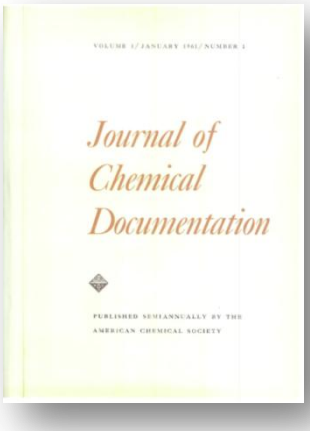
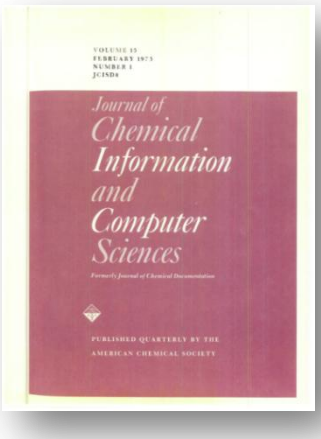
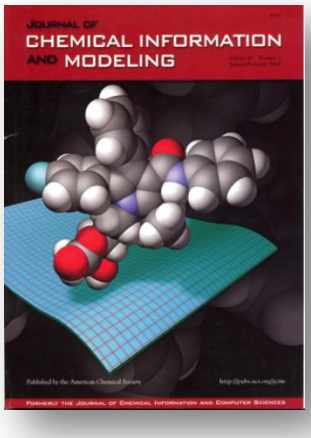
The *Journal of Chemical Information and Modeling*'s 50th Anniversary Symposium, ACS National Meeting, Boston, August 23, 2010

A report by Wendy Warr

The first issue of the *Journal of Chemical Documentation* appeared in 1960. The name of the journal changed to the *Journal of Chemical Information and Computer Sciences* ("*JCICS*") in 1975 and to the *Journal of Chemical Information and Modeling ("JCIM")* in 2005.







This year, 2010, is thus the 50th anniversary of the journal. The anniversary was marked by a symposium shared between CINF and COMP Divisions at the fall 2010 ACS National Meeting and was celebrated by an excellent reception sponsored by ACS Publications, where we all enjoyed useful networking and renewal of old acquaintances (http://pubs.acs.org/page/jcisd8/anniversary/50/index.html).

Over the fifty years of the journal there have been only four main editors (http://pubs.acs.org/page/jcisd8/anniversary/50/editors.html). Herman Skolnik (whose name is honored by the CINF Award) was the editor from 1960 to 1982. Tom Isenhour served as editor from 1982-1989, and he was succeeded by Bill Milne (1989-2004). Associate Editors were appointed in 1989: Pierre Buffet (1989-1997), Reiner Luckenbach (1989-1999), Kenny Lipkowitz (1993-2005), Tony Hopfinger (since 1993), Dušanka Janežič (since 2001) and myself (since 1989). Bill Jorgensen has been editor-in-chief of *JCIM* since 2005. He also edits JCIM's very successful, new sister journal, the *Journal of Chemical Theory and Computation* ("*JCTC*").

The first speaker at the 50th anniversary symposium was Johnny Gasteiger, whose first paper appeared in *JCICS* in 1977. Volume 46, issue 6 of *JCIM* in 2006 was his sixty-fifth birthday present: an issue in his honor. Johnny related how his early publication on the separation of π and σ systems (Gasteiger, G. A Representation of π Systems for Efficient Computer Manipulation. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 111–115) had devolved into a system, RAMSES, overcoming the limits of the connection table (Bauerschmidt, S.; Gasteiger, J. Overcoming the Limitations of a Connection Table Description: a

Universal Representation of Chemical Species. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 705–714). The Molecular Structure Encoding system (MOSES) is a C++ toolkit based on RAMSES. Johnny also outlined his work on reactions, the most current system being THERESA; the calculation of structure descriptors (now available in ADRIANA.Code); 3D structure generation (CORINA); artificial neural networks; and biochemical pathways (BioPath). The research efforts of Johnny's team over the years have led to range of products now marketed (under the aforementioned names) by Molecular Networks (http://www.molecular-networks.com/).

Bill Milne's talk also had a historical perspective. He started by outlining the history of the journal (some of which I have given above), but his main theme was those cheminformatics problems that have been solved and those that are proving intractable. Structure drawing, substructure search, and structure codes fall into the first category. Conversion of names into structures (and *vice versa*) and ligand-protein binding are largely solved. Properties estimation is "somewhat solved". Much of the research in these fields has been published in the journal. Protein-protein binding requires much more exploration, and Markush searching is an open problem, according to Bill, but he did not have time to go into detail about the many papers in *JCICS* on the subject.

