. He concluded that monitoring changes in SAR information content in multiple series could provide some interesting diagnostics in evaluating series progression. Matrices with increased discontinuity are considered to provide rich SAR information. The appearance of new matrices with increased SAR discontinuity or expansion of current matrices provides clear signals to evaluate series progression.

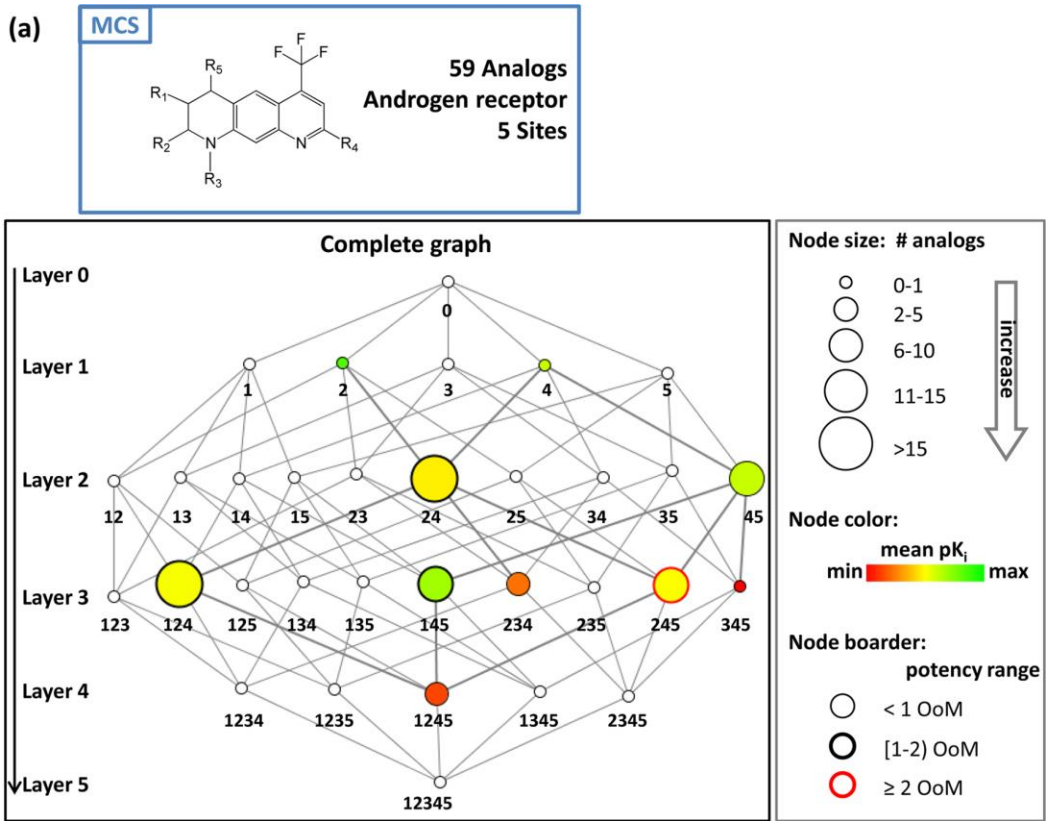
Graphical analysis of analogue series and associated SAR information

Ye (Pauline) Hu, one of Jürgen’s current students in Bonn, presented AnalogExplorer. Analogue series are compounds sharing the same molecular scaffold or maximum common substructure (MCS). The conventional data format for them is a standard R-group table with all substituents and associated potency values. This is difficult to use for large and structurally heterogeneous series, so, as rapidly increasing amounts of SAR data become available, graphical approaches have been introduced to explore structure-activity relationships (SARs) contained in compound data sets. Exemplary MCS-based visualization methods include SAR maps, and the combinatorial analogue graph (CAG). In SAR maps analogous compounds contain substituents at two different sites. In the matrix format each cell represents a compound with corresponding substituents, and cells are colored by potency values.[105] In a CAG,[106] nodes are pairs of compounds with variations at one, two, or maximally three sites, colored by SAR discontinuity scores. Edges are the relationships between substitution sites.

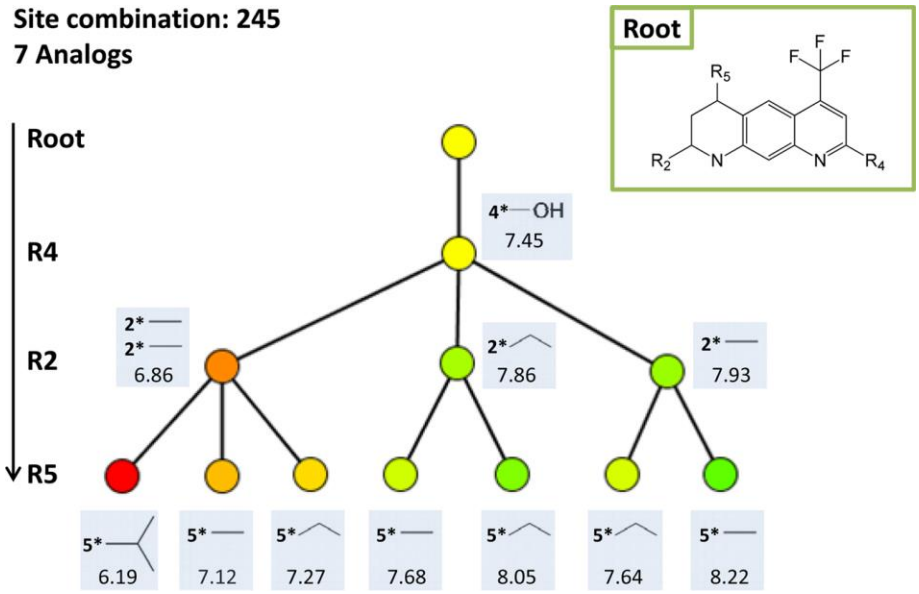
AnalogExplorer[107] uses a compound-based approach, rather than a compound pair or substituent based approach. At a global level it explores substitution sites and site combinations, deconvolutes a series into subsets of analogues having varying R-groups at the same site(s), and prioritizes analogues at specific site(s) that have desired activity. At a local level it represents a subset of analogues at given substitution site(s) on the basis of R-groups.

For graphical analysis, an analogue series is organized into subsets on the basis of the MCS. Each subset comprises compounds having varying R-groups at the same substitution site or site combination. Mapping of an analogue to the MCS of the series determines its subset membership. Each analogue of a series belongs to one and only one subset. An example is given below. In the complete graph, each node represents a substitution site or site combination, and all compounds with

varying R-groups at the given site(s). The root node 0 corresponds to a compound having no R-group at any site. The node 1 represents analogues that only contain R-groups at R1, the node 12 analogues with R-groups at R1 and R2, and so on. Only a subset of these nodes is populated with analogue subsets. Nodes are scaled in size according to the number of analogues of the subset they represent and are colored according to the mean potency values of their analogues. The border thickness of nodes reflects the potency range covered by analogues comprising the corresponding subset.



Nodes are connected via edges according to subset relationships among the substitution sites, that is, if a substitution site defining a node is a subset of other site combinations. Therefore, edges in the graph reflect hierarchical relationships between nodes in adjacent layers. In an AnalogExplorer reduced graph, empty nodes, indicating unexplored sites or site combinations, and edges between them are omitted for ease of interpretation. Another graphical component, termed R-group tree, is designed to represent a subset of analogues with given substitution site(s). An R-group tree of node 245, can be constructed, for example:



Pauline presented graphs for four different applications. The first application was a single analogue series of 52 histamine H4 receptor antagonists, with 5 sites. The majority of site combinations were are associated with subsets of potent analogues and four site combinations were associated with activity cliffs. The second application was a single series of 38 analogues with two targets (tyrosine protein kinase ABL and tyrosine protein kinase SRC) leading to two graphs. Application three was multiple series for the target tyrosine protein kinase SRC. Five graphs were made for five qualifying series that were available in the target set: 22 analogues with four sites, 44 analogues with seven sites, 13 analogues with six sites, 22 analogues with six sites, and 22 analogues with four sites. The fourth application concerned four (out of five) structurally related series targeting tyrosine protein kinase SRC. A matched molecular pair calculation reduced these to a core from a combination of four of the scaffolds. Pauline showed complete and reduced graphs for 101 analogues with four sites. A Java implementation of AnalogExplorer routines is made freely available via the ZENODO open access platform.

Various ways to define molecular similarity

Eugen Lounkine of Novartis described work done with three different types of fingerprints. The concepts of molecular fingerprints and molecular similarity have matured and found innumerable applications, but nowadays Novartis does not just use chemical similarity to find compounds that biologically will behave the same; rather, the company directly builds on biological profiles to assess biosimilarity. From a medicinal chemistry point of view, one of the primary goals of HTS hit list assessment is the identification of chemotypes with an informative SAR. A common way to prioritize them is molecular clustering of the hits. Typical clustering techniques, however, rely on a general notion of chemical similarity or standard rules of scaffold decomposition and are thus insensitive to molecular features that are enriched in biologically active compounds. This hinders SAR analysis because compounds sharing the same pharmacophore might not end up in the same cluster and thus are not directly compared to each other by the medicinal chemist. Similarly, common chemotypes that are not related to activity may contaminate clusters, distracting from important chemical motifs.

Eugen and his colleagues have projected bioactivity onto chemical fingerprints; they have combined molecular similarity and Bayesian models, and introduced an activity-aware clustering approach, and

a feature mapping method for the elucidation of distinct SAR determinants in polypharmacological compounds. They found that activity-aware clustering grouped compounds sharing molecular cores that were specific for the target or pathway at hand, rather than grouping inactive scaffolds commonly found in compound series. Weighted clusters often spread across many conventional clusters and there were large clusters that both methods agreed on.[108]

Eugen and his colleagues have also developed a tool that compares compounds solely on the basis of their bioactivity: the chemical biological descriptor called high-throughput screening fingerprint (HTS-FP). Data are aggregated from 234 Novartis biochemical and cell-based assays and can be used to identify bioactivity relationships among the in-house collection of about 1.8 million compounds. A similarity metric was derived combining both the numerical correlation of the activity *z*-scores, using the Pearson correlation coefficient, and the number of assays in common between the compounds.[109] HTS-FPs have been useful in both virtual screening and scaffold hopping. They are valuable not only because of their predictive power but mainly because they relate compounds solely on the basis of bioactivity. One challenge is the sparse nature of the HTS-FP matrix: the number of biologically annotated compounds still covers only a minuscule fraction of chemical space. To overcome this problem, Novartis has introduced Bioturbo similarity searching[110] that uses chemical similarity to map molecules without biological annotations into bioactivity space and then searches for biologically similar compounds in this reference system.

In addition, capturing the rich descriptions of compound-induced phenotypes from the literature gives yet another molecular fingerprint: the literature fingerprint. A naïve Bayesian model looks for themes around hits. Similarity search around a reference compound finds other compounds mentioned in the same biological or clinical context. By similarity searching a collection of terms, tool compounds for a phenotype of interest could be found. Novartis is carrying out exploratory annotation of phenotypic hits by text mining of abstracts and curated sources (e.g., ChEMBL provides a reference for each compound activity). MeSH terms for PubMed articles are filtered for informative terms. By data mining within and across projects, “signatures” derived from fingerprints can be found. A signature is a fingerprint template endowed with meaning. The meaning is encoded by relating the signatures database to the phenotype database.

Novartis has developed network algorithms to build and navigate heterogeneous similarity networks from the three types of fingerprint. Eugen gave an example starting with a selection of painkillers, connected only by literature relationships. In the first expansion of the network all the seed compounds have neighbors, but they come from different similarity measures: more painkillers (quercetin and doxorubicin) are added from the literature, diclofenac analogues from chemical similarity, and ibuprofen similars from HTS-FP. Next, the neighbors themselves are connected among each other, sometimes with more than one method. Distinct clusters emerge as interesting neighbors of neighbors (connectors) are added: morphine analogues, NSAIDs, and oncology pain management. Pairs that are connected by more than one method can be identified. These voting schemes are intuitive in graphs, and harder to formalize in conventional approaches. Alternatively, one can use degree, number of distinct edge types, etc. A flow algorithm is use to distribute scores. This standard graph neighborhood scoring algorithm is intuitive to carry out and visualize, and easily scalable. Eugen also showed networks of glitazones, antidepressants, and statins and warfarin. Graph representations provide a unique opportunity to combine distinct similarity domains in an interoperable way.

Could "inactive" compounds be good starting points for drug discovery?

Anne Mai Wassermann, now at Pfizer, talked about work done at Novartis while she was one of Jürgen's postdoctoral students. For a long time the paradigm for screening library design has been diversity. Chemical diversity has been used as a surrogate for biodiversity, but biological fingerprints themselves could be used.[111,112] What should then be done about the inactives? A compound inactive in a great many screens might be a good, selective lead from another screen. An analysis across more than 200 Novartis biochemical and cellular HTS assays showed that 112,872 compounds (14%) were consistently inactive in 100 assays. A permutation experiment showed that this was not a random chance effect. NIH Molecular Libraries campaigns also have many genuine inactives. The term "dark chemical matter" (DCM) has been coined for such compounds.

An analysis of 1,273 "dark" and 1,257 active compounds proved that intrinsic compound solution quality is not a factor in the inactivity, but it did reveal bad news about the quality of screening collections. Analysis of the properties of a Novartis set of compounds showed that DCM compounds are more soluble and less hydrophobic than actives, and they are smaller and have fewer rings. When the structural differences, if any, between DCM and actives were studied by multidimensional scaling it turned out that dark compounds and actives are not too different; dark compounds are not outliers in either Novartis or PubChem collections. Active compounds with dark substructures have lower hit rates; the nearest neighbors of actives near DCM tend to be more selective.

Anne Mai displayed some dark substructures; chemists thought that they looked fairly "innocent". She also showed some dark natural products that looked as if they should be active. Could dark compounds be valuable hits and potential tool compounds? Perhaps it could be that they seem inactive at typical screening concentrations. Novartis carried out an analysis of 34 additional high throughput screens in each of which at least 60,000 dark compounds were tested; previously active compounds yielded many more hits in these 34 screens than dark compounds, but, while 87% of the dark hits were hits in only one of the 34 screens, only 57% of previously active compounds showed this (i.e., 43 % of the hits from this compound class hit in more than one of the 34 screens). The difference between active and dark compounds was even greater when natural product compounds were tested at 10 micromolar rather than 1 micromolar in 37 cancer cell lines. This fits with the Novartis hypothesis about concentration.

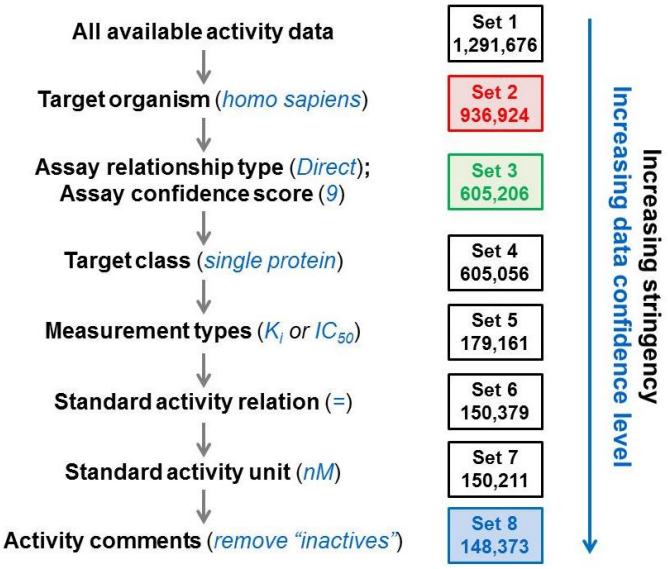
Experiments were next carried out covering broad biology. Of 1,408 compounds (704 dark and 704 active) submitted to 40 reporter gene assays, 92 actives but only 24 dark compounds were hits at 4 micromolar concentration. When a dark compound is active it is more selective. In a gene expression panel, 61 genes were measured for 188 dark compounds and 164 actives, at 1 micromolar and 10 micromolar concentrations. The results again supported the concentration hypothesis. The mechanism of action of a dark compound was elucidated in yeast HIP profiling; 200 dark compounds were tested against 6,000 heterozygous yeast strains, each with a different gene copy deleted. Some initial SAR studies with an antifungal panel have suggested a compound and analogues that were *in vitro* highly potent against *C. neoformans*, which causes fungal meningitis and encephalitis. The compound was clean against a human safety panel.

These experiments demonstrate that, when tested for the right phenotype or target, DCM can elicit strong biological responses. Consequently, Novartis believes that DCM is not generally biologically inert, and concludes that that their reduced promiscuity makes compounds from DCM a valuable resource for selective biological probes, and starting points for drug discovery programs.

Complexity and heterogeneity of data for chemical information science

Finally, Jürgen gave his award address. Similar to the situation in biology a few years ago, we currently witness the advent of the big data era in medicinal chemistry. UniChem, for example, now links 91 million compounds; there are 61 million compounds in PubChem. Increasing amounts of bioactivity data are available. ChEMBL has 13.5 million activity annotations for 1.5 million compounds and 10,774 biological targets. BindingDB has 1.1 million binding records for 495,128 compounds and 7,030 protein targets. PubChem has 60.7 million compounds, 1.15 million assays and 206,541 confirmatory assays. DrugBank records 7,759 drugs, 1,602 approved drugs, and 4,300 protein targets. What an opportunity all these data offer!

Rapidly growing compound numbers and volumes of activity data require elaborate infrastructures for deposition, curation, and organization, but the need for such infrastructures only partly reflects the challenges associated with big data phenomena. The "5Vs" are cited as criteria[113] for big data: volume, velocity, variety, veracity, and value. Jürgen believes that the increasing complexity and heterogeneity of compound data are additional challenges for computational analysis and knowledge extraction, and are probably even greater challenges than mere data volumes. To illustrate his point, Jürgen tabulated some data for trimeprazine and promethazine (closely related anti-allergic agents) in DrugBank, ChEMBL, BindingDB, and PubChem. Data incompleteness also applies in this example. Another criterion that could be added is data confidence. Jürgen took compound datasets from ChEMBL 18 to illustrate varying confidence levels:



Jürgen presented a ligand-centric view of promiscuity and the impact of data confidence. Evidence is mounting that polypharmacological drug behavior is often responsible for therapeutic efficacy, suggesting the consideration of new drug development strategies. Target promiscuity of compounds is at the origin of polypharmacology. For many bioactive compounds, multiple target annotations are available, indicating that compound promiscuity is a general phenomenon, but careful analysis of compound activity data reveals that the degree of apparent promiscuity is strongly influenced by data selection criteria and the type of activity measurements that are considered.[19] The average promiscuity rate of Jürgen's set 1 from ChEMBL was 6.7. The rate fell as confidence level increased;[114] the promiscuity rate of set 8 was only 1.5.

Jürgen's team has also studied compound promiscuity over time. Using sets 2, 3 and 8 from ChEMBL 20, they found that there has only been a minor increase in promiscuity over a great many years.[115] For the years 2004-1014, the promiscuity rate for set 2 has risen from about 1.8 to 2.5; for set 8 it has remained steady at about 1.5. It is interesting that approved drugs are more promiscuous. The promiscuity rate for a set 2 equivalent of approved drugs has risen from 5.9 in 2000 to 24.4 in 2014; for a set 8 it has risen from 1.9 to 3.7. The promiscuity rate of imatinib is particularly interesting: on the basis of low-confidence data, it has risen from 7 in 2004 to 690 in 2014! The high-confidence set 8 figure for 2014 is 27.

Global average promiscuity across five target families, GPCR class A, ion channels, kinases, nuclear receptors, and proteases is only 1.5 for sets of type 8 in ChEMBL 20. For example, one might have expected kinase inhibitors to be more promiscuous but they do not appear to be any more so than average if high confidence data levels are considered. Global average promiscuity does not vary a great deal around 1.5 as molecular weight and lipophilicity are varied, except in the case of compounds with molecular weight less than or equal to 200, where promiscuity is about 2.2.[115]

Ye Hu and Jürgen have also taken a target-centric view of promiscuity, derived from compound activity data.[116] The ability of target proteins to bind structurally diverse compounds and compounds with different degrees of promiscuity was systematically assessed on the basis of activity data and target annotations. Intuitive first- and second-order target promiscuity indices (TPIs) were introduced to quantify these binding characteristics and relate them to each other. TPI_1, the first-order target promiscuity index is calculated as the number of unique scaffolds of all compounds active against a given target; it indicates the ability of a target to interact with structurally diverse compounds. TPI_2, the second-order target promiscuity index, is the average degree of promiscuity of all compounds active against the target; it reflects the tendency of a target to interact with specific and promiscuous compounds.

The average TPI_1 value over all targets is 77 (for K_i data) and 61 (for IC_{50} data). This is not surprising: it is well known that many targets bind structurally diverse compounds. Only about 18% of all targets interact with compounds having no other reported activity ("pseudo-specific" compounds); here the TPI_2 value is 1. Most targets bind varying numbers of promiscuous compounds.

