Full Paper
www.molinf.com
molecular informatics
DOI: 10.1002/minf.201400166

Special Issue United Kingdom

Cheminformatics Research at the Unilever Centre for Molecular Science Informatics Cambridge

Julian E. Fuchs,[a] Andreas Bender,[a] and Robert C. Glen*[a]

Abstract: The Centre for Molecular Informatics, formerly Unilever Centre for Molecular Science Informatics (UCMSI), at the University of Cambridge is a world-leading driving force in the field of cheminformatics. Since its opening in 2000 more than 300 scientific articles have fundamentally changed the field of molecular informatics. The Centre has been a key player in promoting open chemical data and semantic access. Though mainly focussing on basic research, close collaborations with industrial partners ensured real world feedback and access to high quality molecular data. A variety of tools and standard protocols have been developed and are ubiquitous in the daily practice of cheminformatics. Here, we present a retrospective of cheminformatics research performed at the UCMSI, thereby highlighting historical and recent trends in the field as well as indicating future directions.

Keywords: Centre for Molecular Science Informatics · Retrospective analysis · History · Trends · Directions

1 Introduction

In December 2000 the Unilever Centre for Molecular Science Informatics (UCMSI) was opened at the Department of Chemistry of University of Cambridge. Based on an investment by the industrial partner Unilever, a new world-leading research group in the emerging field of molecular informatics was established. The investment included a new building, an established chair in Molecular Science Informatics and three lectureships as well as set up costs (equipment, networking, and software). The Unilever research grants were renewed in year five and year ten. In addition, over the period of the UCMSI's existence, significant additional grants from a variety of industrial, charitable and research council sources were obtained to support the objectives of the UCMSI.

The research centre is located in central Cambridge, thus profiting from a stimulating and exciting research environment. Daily interactions with several other local institutes at the University of Cambridge, the EMBL European Bioinformatics Institute (EBI), and the Cambridge Crystallographic Data Centre (CCDC) create a world class research cluster. Furthermore, several major industrial partners are located in close proximity in Cambridge's science parks and on corporate research sites.

Collaborations with more than twenty industrial partners and especially Unilever allowed access to high quality data sets and formed the basis for state-of-the-art computational modelling.[1] Further industrial partners have included Boehringer Ingelheim, AstraZeneca, BASF, Pfizer, Johnson&Johnson, GSK, Aboca and Eli Lilly, to name but a few. Additional third party funding was attracted at the national level from the UK Engineering and Physical Sciences Research Council, The Medical Research Council, The Wellcome Trust, and the Biotechnology and Biological Sciences Research Council, as well as The National Institutes of Health in the USA. On the European level, major grants from the European Chemical Industry Council (CEFIC) and the European Research Council (ERC), and the Framework-7 program were successfully obtained, funding a variety of international informatics research projects.

The research at the UCMSI covers broad areas of cheminformatics, which coupled with experiments and collaborations with industrial partners in bringing products to the market, ensures real life feedback in several inter-disciplinary research efforts. A main goal of the currently 40 scientists in the UCMSI is the integration of chemistry, biology and materials science through the development and application of molecular informatics. Robert Glen has directed the UCMSI since its opening. He heads an interdisciplinary research group using a broad set of computational methodologies to tackle basic scientific questions in the general area of molecular biosystems. Four additional research groups are focusing on relevant areas on cheminformatics research at the UCMSI. The group of Jonathan Goodman

[a] J. E. Fuchs, A. Bender, R. C. Glen
Centre for Molecular Informatics, Department of Chemistry, University of Cambridge
Lensfield Road, Cambridge CB2 1EW, UK
phone/fax: +44 (0)1223 336472/+44 (0)1223 763076
*e-mail: rcg28@cam.ac.uk

focuses on synthesis, computation and informatics and applies cheminformatics to tackle questions in chemical reactivity and catalysis. Peter Murray-Rust's group focuses on semantic web technologies and develops methodologies and software for intelligent storage and retrieval of chemical data. Andreas Bender's group drives the integration of new large scale data sources (gene expression, biological networks, phylogenetics) in the field of cheminformatics. Recently, Lucy Colwell has established a new group at the Centre, her research aims to identify structural features within large datasets using advanced statistical and data analytics methodologies. Two former group leaders have recently departed from the UCMSI to set up significant research groups. Peter J. Bond recently moved on to a principal investigator position at the Bioinformatics (BII) Institute A*STAR in Singapore and continues his research efforts on multiscale modelling and large scale simulations. Former group leader John Mitchell has moved to the University of St. Andrews as a Reader and continues broad research in cheminformatics from quantum chemistry to molecular simulation technologies. Former group leaders include David Lary, Guy Grant, Dmitry Nerukh, Maxim Federov and Hamse Mussa.