A unique aspect of cheminformatics is that it has been heavily influenced and shaped by the needs of the pharmaceutical industry. Dimitris Agrafiotis reflected on experiences of the past and explored the possibilities he saw for the industry in the future: possibilities lying in the convergence of chemistry, biology, and information technology. First he talked about the world before and after "ABCD" (Agrafiotis, D. K. *et al.* Advanced Biological and Chemical Discovery (ABCD): Centralizing Discovery Knowledge in an Inherently Decentralized World. *J. Chem. Inf. Model.* **2007**, *47*, 1999–2014). Nowadays there are sophisticated tools for SAR analysis (Agrafiotis, D. K.; Wiener, J. J. M. Scaffold Explorer: an Interactive Tool for Organizing and Mining Structure-Activity Data Spanning Multiple Chemotypes. *J. Med. Chem.* **2010**, *53*, 5002–5011) but ABCD goes beyond decision support, and also embraces electronic laboratory notebooks and sequence searching, for example.

The system must also go beyond discovery. Mining of electronic medical records involves handling massive amounts of data usually in SAS datasets. An ABCD plugin will address that problem too. Pharma is an industry in stress. The good times are over; the future will be defined by in-licensing, pre-competitive collaborations, Asian expansion, a surge in biologics, the increasing role of government and academia, translational research, public data, the open source movement, commoditization of medicinal chemistry and other functions, outsourcing, consolidation of software vendors, tougher problems, and return on investment.

In contrast, Val Gillet described some very recent research not yet submitted to *JCIM* (although early results will appear in *Molecular Informatics*). Her team has been working on applications of wavelets in virtual screening, in particular using GRID fields which model the interactions which a small molecule can make with a receptor. These fields are cumbersome to store and compare but they can be compressed using wavelet transformation. This is a technique for representing signals by decomposing

them into components: smoothed, or approximated components, and details or differences which can be ignored. Gillet's team has experimented with Harr compression (as used in JPEG) and Haar thumbnails. They have applied wavelet thumbnails (low-resolution approximations of finely sampled GRID fields) without loss of information. Val also demonstrated other applications, including the development of an alignment method to enable the comparison of the wavelet representations of GRID fields in arbitrary orientation.

Jürgen Bajorath's presentation was remarkable in that it was given remotely using Skype, because Jürgen was prevented from traveling at the last minute. This is not the first time that CINF has used this technique (the first was a presentation by Tony Williams in Salt Lake City in 2009) but it still caused some excitement. Rajarshi Guha changed the slides in Boston while Jürgen presented, with video, from Bonn, Germany.

Jürgen described some research on privileged substructures. Many privileged substructures have been proposed but the existence of truly privileged structural motifs has remained controversial. Many scaffolds thought to be specific to a target class occur in compounds active against other types of targets. Jürgen's team investigated whether molecular scaffolds do exist that exclusively occur in ligands of individual target families. They used Bemis-Murcko scaffolds, carried out systematic data mining of publicly available compound data (BindingDB and PubChem) and defined target communities on the basis of ligand-target networks. The nodes were targets, the edges target pair sets, and the edge width the number of shared compounds. In 18 target communities, 206 diverse hierarchical scaffolds were identified, each represented by at least five compounds, which exclusively bound to targets within one of the target communities. In contrast, most scaffolds that exclusively bind to a single target within a community are only represented by one or two compounds in public domain databases. A subset of community-selective scaffolds displays a notable tendency to produce compounds with different target selectivity. The analysis was extended to ChEMBLdb and it was found that BindingDB and ChEMBLdb contain complementary target and scaffold information.

Peter Johnson spoke next, describing work on automated retrosynthetic analysis carried out by workers in Leeds together with SimBioSys and Pfizer. Many systems for computer aided organic synthesis design were developed in the last century (LHASA, SYNCHEM, IGOR, EROS, WODCA, SynGen etc.) but none has achieved significant user acceptance, partly because such systems required manual creation of reaction knowledge bases, a time consuming task which requires considerable synthetic chemistry expertise. ARChem (a program developed by Peter and his co-workers) circumvents this problem by automated abstraction of transformation rules from very large databases of specific examples of reactions (Law, J. *et al.* Route Designer: a Retrosynthetic Analysis Tool Utilizing Automated Retrosynthetic Rule Generation. *J. Chem. Inf. Model.* **2009**, *49*, 593–602).