Targets that interact with compounds that are structurally diverse (more than 120 distinct scaffolds), but with no other reported activities, have high TPI_1 and low TPI_2. Examples are leukotriene A4 hydrolase and C-X-C chemokine receptor type 3. Targets that interact with compounds that are structurally homogeneous and preferentially promiscuous have low TPI_1 and high TPI_2. Examples are group IID secretory phospholipase A2 and matrix metalloproteinase 16. TPI_2 values establish the promiscuity profiles of target families; Jürgen showed some pie-charts of TPI_2 values for various target families.[116]

We are entering the big data era in chemical information science: compounds and activity data volumes, heterogeneity, and complexity are increasing. Data heterogeneity and inconsistency across databases is observed. Compound data mining offers significant opportunities for pharmaceutical R&D, but ensuring high data confidence and integrity is important. Promiscuity is the molecular basis of polypharmacology. Degrees of promiscuity vary with data confidence. Compound- and target-centric views of promiscuity can be taken.

Conclusion

After Jürgen's award address, Rachelle Bienstock, chair of the ACS Division of Chemical Information, formally presented the Herman Skolnik Award to Jürgen Bajorath:

References

(1) Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism. *J. Am. Chem. Soc.* 1997, *119* (43), 10509-10524.

(2) Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* 1988, *110* (18), 5959-5967.

(3) Andrade, C. H.; Pasqualoto, K. F. M.; Ferreira, E. I.; Hopfinger, A. J. 4D-QSAR: Perspectives in Drug Design. *Molecules* 2010, *15* (5), 3281.

(4) Albuquerque, M. G.; Hopfinger, A. J.; Barreiro, E. J.; de Alencastro, R. B. Four-dimensional quantitative structure-activity relationship analysis of a series of interphenylene 7-oxabicycloheptane oxazole thromboxane A2 receptor antagonists. *J. Chem. Inf. Comput. Sci.* 1998, *38* (5), 925-938.

(5) Rogers, D.; Hopfinger, A. J. Application of Genetic Function Approximation to Quantitative Structure-Activity Relationships and Quantitative Structure-Property Relationships. *J. Chem. Inf. Comput. Sci.* 1994, *34* (4), 854-866.

(6) Santos-Filho, O. A.; Hopfinger, A. J.; Cherkasov, A.; Bicca de Alencastro, R. The receptor-dependent QSAR paradigm: an overview of the current state of the art. *Med. Chem.* 2009, *5* (4), 359-366.

(7) Pan, D.; Liu, J.; Senese, C.; Hopfinger, A. J.; Tseng, Y. Characterization of a Ligand-Receptor Binding Event Using Receptor-Dependent Four-Dimensional Quantitative Structure-Activity Relationship Analysis. *J. Med. Chem.* 2004, *47* (12), 3075-3088.

(8) Pan, D.; Tseng, Y.; Hopfinger, A. J. Quantitative Structure-Based Design: Formalism and Application of Receptor-Dependent RD-4D-QSAR Analysis to a Set of Glucose Analogue Inhibitors of Glycogen Phosphorylase. *J. Chem. Inf. Comput. Sci.* 2003, *43* (5), 1591-1607.

(9) Cramer, R. D. Template CoMFA Generates Single 3D-QSAR Models that, for Twelve of Twelve Biological Targets, Predict All ChEMBL-Tabulated Affinities. *PLoS One* 2015, *10* (6), e0129307.
(10) Otter, T. Toward a New Theoretical Framework for Biology. In *Genetic and Evolutionary Computation Conference (GECCO) Workshop on Self-organization in Evolutionary Algorithms* Seattle, WA, June 26-30, 2004 http://www.cdres.com/content/GeccoTowardANewTheoreticalFrameworkForBiology_Otter_2004.pdf.
(11) Maggiora, G. M. The reductionist paradox: are the laws of chemistry and physics sufficient for the discovery of new drugs? *J. Comput.-Aided Mol. Des.* 2011, *25* (8), 699-708.
(12) Yildirim, M. A.; Goh, K.-I.; Cusick, M. E.; Barabasi, A.-L.; Vidal, M. Drug-target network. *Nat. Biotechnol.* 2007, *25* (10), 1119-1126.
(13) Paolini, G. V.; Shapland, R. H. B.; van Hoorn, W. P.; Mason, J. S.; Hopkins, A. L. Global mapping of pharmacological space. *Nat. Biotechnol.* 2006, *24* (7), 805-815.
(14) Vogt, I.; Mestres, J. Drug-target networks. *Mol. Inf.* 2010, *29* (1-2), 10-14.
(15) Mestres, J.; Gregori-Puigjane, E.; Valverde, S.; Sole, R. V. Data completeness-the Achilles heel of drug-target networks. *Nat. Biotechnol.* 2008, *26* (9), 983-984.
(16) Jalencas, X.; Mestres, J. On the origins of drug polypharmacology. *MedChemComm* 2013, *4* (1), 80-87.
(17) Hu, Y.; Bajorath, J. Promiscuity profiles of bioactive compounds: potency range and difference distributions and the relation to target numbers and families. *MedChemComm* 2013, *4* (8), 1196-1201.
(18) Hu, Y.; Bajorath, J. How Promiscuous Are Pharmaceutically Relevant Compounds? A Data-Driven Assessment. *AAPS J.* 2013, *15* (1), 104-111.
(19) Hu, Y.; Bajorath, J. Compound promiscuity: what can we learn from current data? *Drug Discovery Today* 2013, *18* (13-14), 644-650.
(20) Hu, Y.; Bajorath, J. Activity profile relationships between structurally similar promiscuous compounds. *Eur. J. Med. Chem.* 2013, *69*, 393-398.
(21) Kuhn, M.; Szklarczyk, D.; Pletscher-Frankild, S.; Blicher, T. H.; von Mering, C.; Jensen, L. J.; Bork, P. STITCH 4: integration of protein–chemical interactions with user data. *Nucleic Acids Res.* 2014, *42(Database issue):D401-407*.
(22) Roider, H. G.; Pavlova, N.; Kirov, I.; Slavov, S.; Slavov, T.; Uzunov, Z.; Weiss, B. Drug2Gene: an exhaustive resource to explore effectively the drug-target relation network. *BMC Bioinf.* 2014, *15*, 68.
(23) von Eichborn, J.; Murgueitio, M. S.; Dunkel, M.; Koerner, S.; Bourne, P. E.; Preissner, R. PROMISCUOUS: a database for network-based drug-repositioning. *Nucleic Acids Res.* 2011, *39* (Suppl. 1), D1060-D1066.
(24) Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* 2008, *36* (Database Iss), D901-D906.
(25) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 2012, *40* (D1), D1100-D1107.
(26) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Res.* 2007, *35* (suppl 1), D198-D201.
(27) Wang, Y.; Suzek, T.; Zhang, J.; Wang, J.; He, S.; Cheng, T.; Shoemaker, B. A.; Gindulyte, A.; Bryant, S. H. PubChem BioAssay: 2014 update. *Nucleic Acids Res.* 2014, *42* (D1), D1075-D1082.
(28) Salesse, S.; Verfaillie, C. M. BCR/ABL-mediated Increased Expression of Multiple Known and Novel Genes That May Contribute to the Pathogenesis of Chronic Myelogenous Leukemia. *Mol. Cancer Ther.* 2003, *2* (2), 173-182.

(29) Kim, T. M.; Ha, S. A.; Kim, H. K.; Yoo, J.; Kim, S.; Yim, S. H.; Jung, S. H.; Kim, D. W.; Chung, Y. J.; Kim, J. W. Gene expression signatures associated with the in vitro resistance to two tyrosine kinase inhibitors, nilotinib and imatinib. *Blood Cancer J.* 2011, *1*, e32.
(30) Moffat, J. G.; Rudolph, J.; Bailey, D. Phenotypic screening in cancer drug discovery - past, present and future. *Nat. Rev. Drug Discovery* 2014, *13* (8), 588-602.
(31) Rouvray, D. H. The evolution of the concept of molecular similarity. In *Concepts and Applications of Molecular Similarity;* Johnson, M. A., Maggiora, G. M., Eds.; John Wiley & Sons: New York, 1990; pp 15-42.
(32) Rouvray, D. H. Definition and role of similarity concepts in the chemical and physical sciences. *J. Chem. Inf. Comput. Sci.* 1992, *32* (6), 580-586.
(33) *Concepts and Applications of Molecular Similarity*. Johnson, M. A.; Maggiora, G. M., Eds.; John Wiley & Sons: New York, 1990.
(34) Wilkins, C. L.; Randic, M. A graph theoretical approach to structure-property and structure-activity correlations. *Theor. Chim. Acta* 1980, *58* (1), 45-68.
(35) Crum Brown, A. On the connection between chemical constitution and physiological action. *J. Anat. Physiol.* 1868, *2* (2), 224-242.
(36) Tobler, W. R. A computer movie simulating urban growth in the Detroit region. *Econ. Geogr.* 1970, *46*, 234-240.
(37) McPherson, M.; Smith-Lovin, L.; Cook, J. Birds of a feather: homophily in social networks. *Ann. Rev. Sociol.* 2001, *27*, 415-444.
(38) van Rijsbergen, C. J. *Information Retrieval*; Butterworth: London, 1979.
(39) Harrison, P. J. A method of cluster analysis and some applications. *Appl. Stat.* 1968, *17*, 226-236.
(40) Adamson, G. W.; Bush, J. A. Method for the automatic classification of chemical structures. *Inf. Storage Retr.* 1973, *9* (10), 561-568.
(41) Adamson, G. W.; Bush, J. A. Comparison of the performance of some similarity and dissimilarity measures in the automatic classification of chemical structures. *J. Chem. Inf. Comput. Sci.* 1975, *15* (1), 55-58.
(42) Willett, P.; Winterman, V. A comparison of some measures for the determination of inter-molecular structural similarity: measures of inter-molecular structural similarity. *Quant. Struct.-Act. Relat.* 1986, *5* (1), 18-25.
(43) Willett, P.; Winterman, V.; Bawden, D. Implementation of nonhierarchic cluster analysis methods in chemical information systems: selection of compounds for biological testing and clustering of substructure search output. *J. Chem. Inf. Comput. Sci.* 1986, *26* (3), 109-118.
(44) Willett, P.; Winterman, V.; Bawden, D. Implementation of nearest-neighbor searching in an online chemical structure search system. *J. Chem. Inf. Comput. Sci.* 1986, *26* (1), 36-41.
(45) Ward, J. H. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* 1963, *58*, 236-244.
(46) Jarvis, R. A.; Patrick, E. A. Clustering using a similarity measure based on shared nearest neighbors. *IEEE Transactions on Computers* 1973, *C-22*, 1025-1034.
(47) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: definition and applications. *J. Chem. Inf. Comput. Sci.* 1985, *25* (2), 64-73.
(48) Bawden, D. Browsing and clustering of chemical structures. In *Chemical Structures;* Warr, W. A., Ed.; Springer Verlag: Berlin, 1986; pp 145-150.
(49) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical Similarity Using Physiochemical Property Descriptors. *J. Chem. Inf. Comput. Sci.* 1996, *36* (1), 118-127.
(50) Sheridan, R. P.; Miller, M. D.; Underwood, D. J.; Kearsley, S. K. Chemical Similarity Using Geometric Atom Pair Descriptors. *J. Chem. Inf. Comput. Sci.* 1996, *36* (1), 128-136.

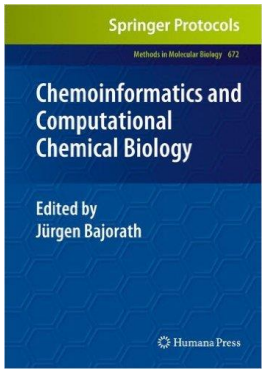
(51) Sheridan, R. P. Chemical similarity searches: when is complexity justified? *Expert Opin. Drug Discovery* 2007, *2* (4), 423-430.
(52) Ginn, C. M. R.; Turner, D. B.; Willett, P.; Ferguson, A. M.; Heritage, T. W. Similarity Searching in Files of Three-Dimensional Chemical Structures: Evaluation of the EVA Descriptor and Combination of Rankings Using Data Fusion. *J. Chem. Inf. Comput. Sci.* 1997, *37* (1), 23-37.
(53) Ginn, C. M. R.; Willett, P.; Bradshaw, J. Combination of molecular similarity measures using data fusion. *Perspect. Drug Discovery Des.* 2000, *20*, 1-16.
(54) Willett, P. Enhancing the effectiveness of ligand-based virtual screening using data fusion. *QSAR Comb. Sci.* 2006, *25* (12), 1143-1152.
(55) Willett, P. Combination of Similarity Rankings Using Data Fusion. *J. Chem. Inf. Model.* 2013, *53* (1), 1-10.
(56) Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. Measuring diversity: experimental design of combinatorial libraries for drug discovery. *J. Med. Chem.* 1995, *38* (9), 1431-1436.
(57) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood behavior: a useful concept for validation of "molecular diversity" descriptors. *J. Med. Chem.* 1996, *39* (16), 3049-3059.
(58) Lajiness, M. S.; Johnson, M. A.; Maggiora, G. M. Implementing drug screening programs using molecular similarity. In *QSAR: Quantitaive Structure-activity Relationships in Drug Design;* Fauchere, J. L., Ed.; Alan R. Liss: New York, 1989; pp 173-176.
(59) Snarey, M.; Terrett, N. K.; Willett, P.; Wilton, D. J. Comparison of algorithms for dissimilarity-based compound selection. *J. Mol. Graphics Modell.* 1998, *15* (6), 372-385.
(60) Allen, F. H.; Davies, J. E.; Galloy, J. J.; Johnson, O.; Kennard, O.; Macrae, C. F.; Mitchell, E. M.; Mitchell, G. F.; Smith, J. M.; Watson, D. G. The development of versions 3 and 4 of the Cambridge Structural Database System. *J. Chem. Inf. Comput. Sci.* 1991, *31* (2), 187-204.
(61) Klebe, G.; Abraham, U.; Mietzner, T. Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules to Correlate and Predict Their Biological Activity. *J. Med. Chem.* 1994, *37* (24), 4130-4146.
(62) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* 1998, *38* (6), 983-996.
(63) Tropsha, A.; Gramatica, P.; Gombar, V. K. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci.* 2003, *22* (1), 69-77.
(64) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* 1996, *39* (15), 2887-2893.
(65) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discovery* 2004, *3* (11), 935-949.
(66) Aruffo, A.; Farrington, M.; Hollenbaugh, D.; Li, X.; Milatovich, A.; Nonoyama, S.; Bajorath, J.; Grosmaire, L. S.; Stenkamp, R.; et, a. The CD40 ligand, gp39, is defective in activated T cells from patients with X-linked hyper-IgM syndrome. *Cell (Cambridge, Mass.)* 1993, *72* (2), 291-300.
(67) Linsley, P. S.; Greene, J. L.; Brady, W.; Bajorath, J.; Ledbetter, J.; Peach, R. Human B7-1 (CD80) and B7-2 (CD86) bind with similar avidities but distinct kinetics of CD28 and CTLA-4 receptors. *Immunity* 1994, *1* (9), 793-801.
(68) Foy, T. M.; Aruffo, A.; Bajorath, J.; Buhlmann, J. E.; Noelle, R. J. Immune regulation by CD40 and its ligand gp39. *Annu. Rev. Immunol.* 1996, *14*, 591-617.
(69) Sica, G. L.; Choi, I.-H.; Zhu, G.; Tamada, K.; Wang, S.-D.; Tamura, H.; Chapoval, A. I.; Flies, D. B.; Bajorath, J.; Chen, L. B7-H4, a molecule of the B7 family, negatively regulates T cell immunity. *Immunity* 2003, *18* (6), 849-861.

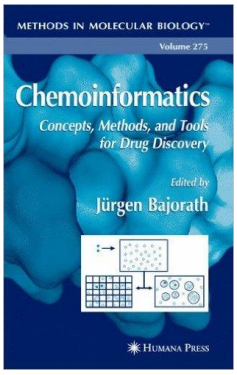
(70) Bishop, C. M.; Svensén, M.; Williams, C. K. I. GTM: The Generative Topographic Mapping. *Neural Comput.* 1998, *10* (1), 215-234.
(71) Chupakhin, V.; Marcou, G.; Gaspar, H.; Varnek, A. Simple Ligand-Receptor Interaction Descriptor (SILIRID) for alignment-free binding site comparison. *Comput. Struct. Biotechnol. J.* 2014, *10* (16), 33-7.
(72) Kireeva, N.; Baskin, I. I.; Gaspar, H. A.; Horvath, D.; Marcou, G.; Varnek, A. Generative Topographic Mapping (GTM): Universal Tool for Data Visualization, Structure-Activity Modeling and Dataset Comparison. *Mol. Inf.* 2012, *31* (3-4), 301-312.
(73) Varnek, A.; Fourches, D.; Hoonakker, F.; Solov'ev, V. P. Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures. *J. Comput.-Aided Mol. Des.* 2005, *19* (9/10), 693-703.
(74) Varnek, A.; Fourches, D.; Horvath, D.; Klimchuk, O.; Gaudin, C.; Vayer, P.; Solov'ev, V.; Hoonakker, F.; Tetko, I. V.; Marcou, G. ISIDA - platform for virtual screening based on fragment and pharmacophoric descriptors. *Curr. Comput.-Aided Drug Des.* 2008, *4* (3), 191-198.
(75) Horvath, D.; Bonachera, F.; Solov'ev, V.; Gaudin, C.; Varnek, A. Stochastic versus Stepwise Strategies for Quantitative Structure−Activity Relationship GenerationHow Much Effort May the Mining for Successful QSAR Models Take? *J. Chem. Inf. Model.* 2007, *47* (3), 927-939.
(76) Gaspar, H. A.; Baskin, I. I.; Marcou, G.; Horvath, D.; Varnek, A. GTM-Based QSAR Models and Their Applicability Domains. *Mol. Inf.* 2015, *34* (6-7), 348-356.
(77) Gaspar, H. A.; Marcou, G.; Horvath, D.; Arault, A.; Lozano, S.; Vayer, P.; Varnek, A. Generative Topographic Mapping-Based Classification Models and Their Applicability Domain: Application to the Biopharmaceutics Drug Disposition Classification System (BDDCS). *J. Chem. Inf. Model.* 2013, *53* (12), 3318-3325.
(78) Gaspar, H. A.; Baskin, I. I.; Marcou, G.; Horvath, D.; Varnek, A. Chemical Data Visualization and Analysis with Incremental Generative Topographic Mapping: Big Data Challenge. *J. Chem. Inf. Model.* 2015, *55* (1), 84-94.
(79) Funatsu, K. Computer-aided synthesis design and reaction prediction. *Kagaku Kogyo* 2007, *58* (2), 124-129.
(80) Masuda, Y.; Kaneko, H.; Funatsu, K. Multivariate Statistical Process Control Method Including Soft Sensors for Both Early and Accurate Fault Detection. *Ind. Eng. Chem. Res.* 2014, *53* (20), 8553-8564.
(81) Kaneko, H.; Funatsu, K. Database monitoring index for adaptive soft sensors and the application to industrial process. *AIChE J.* 2014, *60* (1), 160-169.
(82) Schneider, G.; Fechner, U. Computer-based de novo design of drug-like molecules. *Nat. Rev. Drug Discovery* 2005, *4* (8), 649-663.
(83) Alig, L.; Alsenz, J.; Andjelkovic, M.; Bendels, S.; Benardeau, A.; Bleicher, K.; Bourson, A.; David-Pierson, P.; Guba, W.; Hildbrand, S.; Kube, D.; Luebbers, T.; Mayweg, A. V.; Narquizian, R.; Neidhart, W.; Nettekoven, M.; Plancher, J.-M.; Rocha, C.; Rogers-Evans, M.; Roever, S.; Schneider, G.; Taylor, S.; Waldmeier, P. Benzodioxoles: Novel Cannabinoid-1 Receptor Inverse Agonists for the Treatment of Obesity. *J. Med. Chem.* 2008, *51* (7), 2115-2127.
(84) Hartenfeller, M.; Schneider, G. De novo drug design. *Methods Mol. Biol. (N. Y., NY, U. S.)* 2011, *672*, 299-323.
(85) Hartenfeller, M.; Zettl, H.; Walter, M.; Rupp, M.; Reisen, F.; Proschak, E.; Weggen, S.; Stark, H.; Schneider, G. DOGS: reaction-driven de novo design of bioactive compounds. *PLoS Comput. Biol.* 2012, *8* (2), e1002380.
(86) Spaenkuch, B.; Keppner, S.; Lange, L.; Rodrigues, T.; Zettl, H.; Koch, C. P.; Reutlinger, M.; Hartenfeller, M.; Schneider, P.; Schneider, G. Drugs by Numbers: Reaction-Driven De Novo Design of Potent and Selective Anticancer Leads. *Angew. Chem., Int. Ed.* 2013, *52* (17), 4676-4681.