Over the years of the Centre's existence 67 PhDs have been trained and 36 postdoctoral research associates performed high quality research. The UCMSI hosted 50 conferences/workshops and organized over 200 seminars. Scientists at the UCMSI have won several awards and prizes including the RSC Bader Award to Jonathan Goodman, the Hansch Award to Andreas Bender, and the Novartis Chemistry Lectureship to Robert Glen. In this article we describe the developments in cheminformatics research performed at the UCMSI aiming to identify challenges and opportunities for the field in the future.

2 Methods

2.1 Data Mining in Web of Science

We retrieved all entries related to the UCMSI in all Web of Science databases.[2] Therefore, we used the Web of Science web interface and extracted entries with at least one authors address containing all words "Cambridge", "University", "Molecular", and "Unilever". Furthermore, we discarded entries in the database earlier than the year 2001, to remove three false positive hits of earlier years. Unfortunately, some entries (e.g.[3,4]) are discarded as false negatives at this stage, as different abbreviations in author affiliations have been used. Citing articles and total citations were extracted for all articles and analyzed in terms of self- and non-self citations according to the tools provided in the Web of Science online mask. The Hirsch index (h-index)[5] was calculated to estimate a total impact of the science performed at the UCMSI.

2.2 Analysis of Publications Using Word Clouds

Furthermore, we extracted information on author names and abstracts (where available) from all publications. We identified the leading scientists and their central research topics using word cloud representations ("wordles") generated using WordItOut.[6] We used author surnames only to ensure consistency between different data sources. Full abstracts were used for the generation of topic-related word clouds. To allow for identification of trends over time we split the complete data sets for authors and abstracts into sections of years. Thereby, we analyzed author and thematic contributions for the years 2001–2004, 2005–2007, 2008–2010, and 2011–2014 respectively.

3 Results

We identified in total 325 published items of the UCMSI in the Web of Science database. After an initial growth phase, the article output has reached a stable plateau of between twenty and 40 published items per year (see Figure 1A). The observed decrease for the year 2014 arises from incomplete data for the current year. Linear extrapolation to the complete year increases the number of published articles listed in Web of Science to the average level of approximately 20. Due to lagging indexing in Web of Science linear extrapolation is expected to underestimate values for year 2014. When setting aside the incomplete year 2014, Pearson correlation between years and published items is 0.78, indicating a steadily growing output of research performed at the Centre.

In addition to the number of published items, citations were counted on articles published by the Centre. Soon after publication of the first two articles in 2001, first citations are recorded (see Figure 1B). The amount of citations continues to grow steadily over the years and breaks the barrier of 100 citations in the year 2005. Results are shown on a per-year basis, not as a cumulative count. Nevertheless, an almost linear increase of citations with Pearson correlation coefficient of 0.97 between year and citations is observed when excluding the incomplete year 2014. With 988 citations in the year 2013 only, research from the UCMSI is referenced almost three times per day. Linear extrapolation for year 2014 indicates that the barrier of 1000 citations is likely to be broken for the first time.

In total 5508 citations on the 325 articles were recorded via Web of Science. Only nine percent of these represent self-citations (506 citations), thus reflecting a broad audience and considerable impact within the cheminformatics community. 4109 unique indexed articles reference scientific reports from the UCMSI with less than five percent of self-citations (177 articles). On average, articles from the Centre are cited 16.95 times.