Mapping reactions and finding the initial reaction core are solved problems; the hard part is extending the core to non-reacting atoms. Identifying the precise structural characteristics of each reaction often requires knowledge of the reaction mechanism. Another challenge is minimizing the combinatorial explosion inherent in automated multistep retrosynthesis. One process involved in that is removing interfering functionality. This can be done using statistics on functional groups derived from reaction

databases. Peter's team has been working on optimum constraint of the extended core and reducing, rather than increasing the number of rules derived by ARChem; chemists can be used to generate meta rules for reaction mechanisms. Peter concluded by illustrating some other improvements to ARChem, such as ordering of search results, which matter to the user but do not represent great technical advances.

Like many of the speakers, Michael Gilson has been on the editorial advisory board of *JCIM* but Michael's talk in the anniversary symposium was more related to *JCTC* than to *JCIM,* and indeed the work presented has been published in *JCTC (*Gilson, M. K. Stress Analysis at the Molecular Level: a Forced Cucurbituril-Guest Dissociation Pathway. *J. Chem. Theory Comput.* **2010**, *6*, 637–646). Michael presented molecular dynamics simulations consistent with long-ranged entropy effects throughout a protein upon binding a peptide, and explained why the concept of mechanical stress may be useful in thinking about such effects. His results suggest that computational stress analysis can provide mechanistic insight into supramolecular systems.

Elizabeth Amin also presented the results of some recently published research, this time published in *JCIM* (Chiu, T.-L. *et al.* Identification of Novel Non-Hydroxamate Anthrax Toxin Lethal Factor Inhibitors by Topomeric Searching, Docking and Scoring, and in Vitro Screening. *J. Chem. Inf. Model.*, **2009**, *49*, 2726–2734). The lethal factor (LF) enzyme is secreted by *Bacillus anthracis* as part of the anthrax lethal toxin. To date, no LF inhibitor is available as a therapeutic or preventive agent. Amin's team has identified five promising novel LF inhibitor scaffolds with low micromolar inhibition, using topomeric shape-based searching techniques.

Tudor Oprea addressed the issue of "druglikeness". He and Oleg Ursu have used extended connectivity descriptors computed by the Morgan algorithm and extracted them as SMARTS queries. In a method rooted in the information gain concept, already applied to derive selection rules in decision trees, they aimed at a better separation between drugs and non-drugs (Ursu, O; Oprea, T. I. Model-Free Drug-Likeness from Fragments. *J. Chem. Inf. Model.* **2010**, *50*, 1387–1394). The most discriminating atom environments (having the highest information gain) were selected as model-free druglike filters.

Tudor concluded, however, that there is a danger in relying indiscriminately on machine learning techniques that artificially separate drugs from non-drugs, especially in regard to the Available Chemicals Directory (ACD). This is likely to influence the usefulness of such classifiers negatively, as 40% of the "non-drugs" are similar to drugs. Oprea uses the term "model-free" to emphasize the fact that his method does not use kernel functions and does not force ACD into a negative label, but he admits that any learning process actually relies on models. Ultimately, "druglikeness" is defined by regulatory agencies and cannot be predicted. Oprea defined three difficulties: the drug dataset is small (some people claim that there are 8,000 drugs but Tudor can find only 3,800); the drug character of molecules changes over time as drugs are withdrawn from the market, and drugs have high heterogeneity (from lithium to cyclosporine).

Alex Tropsha talked about "chemocentric informatics", or enabling bioactive compound discovery through structural hypothesis fusion. The information resources available to us have broadened

dramatically including large chemical genomics databases (e.g., ChEMBL, PubChem, PDSP, ToxCast), digital libraries (e.g., PubMed), gene expression profiles (e.g., cmap), and others. To address some of the limitations of QSAR models Alex suggests adding cheminformatics to "omics". He described the use of digital libraries (Baker, N.C.; Hemminger, B. M. Mining connections between chemicals, proteins, and diseases extracted from Medline annotations. *J. Biomed. Inform.* **2010**, *43*(4), 510-519) for establishing new datasets to analyze the relationships between chemical structure and biological activity. Alex' team has transformed assertional metadata into a database for modeling toxicity (Rodgers, A. D. *et al.* Modeling Liver-Related Adverse Effects of Drugs Using kNearest Neighbor Quantitative Structure-Activity Relationship Method. *Chem. Res. Toxicol.* **2010**, *23*(4), 724-732).

Data curation, however, is vital (Fourches, D.; Muratov, E.; Tropsha, A. Trust, but Verify: on the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research *J. Chem. Inf. Model.* **2010**, *50*, 1189–1204). Alex illustrated how computational models help in detecting and correcting erroneous data and he described a study combining QSAR modeling, virtual screening, text mining, and gene expression profiling for identifying novel, experimentally confirmed, high-affinity GPCR ligands as potential anti-Alzheimer drug candidates. He concluded that both chemical and biological data in integrated databases should be carefully curated, that QSAR models have the power of correcting erroneous biological data, and that structural hypothesis fusion and focused experimental validation afford opportunities for drug (re)profiling.