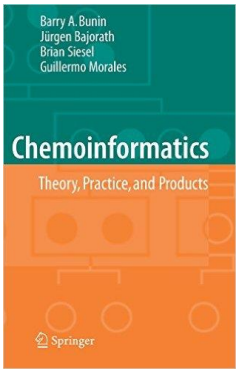
(87) Rupp, M.; Proschak, E.; Schneider, G. Kernel approach to molecular similarity based on iterative graph similarity. *J. Chem. Inf. Model.* 2007, *47* (6), 2280-2286.
(88) Rupp, M.; Schneider, G. Graph Kernels for Molecular Similarity. *Mol. Inf.* 2010, *29* (4), 266-273.
(89) Reutlinger, M.; Rodrigues, T.; Schneider, P.; Schneider, G. Multi-Objective Molecular De Novo Design by Adaptive Fragment Prioritization. *Angew. Chem., Int. Ed.* 2014, *53* (16), 4244-4248.
(90) Reutlinger, M.; Rodrigues, T.; Schneider, P.; Schneider, G. Combining On-Chip Synthesis of a Focused Combinatorial Library with Computational Target Prediction Reveals Imidazopyridine GPCR Ligands. *Angew. Chem., Int. Ed.* 2014, *53* (2), 582-585.
(91) Rodrigues, T.; Schneider, P.; Schneider, G. Accessing New Chemical Entities through Microfluidic Systems. *Angew. Chem., Int. Ed.* 2014, *53* (23), 5750-5758.
(92) Rodrigues, T.; Hauser, N.; Reker, D.; Reutlinger, M.; Wunderlin, T.; Hamon, J.; Koch, G.; Schneider, G. Multidimensional De Novo Design Reveals 5-HT2B Receptor-Selective Ligands. *Angew. Chem., Int. Ed.* 2015, *54* (5), 1551-1555.
(93) Reker, D.; Rodrigues, T.; Schneider, P.; Schneider, G. Identifying the macromolecular targets of de novo-designed chemical entities through self-organizing map consensus. *Proc. Natl. Acad. Sci. U. S. A.* 2014, *111* (11), 4067-4072.
(94) Reutlinger, M.; Koch, C. P.; Reker, D.; Todoroff, N.; Schneider, P.; Rodrigues, T.; Schneider, G. Chemically Advanced Template Search (CATS) for Scaffold-Hopping and Prospective Target Prediction for Orphan' Molecules. *Mol. Inf.* 2013, *32* (2), 133-138.
(95) Grabowski, K.; Baringhaus, K.-H.; Schneider, G. Scaffold diversity of natural products. Inspiration for combinatorial library design. *Nat. Prod. Rep.* 2008, *25* (5), 892-904.
(96) Reker, D.; Perna, A. M.; Rodrigues, T.; Schneider, P.; Reutlinger, M.; Monch, B.; Koeberle, A.; Lamers, C.; Gabler, M.; Steinmetz, H.; Muller, R.; Schubert-Zsilavecz, M.; Werz, O.; Schneider, G. Revealing the macromolecular targets of complex natural products. *Nat. Chem.* 2014, *6* (12), 1072-1078.
(97) Rodrigues, T.; Reker, D.; Kunze, J.; Schneider, P.; Schneider, G. Revealing the Macromolecular Targets of Fragment-Like Natural Products. *Angew. Chem., Int. Ed.* 2015, *54* (36), 10516-10520.
(98) Tyrchan, C.; Bostroem, J.; Giordanetto, F.; Winter, J.; Muresan, S. Exploiting Structural Information in Patent Specifications for Key Compound Prediction. *J. Chem. Inf. Model.* 2012, *52* (6), 1480-1489.
(99) Wawer, M. J.; Jaramillo, D. E.; Dancik, V.; Fass, D. M.; Haggarty, S. J.; Shamji, A. F.; Wagner, B. K.; Schreiber, S. L.; Clemons, P. A. Automated structure-activity relationship mining: connecting chemical structure to biological profiles. *J. Biomol. Screening* 2014, *19* (5), 738-748.
(100) Peltason, L.; Hu, Y.; Bajorath, J. From Structure-Activity to Structure-Selectivity Relationships: Quantitative Assessment, Selectivity Cliffs, and Key Compounds. *ChemMedChem* 2009, *4* (11), 1864-1873.
(101) Iyer, P.; Hu, Y.; Bajorath, J. SAR Monitoring of Evolving Compound Data Sets Using Activity Landscapes. *J. Chem. Inf. Model.* 2011, *51* (3), 532-540.
(102) Iyer, P.; Stumpfe, D.; Bajorath, J. Molecular Mechanism-Based Network-like Similarity Graphs Reveal Relationships between Different Types of Receptor Ligands and Structural Changes that Determine Agonistic, Inverse-Agonistic, and Antagonistic Effects. *J. Chem. Inf. Model.* 2011, *51* (6), 1281-1286.
(103) Wassermann, A. M.; Haebel, P.; Weskamp, N.; Bajorath, J. SAR Matrices: Automated Extraction of Information-Rich SAR Tables from Large Compound Data Sets. *J. Chem. Inf. Model.* 2012, *52* (7), 1769-1776.
(104) Gupta-Ostermann, D.; Shanmugasundaram, V.; Bajorath, J. Neighborhood-Based Prediction of Novel Active Compounds from SAR Matrices. *J. Chem. Inf. Model.* 2014, *54* (3), 801-809.

(105) Agrafiotis, D. K.; Shemanarev, M.; Connolly, P. J.; Farnum, M.; Lobanov, V. S. SAR Maps: A New SAR Visualization Technique for Medicinal Chemists. *J. Med. Chem.* 2007, *50* (24), 5926-5937.
(106) Peltason, L.; Weskamp, N.; Teckentrup, A.; Bajorath, J. Exploration of Structure-Activity Relationship Determinants in Analogue Series. *J. Med. Chem.* 2009, *52* (10), 3212-3224.
(107) Zhang, B.; Hu, Y.; Bajorath, J. AnalogExplorer: A New Method for Graphical Analysis of Analog Series and Associated Structure-Activity Relationship Information. *J. Med. Chem.* 2014, *57* (21), 9184-9194.
(108) Lounkine, E.; Nigsch, F.; Jenkins, J. L.; Glick, M. Activity-Aware Clustering of High Throughput Screening Data and Elucidation of Orthogonal Structure-Activity Relationships. *J. Chem. Inf. Model.* 2011, *51* (12), 3158-3168.
(109) Petrone, P. M.; Simms, B.; Nigsch, F.; Lounkine, E.; Kutchukian, P.; Cornett, A.; Deng, Z.; Davies, J. W.; Jenkins, J. L.; Glick, M. Rethinking Molecular Similarity: Comparing Compounds on the Basis of Biological Activity. *ACS Chem. Biol.* 2012, *7* (8), 1399-1409.
(110) Wassermann, A. M.; Lounkine, E.; Glick, M. Bioturbo Similarity Searching: Combining Chemical and Biological Similarity To Discover Structurally Diverse Bioactive Molecules. *J. Chem. Inf. Model.* 2013, *53* (3), 692-703.
(111) Petrone, P. M.; Wassermann, A. M.; Lounkine, E.; Kutchukian, P.; Simms, B.; Jenkins, J.; Selzer, P.; Glick, M. Biodiversity of small molecules - a new perspective in screening set selection. *Drug Discovery Today* 2013, *18* (13-14), 674-680.
(112) Wawer, M. J.; Li, K.; Gustafsdottir, S. M.; Ljosa, V.; Bodycombe, N. E.; Marton, M. A.; Sokolnicki, K. L.; Bray, M.-A.; Kemp, M. M.; Winchester, E.; Taylor, B.; Grant, G. B.; Hon, C. S.-Y.; Duvall, J. R.; Wilson, J. A.; Bittker, J. A.; Dancik, V.; Narayan, R.; Subramanian, A.; Winckler, W.; Golub, T. R.; Carpenter, A. E.; Shamji, A. F.; Schreiber, S. L.; Clemons, P. A. Toward performance-diverse small-molecule libraries for cell-based phenotypic screening using multiplexed high-dimensional profiling. *Proc. Natl. Acad. Sci. U. S. A.* 2014, *111* (30), 10911-10916.
(113) Lusher, S. J.; McGuire, R.; van Schaik, R. C.; Nicholson, C. D.; de Vlieg, J. Data-driven medicinal chemistry in the era of big data. *Drug Discovery Today* 2014, *19* (7), 859-868.
(114) Hu, Y.; Bajorath, J. Influence of Search Parameters and Criteria on Compound Selection, Promiscuity, and Pan Assay Interference Characteristics. *J. Chem. Inf. Model.* 2014, *54* (11), 3056-3066.
(115) Hu, Y.; Jasial, S.; Bajorath, J. Promiscuity progression of bioactive compounds over time. *F1000Res* 2015, *4* (Chem. Inf. Sci.), 118.
(116) Hu, Y.; Bajorath, J. Quantifying the tendency of therapeutic target proteins to bind promiscuous or selective compounds. *PLoS One* 2015, *10* (5), e0126838.

Wendy Warr, Symposium Reporter







Scientific Integrity: Can We Rely on the Published Scientific Literature?

Ever since the internet became the primary means of disseminating scientific research, scientific publishing has lurched from crisis to crisis, with seemingly increasing rates of fraud, plagiarism, retraction and selective publications. There are many reasons behind this, including the pressure on researchers to publish, and easier ways for publishers and academics to identify plagiarism. At a time when the ACS has created its statement “Scientific insight and integrity in public policy,” it seemed appropriate to organize a symposium on scientific integrity, with a particular focus on the degree to which we can rely on the published scientific literature. This symposium is one of a series on related topics and follows the symposium on data reproducibility organized by Martin Hicks at the spring ACS National Meeting in Denver.

The first speaker of the morning was Chris Leonard from QScience, who was speaking in his capacity as a member of the Committee on Publishing Ethics (COPE). His talk was entitled “Integrity, ethics and trust in scientific research literature.” He explained that the new publishing landscape is often defined in terms of technology and the “bells and whistles” that can improve many aspects of a manuscript, but it also offers the potential of a new era of ethics, integrity and reproducibility, especially in peer review and emerging publication areas such as datasets. The price is increased vigilance at the authoring and reviewing stages, but the cost of not doing so is incalculable. Chris explained COPE’s role in assisting publishers, editors, and authors in this task.

Our second speaker was Chris Proctor from British American Tobacco (BAT) who spoke on “Policy making at the American Chemical Society: developing a statement on scientific integrity.” His presentation examined, from the perspective of a member of the writing committee, how the American Chemical Society created its statement “Scientific insight and integrity in public policy.” Acting with integrity is absolutely critical to the scientific process, so much so that this concept has been embedded in government policies around the world. Leading scientific membership associations have in turn created their own policies to support these initiatives and their members in best practice. The second half of his talk focused on the importance of integrity in controversial areas of science, in particular the case of tobacco harm reduction.

The last speaker before the interval was Martin Hicks of the Beilstein Institute whose subject was “publishability.” He explained that publishing scientific papers has several functions: registration, certification, dissemination, inspiring innovation and archiving. The scientific community is then expected over time to verify the ideas and results and in the end uncover the objective truth. It is assumed that authors make best efforts to ensure that their submissions are correct, and that the scientific community has the time and resources to concern itself with review, reproduction, and validation. Nowadays, most people are, in principle, able to get access to individual articles that they need, but the amount of information has increased so much that scientists cannot keep up with all the publications in their own area of expertise, except for just skimming through the tables of contents. To maximize their reputation, scientists are expected to publish a certain number of papers per year in journals having a certain minimum Impact Factor. Having a significant new idea is no longer sufficient: the numbers are what matters, behavior adapts to the system, and output is adjusted accordingly. It seems likely that this is also leading to the increase in plagiarism and other unethical behavior. Martin’s presentation discussed the effects of this increasingly problematical “publish or perish” paradigm and was illustrated with experiences gained during the publishing of the journals of the Beilstein Institute.

After the interval, Na Qin of Michigan State University gave a librarian's perspective with a presentation entitled "What is the role of peer review in protecting the integrity of scientific research?" Scientific misconduct, such as plagiarism, data manipulation, data fabrication or duplicate publications, occurs with distressing frequency in science communities. The peer review system has long been used as a self-regulation approach to maintain the standards of quality, improve performance, and provide credibility, but fraudulent and flawed research has been published even in peer-reviewed journals, and the number of articles retracted for fraud or error has risen dramatically in the last decade. This presentation considered questions such as "Is detecting scientific misconduct or errors a primary goal of peer review?" and "What is the role of peer review in ensuring the responsible conduct of scientific research?" and addressed the difficulties and limitations of anonymous peer review in detecting irresponsible conduct in scientific research. These include the naturally conflicting concepts from peer review: expertise and objectivity, and the capacity to expose or minimize legal or ethical issues. Na's intention was to start a conversation on whether peer review is, by its nature, ill-equipped to detect scientific misconduct. The practice of sciences involves its own self-corrections, and the peer review system does not replace that. Understanding this can reduce the expense to researchers who try to use or replicate fabricated results.

Next, Rajeev Voleti, ScienceOpen, who was a last-minute replacement for Stephanie Dawson, presented Stephanie's talk entitled "An open, network-based answer to the reproducibility crisis: the ScienceOpen peer review concept." Spectacular failures of the anonymous peer review system, even in highly prestigious journals, paired with research demonstrating extremely low levels of reproducibility in landmark studies, have called the present system of scientific quality assurance into question. To create a more effective, transparent and fairer system that begins to address the question of reproducibility, ScienceOpen was developed to provide a networking and publishing platform. A researcher network forms the basis for public post-publication peer review, and as a transparent network approach, provides more rigorous quality control than two anonymous referees.

Articles submitted to ScienceOpen are published rapidly after an editorial check, followed by an open peer review process. A unique versioning concept allows researchers to continue to improve their published work, based on comments and reviews by scientists in the field. Papers are not marked as approved because information on the reproducibility of experiments comes later than the first expert check, and thus the status of a paper may change. An article published on ScienceOpen is also placed within the wider context of all Open Access publications in its field, as ScienceOpen aggregates content from a variety of sources, opening them up to discussion with the same tools for commenting, sharing, and discovery. With this holistic concept, ScienceOpen provides high-quality open access publishing services, while redefining publishing as one element in a whole suite of communication tools available to the researcher. Scholarly publishing is not an end in itself, but the beginning of a dialogue to move the whole field forward.

The last speaker of the morning was my co-chair and co-organizer of the symposium, Judith Currano, chemistry librarian at the University of Pennsylvania, who discussed "Managing new threats to the integrity of the scientific literature." This paper, co-authored by a professor who edits an online journal, framed the challenges facing scientists at all levels, as a result of the highly variable quality of the scientific literature resulting from the introduction of a deluge of new open access online journals, many from previously unknown publishers with highly variable standards of peer review. The problems are so pervasive that even papers submitted to well-established, legitimate journals may include citations to questionable or even frankly plagiarized sources. Judith suggested ways in which science librarians can work with students and researchers to increase their awareness of these new threats to the integrity of the scientific literature, and to increase their ability to evaluate the reliability

of journals and individual articles. Traditional rules of thumb for assessing the reliability of scientific publications (peer review, publication in a journal with an established Thomson-Reuters Impact Factor, and a credible publisher) are more challenging to apply given the highly variable quality of many of the new open access journals, the appearance of new publishers, and the introduction of new impact metrics, some of which are interesting and useful, but others of which are based on citation patterns found in poorly described data sets or nonselective databases of articles. The authors suggested that instruction of research students in responsible conduct of research be extended to include ways to evaluate the reliability of scientific information.

After the lunch break, Cesar Berrios of Faculty of 1000 gave a talk entitled “Towards a more reproducible corpus of scientific literature.” Several recent reports and high profile retractions have added to a growing chorus of concern among scientists and laypeople clamoring for a restructuring of the system necessary for reproducibility in science, but the problem is a complex one for which there is no single solution. Some factors that have contributed include: poor training in proper experimental design; increased emphasis on making outlandish statements; and an overreliance on publishing papers in peer reviewed journals with high impact factors for purposes of career progression and tenure. The availability and accessibility of all underlying data necessary to reproduce a study has been identified as integral to solving these issues, yet most traditional journals often have limited space available for each paper. Furthermore, there are numerous technical obstacles in making datasets truly accessible. These issues combine to create a scientific culture where sharing and publishing data ends up low on a researcher’s list of priorities, impeding further progress towards reproducible research. F1000Research are addressing some of these challenges. They have implemented several initiatives to provide methods and tools to capture the production of scientific data, and to establish this as an important output of research activity in itself. All F1000Research articles include the underlying data to enable others to attempt to reproduce the findings, and even to reuse the data. Authors are also offered the option to publish data-only papers that include just the data, together with a detailed description of the protocol used to generate the data. In addition, all articles are openly peer reviewed, post-publication, and previous versions of each article are archived. Cesar described how F1000’s data policy and transparency in the peer review process allows reviewers and readers to scrutinize the data underlying the conclusions carefully, and to follow the full provenance of each paper, ultimately leading to a more trustworthy corpus of scientific literature.

The next speaker, James Solyst of Swedish Match, described “Extraordinary public access to scientific evidence in the FDA modified risk tobacco product process.” Section 911 of the Federal Food, Drug and Cosmetic Act - Modified Risk Tobacco Product (MRTP) provides a process for a company to submit scientific evidence demonstrating a product is of lower risk (modified risk) than another tobacco product. A MRTP application must demonstrate that by switching from one product (cigarettes for example) to another product (Swedish snus, for example), a user reduces his or her individual risk, and the switch benefits the health of the overall population. Reducing harm is generally accepted as a good thing, but tobacco use is widely viewed as something to be eliminated. In addition, the tobacco industry has a challenging history of scientific integrity. Thus, the MRTP process is filled with difficult public health issues, and the FDA, which implements the Tobacco Control Act, has had to manage a process that ensures scientific integrity and is consistent with public health goals. One way that this has been accomplished is through extraordinary transparency: specifically, by making all information in an MRTP publicly available (except for confidential business information). This is not the case with other FDA product applications. A current MRTP application for Swedish snus provides a case study that was reviewed by James.

The next two talks before the afternoon interval discussed fraud and integrity in crystallographic journals and databases. Sean Conway (International Union of Crystallography, IUCr) presented "Validation and fraud in small-molecule crystallography." Publishing in crystallography is underpinned by a wealth of structural data. In IUCr journals, submitted data are rigorously checked for correctness and consistency. Even so, the journals have experienced cases of fraud in small-molecule reports. In-house validation software is constantly evolving to guard against a growing variety of egregious errors and fraudulent practice. Ian Bruno of Cambridge Crystallographic Data Centre (CCDC) discussed "Scientific integrity: a crystallographic perspective." The Cambridge Structural Database (CSD) contains over 750,000 experimental determinations of small molecule crystal structures, the majority of which are made available by researchers to support the science published in journal articles. The crystallographic community has, over the years, developed tools that support evaluation of the scientific integrity of crystal structure data and some publishers and journals pay particular attention to this during the peer review process. Once structures are published, CCDC undertakes further scientific processing of the data before including structures in the CSD. This presentation offered a perspective on scientific integrity based on crystal structure data collected over the last half century, and experiences encountered during this time. It also looked at the role domain-specific data centers such as the CCDC can play now, and in the future, to help ensure trust in the results of scientific research.

The last three presentations considered the important roles that publishers play. Ray Boucher from Wiley described "The ways publishers help, maintain, and support responsible research." At all stages of the publishing process, including pre-publication and post-publication, the publisher is engaged in helping to maintain the integrity of the scientific record. The talk covered areas where the publisher is involved, such as publishing workshops and how the next generation are trained; plagiarism and other pre-publication software packages for specific communities; maintaining the quality of peer review; ethics guidelines; bodies such as the Committee on Publication Ethics (COPE); and how issues are dealt with; and an analysis of the process and practice of retractions. The talk illustrated how the publisher supports and promotes the publication of responsible research, and how interaction with the community is key to this process.

The second speaker in this group, Guido Hermann of Thieme, gave a talk entitled "Integrity, trust and reproducibility: how scientific publishers can contribute." Thieme publishes scientific information in various formats: journals, reference works, encyclopedias, monographs and textbooks. Scientists have to rely on the validity of the published information and Guido described how Thieme addresses this issue, the internal procedures and mechanisms to safeguard the quality of their publications, and Thieme's experiences with fraud and plagiarism. The talk considered questions such as "How do we engage our authors, editors, advisors and readers in this process?" and "Are there differences between original research articles, reference works and textbooks?" Guido presented background information, and highlighted some of Thieme's key findings and best practices.