No single article accounts for the large average citation count. By contrast, eight articles are individually cited over

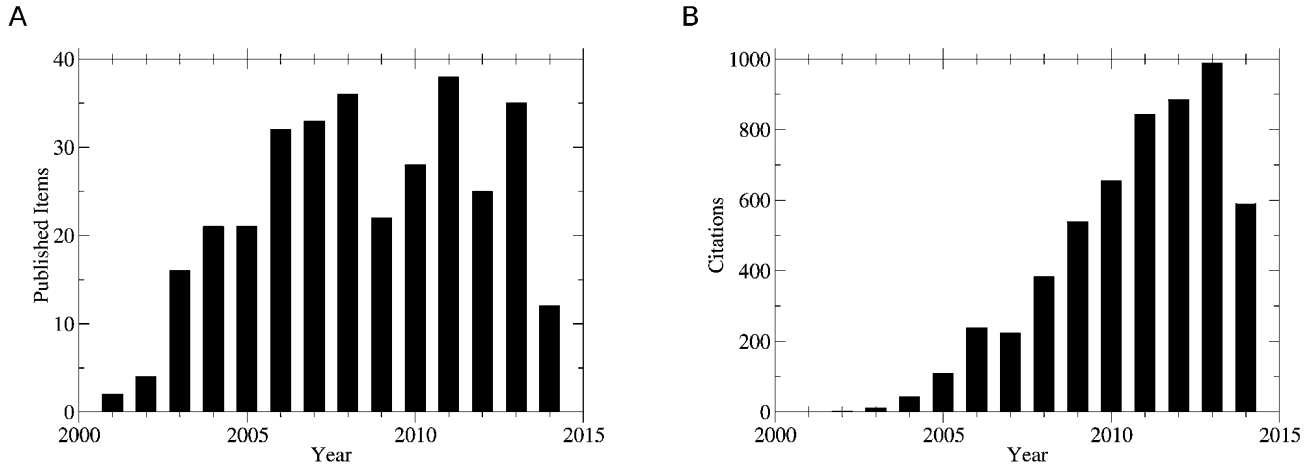

Figure 1. Scientific publications at the UCMSI: A) Published items at the UCMSI are shown over the years of its existence since 2000. An increase to a stable plateau of around two publications per month is visible. B) Citations related to items published at the UCMSI. The increasing importance of research performed in the field of cheminformatics is reflected by steadily growing citations per year.

100 times, all of them have been published in 2004 or later.[7–14] This broad impact is reflected by an h-index of 39 after 14 years of existence of the Centre. Eight articles are indicated as "highly cited articles" within Web of Science,[9,11–13,15–18] thus representing the top one percent publications in the respective field. These articles only partially overlap with articles already cited more than 100 times. This indicates that they will most likely be the next ones crossing the barrier of 100 citations.

By creating word clouds from author lists of all published items we identified key scientists at the UCMSI in Cambridge (see Figure 2A). Robert C. Glen has published 83 items over the years, closely followed by the group leaders Peter Murray-Rust, Andreas Bender, and Jonathan Goodman, each contributing between 60 and 70 articles. Splitting author contributions according to intervals of years allows the development of the UCMSI to be followed. In earlier years, lecturer John Mitchell was a highly active researcher at the UCMSI, as was Henry Rzepa in collaboration with Peter Murray-Rust (see Figure 2B). In later years the name "Andreas Bender" emerges more and more with a short break during his time at Novartis and the University of Leiden (see Figure 2C and Figure 2D). In recent years, the group of Peter Bond focusing on simulations added additional output to the UCMSI (see Figure 2E).

Wordles created on the basis of published abstracts identify key topics and methods applied at the UCMSI (see Figure 3A). "Data" is the most abundant single phrase and appears mostly in connection with the second most occurring words "molecular" and "chemical". Therefore, chemical data form the foundation of all research performed at the Centre. "Model" and "models" appear prominent in the list, giving hints how chemical data is utilized in the generation of computational models for a variety of mostly chemical and biological properties.