Bobby Glen's talk also covered drug discovery. He first used Zomig to exemplify some of the issues of drug delivery, safety, and efficacy. One is solubility. Both predicting and measuring solubility are difficult problems: Bobby illustrated this fact with literature examples and with some work of his own team using random forest. So, they adopted a reliable, reproducible method to create a "standard" dataset of solubilities. Bobby described the protocol in some detail. This work was so important that they collaborated with *JCIM* to produce a solubility challenge the results of which were published in Hopfinger, A. J. *et al.* Findings of the Challenge to Predict Aqueous Solubility. *J. Chem. Inf. Model*. **2009**, *49*, 1–5.

A second challenge is metabolism. Understanding pharmacokinetics of drugs is very important. Metabolism can alter activity (for example, from antagonist to agonist), deactivate drugs, convert pro-drugs into active forms, produce toxic compounds, and create environmental toxins. Bobby developed a method (MetaPrint2D, available on the Web) using circular fingerprints to predict the sites and products of metabolism.

Bobby's third topic was new targets. One that interests him is apelin, a GPCR which is a difficult target and has interesting pharmacological effects. The group replaced each of the amino acids by alanine and looked at the changes in the biological activity. They also constructed cyclic peptides and used NMR to study the shape of the peptides. Analysis was done with replica exchange molecular dynamics. A beta-turn at the RPRL motif was important for binding affinity (Macaluso, N. J. M.; Glen, R. C. Exploring the RPRL' Motif of Apelin-13 through Molecular Simulation and Biological Evaluation of Cyclic Peptide Analogues. *ChemMedChem* **2010**, *5*(8), 1247-1253). Analogues were synthesized, pharmacophores were generated, and molecular dynamics was used to study them. The group then attempted to make an

antagonist by stabilizing the antagonist conformation and they designed linkers to the allosteric binding site. A competitive antagonist is currently being evaluated in disease models.

The symposium concluded, appropriately, with a presentation by *JCIM*'s most prolific author Peter Willett (http://pubs.acs.org/page/jcisd8/anniversary/50/most-prolific.html). Peter talked about weighting and fusion methods for similarity-based virtual screening. These techniques were used to search the MDDR and WOMBAT databases. Binary fingerprints work well but it was hoped that use of fragment frequency information might produce even better results. It is assumed that if two molecules have multiple occurrences of a fragment in common they are more similar than if they have just a single occurrence in common, and if two molecules share a very rare fragment, they are more similar than if they share a very common fragment. Experiments show that the former assumption is correct but that there is much less evidence for the latter (Arif, S. M.; Holliday, J. D.; Willett, P. Analysis and use of fragment occurrence data in similarity-based virtual screening. *J. Comput.-Aided Mol. Des.* **2009**, **23**, 655-668; Arif, S. M.; Holliday, J. D.; Willett, P. Inverse frequency weighting of fragments for similarity-based virtual screening. *J. Chem. Inf. Model.* **2010**, *50*, 1340–1349).

Experiments in text retrieval show that documents retrieved by multiple search engines are more likely to be relevant to a query than if they are retrieved by a single search engine. To see if this effect also applies in cheminformatics, researchers at Sheffield have carried out extensive virtual screening experiments to investigate whether structures retrieved by multiple virtual screening methods are more likely to be active than if they were retrieved by a single method. Sets of 25 searches for a reference structure were carried out using five different similarity coefficients and five different fingerprints. As the number of searches increases from 1 to 25, there is a rapid decrease in the numbers of molecules retrieved in all of the searches, and a rapid increase in the percentage of those retrieved molecules that are active. This provides an empirical rationale for the use of data fusion, where multiple rankings of a database are combined to give a single, fused ranking. The Sheffield team has experimented with a whole range of different combination rules, some used previously and some novel. Their results show conclusively that one of the new rules, called CombRKP, is by far the most effective in virtual screening, this arising from the rule approximating molecular probabilities of activity (Chen *et al*. Combination rules for group fusion in similarity-based virtual screening. *Molecular Informatics* **2010**, *29*, 533-541).

The symposium presented an interesting mixture of history, philosophy, strategy and up-to-date research. Symposia in honor of people or journals can tend to lean towards nostalgia and self-congratulation, so it was a pleasure on this occasion to hear some recent results as well as the historical perspectives. The number of citations to the journal itself was impressive, even allowing for the fact that the speakers would be biased. *JCIM* is the foremost journal in cheminformatics (Willett. P. A bibliometric analysis of the literature of chemoinformatics. *Aslib Proceedings*, **2008**, *60*(1), 4-17) and I hope that it will continue in that role for the next 50 years.