Finally, Richard Kidd of the Royal Society of Chemistry (RSC), presented "The write stuff: scientific integrity and publishing," and considered the responsibilities of a publisher in addressing the questions of scientific integrity, and how is this changing. He described the RSC's principles and practices, and how RSC works with our community worldwide to evolve their approaches. He considered how the increasing push towards the availability of original data makes validation easier, and the exact meaning of "reproducible."

William Town, Symposium Organizer

A Joint CINF-CSA Trust Symposium

The symposium, chaired by Grace Baysinger (Stanford University) and Jonathan Goodman (University of Cambridge), opened with a presentation by Charles Huber (University of California, Santa Barbara) on the "Chemical Information Sources Wikibook: the open source created by chemical information professionals for chemical information professionals." He related the origins of the resource in the printed reference created by Gary Wiggins (Indiana University) in 1991, and its subsequent migrations to the Internet and to the Wikibook platform, concluding with a description of its current format and plans for the future, with a call to interested parties to contribute.

Donna Wrublewski (Caltech) and Neelam Bharti (University of Florida) presented a paper "Soft skills of chemical research: academic integrity and research ethics." They defined the concepts involved: academic integrity, which includes areas of authorship, plagiarism, responsible conduct of research and conflicts of interest; and research ethics, including honestly, objectivity, integrity, openness, respect for intellectual property, and responsible publication, among others. They then described a program created at the University of Florida, by the library, in partnership with the Office of Research and Office of Undergraduate Research, that reached out to undergraduate, graduate in master's and Ph.D. programs, and to distance learning students, to introduce these concepts.

"Integrating bibliographic management tools in chemical information literacy instruction" by Svetla Baykoucheva and Joseph Jouck (both University of Maryland), related to their incorporation of bibliographic management tools (EndNote Online, Mendeley, and Zotero) into a chemistry research assignment, using Web of Science, PubMed and SciFinder. Teaching assistants for the large classes were trained on the use of the tools, and students also had access to online tutorials and guided assignments. (Note: Svetla Baykoucheva is the author of *Managing Scientific Information and Research Data*, Elsevier, 2015. ISBN: 978-0-08-100195-0.)

Vincent Scalfani (Universtiy of Alabama) spoke on "Replacing the traditional graduate chemistry literature seminar with a chemical information literacy course" (co-authored with Stephen Woski and Patrick Frantom, both also University of Alabama). They developed a course (CH584: Literature and Communication in Graduate Chemistry) for second year graduate students that incorporated the presentation of a literature seminar with instruction on chemistry information resources, critical analysis, scientific writing and presentation, and peer review, among others. The students learned more about the basic skills, and did a superior job on their presentations. The course was deemed so successful that they are developing an equivalent course for the Chemical and Biological Engineering program.

Elaine Cheeseman of the ScienceIP service of Chemical Abstracts Service described the work of a professional chemical information searcher entitled "Chemical information skills: a searcher's perspective." Professional searching requires not only expertise in sophisticated search tools, but also skills in analyzing a client's needs and in generating reports that can distill a huge mass of data into a form that the client finds useful. Collaboration at every stage of the process is crucial. Elaine stressed the two key qualities of the professional searcher, namely: meticulous attention to detail, and a desire to learn.

The next presentation "Patents: the essential multifunctional tool for science, business and intellectual property information" by Edlyn Simmons (Simmons Patent Information Service, LLC) described the

nature, content, and uses of patents. She likened patents to a Swiss army knife, able to reveal not only technical, but also competitive intelligence and legal information to the user. The description of the invention allows users to build on the substances and techniques described, while the claims delineate what intellectual property is legally protected, and the information about inventors and assignees can be valuable to analyzing trends in industry, as well as possible licenses, acquisitions or recruitments. Edlyn also described some of the key tools that the patent searchers in both governmental and commercial sectors use in their trade.

Grace Baysinger spoke on “Career information resources for graduate students and postdocs.” There is a wide variety of tools that chemists in the early stages of their careers can use to their advantage. Professional societies, such as ACS and RSC, offer career information as well as job listings. Funding agencies, such as NSF and NIH, provide guidance for grant writing. Libraries supply guides for job seeking and resume writing, as well as newer skills such as establishing your research identity through ORCID, developing data plans, and establishing a professional presence on social networks.

Some of the “do’s” and “don’ts” of dealing with chemical structures for cheminformatics were discussed in “So I have an SD file…what do I do next?” by Rajarshi Guha (National Center for Advancing Translational Sciences) and Noel O’Boyle (NextMove Software). Their principal message was: *avoid ambiguity in chemical structure description*! Some file formats for chemical structure description have ambiguities, or they lose data via data compression. Users should stick to the formats that have a unique identifier for each structure and a unique structure for each identifier. Stereochemical descriptors, such as R/S and +/-, are also frequently ambiguous and must be used with care. Verification of structures is key. ChemSpider provides a useful service for this purpose.

Leah McEwen (Cornell University) spoke to a related topic in her paper entitled “Chemical literacy for the ages: essential skills in 2D chemical representation.” She discussed the history of 2D chemical structure notation and chemical nomenclature. Both have had evolving standards from a variety of sources, such as IUPAC and Chemical Abstracts Service. The electronic era of chemical information retrieval has made, if anything, knowledge of the grammar and vocabulary of this basic language of chemistry even more important for chemists.

Neelam Bharti returned in the afternoon session with “From the lab to the library: a new journey.” She described the skills required of a chemistry librarian, and how research chemists can make the transition to librarianship using many of the key skills that they bring with them to their advantage.

In “Experiments with chemists and information,” Jonathan Goodman described how chemical information instruction at the University of Cambridge has changed over time. The focus has shifted from the print literature of handbooks and journals to open data and social media. In some areas, the students now have greater familiarity with the tools than their faculty. The challenge for instructors is in directing the students to apply these skills to chemical research (and to enlighten their faculty in the process).

The next few presentations looked at specific tools and their uses to enhance skills in chemistry. Stuart Chalk (University of North Florida) discussed “ChemData: a web application for learning chemical informatics.” ChemData is a prebuilt website application framework which he is using in his Chemical Information Science course. It allows students to learn basic concepts in manipulating chemical data. It is now being incorporated into an OLCC course (see: olcc.ccce.divched.org) where students will use datasets from the NIST Reference Data collections for their projects, learning how to deal with metadata, XML, Scientific Markup Language, and the Semantic Web.

Andras Stracz (ChemAxon) discussed the use of Marvin Live in his talk, “Improving geographically distributed research with real time collaboration.” Many organizations have widely scattered research teams who must find ways to share ideas and data. Marvin Live allows users to automatically capture ideas from project meetings and brainstorming sessions automatically, and provides a framework to connect to other cheminformatics applications which the participants may wish to call upon in their discussions. By automating these processes, Marvin Live saves time and ensures that no important ideas will be lost by poor documentation.

Joshua Bishop (PerkinElmer Informatics) discussed “Chemical research toolkit: an end-to-end solution.” He focused on the problems faced by biomedical researchers in dealing with huge numbers of compounds, their properties, and the structure-activity relationships among them. He described a variety of tools that can aid in collecting and analyzing the data at each step of the research process.

Rajeev Hotchandani (Scilligence) presented “ELN, RegMol and Inventory: from synthesis to registration to inventory,” describing three interconnected software solutions from Scilligence that include an electronic lab notebook package for collecting data. RegMol can capture molecule data from the ELN to create a compound registry, and transfer the information to Inventory for tracking samples and lots of chemicals. Single-click connectivity makes the system simple, quick, and effective for the researcher.

Charles Huber, Symposium Presenter

Jean-Claude Bradley Memorial Series
Edited by: Andrew Lang, Antony Williams
Journal of Cheminformatics
Collection published March 22, 2015; Last updated August 8, 2015
http://www.jcheminf.com/series/j-cbradley_memorial_series

Bi-Society Symposium on Laboratory Safety Information

The Chemistry Division of the Special Library Association (SLA DCHE) and the Chemical Information Division of the American Chemical Society (ACS CINF) co-hosted a bi-society symposium on Laboratory Safety Information during the SLA Annual Conference and Info-Expo on June 15, 2015. Librarians and information professionals from academia, industry, and government, cheminformaticians, educators, and chemical safety professionals from both divisions and beyond all gathered in Boston to explore this timely topic. Slides used for the presentations are available on SLA DCHE website at http://chemistry.sla.org/2015/slides-from-chemistry-division-sessions-at-sla-2015/.

Due to a few recent high-profile incidents in academic and corporate research labs involving chemicals, creating a safer research environment and culture becomes a priority in many organizations. Providing services and access to chemical safety information more efficiently and effectively could be contributions from librarians and other information professionals to this crucial initiative. The bi-society symposium sessions are dedicated to explore the roles information professionals can play from three perspectives: enhancing access to safety information resources, developing educational materials and sessions, and integrating information systems into lab workflows.

The kick-off session of the bi-society symposium was the DCHE Breakfast and Academic / Corporate Roundtable on Laboratory Safety Information and Practices. Stephanie Publicker (TOXNET), Evan Bolton (PubChem), and Steve Dueball (Elsevier) overviewed three different databases that disseminate chemical safety information. Publicker introduced a suite of resources regarding chemical safety provided by the NIH Environmental Health and Toxicology Information Program, including *Guide to Web Resources in Laboratory Safety*, TOXNET, Hazardous Substances Data Bank (HSDB), ChemIDplus, Haz-Map, Disaster Information Manager Research Center, and Wireless Information System for Emergency Responders, etc. Content of these resources is either collected by the program or carefully curated, and the resources are designed for a variety of user groups. Bolton demonstrated how PubChem as an archive of the biological activities of chemical substances provided a central location for chemical safety information and data with clear provenance. One of the goals for PubChem in this direction is to provide concise data views for safety information of highly used lab chemicals. Currently, the PubChem team is collaborating with chemistry librarians and chemical safety professionals to create Laboratory Chemical Safety Summary (LCSS) to meet this goal. Dueball mentioned that lab safety information could be found in many Elsevier products, such as Reaxys, Pathway Studio, Pharmapendium, and EMBASE. He focused on Reaxys, which provides easy access to flash points, toxicity, and other lab safety information of chemicals. It can also be used to search specific safety topics such as the right lab coat to choose.

After the talks, participants discussed questions around the librarians' role in the safety information realm. People mostly shared their experiences with providing access to lab safety information and some librarians mentioned that they participated in reviewing grants for animal research and lab protocols. The importance of chemical identifiers and the threat of their inconsistency were also discussed. CAS Registry Numbers usually get associated with chemicals when people obtain them and they could be a good point of entry to deliver safety information; but these numbers are not always available for other safety information providers to incorporate into the resources. The disconnection is an issue and may be solvable through other machine-readable identifiers of chemicals if they are adapted broadly. The breakfast session warmed up attendees with brain exercises on issues around lab safety information.

The second session, which was entitled “Exploring Safety Information Literacy: Towards a Safer Research Environment,” drew a big crowd. This safety information literacy session provided us diverse perspectives of librarians, chemistry faculty, and environmental health and safety officers, towards the safety information resources and education. Grace Baysinger (Stanford University) introduced a wide variety of resources providing lab safety information and how they can be used in research settings, from chemistry librarians’ perspective, including databases openly available online or subscription based, and traditional handbooks. Baysinger pointed out that data provenance is crucial in determining how useful the data would be and suggested that piping safety data properly into Electronic Lab Notebooks is an emerging need and should be developed in the near future.

Martin Walker (SUNY at Potsdam) shared his perspective as an educator in bridging the hazard information gap between experts and students in teaching labs. He discussed examples of how different types of lab hazard information could become useful in labs. Walker suggested that an open and comprehensive database, that was searchable for different types of hazards, and presented clear information for working chemists and students, had been much needed to fill the gap of hazard information. A predictive tool for prophetic substances and working procedures would be a plus for such systems.

Robin Izzo (Princeton University) spoke on the training needs of graduate students, postdoc, faculty, and research staff from the Environment Health and Safety officers’ perspective. She used several anecdotes to illustrate the variety of issues requiring different types of safety information for different researcher and staff groups. Izzo also highlighted how collaboration between lab safety professionals and librarians could deliver information and solutions efficiently and effectively.

The last, but not least, session of the one-day symposium was entitled “Enriching research management systems with point-of-need information delivery: case studies with laboratory safety information.” The four presentations in this session demonstrated the potentials for chemical information specialists to work with other stakeholders in this area, to create point-of-need information systems and facilitate best practices in the academic, corporate, and government sectors.

Damien Hammond (DuPont Protection Technologies) gave an overview of the DuPont SafeSPEC online product selector tool and guided the audience through use cases of the system. The system incorporated dynamic product data and accessories, literature and videos, technical information, and chemical resistance databases, as well as purchase guides, and FAQs. The database can be searched and browsed by hazards, industry, and existing guide. Users can search up to five hazards at a time. Hammond emphasized that the system was made for easy access so that end-users could directly specify their circumstance to obtain the recommended protection. The system even could remind users to check additional hazards relevant to the circumstance described.

Ralph Stuart (Keene State College), a long-time Chemical Hygiene Officer with experiences in institutions with different sizes, introduced his perspectives on how an information system could make and break lab chemical safety cases. He focused on the collaborative initiative called RAMP (Recognize hazards, Assess the risk, Manage the risk, and Plan for emergencies/Protect the environment) and how a variety of information and information systems could help with each step. Stuart pointed out the next steps of the collaboration between ACS CINF and ACS CHAS (Division of Chemical Health & Safety). The team is organizing chemical safety information on the web using the RAMP paradigm, developing better process descriptors to identify specific points when protections are required, and supporting chemical information literacy through the development of safety rubrics.

Jeffrey Whitford (Sigma-Aldrich) introduced his company's platform for selecting greener alternatives and showed how it leverages green chemistry principles, organizational knowledge base, and local expertise with the system, to enable reengineering of products in a greener way. Quantitative evaluation of the processes before and after reengineering gave concrete evidence for researchers to choose a greener and safer alternative.

Evan Bolton (PubChem) discussed how the rich data with clear provenance in PubChem could be used to assemble concise health and safety information for easy access, create a knowledge map, and add structure to textual data, and classify chemical relativities. The collaboration among PubChem, ACS CINF, and ACS CHAS on the RAMP projects will take these steps to enable and/or create a series of information systems useful for lab safety professionals and lab researchers. Last, the session moderator Leah McEwen (Cornell University) led a panel discussion on the challenges of building information systems that can be incorporated into the workflow of individual labs, through ELNs or other tools.

The bi-society symposium hosted by SLA DCHE and ACS CINF highlighted a variety of information resources for lab safety information, needs in safety information education, and the current status of building information systems. It also planted seeds for future collaborations among chemical information specialists, cheminformaticians, chemical safety professionals, and educators. Many of the speakers and attendees of the symposium participated in the ACS National Meeting in August 2015 and went further with exploring these topics in the CHAS symposium sponsored by CINF.

Ye Li, Leah McEwen, and Amanda Schoen, Bi-Society Symposium Organizers

The iRAMP Chemical Safety Information Project (http://www.irampp.org), jointly funded by CINF and CHAS, was honored with a ChemLuminary Award at the ACS National Meeting in Boston.

One of the early fruits of this professional community project is the formulation of data views for chemical safety information in the PubChem database (https://pubchem.ncbi.nlm.nih.gov/), now in production for over 3,000 chemicals. This data view is based on the Laboratory Chemical Safety Summaries (LCSS) format described in the *Prudent Practices in the Laboratory: Handling and Management of Chemical Hazards*.

The announcement of the LCSS availability in PubChem is at:
http://pubchemblog.ncbi.nlm.nih.gov
The splash screen for the LCSS views can be found at:
https://pubchem.ncbi.nlm.nih.gov/lcss/

Thanks to our collaborators in the CINF and CHAS communities, and especially the PubChem team, for their help in moving this idea forward!

Leah McEwen and Ralph Stuart

Editors' Corner

As I begin to write this column, the fall 2015 American Chemical Society Meeting in Boston is underway. The chosen theme for this meeting is "Innovation from Discovery to Application." One could write about how information is used throughout the innovation process, but I have chosen a different topic, though one still related to the theme. Accordingly, here we will take a look at innovation within the field of chemical information and cheminformatics.

To do this, I have searched for patents related to chemical information and cheminformatics. At first glance, it may seem that there are not that many. In the U.S. Patent and Trademark Office database of granted patents, the phrase "chemical information" occurs only twice in a patent title, twenty times in a patent abstract, and forty-seven times in claims. Moreover, most of the retrieved patents are about analytical instruments retrieving "chemical information" about a sample. "Cheminformatics" and its variants "chemoinformatics" and "chemiinformatics" are even more rarely used: once in a title, twice in claims, and zero times in U.S. patent abstracts. By combining several other search strategies and different databases, however, I was able to find 345 U.S. patents, and U.S. published, but ungranted applications related to our topic.

Of course, the usual disclaimers of prior art searching apply here. Since the search strategies are based on keywords and classifications, it is possible for the search to miss relevant documents that use different terminology or have been classified into a different classification. Also, for convenience I have chosen to examine U.S. patent documents only; contributions to chemical information and cheminformatics that were not patented in the United States are therefore ignored. Finally, the 345 U.S. patent documents (granted patents plus published abstracts) have been hand-selected for relevance, since the search strings used favored recall over precision. This adds a subjective element to the search, as does the fact that cheminformatics as a discipline does not have precise boundaries. Thus, it is entirely possible that a different searcher would have found a different set of patent documents to analyze. However, I believe that the set of patent documents used here is representative of the total.

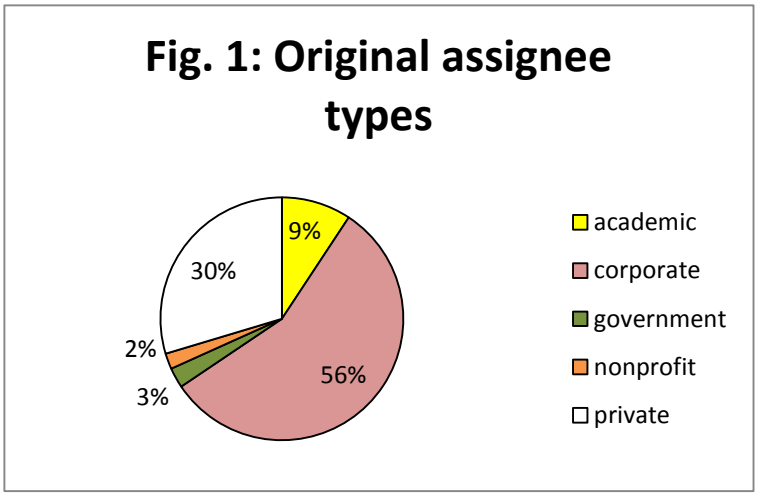
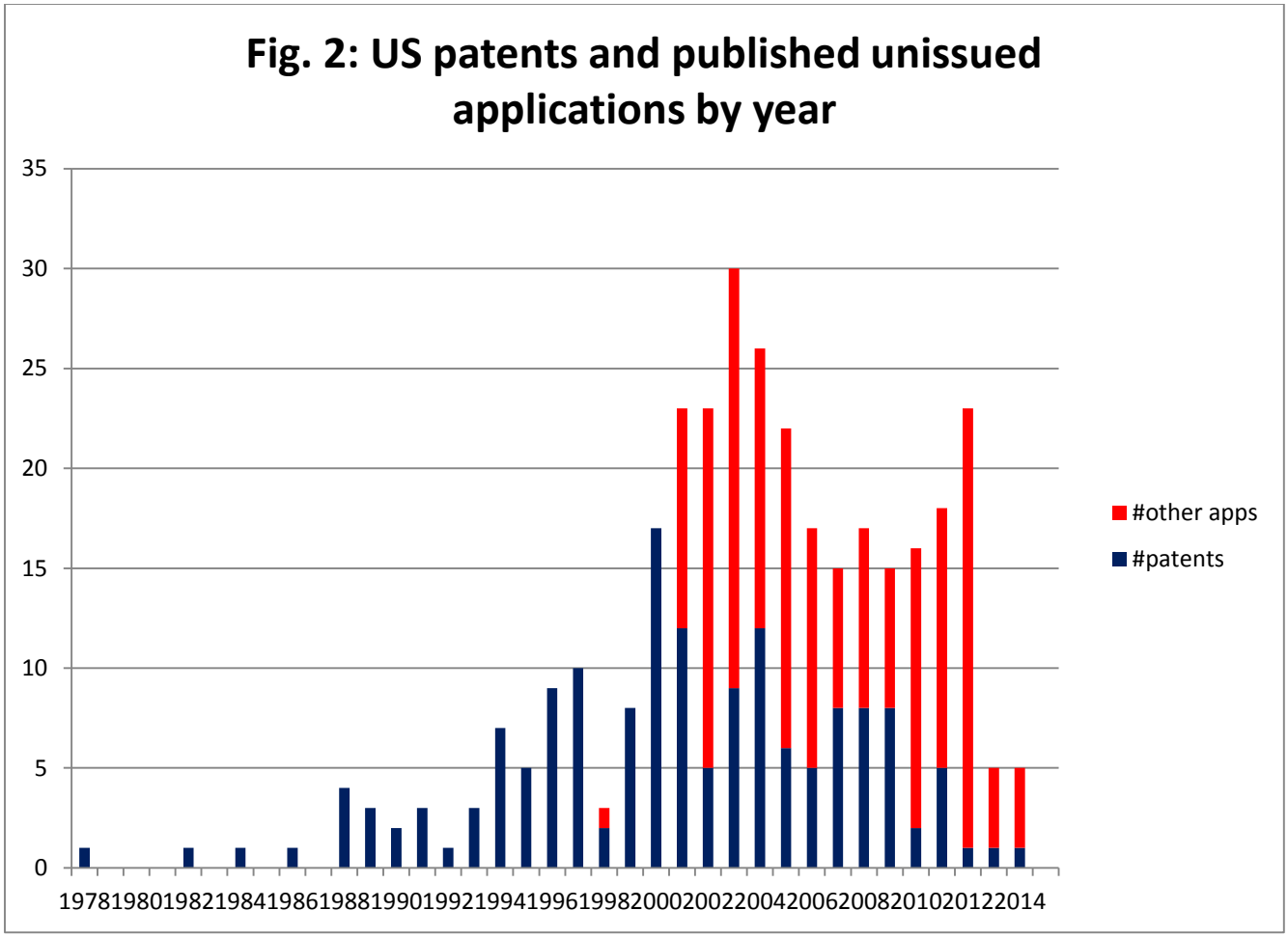


Figure 1 is a pie chart showing the distribution of original assignee types of all U.S. patents and published patent applications in the dataset. Assignees have been classified into five types: academic, corporate, government, nonprofit, and private (meaning one or more individuals, usually the inventors). Note that ownership of a patent can change hands after the initial assignment.