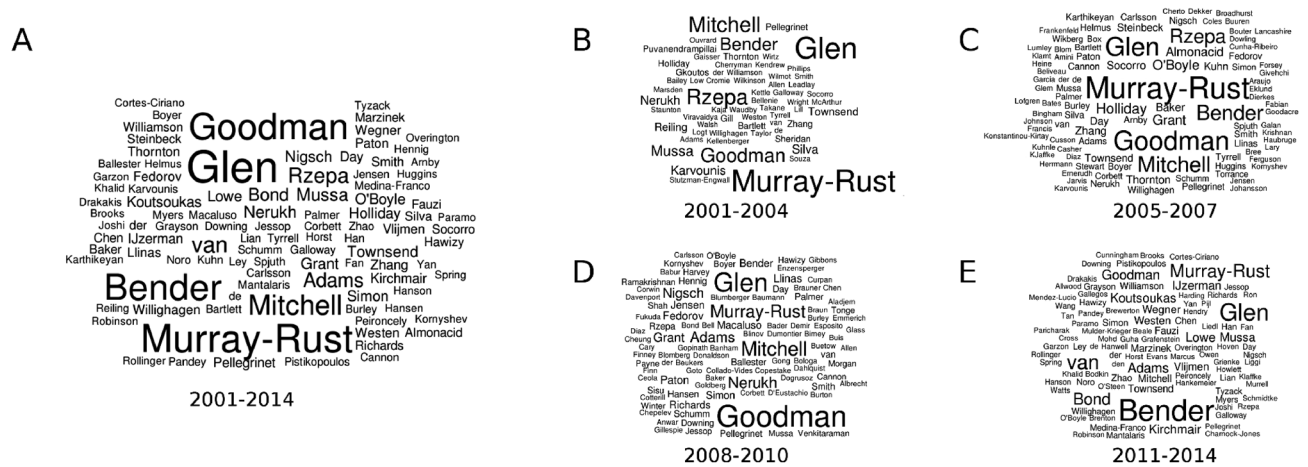

Figure 2. Word clouds of co-authors listed on publications of the UCMSI: A) Word cloud over all years of existence. B–E) Author occurrences are split according to years 2004–2004, 2005–2007, 2008–2010, and 2011–2014 respectively.

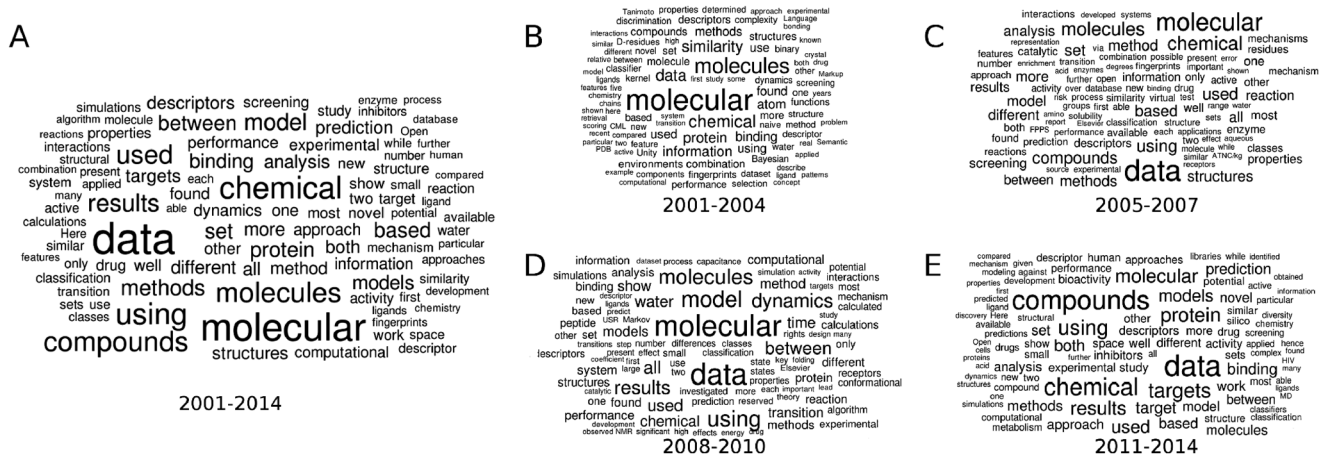

Figure 3. Wordles generated from indexed abstracts of publications from the UCMSI: A) Word cloud over all years of existence of the Centre. B-E) Word occurrences are analysed in terms of chronological development over the years 2001–2004, 2005–2007, 2008–2010, and 20011–2014.

Splitting abstracts according to publication years does not reveal any consistency in the research efforts on-going at the UCMSI (see Figures 3B–3E). Contributions of "molecular" and "data" are constantly high. Some words linked to the observables predicted in established models appear transiently, e.g. "target" for recently established target prediction tools. Nevertheless, the low frequency of individual modelled molecular properties reflects the broadness of topics covered. Furthermore, structure-based modelling approaches[19] appear more prominent in recent years as indicated by the increasing occurrence of the keywords "protein", "water", and "dynamics".

4 Discussion

Analysis of citations has allowed the identification of key topics and methods applied at the UCMSI. Science is often centred around method development which is reflected by several publications in leading journals in the field of cheminformatics and modelling. Additionally, application of the newly developed methods ensures real life feedback and enabled publications in fields ranging from drug discovery to synthesis, materials science to electrochemistry. In the following paragraphs we will select particular fields of research where major advances have been made at the Centre.