Figure 2 is a bar chart showing the year of filing of U.S. patents and published, unissued patents in the field of chemical information and cheminformatics. It appears that applications for chemical information and cheminformatics are rare until 1988, which is approximately when personal computers with graphical user interfaces became popular. There is a possible peak in activity in the early 2000's, but it's difficult to state that with certainty, since recent years are always underrepresented in this type of chart due to the time between filing an application and publishing it as either a published application or an issued patent.



As for other evidence of innovation in the cheminformatics field, the simple fact that the ACS Division of Chemical Information was able to field a technical program with a scheduled 218 presentations in 22 topics (including the co-sponsored symposia) speaks for itself. (Admittedly, these numbers are much higher than average). Other evidence is the existence of at least three journals on the subject: *Journal of Chemoinformatics, Journal of Chemical Information and Modeling,* and *Molecular Informatics;* and others, if one is allowed to encroach into the related field of computational chemistry. So congratulate yourselves: you are part of an innovative group of people!

David Shobe, Assistant Editor, Chemical Information Bulletin

The Chemical Structure Association Trust: Advancing Scientific Discovery for Fifty Years

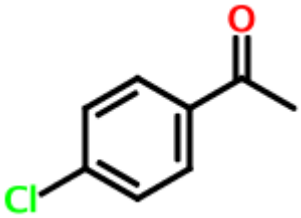
The Chemical Structure Association Trust (CSA Trust) is an internationally-recognized, registered charity that promotes and supports the advancement of scientific discovery through the application of computer technologies in the management and analysis of chemical structure information.

In support of its Charter, the Trust provides grants specifically to nurture young scientists, ages thirty-five or younger, who have demonstrated excellence in research related to the storage, retrieval, and analysis of chemical structures, reactions, and compounds. Since its inception in 1988, almost one hundred students and researchers worldwide have benefited from travel bursaries and the CSA Trust Grant Program to further their education and research work, but the organization has a rich history that predates the formalization of its charity status. Its roots were planted half a century ago in 1965 when the Chemical Notation Association (CNA) was formed in the United States. It has been an interesting journey from the CNA to the CSA Trust and I have been blessed to have been a part of it almost from the beginning along with other members of the American Chemical Society's Division of Chemical Information. In honor of the organization's 50th Anniversary, I'd like to give you a brief overview of its past and its present activities.

The Beginning

The concept of "chemical structures" emerged during the first International Chemical Congress held in Karlsruhe, Germany in 1860 where the leading chemists of the time met to resolve their ideas about atoms, molecules, and equivalents. At the end of the meeting, Alexander Butlerov predicted that it would be the future task of chemists to determine the atomic arrangements of these molecules[1] and within a few years the development of descriptive line notations began.[2] The emergence of punch-card technologies during the middle of the last century renewed interest in these notations, and in 1949 the International Union of Pure and Applied Chemistry (IUPAC) invited the submission of simple notations that would be suitable for international adoption. They ultimately chose a notation submitted by G. Malcom Dyson, but it was one of the other seven notations that were submitted that caught the attention of those working in the field.[i] This notation, based on Zipf's "Principle of Least Effort"[3] was introduced in 1950 by a chemist, William J. Wiswesser, then working at the U.S. Department of Agriculture in Frederick, MD.[4] For the notation, Wiswesser used the numbers one through ten, twenty-six capital letters of the alphabet, three punctuation marks (&, -, and /) and a blank space. Thus WLN could be used on any typewriter as well as on computers and punch-card accounting systems.[5] When WLN symbols, each representing a specific structural fragment, were connected in a specific linear format, the result was a unique and unambiguous chemical structure formula. Below is an example taken from a 2007 blog that is definitely worth a quick read.

[i] It should be noted that the selection of the Dyson notation was criticized, and a petition was signed by about 1,000 chemists, including several who had submitted notations for consideration, stating that the Wiswesser Notation had not been given adequate consideration. The appeal was taken to the American Chemical Society and the National Academy of Sciences - National Research Council who requested that the National Science Foundation do a study, the results of which showed that more testing of both notations should be done before any decision was made. This was not done and the Dyson Notation was selected. A cloud hung over the decision because Dyson was the chair of the IUPAC Commission that called for the submission of notations (see *Survey of Chemical Notation Systems: A Report,* a Report of the Committee on Modern Methods of Handling Chemical Information, National Academy of Sciences - National Research Council, Publication 1150, Washington, D.C., 1964 (see p. 442-3).



GR DV1

The “R” stands for benzene, the “G” stands for chlorine, the "DV1" stands for the 4-acyl substituent. Here, the "D" denotes the 4-postion. The 3- position would result in "CV1," and the 2- position would result "BV1." The space character means that the character following it should be interpreted as a ring locant.[6]

The notation was simple and gradually became very popular, as it could be read and written by humans as well as by computers (comments on the above-mentioned blog reinforces WLN’s simplicity). In fact, Dr. Elbert George Smith, an associate professor at Mills College in Oakland, CA, used WLN to encode thousands of structures contained in chemistry reference books such as the *Merck Index* and Lange’s *Handbook of Chemistry*, ultimately developing a manual, *The Wiswesser Line-formula Chemical Notation*, published by McGraw-Hill in 1968.

Because of its utility and simplicity, pharmaceutical companies began to use WLN to create files of their in-house compounds for computer manipulation and analysis. In 1965 the Chemical Notation Association (CNA) was established so that the chemical community could collaborate to enhance the utility of the system. A UK chapter was established in 1969 and at one point the CNA had more than two hundred members representing more than eighty international organizations that had adopted chemical notations to manage their respective chemical structure files.[7] According to Wendy Warr, WLN was widely taught around the world and had been adapted to French, German, and even Japanese.[8]

Needless to say, chemical notations eventually found their way into commercial products. One of the earliest was the Index Chemicus Registry System (ICRS) offered by the Institute for Scientific Information (ISI, acquired by Thomson Reuters in 1992).[9] This product was launched in January of 1970 and provided the WLNs for the new organic compounds published in ISI’s Current Abstracts of Chemistry and Index Chemicus. The product was a major breakthrough for substructure searching (remember, the technology for drawing chemical structures had not yet emerged) and it was embraced by major pharmaceutical companies worldwide. ISI was a key force behind the spread of WLN and the staff who worked on ISI’s publications (of which I was one) were trained directly by Dr. Wiswesser. We were encouraged to become active in the CNA and participate in WLN development. This was an intense effort, but the CNA ultimately produced a revised manual that was published in 1975.[10] I have very fond memories of lengthy meetings debating WLN rules with colleagues, both in the United States and abroad. It was both exciting and intellectually stimulating to be part of something so new and innovating, and the friends made back then remain friends to this day!

During these early days the parent CNA in the USA and those in the UK Chapter worked very closely together. They were a vibrant, active, dedicated, and very collaborative community. Those in the USA focused primarily on training users of WLN and monitoring rule changes. The CNA(UK) was more broadly focused and even was assigned the rights to the Dyson Notation in 1980, following Dr. Dyson’s death. That is not to say that the CNA(UK) did not offer training. They held many tutorials on how to use WLN, but they also organized a series of conferences in the area now termed

"cheminformatics," covering topics ranging from substructure searching and the design of information systems, through to integrated databases and searching the chemical literature and patents. The valuable content of many of these conferences was captured and published and are referred to in the book, *Chemical Structures: The International Language of Chemistry*, edited by Wendy Warr.[11]

The CNA(UK) was also involved in the organization of the NATO/CNA Advanced Study Institute on Computer Representation and Manipulation of Chemical Information that was held in June 1973 in Noordwijkerhout, The Netherlands. This was the predecessor to the successful series of International Conferences on Chemical Structures (ICCS) that have been held at the same venue every three years since 1987. In the 1970's the CNA(UK) also established a newsletter as a vehicle for members to share experiences, distribute announcements, etc. This newsletter continues today under the strong editorship of Grace Baysinger, Head Librarian and Bibliographer of the Swain Chemistry and Chemical Engineering Library at Stanford University, and an active member of the CSA Trust Board (see the most recent issue at: http://www.csa-trust.org/?page_id=21).

Times are a-Changing

Up until the early 1980's chemical notations played a major role in the analysis of chemical structures. But technology was changing and two major advances would ultimately diminish their role. First, programs were written to convert WLNs to computer-manageable connection tables that were then amenable to structure and substructure searching. The availability of connection tables made it possible to do more with structures, and tools were developed to run similarity searches, generate and search 3D structures, calculate properties, generate names, map reactions, and exchange structure files with collaborators, to mention just some of the features of structure-handling systems.

The second advance was the availability of affordable graphics terminals coupled with software to convert between graphical representations of compounds and their connection tables. In 1979, Molecular Design Ltd. (MDL, acquired by Maxwell Communications Corporation in 1987, and by Reed Elsevier in 1997) offered their Molecular Access System (MACCS) for interactive graphical registry and for both full structure and substructure retrieval. CAS introduced its own online service, CAS Online, in 1980. This began as a pilot version made available to a limited group of customers. About 500,000 substance records were available and could be searched only by screen numbers representing specific molecular structural features. Searching by screens was not the most convenient method for information users, and yet many found the new system useful. When CAS Online was introduced to the general public, it provided access to 1.8 million substance records, about one-third of the total Registry database. Other segments of the Registry were added to CAS Online in increments as the search capacity was increased at CAS. In November 1981, CAS introduced searching by structure or substructure diagram. Users with a specific model of intelligent graphics terminal, the Hewlett-Packard 2647A, could select structure features from a menu and then assemble them on the terminal monitor by using a graphics tablet and stylus. These terminals could display answers with well-drawn structure diagrams. True structure-based searching was now possible for chemists rather than their information scientist intermediaries and notations began to take a back-seat.

In July of 1981, members of the CNA(UK) agreed to create an organization that would go beyond a narrow focus on notations to addressing a broader spectrum of chemical structure and data handling issues. Thus the Chemical Structure Association (CSA) was created. The CNA(UK) continued as a sub-group for those still involved with WLN. The CSA organization was officially launched on September 6, 1982 at the University of Exeter, UK, when the Executive Committee of the CNA(UK)

became the first Executive Committee of the CSA. Dues continued to be paid to the CNA parent body in the United States by the WLN sub-group. The CSA promoted educational activities and enjoyed an international membership, many of whom had been members of CNA(UK). It became a very active global organization and it was involved in organizing conferences and meetings as well as continuing with the Newsletter. In December 1983, the CNA(UK) was closed and its funds were passed on to the CSA.

The success of CSA conferences, courses and seminars led to a surplus of funds, and it was suggested that it would be beneficial to set up a UK registered charity so that rather than the Trust paying excessive corporation taxes, the funds could be used for charitable purposes, for example, for awards and grants. Following extensive discussions with the UK Charities Commission, the Declaration of Trust was made on December 5, 1988 and the CSA Trust was declared a UK Registered Charity No: 328042. The CSA and the CSA Trust continued to operate successfully side-by-side, but in the 1990s the CSA membership began to decline. In parallel, the Trust was having difficulty finding Trustees willing to put in the level of effort required to run it effectively. By 2001 this situation prompted both organizations to propose that they merge. The Trust could then take on the role of conference organization and newsletter production, in addition to allocating funds for grants and bursaries.

After several communications with the Charities Commission the merger was successfully achieved in 2001 and the CSA was formally closed, transferring all its funds to the CSA Trust. New committees were set up by the Trust to promote its work and oversee the activities formerly carried out by the CSA, including: Public Relations (Newsletter, Website), Meetings and Training, Fundraising, Finance, Grants, and Awards.

The Present

The above activities carry on today, managed and overseen by a board of twenty trustees drawn from academic, government, and commercial organizations in countries around the world. The Trust has awarded more than £25,000 in bursaries and grants to support travel and research work.

The Grant program is a boon to young researchers. Here is a comment from Dr. Noel O'Boyle, NextMove Software, Cambridge, UK, a 2010 grant recipient: "Young researchers need all the help they can get, as the odds are stacked against them. The CSA Trust Grants are a lifeline and an encouragement during some difficult years. The grant allowed me to attend and present my work at an international conference, the German Conference on Chemoinformatics. Perhaps more importantly, it enhanced my CV at a time when I was establishing my research career, as it showed that the quality of my work was highly-regarded on an international level." The call for the 2016 Grant proposals appears immediately following this article, and proposals are due by March 25, 2016. Please feel free to circulate this information widely.

While Grants remain a major focus of the Trust, support of symposia and workshops also continues. Each year the Trust develops a symposium jointly with the Division of Chemical Information of the American Chemical Society (ACS) that is held at one of the ACS National Meetings. The Trust also supports the Sheffield Conference on Cheminformatics that is held every three years at the University of Sheffield, UK (see programs at: http://cisrg.shef.ac.uk/shef2013/default.htm#prev). The next Sheffield conference will be held on July 6, 2016 at The Edge, University of Sheffield, UK (http://cisrg.shef.ac.uk/shef2016). In addition, the Trust is a founder and continuing supporter of the International Conference on Chemical Structures that has been held every three years since 1987

(see http://www.int-conf-chem-structures.org for information on the 2014 conference). The next conference is scheduled to begin on June 4, 2017, in Noordwijkerhout, the Netherlands.

Today, the CSA Trust continues the dedicated efforts of its original incarnation, the Chemical Notation Association. While the organization no longer has hands-on involvement in the development of tools for the creation and analysis of chemical structure information, it is committed to providing financial support for each new generation of young researchers whose passion and knowledge may ultimately unlock the secrets of chemical structures. The goal of the CSA Trust is to shine a light on the essential importance of chemical structure information to the advancement of scientific discovery. If you are interested in supporting the Trust's goal as an active participant or as a financial supporter, please do not hesitate to contact me at chescot@aol.com.

Acknowledgements

This article could not have been written without the input of Phil McHale, Janet Ash, and many past CNA members and current CSA Trustees. Phil created a poster on the History of the Chemical Structure Association Trust that was presented at a joint meeting of the Royal Society of Chemistry's Chemical Information and Computer Applications Group, the RSC Historical Group, and the CSA Trust on November 29, 2010 (see: http://www.csa-trust.org/files/CSAT_History.pdf). Janet compiled a history of the Trust that is included in the Trust's Procedures Manual and is based upon input from those who have been active, for many years, in both the Trust and the Chemical Notation Association.

References

(1) Butlerov, A. M. *Zeitschrift fur Chemie und Pharmacie*, **1861**, *4*, 549-60.
(2) Wiswesser, W. J. 107 Years of Line-Formula Notations (1861-1968). *Journal of Chemical Documentation*, **1968,** *8* (3), 146.
(3) Survey of Chemical Notation Systems: A Report, a Report of the Committee on Modern Methods of Handling Chemical Information, National Academy of Sciences - National Research Council, Publication 1150, Washington, D.C., **1964** (see p. 440).
(4) Garfield, E. Is Shorthand the Route to Success in Science or Anything Else? Part 1. History and Evolution of Stenographic Languages. *Essays of an Information Scientist*, **1986**, *8*, 9.
(5) Garfield, E. The Retrieval & Dissemination of Chemical Information. II. The Wiswesser Line Notation. *Essays of an Information Scientist*, **1977,** *1*, 111.
(6) Apodaca, R. Everything Old is New Again - Wiswesser Line Notation (WLN). *Depth-First.* Published online: July 20, 2007. http://depth-first.com/articles/2007/07/20/everything-old-is-new-again-wiswesser-line-notation-wln/ (accessed Sep 12, 2015).
(7) Gelberg, A. In Memoriam: William Joseph Wiswesser: 1914-1989. *Chemical Information Bulletin*, **1990**, *42* (1), 2.
(8) Warr, W. A. Diverse uses and Future Projects for Wiswesser Line-formula Notation. *Journal of Chemical Information and Computer Sciences*, **1982**, *22* (2), 98-101.
(9) Garfield, E. The Retrieval & Dissemination of Chemical Information. III. The Index Chemistry Registry System ICRS). *Essays of an Information Scientist*, **1977**, *1*, 113.
(10) Smith E. G., Baker, P. A., in collaboration with the members of the Chemical Notation Association, *The Wiswesser Line-formula Chemical Notation (WLN)*, 3rd ed.; Chemical Information Management: Cherry Hill, NJ, 1975.
(11) Warr, W., Ed. *Chemical Structures: The International Language of Chemistry;* Springer-Verlag: Berlin Heidelberg, 1988.

Bonnie Lawlor, CSA Trust Secretary

Chemical Structure Association Trust Grant: Applications Invited for 2016

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 2016.

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). 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. 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, to gain access to special computational facilities, or to acquire unique research techniques in support of one's research.

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, for example, cost of equipment, travel expenses if the request is for financial support of meeting attendance, etc. The relevance of the above-stated purpose to the Trust's objectives, and the clarity of this statement, are essential in the evaluation of the application;
3. A brief biographical sketch, including a statement of academic qualifications;
4. Two reference letters in support of the application.

Additional materials may be supplied at the discretion of the applicant only if relevant to the application, and if such materials provide information not already included in items 1 - 4.

Deadline for Applications:

Applications for the 2016 Grant are due by March 25, 2016. Successful applicants will be notified no later than May 2, 2016.

Address for Submission of Applications:

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

2015 Grant Awardees:

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

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

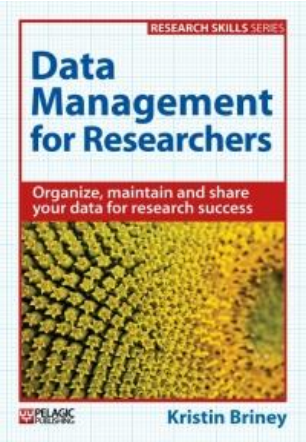
Dr. Oxelandr Isayev (Division of Chemical Biology and Medicinal Chemistry, UNC Eshelman School of Pharmaacy, University of North Carolina at Chapel Hill) was awarded a grant to attend summer classes at the Deep Learning Summer School 2015 (University of Montreal) to expand his knowledge of machine learning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

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

A complete list of the previous grant awardees is at http://bulletin.acscinf.org/node/786.

Bonnie Lawlor, Chair, CSA Trust Grant Committee

Book Reviews



Briney, K. *Data Management for Researchers: Organize, Maintain and Share your Data for Research Success;* Pelagic Publishing: Exeter, UK, 2015. 191 p. + x. ISBN 978-1-78427-012-4 Hardcover, £ 49.99. ISBN 978-1-78427-011-7 Paper, £ 24.99.

An excellent practical treatise on the art and practice of data management, this book is essential to any researcher, regardless of subject or discipline. Each of the eleven chapters begins with a recounting of a real life encounter with data management, some favorable, some disastrous. Data are defined broadly, as anything one performs analysis upon and specific examples are discussed. Data management is described in detail as those practices necessary for efficient use of data before, during, and after the research is performed. Each chapter has a concluding summary and references, and the text concludes with an index.

Chapter 1 covers the importance of data management in modern research. Funding agencies now require data management plans and data sharing, reproducibility concerns highlight data management issues, and researchers cannot manage their increasing amounts of digital data the same way as physical samples. The difference between doing data management and writing a data management plan is also discussed.

Chapter 2 describes the “new” circular lifecycle of data (as opposed to the “old” lifecycle that was linear): see the figure below.

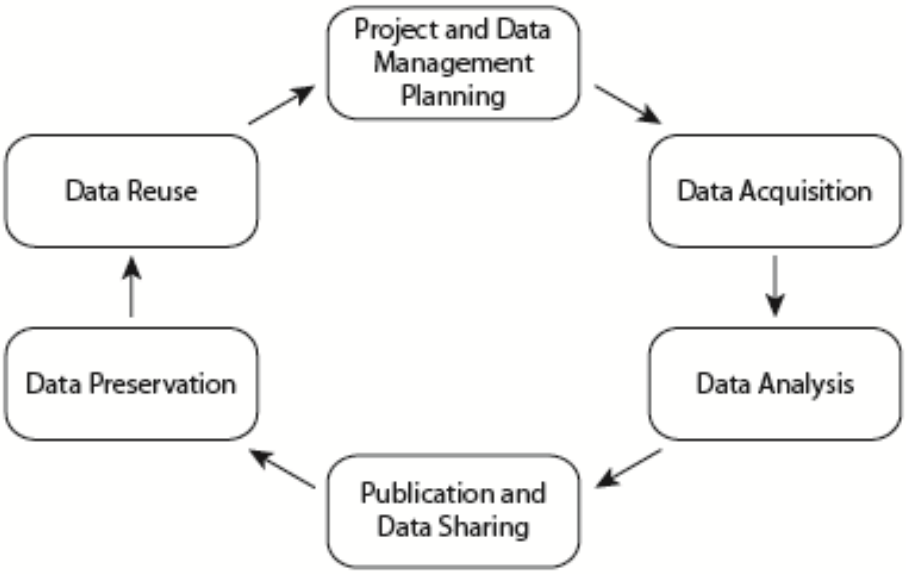


Figure reprinted with permission by Kristin Briney

This lifecycle defines the organization of chapters 3-6 and 10-11 while chapters 7-9 come under the category of storage, covering data security, storage and backups, and long-term preservation.

Chapter 3 covers data management plans and data policies. These policies come from granting agencies, government, and institutions and cover issues such as data retention (including policies), ownership, and copyright. Notebooks, electronic and paper, are covered in depth in chapter 4, and the advantages and disadvantages of each are discussed. The chapter also reviews other types of documentation such as methods, metadata, and standards from publishers and professional societies. File organization, including naming, documentation, and databases, is described in chapter 5. Data analysis is discussed in detail in chapter 6, including the retention of both raw and analyzed data, and analysis methods.

Chapters 7-9 digress from the roadmap outlined in chapter 2 and treat the topics of data security and storage in depth. Managing sensitive data is an important aspect of data security and responsibility, ethics, and methods (including encryption) are described (chapter 7) as well as cloud versus local storage issues. Storage and backup methods (chapter 8) are essential aspects, including long-term versus short term, hardware and software, and storage of non-digital data. Long-term storage is discussed in detail (chapter 9), including retention times (regulated or not), selection of data to be retained or culled, and more on hardware and software including obsolescence. Data ownership, personal copies, and outsourcing in repositories are also essential considerations.

Chapter 10 covers data sharing, (including sharing with a research group), organization, publication, and public access. The last includes Open Access. A brief description of intellectual property (IP), that is, copyright, trade secrets, and patents, is included, although for patents additional sources should be consulted. Licensing is recommended for all data sharing, including collaboration and copyrightable material. Citations and altimetrics are discussed, as well as repositories and their locations. Librarians are cited as resources for data management support.

Chapter 11 covers data reuse and restarts the data lifecycle. Sources of data include libraries and published articles. Reuse rights vary and some exclude use for commercial research. Error treatment and citation practices are discussed with examples.

I noticed that Table 4.3, “Different Representations of the Molecule Acetone,” (p. 60) has the InChI code, but not the InChIKey. Only the CAS Registry Number is listed for CAS, but CAS also has systematic names. (CAS systematic nomenclature is a dialect of IUPAC nomenclature.) Also, to turn to another issue, I have often wondered about the extent of the embargo on reuse of data for “commercial purposes.” MEDLINE had such an embargo, but did that cover the contents of literature searches performed for commercial enterprises, or by consultants to commercial enterprises? Does that embargo also apply to the use of PubMed information? It would seem to be even harder to enforce (if it ever were enforceable).

Readers of this *Bulletin* will see a continuation of a theme on information management[1] covering issues essential to the effective performance of any kind of scientific research. Although it’s been decades since this reviewer generated any laboratory data, he does continue to perform literary research for publication and he is prompted to improve his data management.

(1) Baykoucheva, S., *Managing Scientific Information and Research Data,* Chandos Publishing, Amsterdam, Boston, 2015. Reviewed in *Chemical Information Bulletin* **2015,** *67* (3), p. 20-22.

Robert Buntrock, Member, CINF Communications and Publications Committee

The Merck Index

Many readers will be aware that the Royal Society of Chemistry acquired *The Merck Index** from Merck & Co. in 2012. *The Merck Index* is an incredibly useful resource that has gained legendary status among chemists, and so we were delighted to take on its stewardship and future development.

We published the 15th print edition in 2013, but our main ambition was to create a modern, user-friendly online home for the same content. *The Merck Index Online* enables users to search by chemical structure, physical properties and text, or a combination of these. Free from the space restrictions of a single printed volume, we have also reintroduced about 1,500 entries that were previously cut from print editions.

The Merck Index Online team here in Cambridge, UK comprises the Editor, Serin Dabb, who provides editorial oversight, and two Data Content Editors, Michael Townsend and Mark Archibald, who investigate new scientific areas and research and write the content. Our scientific backgrounds are spread across organometallic chemistry and catalysis (Serin), physical organic chemistry (Michael) and synthetic organic chemistry (Mark). We're always keen to hear comments, suggestions, and corrections from readers, either regarding content or features of the website; you can reach us at rscindex@rsc.org.

One of the things we love about *The Merck Index* is the quirky nature of some of the older content. Perhaps the most famous example is caproic acid's eponymous "goat-like odor." This gives a flavour, often quite literally, of a former age of chemistry when smell and taste were routine methods of product characterization. When we bring old records up to date, we try to add the relevant modern science without removing this link to the past.

In keeping with *The Merck Index's* tradition, our primary focus is on substances of pharmaceutical interest, with the aim of including every newly approved drug. That said, the scope of the existing content spans all of chemistry and we have already created new entries in the fields of materials chemistry, agrochemicals, and synthetic chemistry. We plan to carry on in this vein: if a chemical substance is of significant interest and importance, for medicine or technology or anything else, then it belongs in *The Merck Index*.

This continued survey of the broad scope of chemistry goes hand-in-hand with continued technical development of the online platform. We recently added a browsing feature to complement the usual search-based approach, which restores the serendipitous discovery that was always possible with the printed book. Further development is planned to present important information as clearly as possible, and to provide clear links to related external content. The combination of modern database technology with expert curation of the overwhelming mass of chemical data offers readers easy access to the relevant, authoritative information they need.

Mark Archibald, Royal Society of Chemistry

**The name THE MERCK INDEX is owned by Merck Sharp & Dohme Corp., a subsidiary of Merck & Co., Inc., Whitehouse Station, N.J., U.S.A., and is licensed to The Royal Society of Chemistry for use in the U.S.A. and Canada.*

Committee Reports

CINF Communications and Publications Committee

The CINF Communications and Publications Committee held a virtual meeting at the end of July, and followed up with an in-person meeting related to the CINF website at the ACS meeting in Boston. Our committee includes: Svetla Baykoucheva, Bob Buntrock, Stuart Chalk, Judith Currano, Graham Douglas, Belinda Hurley, Erja Kajosalo, Svetlana Korolev, Bonnie Lawlor, Dave Martinsen, Patti McCall, Carmen Nitsche, Vin Scalfani, David Shobe, Teri Vogel, and *ex officio* Rachelle Bienstock, CINF Chair, and Donna Wrublewski, Membership Chair.

If one word could be used to summarize the current status of the Committee it is *transition*. Patti McCall, Webmaster, and Erja Kajosalo, Assistant Webmaster, have done a great job in keeping our website up to date, and we thank them both. As Erja looks for other opportunities within and outside CINF, Patti will continue as Webmaster, and Stuart Chalk will take over as Assistant Webmaster. The CINF website has experienced some infrastructure problems, and so we are looking for a new solution. A small team will evaluate several options and make a recommendation to the Executive Committee. Be on the lookout for an updated website early next year (hopefully).

Belinda Hurley and Carmen Nitsche have done a wonderful job planning and hosting the CINF Webinar Series. They have brought us perspectives from a variety of people who would not normally be seen at an ACS meeting, with two very different looks at altmetrics, an overview of CHORUS, and thoughts on Net Neutrality. One more webinar with John Regazzi speaking on “Infonomics and the Business of Free” is scheduled on September 30 this year. After two years of managing the webinars, both coordinators feel the need to turn over the responsibility in order to bring a fresh perspective to the program.

NEEDED: We are looking for two volunteers to take over that task. Access to a webinar platform, such as AdobeConnect or GoToMeeting, would be helpful.

The *CIB* has been produced through the efforts of our two editors, Svetlana Korolev (editor of issues 2 and 4) and Vin Scalfani (editor of issues 1 and 3), and their assistant editors, Teri Vogel and David Shobe. With the previous issue, we decided to release the initial ASAP version in PDF of the bulletin as soon as it was ready for publication, rather than wait for the Webmaster to convert the entire issue to HTML. Since the HTML version usually results in some additional corrections, a final PDF version will be released after the HTML version has been completed. This is most critical on issues 1 and 3, which contain the CINF Technical Program.

A special thanks to Svetlana, who has been working on the *CIB* for 6 years, and has requested to step down after this issue. We are grateful for her efforts for so many years. We welcome Judith Currano, who has agreed to join a team of editors. Together they will work out how to divide and conquer the 2016 issues of the Bulletin.

Finally, my own term as Chair of the committee will end at the end of 2015. It has been a pleasure to work with such a wonderful group of people: hard working, enthusiastic, and willing to try new things, and step in for new tasks. Graham Douglas, who has been Assistant Chair, will be taking over as Chair in 2016, and I wish him the best in that new role.

David Martinsen, Chair, CINF Communications and Publications Committee

CINF Education Committee

The CINF Education Committee met on Saturday, August 15, 2015 from 1:00 pm - 3:00 pm in the Boston Convention and Exhibition Center, Room 107C.
Attendees: Grace Baysinger (Chair), Chuck Huber, Ye Li, Teri Vogel, Martin Walker, Donna Wrublewski. Consultant: Adrienne Kozlowski. Guests: Judith Currano, Jeremy Garritano.
Absent: Christina Keil, Marion Peters, and Susanne Redalje.

Announcements:

- "Chemical Information Skills: The Essential Toolkit for Chemical Research - A Joint CINF-CSA Trust Symposium" had 15 speakers and was held on Wednesday, August 19, 2015. ACS videotaped this symposium.
- A BCCE Meeting or an ACS Regional Meeting are good outreach venues for CINF attendees to connect with high school teachers. The 2016 Biennial Conference on Chemical Education (BCCE) will be held July 31 - August 4, 2016 at the University of Northern Colorado, Greely Colorado. The deadline for submitting a symposium or workshop is December 6, 2015. We need to decide if we want to participate.

SOCED Report and K-12 Education:

Jeremy Garritano, who is a member of the Society Committee on Education (SOCED), gave a report of topics covered in the SOCED Meeting on Friday, August 14, 2015. Excerpts from the 2015 fall SOCED report to ACS Council, followed by comments by CINF Education Committee members are below.

- SOCED voted to approve revisions to the *ACS Guidelines for Chemistry in Two-Year Colleges*. Many guidelines on chemical information are similar to the guidelines for 4-year schools.
- The ACS Committee on Professional Training (CPT) has published the *2015 Guidelines for Undergraduate Professional Education in Chemistry*. (CPT is revising the supplemental guidelines, one third per year for three years.)
- The committee also voted to make the pilot program of ACS International Student Chapters a permanent feature of the student chapters program.
- SOCED was informed of recent developments related to the new American Association of Chemistry Teachers (AACT), including $50,000 in grants from the Camille and Henry Dreyfus Foundation and the Ford Motor Company. AACT has over 78,000 members, mainly high school.
- SOCED provided input on a coordinated strategy to address the needs of faculty better at various stages of their careers at two- and four-year institutions. Approximately 500 faculty members responded to a survey that assessed faculty development needs. Survey respondents expressed interest in networks that would allow them to explore and share teaching ideas, and exchange views on mentorship. Interest in resources to improve laboratory safety was also very high.
- SOCED also received updates from Education Division staff on the 10th anniversary of ACS ChemClubs, Division strategic planning, and the launch of an Individual Development Plan (IDP) tool for graduate students. The online IDP resource is expected to launch at the end of September.
- SOCED is revising Chemistry in Context (for undergrads) and the high school textbook *Chemistry in the Community*.

The CINF Education Committee wants to learn more about guidelines and standards to see what skills students are expected to learn before they reach college. For example, are students being

asked to read, filter, think critically, and learn Boolean logic? There's a lot more inquiry-based learning in high school these days, so they have to look up a lot more now. Judith Currano asked each member of the Education Committee to send her a few sentences on "What do you want your incoming college freshmen to know." Adrienne Kozlowski suggested lab reports and Martin Walker suggested term papers as examples of the types of written documents students need to know how to create. After the ACS Meeting, Susanne Redalje suggested creating a list of free resources and having the committee be a source of support for smaller schools that don't have a science librarian.

The Chemical Information Sources (CIS) Wikibook Project

The CINF Education Committee has agreed to spearhead efforts and coordinate the updating of the Chemical Information Sources Wikibook. Chuck Huber and Grace Baysinger will focus on the content, and Martin Walker and Ye Li will focus on technical aspects, including the organization and structure. For archival purposes a permanent snapshot has been generated. A metadata template was created for commonly cited resources (e.g., SciFinder). Martin will suggest and share some navigation templates to adopt as part of the revision process.

One key question is *do we use the CIS Wikibook as the repository or place to link content and shut down Explore Chemical Information Teaching Resources (XCITR)?* The plan is to merge the Chemical Information Instructional Materials (CIIM) content into the appropriate section of the CIS Wikibook, rather than have it be a separate location. When Gary Wiggins created the CIIM, the scope excluded commercially produced content. We should broaden the scope of CIIM to include contributions from information providers, instructors, and librarians. We need to decide what types of materials to accept (for example, videos, tutorials, or PowerPoint presentations). A private group called the CINF Chem Wikibooks was created in the ACS Network to support this project.

Outreach, future symposium, other topics

- In support of National Chemistry Week 2015, Grace Baysinger compiled a list of electronic resources. The theme is "Chemistry Colors Our World" and topics for the electronic resource lists include: About Color, Visible Spectrum, Absorption/Reflection, Rainbows, Food Colors, Natural Dyes and Pigments, Fireworks, and Chemistry and Art.
- The new ACRL Framework for Information Literacy for Higher Education is a useful document to add to the CINF Chemical Information Literacy page.
- The Cheminformatics OLCC (http://olcc.ccce.divched.org/) will have a more traditional library instruction module added. Leah McEwen and Ye Li have been helping with this effort. Should the Education Committee be doing more with Robert Belford?
- Having an "education" symposium at a fall ACS national meeting is preferred because more faculty attend national meetings in the fall than in the spring, Should the Education Committee reduce the amount of programming and focus on other things because of our limited resources?
- Guidelines (drafted by Judith Currano) on what information skills are needed by successful graduate students in chemistry still needs to be finished. The draft document is in the CINF Education Committee's private group on the ACS Network.

Many thanks to Martin Walker for taking notes for the CINF Education Committee Meeting!

Grace Baysinger, Chair, CINF Education Committee

ACS Council Meeting

The Council of the American Chemical Society met in Boston, MA on Wednesday, August 19, 2015 from 8:00am until approximately 11:30am in the Grand Ballroom of the Sheraton Boston Hotel. There were a number of items for Council Action and they are summarized below.

Nominations and Elections

Committee on Committees: Council voted to fill six slots on the Committee on Committees. There were twelve nominees as follows: Christopher J. Bannochie, Fran K. Kravitz, Michelle V. Buchanan, Patricia A. Redden, Alan B. Cooper, Carolyn Ribes, Jetty Duffy-Matzner, Sharon P. Shoemaker, Donna G. Friedman, Julianne M. D. Smist, Lynn G. Hartshorn, and Stephanie J. Watson. The five candidates receiving the highest numbers of votes were declared elected for the 2016-2018 term. These were: Christopher J. Bannochie, Michelle V. Buchanan, Alan B. Cooper, Carolyn Ribes, and Donna G. Friedman. The candidate receiving the sixth highest vote, Jetty Duffy-Matzner, was declared elected for the remainder of the 2016-2017 term.

Council Policy Committee: Council voted to fill four slots on the Council Policy Committee. There were eight nominees as follows: John R. Berg, Lisa Houston, Frank D. Blum, Lee H. Latimer, Mary K. Carroll, Doris I. Lewis, Dwight W. Chasar, and Barbara P. Sitzman. The four candidates who received the highest numbers of votes were declared elected for the 2016-2018 term. These are as follows: Lisa Houston, Frank D. Blum, Lee H. Latimer, and Mary K. Carroll,

Committee on Nominations and Elections: Council voted to fill five slots on the Committee on Nominations and Elections. There were ten nominees as follows: V. Dean Adams, Roland F. Hirsch, Matthew K. Chan, C. Marvin Lang, David A. Dixon, Les W. McQuire, Mary K. (Moore) Engelman, Donivan R. Porterfield, Joseph A. Heppert, and Ralph A. Wheeler. The five candidates who received the highest numbers of votes were declared elected for the 2016-2018 term. These are as follows: Roland F. Hirsch, C. Marvin Lang, Les W. McQuire, Mary K. (Moore) Engelman, and Donivan R. Porterfield.

New Procedures for Balloting and Preferential Voting

The Committee on Nominations and Elections (N&E) put forth new balloting and preferential voting procedures for the elections of President-Elect, District Directors, and Directors-at-Large. They passed and will be effective pending approval by the ACS Board of Directors. The procedures are as follows:

General Policy:

1. These detailed procedures add specificity to the provisions in Bylaw V of the ACS Governing Documents (www.acs.org/bulletin5). They have been developed by the Committee on Nominations and Elections (N&E) in cooperation with the ACS Secretary and General Counsel, and approved or subsequently amended by the ACS Council. They reflect the procedures currently used by N&E in national elections, with the exception of adding preferential voting requirements and procedures when electing two or more individuals to fill Director-at-Large positions. Nothing in these procedures is to be construed in a manner that is inconsistent with the Bylaw V, Sec. 11, c balloting procedures, established by N&E and approved by the Council Policy Committee, which address fair balloting,

anonymity, protection against fraudulent balloting, ballot archiving, and the timely reporting and archiving of balloting results.

2. Wherever possible, elections should result in the winning candidate (or candidates for certain elections as described below) receiving a majority of the valid votes cast. Depending upon the number of candidates on the ballot, this may result either by receiving a majority of single-choice ballots or by preferential voting as a result of receiving a majority of the remaining votes by having the second-preference votes of eliminated candidates added to their first-preference votes (recalculation of votes). Where multiple candidates for the same office are to be selected, a majority shall consist of more than half of the total number of ballots that remain valid at that step in the elimination process.

General Procedures:

1. Preferential voting will be used in elections whenever there are more than two candidates for a single position, or whenever there are multiple positions to be filled for the same office, such as typically occurs when electing Directors-at-Large or when selecting two candidates for District Director or President-Elect from among four (or more) nominees.