4.1 Linking Cheminformatics and Biology

Starting from innovative ways to encode and compare chemical environments in molecules,[7,8,10] several steps have been taken to link chemical with biological properties of molecules[20] based on statistical modelling techniques.[21,22] Thereby, the fields of cheminformatics and bioinformatics increasingly overlap and even fuse, leading to approaches like proteochemometric modelling (PCM).[23,24] With the establishment of target prediction tools based on statistical modelling,[25,26] novel approaches to cluster molecules using predicted biological effect can be implemented.[27] Furthermore, these novel methodologies proved to be helpful in multi target drug design.[28] Recent directions in the area comprise the inclusion of further biological data sources, e.g. from biological networks and phylogenetics[29,30] as well as gene expression data.[31] Several successful applications of target prediction algorithms[32–34] underline the increasing accuracy of such data-driven approaches and point towards a bright future given the increase in available data sources.[35,36] In a recent success story cheminformatic tools were applied to identify potential targets of a series of synthetic biscoumarins showing anti-cancer activity in vitro and in vivo.[37] After in depth computational characterization of predicted protein-ligand interactions (see Figure 4A), the in silico predicted protein target tumour necrosis factor α was verified by experimental techniques. An emerging future direction in this field is the prediction of biological effects of compound combinations that is financed via an ERC Starting Grant to Andreas Bender. Further stimuli in this area expected from statistical analyses of genomic sequence data.[38,39]

4.2 Data Semantics and Accessibility

Growth of the available ("Big") data brings new challenges to the field of cheminformatics. Millions of chemical substances have been characterized, many hundreds of thousands of three-dimensional structures have been deposited in databases, and associated biological data is spread over millions of publications and patents.[40] Therefore, the development of standards for chemical identification, indexing and storage is crucial for future successful data retrieval. Usage of the IUPAC standard International Chemical Identifiers (InChIs) allows the encoding of chemical information using a unique text descriptor, e.g. for database