2. The preferential voting method that will be used under these procedures is known as the instant run-off method.

a. The preferential ballot shall afford the voter an opportunity to rank the candidates in order of preference. After the initial vote, if a candidate receives a majority of the first-preference votes cast, then that candidate shall be declared elected. If no candidate receives a majority of the first-preference votes cast, the candidate with the fewest number of first-preference votes is eliminated. The eliminated candidate's second-preference votes (i.e. the second-preference votes of those who cast their first-preference vote for the eliminated candidate) are redistributed to the remaining unelected candidates. When recalculating vote totals following the elimination of a candidate, those ballots on which no preference is indicated for any of the remaining candidates shall be deemed invalid in that and any subsequent rounds. In each of those rounds, a majority shall consist of a majority of the number of valid ballots that remain at that step in the elimination process. This procedure continues until a candidate receives a majority of the votes.

b. If there are multiple positions for the same office to be filled and only one candidate receives a majority of first-preference votes cast, then the candidate receiving the majority shall be declared elected. The elected candidate's second-preference votes are redistributed to the remaining unelected candidates. Additionally (or if no candidate receives a majority of first-preference votes cast), the candidate with the fewest first-preference votes is eliminated from further consideration; the second-preference votes of the eliminated candidate are redistributed to the remaining unelected candidates. The procedure continues until one candidate receives a majority. When a candidate receives a majority and is declared elected, the elected candidate's second-preference votes are redistributed to the remaining unelected candidates. After the second-preference votes are redistributed and no remaining candidate receives a majority, then the candidate with the lowest number of votes is eliminated and the eliminated candidate's second-preference votes are redistributed to the remaining unelected candidates. The process is repeated until the number of elected candidates equals the number of positions available.

c. When recalculating vote totals following the elimination of a candidate, those ballots on which no distinct preference is indicated for any of the remaining candidates shall be deemed invalid for this and any subsequent candidate elimination rounds.

d. Preferential voting as described in these procedures shall be used whether the ballot is paper or electronic. Hand-marked (or paper) ballots will be counted only if the voter's intent to vote in favor of a particular candidate or candidates can be reasonably determined from the marking(s) on the ballot. All valid votes are tallied. Where a determination of intent is needed, it will be made by the Chair of the Committee on Nominations and Elections, or his or her designee.

e. In the event of a tie for last place in the first round (i.e. two candidates have the same number of first preference votes), the candidate with the fewest second choice preferences will be eliminated and their second preference votes will be redistributed. In the event of a tie for last place in the second or succeeding rounds, the candidate with the lowest first round preference votes is eliminated. The eliminated candidate's second preference votes are redistributed in the next round. If the two candidates that are tied are not in last place, then the candidate with the lowest vote total in that round will be eliminated. The eliminated candidate's second preference votes are redistributed in the next round. In the event of a tie in the final round, Council balloting will break the tie per the ACS Bylaws.

Election Procedures:

1. President-Elect:

a. When there are two candidates, a single-choice ballot shall be used, and the candidate receiving the greater number of votes shall be declared elected.

b. When there are more than two candidates, or when Councilors are selecting two candidates from among several nominees, a preferential ballot shall be used as described (in the General Procedures) above.

2. Director-at-Large:

a. If there is only one position to be filled and there are two candidates, a single-choice ballot shall be used and the candidate receiving the greater number of votes shall be declared elected.

b. If there is only one position to be filled and there are three or more candidates, a preferential ballot shall be used as described (in the General Procedures) above.

c. If there are two or more positions to be filled and three or more candidates, a preferential ballot shall be used as described above. However, where two candidates must be selected, the preferential voting method as described above continues so that the first candidate receives a majority of the votes and then the second candidate receives the next majority of the votes.

3. District Directors:

a. Where there are two candidates, a single-choice ballot shall be used, and the candidate receiving the greater number of votes shall be declared elected.

b. When there are more than two candidates, or when Councilors are selecting two candidates from among several nominees, a preferential ballot shall be used as described above.

Pending approval by the ACE Board, the ACS Bylaws will be changed (Bylaw V, Sec. 2,d; Sec. 3,c; and Sec. 4,d and f.) The petition was approved by the ACS Committee on Constitution and Bylaws and they concluded that it will have a minor positive impact on ACS finances ($0 - $100K).

New Procedures for Member Expulsion

The Committee on Membership Affairs (MAC) put forth new procedures for the expulsion of ACS members for Council vote. They passed and will be effective pending approval by the ACS Board of Directors. The procedures are as follows:

Procedure for Expulsion of a Member

Section 1. Members of the American Chemical Society (hereinafter referred to as "SOCIETY") may be dropped from the rolls of the SOCIETY (hereinafter referred to as "expelled" or "expel") for conduct that in any way tends to injure the SOCIETY or to affect adversely its reputation, or that is contrary to or destructive of its objects as described in the SOCIETY's Constitution.

Section 2. The procedure to expel a member shall be initiated when the specific charges and reasonable substantiating evidence are submitted in writing to the Secretary of the SOCIETY and signed by at least five members of the SOCIETY (secretary@acs.org).

Section 3. The Secretary shall, without delay, forward the documented charges to the Chair of the Council Committee on Membership Affairs (MAC), who shall determine that the members submitting the charges are aware of the gravity of the actions and the procedures to be followed. If the Chair of MAC is the accused party, the Vice-Chair will act as substitute for the Chair. Within thirty days of receipt, the Chair of MAC shall appoint and call a meeting of a Hearing Subcommittee. The Hearing Subcommittee shall consist of not more than six members including those on the MAC Executive Committee and one or more other members of MAC with the longest tenure on the committee. Members of MAC who signed the request for expulsion, as described in Section 2 may not serve on the Hearing Subcommittee. In addition, any member of MAC who has a financial interest with either the accused or any of the accusers must recuse himself or herself from the proceedings. The Chair of MAC shall chair the Hearing Subcommittee, unless the Chair signed the request for expulsion, as described in Section 2. If the Chair of MAC cannot serve, the Subcommittee will elect its own Chair.

a. Within thirty days, the Hearing Subcommittee shall continue the expulsion process, dismiss the charges as ill founded, or find an alternative solution to the issue, and the Chair shall inform the accused member and those who brought the charges of the decision of the Hearing Subcommittee.

b. If the proceedings continue, the accused member shall be offered an opportunity to answer the allegations of the charges before the Hearing Subcommittee. Every reasonable effort shall be made to contact the accused member throughout this procedure. That effort shall include a certified letter to the last known address on the official SOCIETY membership rolls. This letter shall offer the accused member choice of one of the following options:

(1) The accused member may resign.

(2) The accused member may request a hearing by the Hearing Subcommittee. A two-thirds (2/3) vote of the members of the Hearing Subcommittee shall be required to expel the accused member.

(3) The accused member may choose not to respond and thus forfeit his/her membership in the SOCIETY.

c. Upon notification, the accused member shall have thirty days to make a written response to the allegations from the date of issuance of the notice described in Section 3, b, above. The Hearing Subcommittee shall decide how to proceed after reviewing the member's response. The Chair shall inform the member and those who brought the charges of the decision of the Hearing Subcommittee.

If no contact with the accused member can be made after a reasonable effort, the Hearing Subcommittee may expel the member in question with a two-thirds (2/3) vote of its members.

Section 4. Within thirty days, the accused member may appeal an adverse decision of the Hearing Subcommittee to the Council Policy Committee, which shall consider the appeal at its next regularly scheduled meeting, or at an earlier meeting specially called for the purpose of considering the appeal. Decisions of the Council Policy Committee are final, as of the date of the decision.

Section 5. An application for readmission by the charged member after an expulsion or resignation or after the initial statement of charges is received will only be considered by MAC after a minimum of five years have passed from the original expulsion or resignation. Members of MAC who signed the request for expulsion as described in Section 2, must recuse themselves from the readmission vote. The application must be approved by a two-thirds (2/3) vote of MAC. Effective TBD.

Pending approval of the ACS Board of Directors, the ACS Bylaws will need to be changed (Bylaw I, Sec. 5) MAC submitted a petition for this purpose and Council did vote on it at this meeting. The petition has been approved by the ACS Committee on Constitution and Bylaws and they have concluded that it will not have any impact on ACS finances ($0).

Five New International Chemical Sciences Chapters

Five legal applications were received for the formation of International Chemical Sciences Chapters. These are from the United Arabs Emirates (UAE), Peru, Nigeria, Brazil, and Australia. The petitions were initiated and approved by ACS members in good standing and residing in the relevant territories and meet all of the requirements of Bylaw IX of the Society. Each application includes a proposed budget for the operations of the Chapter, which includes no allotment of funds from the Society. The petitions have been reviewed by the ACS Committee on International Activities (IAC). Council also approved, but operations of the Chapters are contingent upon the approval of the ACS Board of Directors.

Bylaw Changes for Consideration Only

No petitions to change the Bylaws were presented for consideration at this meeting.

Reports of Elected Committees (Highlights)

Nominations and Elections (N&E)

N&E announced the candidates for the fall 2015 ACS national election:

Candidates for President-Elect, 2016: G. Bryan Balazs, Associate Program Leader, Lawrence Livermore National Laboratory, Livermore, CA and Allison A. Campbell, Associate Laboratory Director, Pacific Northwest National Laboratory, Richland, WA.

Candidates for Directors-at-Large, 2016-2018: Lee H. Latimer, Head of Chemistry, NeurOp, Inc., Oakland, CA; Willem R. Leenstra, Associate Professor, University of Vermont, Burlington, VT; Ingrid Montes, Professor, University of Puerto Rico-Rio Piedras Campus, San Juan, PR; Mary Jo Ondrechen, Professor of Chemistry and Chemical Biology, Northeastern University, Boston, MA; and Thomas W. Smith, Professor, Chemistry & Microsystems Engineering, School of Chemistry and Materials Science, Rochester Institute of Technology, Rochester, NY.

Candidates for District I Director, 2016-2018: Thomas R. Gilbert, Professor, Northeastern University, Boston, MA; Laura E. Pence, Professor of Chemistry, University of Hartford, West Hartford, CT.

Candidates for District V Director, 2016-2018: John E. Adams, Curators' Teaching Professor, University of Missouri-Columbia, Columbia, MO and Kenneth P. Fivizzani, Retired, Nalco Company, Naperville, IL

Council Policy (CPC)

The CPC Long-Range Planning Subcommittee was asked to review the way local sections and divisions are currently represented on Council. The Task Force is examining issues that affect the Divisor formulae set out in the Bylaws; for example, how sacrosanct is the rule that twenty percent of elected Councilors shall be elected by divisions, and eighty percent shall be elected by local sections? What would Council look like if the ratio were changed, for example to 70/30? Should there be more division representation on Council and what would be the impact? Would this result in more resources for divisions? Should a cap be placed on the number of Councilors per local section and divisions to ensure more balance? What does representation look like for international members of local sections and divisions? Comments on these questions can be submitted to President@acs.org.

Reports of Society Committees and the Committee on Science (Highlights)

Budget and Finance (B&F)

B&F reviewed the Society's 2015 probable year-end financial projection which expects a Net Contribution from Operations of $15.5 million, or $2.1 million higher than the Approved Budget. Total revenues are projected at $512.1 million, which at $481,000 favorable is essentially on Budget. Total expenses are projected at $496.6 million, which is $1.6 million or 0.3% favorable to the Approved. This variance is the result of lower-than-budgeted expenses across almost all major expense categories. Net assets are estimated to reach $185M (up from $145M), but need to be at least $250M to be in compliance with ACS guidelines.

Education (SOCED)

SOCED voted to approve revisions to the ACS Guidelines for Chemistry in Two-Year Colleges. The committee voted to make the pilot program of ACS International Student Chapters a permanent feature of the student chapters program. ACS has chartered 15 International Student Chapters since the pilot launched last year.

Science (ComSci)

ComSci voted to recommend approval of the draft ACS policy statement on energy, a notably improved statement on this critical economic and environmental issue. At this meeting, the committee sponsored a roundtable discussion with leaders of divisions, journals, and outside experts on moving advanced materials from discovery to application.

Divisional Activities (DAC)

DAC recently completed a review of a white paper to help divisions identify, evaluate, and pursue international engagement opportunities; received an update on several changes to the Meeting Abstracts Programming System (MAPS); was briefed on a recently created task force that seeks to enhance the content and functionality of the acs.org web pages that help division and local section volunteers execute their volunteer duties; and voted to fund 14 Innovative Project Grants (IPG) totaling $77,050. The Multidisciplinary Program Planning Group is proposing the following 2018 national meeting themes to the divisions for their consideration: Nexus of Food, Energy and Water (spring/New Orleans), and Nanotechnology (fall/Boston).

Local Section Activities (LSAC)

LSAC will award twenty Innovative Project Grant (IPG) grants totaling $39,886. This brings the total for 2015 to 34 IPG awards totaling over $75,000. Since the inception of the program, a total of 166 local sections have received at least one award. The committee voted to keep the current local section allotment formula in place for the next three years, and developed a new process for managing the annexation of unassigned territories by multiple sections.

Membership Affairs (MAC)

MAC reported that as of July 31, the ACS membership was 156,561; 2,055 fewer than on the same date in 2014. The number of new members who have joined this year is 14,457; 147 fewer than this time last year. The Society's overall retention rate is 84%. The number of international members has increased to 25,989; 1,014 higher than in July, 2014. The international retention rate is 85%. The committee intends to submit a petition for consideration in San Diego to permanently extend the Unemployed Member Dues Waiver Benefit period from two years to three years.

Economic and Professional Affairs (CEPA)

The committee reported that ACS ChemCensus data showed that Domestic Unemployment among ACS member chemists edged slightly upwards in the last year from 2.9% to 3.1%. Still, the current unemployment rate is lower than it was from 2009 to 2013. The ChemCensus also showed a modest salary increase year-over-year. For the first year since 2004, the percentage of ACS members

working in manufacturing increased. These trends are mirrored by a slight decline in the percentage of members in academia. Other workforce categories remained relatively flat.

Meetings and Expositions (M&E)

M&E accepted 9,271 papers for the Boston meeting. As of the Council meeting, total attendance for the meeting was 13,888, with the breakdown as follows:

Regular attendees:	8,129
Students:	3,462
Guests:	426
Exhibitors:	1,278
Exhibit-only:	593

Attendance at the fall national meetings since 2004 is as follows:

2004: Philadelphia, PA	14,025
2005: Washington, DC:	13,148
2006: San Francisco, CA	15,714
2007: Boston, MA:	15,554
2008: Philadelphia, PA:	13,805
2009: Washington, DC:	14,129
2010: Boston, MA:	14,151
2011: Denver, CO:	10,076
2012: Philadelphia, PA:	13,251
2013: Indianapolis, IN:	10,840
2014: San Francisco, CA:	15,761
2015: Boston, MA:	13,888

The Exposition had 475 booths with 325 exhibiting companies. There were nearly 5,500 downloads of the Boston Mobile App. The committee established a new Operations Subcommittee, responsible for monitoring the financial success of the national meetings, monitoring compliance with the National Meeting Long Range Financial Plan and the recommendations of the 2015 Task Force on Implementing National Meeting Financial Targets.

Constitution and Bylaws (C&B)

C&B certified 14 bylaws in 2015, and has reviewed bylaws for 9 local sections and 2 divisions since the spring meeting in Denver. The use of C&B's new bylaw templates and expedited bylaw process enables faster bylaw reviews than in previous years. New petitions to amend the Constitution or Bylaws must be received by the Executive Director no later than November 25, to be included in the Council agenda for consideration at the spring 2016 meeting in San Diego. Contact bylaws@acs.org for more information.

Reports of Other Committees (Highlights)

Chemical Safety (CCS)

CCS reported that it had been requested by the Chemical Safety Board (CSB) to assist in developing guidance with methods to recognize, assess, and control hazards in research laboratories. The CCS released its final report, “Identifying and Evaluating Hazards in Research Laboratories,” in 2015. Councilors and their institutions who are engaged in research were urged to consider using this guide to help keep their laboratories safer. This report is available at www.acs.org/safety.

Chemists with Disabilities (CWD)

In June 2015, CWD held a strategic planning workshop. Accomplishments include refined mission and vision statements, as well as 4 goals and strategies to achieve these goals. CWD and the ACS Standard Exams Institute initiated conversations regarding collaboration to address issues such as the unavailability in Braille of ACS Standard Exams, Practice Tests and Study Guides.

Community Activities (CCA)

CCA reported that one of its most popular programs is the nation-wide Illustrated Poem Contest (IPC). Local sections submit outstanding entries based on the theme of National Chemistry Week or Chemists Celebrate Earth Day. In the spirt of the IPC, the chair used various styles of verse to report on CCA’s public outreach event held at the Boston Museum of Science, and to announce the theme of National Chemistry Week 2016, "Solving Mysteries through Chemistry."

Committee on Ethics (ETHX)

ETHX requests that the committee be kept informed of any ethics-related activity sponsored by ACS entities such as committees, local sections and divisions. ETHX will provide this information (with proper credit) to national meeting attendees. Please contact the committee staff liaison at E_slater@acs.org.

International Activities (IAC)

IAC welcomed dignitaries from our sister societies in Canada, Cuba, India, Germany, Taiwan, the United Kingdom, and leadership of the Organization for the Prohibition of Chemical Weapons (OPCW), the US National Academies of Science, the Iraqi Chemical Society, ACS International Chapters, and ACS International Student Chapters. The committee approved the ACS Global Innovation Initiatives (Gii) Singapore White Paper and chose South America and Mexico for the 2017 joint ACS-Pittcon program to foster exchange and research collaboration in analytical chemistry.

Committee on Minority Affairs (CMA)

CMA focused its activities at this meeting on the 20th anniversary celebration of the ACS Scholars Program. The program has awarded more than $17 million in scholarship assistance since 1995 to enable 2,500 talented minority students to pursue their dreams of a degree in the chemical sciences. The new Scholars Endowment Fund now has commitments of more than $2 million. Nominations are being sought for the Stanley C. Israel Award. Instructions for nominations can be found at http://www.acs.org/stan-israel-award.

Committee on Patents and Related Matters (CPRM)

CPRM continues to monitor legislative and regulatory developments influencing intellectual property in ways that impact the chemical enterprise. The committee website is updated frequently and contains a wealth of helpful information about intellectual property matters relevant to those in the chemical enterprise.

Project SEED (SEED)

SEED announced another successful SEED program with the participation of 411 high school students. These students are currently placed in over 100 laboratories across the nation, under the supervision of over 400 volunteer scientists and coordinators in 39 states, the District of Columbia, and Puerto Rico. The committee awarded 32 first year non-renewable College Scholarships to SEED alumni in 17 states and Puerto Rico.

Committee on Public Relations and Communications (CPRC)

CPRC co-sponsored a number of events in Boston to showcase ways to increase public appreciation for chemistry: the PBS preview of “Mystery of Matter: Search for the Elements”; a symposium on the public perception of chemistry co-sponsored with *Chemical & Engineering News*, and the ACS Office of Public Affairs; ChemChamps; and Wikipedia edit-a-thon, co-sponsored with the Division of Chemical Information.

Senior Chemists Committee (SCC)

This meeting marked the third anniversary of the formation of the SCC at the Philadelphia National Meeting. SCC has been able to establish a number of initiatives through its provision of mini-grants to local sections to sponsor senior-related activities, several of which were recognized by the initial ChemLuminary awards at Boston. A committee retreat is being planned for this fall to identify priorities that will serve the SCC constituency as well as meeting the strategic goals of the committee.

Committee on Technician Affairs (CTA)

CTA is now accepting nominations for the 2016 National Chemical Technician Award. This annual award is presented in recognition of outstanding technical and communication skills, reliability, leadership, teamwork, publications, and presentations. For more information about the award, please visit the committee website at http://www.acs.org/cta.

Younger Chemists Committee (YCC)

The Program-in-a-Box effort continues to grow rapidly with a 43% increase in the number of disseminated boxes between the fall 2014 and February 2015, when 181 boxes were delivered to local sections and international chapters. At this meeting YCC participated in the 5th Younger Chemists Crossing Borders, an exchange which brings younger chemists from parts of Europe to the meeting. YCC is currently in discussions with N&E, ACS Webinars, ACS Office of Public Affairs, and the presidential candidates about holding a roundtable webinar, “Catalyze the Vote,” where the candidates can speak to the younger constituency about their vision for the Society in the future.

Actions of the Board of Directors

The Board's Committees (Executive Session):

The Board of Directors received reports from its Executive Committee, the Committee on Grants and Awards (G&A), the Society Committee on Budget and Finance (B&F), and the ACS Governing Board for Publishing.

On the recommendation of the Committee on Grants and Awards, and of the Committee on Public Relations and Communications, the Board voted to approve a Society nominee for the National Science Board Public Service Award.

On the recommendation of the Committee on Grants and Awards and of the Committee on Younger Chemists, the Board voted to approve a Society nominee for the 2016 Alan T. Waterman Award.

On the recommendation of the Committee on Budget and Finance, the Board voted to approve an advance member registration fee of $415 for national meetings held in 2016; to authorize a new program funding request for the ACS Festival Series program; and to reauthorize funding for the ACS Science Coaches program.

The Executive Director/Chief Executive Officer's Report

The Executive Director/CEO and his direct reports updated the Board on the activities of Chemical Abstracts Service (CAS), the ACS Publications Division, and the Society's Secretary and General Counsel.

On the recommendation of the Joint Board-Council Committee on Publications, the Board voted to reappoint an Editor-in-Chief of an ACS journal.

Other Society Business

The Board also:

- Held a discussion on strategic questions related to the health and strength of local sections and divisions.
- Received reports from the Presidential Succession on their symposia and events in Boston, and planned activities for 2016.
- Approved a resolution to extend sincere congratulations to the Sociedad Química de México on the occasion of the 50th Congreso Mexicano de Química, 7-10 October 2015, in Querétaro, Querétaro México.

The Board's Regular (Open) Session

The Board held an overflow open session on Sunday, August 16, that featured George Whitesides. Professor Whitesides' topic was "Reengineering Chemistry." Following the presentation, members of the presidential succession and the Executive Director and CEO offered brief reports on their activities. (The officers provided more extensive reports as part of their reports to the Council.)

Andrea Twiss-Brooks and Bonnie Lawlor, CINF Councilors

Joint Board-Council Committee on CAS

The Committee on Chemical Abstracts Service (CCAS) met in Executive Session on August 14, 2015, where CAS management reported on highlights from the first half of 2015, including updates on the portfolio of new solutions designed to enable discovery, and advance workflows for scientific researchers and patent professionals around the world. CAS President, Manuel Guzman, reported that CAS continues to execute on its Strategic Plan for Growth and Optimization with the launch of three new products thus far in 2015 (PatentPak, NCI Global, and upgraded CHEMCATS) while sustained financial performance continues to support ACS initiatives.

- PatentPak is a robust, new patent workflow solution available in SciFinder. Designed to radically reduce time spent acquiring and searching through multiple patents to find vital chemistry, PatentPak saves users up to half the time they spend researching patents by providing instant access to hard-to-find chemistry in patents and patent families in languages users know.

- NCI Global, an online regulatory solution, provides access to inventories and regulatory information essential for any organization that manufactures, imports, exports, or transports chemicals. With inventories and regulatory lists organized by country, users get right to the information they need.

- CHEMCATS (Chemical Catalogs) is a catalog database containing information about commercially available chemicals and worldwide suppliers. CHEMCATS is updated at least two times per week with new and revised catalog information. New opportunities for CHEMCAT suppliers (e.g., displaying company logo, and featured listings) are now available.

Committee members were pleased to learn that development on new technology platforms for the CAS flagship products, SciFinder and STN continues, with another release of STN launched in August, and completion of the product roadmap, along with defined product capabilities for revitalizing SciFinder. In addition, CAS assigned the 100 millionth CAS Registry Number to a substance designed to treat Acute Myeloid Leukemia, in the 50th anniversary year of the CAS REGISTRY, the world's largest database of unique chemical substances.

A preview demo of MethodsNow delighted committee members. This new solution for analytical scientists, coming later in 2015, provides access to the largest collection of methods and will save researchers time, making method selection and optimization simple and efficient.

CCAS held a lively discussion and gave CAS management useful input on a range of topics and issues, including new product concepts. CCAS continues in its important role as a conduit of information between Society members, the ACS Governing Board for Publishing, and CAS management, providing valuable feedback on current and future initiatives.

Grace Baysinger, Chair, Joint Board-Council Committee on CAS

"Beyond CASSI" is a newly released document that contains short journal title abbreviations from the early chemical literature, and other historical reference sources that may not be listed in the print version of CAS Source Index (CASSI), or the free online CASSI Search Tool (see "About" section).

Joint Board-Council Committee on Publications

The Joint Board-Council Committee on Publications (JBCCP) met in Boston and discussed the following.

The monitoring reports for *Biochemistry, Journal of Proteome Research* and *Journal of the American Chemical Society* were presented, discussed thoroughly, and accepted with thanks. Editor reappointments were reviewed and recommendations were made.

Staff presented an overview of *C&EN*'s significant editorial and marketing initiatives. *C&EN*'s "The Talented 12" issue (http://talented12.cenmag.org) highlighted twelve path-paving young researchers and entrepreneurs who are using chemistry to solve global problems. *C&EN* is collaborating with the Spanish organization, Divulgame.org, to translate a selection of content into Spanish, and offer it to readers via *C&EN*'s website and social media channels. *C&EN* also created *ACS Chemoji*, a mobile keyboard app for Android and iPhone that allows users to share chemistry-themed "emojis" via text message and social media. This app was built in collaboration with ACS Publications and Chemical Abstracts Service, and was launched at the ACS National Meeting in Boston. Marketing initiatives in support of *C&EN*'s advertising sales goals were summarized for the committee.

Journals Publishing staff presented the early results of a systematic survey of ACS corresponding authors. Initial responses indicate that authors choose a journal for submission on the basis of scope, impact, citations, and speed to decision. Authors are generally satisfied with the current peer review system, and rank interactions with ACS editors very highly. The committee suggested that the survey comments be returned to the respective journals to inform both operational and strategic planning.

Staff will ask the Editors-in-Chief of ACS Publications to refine their respective policies of what constitutes "Supporting Information" with respect to both data and to references at its 2016 Conference of Editors, and to clarify such policies in published information for authors.

The committee and staff expressed their sincere appreciation to Dr. Stephanie Brock for her outstanding service as chair of the Joint Board-Council Committee on Publications from 2013 to 2015.

Stephanie Brock, Chair, Joint Board-Council Committee on Publications

C&EN
CHEMICAL & ENGINEERING NEWS

How the Internet Changed Chemistry
Volume 93, Issue 32
Issue Date: August 17, 2015
http://cen.acs.org/articles/93/i32/Internet-Changed-Chemistry.html

Another Committee, Another Acronym: Demystifying SOCED

Since 2013, I have been an Associate Member of the Society Committee on Education (SOCED, http://www.acs.org/content/acs/en/about/governance/committees/education.html). I wanted to take this opportunity to describe the work of SOCED and hopefully encourage others to investigate ways that they can participate within ACS beyond CINF, but yet still support our mission and goals as a division.

Simply, the mission of SOCED is to support the development and improvement of ACS educational programs from kindergarten through graduate school and beyond. The Committee works very closely with a number of staff from the ACS Education Division (the staff side of ACS, not the technical division), including the Director of the Education Division, Dr. Mary Kirchhoff, who is the staff liaison to SOCED.

SOCED is also separate from the Committee on Professional Training (CPT), as well from the ACS Division of Chemical Education (CHED). We communicate across our groups, but each has a slightly different focus and mission. Some more information is taken directly from the "About" section of the SOCED web page.

SOCED's responsibilities include:

- Implementing ACS policies in chemical education
- Developing reports and recommendations for the ACS Board and Council on ACS policies and programs related to chemical education
- Receiving, reviewing, and making recommendations to the Board and Council on proposals for policies and programs in chemical education
- Acting in an advisory capacity on matters relating to chemical education
- Recommending approval or disapproval of requests for the funding of new or unbudgeted items related to chemical education
- Drafting statements for ACS Board approval on the annual budgets for both the National Science Foundation's education programs and the U.S. Department of Education.

As you can see, SOCED deals with a variety of high-level issues related to chemical education. If you are not aware that ACS has positions on policy issues, I suggest you browse the following page: http://www.acs.org/content/acs/en/policy/publicpolicies.html.

For more concrete examples, at the most recent ACS National Meeting in Boston the following topics that SOCED considered were:

- Discussing the Education Division's efforts to create any online Individual Development Plan (IDP) system to be rolled out this fall to graduate students and post-docs
- Voting to approve the *ACS Guidelines for Chemistry in Two-Year Colleges*
- Providing feedback on strategic planning for SOCED
- Discussing the need for developing standard student learning outcomes for General Chemistry and initiating a taskforce to investigate this further
- Hearing updates from the American Association of Chemistry Teachers (AACT), Office of Public Affairs (OPA), the Committee on Professional Training (CPT), and the Committee on Chemical Safety.

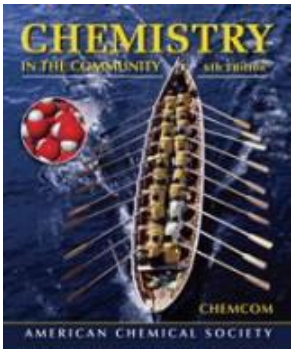
How is SOCED organized?

SOCED consists of fifteen members, ten of whom must be ACS councilors. There are usually ten additional associates. One or two consultants may also be appointed. The Committee Chair, who must be an ACS councilor, is appointed each year by the ACS President and Board Chair. The SOCED Chair may serve no more than three years in this capacity. SOCED members may serve up to nine consecutive years.

SOCED has three subcommittees: Subcommittee A (Precollege), Subcommittee B (College/University) and Subcommittee C (Olympiad). As an Associate Member of SOCED, I can speak, but I am not allowed to vote when SOCED meets as a whole. However, I am a full voting member of Subcommittee B. Subcommittee B takes the lead on issues related to chemical education in higher education and brings appropriate initiatives, feedback, or ideas to the full SOCED group. Topics Subcommittee B has considered at the last several meetings include:

- Providing feedback to the ACS Education Division staff on ways to promote the new Individual Development Plan (IDP) system, as well as what faculty are looking for when it comes to professional development opportunities
- Discussing the implementation of international ACS student chapters
- Hearing reports from the Undergraduate Programs Advisory Board and the Graduate Education Advisory Board (GEAB)
- Choosing judges for the ChemLuminary ACS Student Chapter Interaction Award.

Many of the programs coordinated and offered by the ACS Education Division can be found here: http://www.acs.org/content/acs/en/education.html.



Besides my general work on SOCED and Subcommittee B, I have also participated in a couple of task forces over the years. On one, I helped a small group discuss the future of *Chemistry and the Community*, a high school level textbook.

Another task force, of which I am now currently the chair, has been brought together to help draft a practitioner statement on the importance of hands-on laboratory activities. The ACS has a position statement on the importance of hands-on laboratory activities to help guide policy at the national, state and local levels (http://www.acs.org/content/acs/en/policy/publicpolicies/invest/computersimulations.html), but SOCED agreed that another, more detailed document for practitioners (teachers, professors, etc.) is needed.

If you have any further questions about SOCED, please feel free to contact me at jgarrita@umd.edu.

Jeremy Garritano, Associate Member, Society Committee on Education

Sponsor Announcements

CINF Social Networking Events

Thanks to all who helped the Division of Chemical Information to put together a great set of social events for the fall ACS National Meeting in Boston. As always, it started with the Welcoming Reception on Sunday evening. We had excellent attendance at the reception, more than we anticipated. There was a nice selection of food and drinks, and a good time was had by all. The division is grateful to the sponsors for this event: Springer/*Journal of Cheminformatics*, Optibrium, PerkinElmer, AAAS/Science, and Thieme Chemistry.

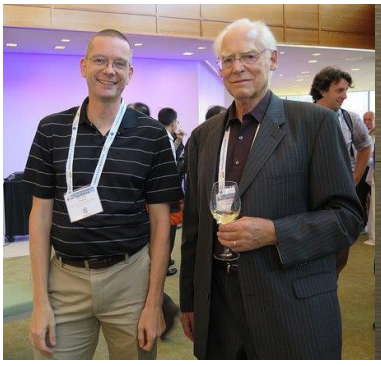
Tuesday's first event was the Division of Chemical Information Luncheon sponsored exclusively by the Royal Society of Chemistry. This event sold out once again with capacity at 95. A great luncheon hosted a very entertaining talk by our invited speaker, Michele Derrick, who spoke about "CAMEO: a database for technical information on materials in museums."

Tuesday evening was the Herman Skolnik Award Reception. This year's recipient was Dr. Jürgen Bajorath. The reception was very well attended with something to eat for everyone, from sushi to carving stations. Many thanks to ACS Publications, Pfizer, Novartis, Cresset, and Schrödinger for sponsoring this event.

Special thanks to our division colleagues, Graham Douglas, Michael Qiu, Erin Davis, and Rachelle Bienstock, who put in time and effort fundraising, and making arrangements for rooms, menus and speakers for our social events. We would also like to thank Bio-Rad, whose marketing department handles all our graphic arts and banner production. And, of course, we all greatly appreciate our generous sponsors, without whom we would not have been able to organize such a great line-up of events in Boston for everyone to enjoy.

We are now planning events for San Diego spring 2016 and look forward to seeing all of you there!

Philip Heller, Chair, CINF Fundraising Committee



CINF photos from the fall 2015 ACS National Meeting are at:
https://www.flickr.com/photos/cinf/collections/
The photos were taken by Wendy Warr and Stuart Chalk.

ACS Publications on your Tablet or Smartphone

The Publications Division of the American Chemical Society recently released ACS2Go, an improvement to its mobile platform that offers readers enhanced features and an improved reading experience. It is ACS Publications' mobile platform optimized for tablets and smartphones, allowing readers to pair their mobile device easily to an organization, to access full-text articles both on and off campus. Library patrons have access to cutting-edge ACS research, anytime and anywhere, with ACS2Go. They can access full-text articles from journals that have been licensed by the institution's library, which means that patrons can visit http://pubs.acs.org/ from their mobile device while on an authenticated institutional wi-fi network. Their mobile device will be automatically paired with the institution's access rights. The pairing will then enable them to access content while offsite, for up to four months, and will be refreshed each time they access ACS2Go while within the institution's network. Learn more at pubs.acs.org/acs2go.

Ideas, Insights and Advices for the Scientific Community

ACS Axial is the newest initiative from ACS Publications. It is an interactive blog designed to be an indispensable resource for authors, researchers, librarians and customers; not only does it present, but it also seeks stories covering the latest innovations in scientific research, publishing and beyond.

Visit http://axial.acs.org/ to read some of the latest pieces, authored by ACS Publications' staff, editors, authors, and librarians.

Software and Services for Small Molecule Discovery and Design

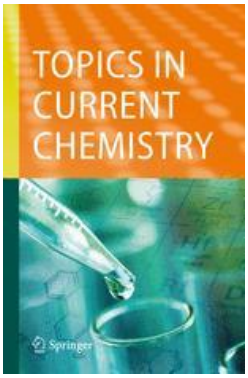
Cresset's patented methods deliver novel, realistic results that help chemists discover, design and optimize the best small molecules for their projects.

Our software and consulting services are used on a daily basis by computational, medicinal, and synthetic chemists from the world's leading research organizations. Eight out of the top 10 pharmaceutical companies use Cresset technology in their drug discovery research.

Our robust scientific methods use 3D molecular shape and electrostatics to understand the chemical interactions that underpin biological activity.

See our presentations and posters from ACS Boston 2015 and find out how to get a free software evaluation www.cresset-group.com.

Book series *Topics in Current Chemistry* to be re-launched as online journal



Springer is re-launching the book series *Topics in Current Chemistry* (*TCC*) as an online journal. The individual contributions will appear in six electronic issues per year with the first articles to be published in the fall of 2015. *Topics in Current Chemistry* is well known for presenting high-quality reviews of the present position and future trends in modern chemical research.

Nobel laureate and series editor, Jean-Marie Lehn, said: “The transformation of *Topics in Current Chemistry* into a journal of reviews is an exciting development. The scientific community values the book series greatly and, now that *TCC* is a journal, authors will have their contributions indexed, found, read, and cited faster.”

The journal will continue to cover all areas of chemical science, including the interfaces with related disciplines such as biology, medicine, physics and materials science. The new journal format will allow editors and authors to use tools facilitating manuscript preparation, submission, review and tracking. Moreover, authors will have the opportunity to publish their articles using the open access publishing model under the Springer Open Choice program. The editorial board of world-leading chemists will transition from the book series to the journal. Wai-Yeung Wong of the Hong Kong Baptist University has been newly appointed as editor for the journal.

“At this time, we aim to further strengthen the reputation and impact of this successful book series by publishing it as a journal. Topical review articles covering present and future trends in modern chemical research will be featured. *TCC* will benefit from this endeavor in terms of academic impact, journal citations and visibility,” said Wong. “I am also looking forward to working with thematic guest editors and editorial board members to develop *TCC* into a world-leading academic journal of chemical sciences."

Topical collections of in-depth reviews will provide comprehensive insights into areas where new research is emerging and will be viewable online as they develop. Articles belonging to a topical collection will be highlighted with a button “Topical Collection” directing the reader to a listing of all articles in the same collection. Once completed, topical collections will become available in hardcover print format.

Elizabeth Hawkins, Senior Editor at Springer, said: "The journal format of *Topics in Current Chemistry* allows us to disseminate new material better and meet the expectations and demands of 21st-century chemists. At the same time, we are able to build on the historical significance of the series, which dates back to 1949."

Topics in Current Chemistry, ISSN 2364-8961 (digital version). Homepage: www.springer.com/41061

Thieme Chemistry

Science of Synthesis 4.1 new release and video

In June of this year, Thieme Chemistry released Science of Synthesis (SOS) 4.1, the latest version of its unique full-text resource for methods and experimental procedures in synthetic organic chemistry. Included are the latest Knowledge Updates and additions from the Reference Library: a total of approximately 1,650 printed pages of new material. An enhanced interface design and increased content linking through Digital Object Identifiers further enrich the user experience.

Five hundred new SOS Knowledge Updates pages include an entirely new chapter on five-five-fused hetarenes featuring examples of more unusual selenium and tellurium systems. The use of supercritical carbon dioxide as a reaction medium for organic synthesis is another focus.

The available content from the Science of Synthesis Reference Library has also been expanded to include two new volumes comprising a total of 1,168 printed pages. C-1 Building Blocks in Organic Synthesis (2 vols.), edited by Piet W. N. M. van Leeuwen and written by 54 experts, reviews a wide range of reactions to form C-C bonds, including reactions involving catalytic methods, an area that has seen significant developments in recent years. The authoritative overview includes contributions on the first catalysts to enable the introduction of fluoromethyl groups in aromatics.

Science of Synthesis 4.1 also comes with an enhanced interface design that features book covers with zoom functionality to facilitate navigation and allow for a quick overview of volume editors. Also included are improved linking to the primary literature through Digital Object Identifiers, a number of bug fixes, and general software improvements.

A new video introduction shows in an entertaining way how researchers can benefit from using Science of Synthesis: https://www.youtube.com/watch?v=rzDru_VLuaQ.

To get access to Science of Synthesis 4.1 or a free trial please visit: http://sos.thieme.com.

For more information about Science of Synthesis please visit the website at: www.thieme-chemistry.com/sos/.

CINF Officers and Functionaries

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

Chair Elect:
See Chair

Past Chair/Nominating Chair:
Ms. Judith Currano
University of Pennsylvania
currano@pobox.upenn.edu

Secretary:
Ms. Leah McEwen
Cornell University
lrm1@cornell.edu

Treasurer/Finance Committee Chair:
Dr. Rob McFarland
Washington University
mcfarland@wustl.edu

Councilor:
Ms. Bonnie Lawlor
Chescot Publishing, Inc.
chescot@aol.com

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

Alternate Councilor:
Mr. Charles Huber
University of California, Santa Barbara
huber@library.ucsb.edu

Alternate Councilor:
Dr. Guenter Grethe
Scientific Research Consultant
ggrethe@att.net

Program Committee Chair:
Dr. Erin Davis
Cambridge Crystallographic Data Centre
erinbolstad@gmail.com

Membership Committee Chair:
Dr. Donna Wrublewski
California Institute of Technology
dtwrub@caltech.edu

Archivist/Historian:
Ms. Bonnie Lawlor
See Councilor

Audit Committee Chair:
TBD

Awards Committee Chair:
Ms. Andrea Twiss-Brooks
University of Chicago
atbrooks@uchicago.edu

Careers Committee Co-Chairs:
Ms. Susan Cardinal
University of Rochester
scardinal@library.rochester.edu

Ms. Pamela Scott
Pfizer
pamela.j.scott@pfizer.com

***Chemical Information Bulletin* Editors:**
Ms. Svetlana Korolev (summer & winter)
University of Wisconsin, Milwaukee
skorolev@uwm.edu

Dr. Vincent Scalfani (spring & fall)
University of Alabama
vfscalfani@ua.edu

Assistant Editors:
Dr. David Shobe (summer & winter)
Patent Information Agent
avidshobe@yahoo.com

Ms. Teri Vogel (spring & fall)
University of California, San Diego
tmvogel@ucsd.edu

Communications and Publications Committee Chair:
Dr. David Martinsen
American Chemical Society
d_martinsen@acs.org

Constitution, Bylaws & Procedures:
Ms. Bonnie Lawlor
See Councilor

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

Fundraising Committee Chair:
Mr. Philip Heller
Thieme Publishers
philip.heller@thieme.com

Tellers Committee Chair:
Ms. Susan Cardinal
University of Rochester
scardinal@library.rochester.edu

Webmaster:
Ms. Patti McCall
University of Central Florida
patti.mccall@ucf.edu

Contributors to this Issue of *CIB*

Articles & Sponsor Announcements

Grace Baysinger
Rachelle Bienstock
Bob Buntrock
Christine Casey
Susan Cardinal
Stuart Chalk
Debra Davis
Erin Davis
David Evans
Guenter Grethe
Philip Heller
Svetlana Korolev
Bonnie Lawlor
Ye Li
Daniel Lowe
David Martinsen
Leah McEwen
Steffen Pauly
Sue Peffer
Steven Rosenthal
David Shobe
Keith Taylor
William Town
Andrea Twiss-Brooks
Wendy Warr
Erin Wiringi

Editing & Production

Svetlana Korolev
Bonnie Lawlor
Patti McCall
David Shobe
Wendy Warr

A CINF directory including mail addresses, fax and phone numbers of the Executive Committee, Committee Chairs, Divisional Representatives, and other Functionaries is at the CINF website.