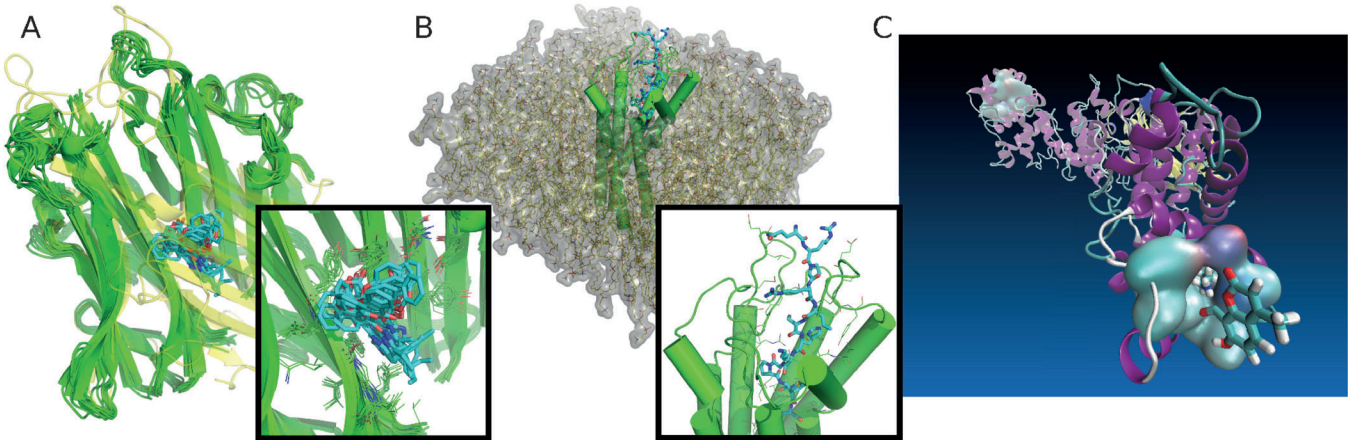

Figure 4. Examples of cheminformatics research currently performed at the UCMSI: A) Ensemble of protein-ligand poses extracted from a molecular dynamics simulation of tumor necrosis factor α (TNFα, green cartoon) in complex with a biscoumarin (cyan sticks). The native trimeric structure of TNFα (semi-transparent yellow cartoon, PDB: 1TNF) is disrupted by ligand binding. A zoom on the binding site highlights the hydrophobic environment of the ligand, allowing for several binding modes. B) Model of the apelin receptor (green cartoon) in complex with the apelin-13 peptide (cyan sticks). Starting coordinates of a molecular dynamics simulation box are depicted including a model lipid bi-layer (yellow lines and semi-transparent surface) along with explicit water molecules and counter ions (both not shown). The predicted binding mode of the apelin-13 peptide and surrounding residues are highlighted in the included zoom on the binding cavity. C) Illustration of the simulation approach followed to investigate selectivity of Schiff base forming covalent IRE-1 inhibitors. IRE-1 is shown as cartoon with the reactants Lys-907 and inhibitor 4μ8C highlighted as sticks.

searches.[41] Based on InChIs several tools have been developed, e.g. to convert structures to chemical names in a fully automated way[42] or chemistry-aware text mining.[43] Recently, InChIs have been extended to RinChIs to depict chemical reactions unambiguously.[44] Chemical Markup Language (CML) introduces and specifies particular data fields to efficiently store and retrieve chemical data.[45] Based on CML the World-Wide Molecular Matrix (WWMM) has been introduced to collect and connect chemical information of various sources.[46] Chem4Word has been created in a collaboration between Microsoft Research and the UCMSI to facilitate the handling of chemical information within text processing software. To date, the plug-in has been downloaded more than 400 000 times. Over the last decade, the UCMSI has been a key player in setting standards for open data and data quality standards in chemistry.[47–49] In recent years, technologies for data semantics and natural language processing have been employed attempting to directly extract the scientific context of published chemical data.[50]

4.3 Modelling of Physicochemical Properties and ADMET Parameters

The increase in data sources over recent years allowed the establishment of more accurate computational models, even for complex biological phenomena, e.g. absorption, distribution, metabolism, excretion and toxicity (ADMET) of xenobiotics.[51] Dozens of computational methods for prediction of metabolic reactions on different levels of complexity have been published in the literature.[16] Thorough classification of annotated biotransformations[52] facilitated the development of novel predictive models for general metabolic reactivity (Metaprint2D),[53,54] P-glycoprotein transport,[55] solubility,[56] and cytochrome P450-catalyzed metabolic reactions.[57–59] Thereby, innovative computational methods making use of recent advances in graphics processing unit (GPU) based computing have been employed. Using these approaches, new levels of throughput and thus modelling accuracy are in range. Furthermore, the UCMSI ran a solubility competition (with over 100 entries) for the cheminformatics community and provided high quality experimental data to facilitate further method developments in the area.[60–62]

4.4 Bioactive Compound Discovery

The UCMSI has been a driving force of computer-aided drug design over the past decades. In addition to support of external drug design efforts, local lead discovery projects have been guided by computational technologies. Novel ligands for the G-protein coupled receptor (GPCR) apelin have been successfully identified and biologically characterized.[63] Based on modelled structures of the apelin receptor (see Figure 4B), optimization of the peptide-derived compounds is on-going and shows enormous potential for further development and recently the first human study of apelin biased agonists has been completed in Addenbrooke's hospital in Cambridge. (Design, characterization and first-in-human study of the vascular actions of a novel 'biased' apelin receptor agonist. Anthony Davenport, Aimee Brame, Janet Maguire, Peiran Yang, Alex Dyson, Rubben Torella, Joseph Cheriyan, Mervyn Singer, Robert Glen, Ian Wilkinson. Hypertension, in press, 2015). Furthermore, small

molecule antagonists of the 5-HT1B GPCR have been identified and optimized for development as a potential treatment for Pulmonary Hypertension.[64] Several newly designed and synthesized compounds are currently in clinical studies in Addenbrooke's Hosptial, Cambridge. In another compound discovery effort large scale simulation approaches (see Figure 4C) have been used to identify covalent binders of the endonuclease domain of IRE-1.[65,66] In a follow-up study selectivity of the Schiff base forming compounds has been investigated by advanced computational techniques.[67] Research at the UCMSI also lead to joint patents with Unilever on the CB1/2 receptors and NCKX ion channels showing benefits in skin for use in home and personal care products. To assist drug discovery several analyses of molecular diversity in the context of diversity-oriented chemical synthesis have been performed in collaboration with experimental groups.[68–70]

In addition to these four areas the UCMSI has been particularly active in development of innovative simulation and analysis methodologies[71,72] and cheminformatic support for organic synthesis.[73,74] Industrial partners emphasize the productivity of collaborative research efforts with the Centre for Molecular Informatics. Ola Engkvist, team leader of Computational Chemistry at AstraZeneca and involved in several joint research projects, highlights: "The collaboration with the UCMSI provides AstraZeneca with the opportunity to work closely with one of the world leading groups in cheminformatics. The combination of the UCMSI's outstanding scientific knowledge with AstraZeneca's industrial experience provides a platform for state-of-the-art research in cheminformatics. The proximity to one of the AstraZeneca science hubs is an additional plus that facilitates smooth collaborations." Jim Crilly, senior vice-president of the Strategic Science Group at Unilever, adds: "Unilever is very proud to have instigated in partnership with the University of Cambridge the Centre for Molecular Informatics under the Leadership of Professor Robert Glen and pleased that it has developed into a global centre of excellence in a hugely important field of research. Collaboration with the centre has brought a new way of working into our own research endeavours and accelerated our discovery process."

Following a clear mission statement published in 2002[75] the UCMSI is developing tools and standards in molecular informatics. With the increase in computer power and data accessibility, molecular modelling allows chemical and biological effects of increasing complexity and size to be captured. Available data sources range from chemistry and related bioactivity (PubChem,[76] ChEMBL,[35] DrugBank[77]) via protein sequence (UniProt[78]) and structure (Protein Data Bank,[79] Pfam[80]), to cellular pathways (KEGG[81]) and responses (LINCS[82]). A collection of online molecular biology databases has recently been published with the database issue of Nucleic Acids Research.[83]

With the advent of microsecond simulations of biological macromolecules[84] and sophisticated enhanced sampling methods,[85] dynamic processes underlying macromolecular recognition processes can be modelled with a new level of accuracy.[86] Protein-ligand binding processes may be studied at atomistic resolution, reports include full sampling of binding and unbinding of both fragments[87] and small molecules.[88] With an increasingly accurate description of protein dynamics, its role in biomolecular recognition processes can be studied, thereby allowing pharmaceutically relevant properties like binding specificity[89] or binding kinetics[90] including allosteric mechanisms[91] to be probed. On the other hand, integration of innovative data sources from emerging "omics" fields such as proteomics,[92] metabonomics/metabolomics,[93] or lipidomics[94] will allow the capture of novel biological properties which have limited or no direct chemical leads to their mechanism and function.

5 Conclusions

The data presented underline the central role in the cheminformatics world that is occupied by the UCMSI. The research institute sets world-wide standards and publishes technologies and software that is widely used and cited. The broad range of topics covered at the UCMSI ensures its broad impact ranging from basic cheminformatics, data mining and machine learning techniques to complex simulations, to chemical reactivity and synthesis. The connection of all these areas in a single research centre offers unique opportunities for scientists involved, industrial partners, as well as the whole cheminformatics community.

Acknowledgements

J. E. F thanks the *Medical Research Council* for funding (Grant Number MR/K020919/1). Furthermore, the UCMSI acknowledges all funding sources for continuous support over the past 15 years. We thank *Ola Engkvist*, *AstraZeneca*, and *Jim Crilly*, *Unilever*, for their encouraging feedback on joint research efforts and current and former members of the UCMSI for their valuable input on the manuscript.

References

[1] F. Nigsch, N. J. Macaluso, J. B. Mitchell, D. Zmuidinavicius, *Expert Opin. Drug Metab. Toxicol.* **2009**, *5*, 1–14.
[2] www.isiknowledge.com, database access 20.08.2014.
[3] S. E. Adams, J. M. Goodman, R. J. Kidd, A. D. McNaught, P. Murray-Rust, F. R. Norton, J. A. Townsend, C. A. Waudby, *Org. Biomol. Chem.* **2004**, *2*, 3067–3070.
[4] J. A. Townsend, S. E. Adams, C. A. Waudby, V. K. de Souza, J. M. Goodman, P. Murray-Rust, *Org. Biomol. Chem.* **2004**, *2*, 3294–3300.
[5] J. E. Hirsch, *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 16569–16572.
[6] www.worditout.com, accessed 20.08.2014.
[7] A. Bender, R. C. Glen, *Org. Biomol. Chem.* **2004**, *2*, 3204–3218.

[8] A. Bender, H. Y. Mussa, R. C. Glen, S. Reiling, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1708 – 1718.
[9] N. Adams, U. S. Schubert, *Adv. Drug Deliv. Rev.* **2007**, *59*, 1504 – 1520.
[10] A. Bender, H. Y. Mussa, R. C. Glen, S. Reiling, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 170 – 178.
[11] M. V. Fedorov, A. A. Kornishev, *J. Phys. Chem. B* **2008**, *112*, 11868 – 11872.
[12] M. V. Fedorov, A. A. Kornishev, *Electrochim. Acta* **2008**, *53*, 6835 – 6840.
[13] E. Demir, M. P. Cary, S. Paley, K. Fukuda, C. Lemer, I. Vastrik, G. Wu, P. D'Eustachio, C. Schaefer, J. Luciano, F. Schacherer, I. Martinez-Flores, Z. Hu, V. Jimenez-Jacinto, G. Joshi-Tope, K. Kandasamy, A. C. Lopez-Fuentes, H. Mi, E. Pichler, I. Rodchenkov, A. Splendiani, S. Tkachev, J. Zucker, G. Gopinath, H. Rajasimha, R. Ramakrishnan, I. Shah, M. Syed, N. Anwar, O. Babur, M. Blinov, E. Brauner, D. Corwin, S. Donaldson, F. Gibbons, R. Goldberg, P. Hornbeck, A. Luna, P. Murray-Rust, E. Neumann, O. Ruebenacker, M. Samwald, M. van Iersel, S. Wimalaratne, K. Allen, B. Braun, M. Whirl-Carrillo, K. H. Cheung, K. Dahlquist, A. Finney, M. Gillespie, E. Glass, L. Gong, R. Haw, M. Honig, O. Hubaut, D. Kane, S. Krupa, M. Kutmon, J. Leonard, D. Marks, D. Merberg, V. Petri, A. Pico, D. Ravenscroft, L. Ren, N. Shah, M. Sunshine, R. Tang, R. Whaley, S. Letovksy, K. H. Buetow, A. Rzhetsky, V. Schachter, B. S. Sobral, U. Dogrusoz, S. McWeeney, M. Aladjem, E. Birney, J. Collado-Vides, S. Goto, M. Hucka, N. Le Novère, N. Maltsev, A. Pandey, P. Thomas, E. Wingender, P. D. Karp, C. Sander, G. D. Bader, *Nat. Biotechnol.* **2010**, *28*, 935 – 942.
[14] L. Simon, J. M. Goodman, *J. Am. Chem. Soc.* **2008**, *130*, 8741 – 8747.
[15] T. Scior, A. Bender, G. Tresadern, J. L. Medina-Franco, K. Martinez-Mayorga, T. Langer, K. Cuanalo-Contreras, D. K. Agrafiotis, *J. Chem. Inf. Model.* **2012**, *52*, 867 – 881.
[16] J. Kirchmair, M. J. Williamson, J. D. Tyzack, L. Tan, P. J. Bond, A. Bender, R. C. Glen, *J. Chem. Inf. Model.* **2012**, *52*, 617 – 648.
[17] P. J. Ballester, J. B. O. Mitchell, *Bioinformatics* **2010**, *26*, 1169 – 1175.
[18] U. Grienke, M. Schmidtke, S. von Grafenstein, J. Kirchmair, K. R. Liedl, J. M. Rollinger, *Nat. Prod. Rep.* **2012**, *29*, 11 – 36.
[19] R. C. Glen, S. C. Allen, *Curr. Med. Chem.* **2003**, *10*, 763 – 767.
[20] F. M. Fauzi, A. Koutsoukas, R. Lowe, K. Joshi, T. P. Fan, R. C. Glen, A. Bender, *J. Chem. Inf. Model.* **2013**, *53*, 661 – 673.
[21] H. Y. Mussa, L. Hawizy, F. Nigsch, R. C. Glen, *J. Chem. Inf. Model.* **2011**, *51*, 4 – 14.
[22] H. Y. Mussa, J. B. Mitchell, R. C. Glen, *J. Cheminf.* **2013**, *5*, 37.
[23] G. J. P. van Westen, R. F. Swier, J. K. Wegner, A. P. IJzerman, H. W. T. van Vlijmen, A. Bender, *J. Cheminf.* **2013**, *5*, 41.
[24] G. J. P. van Westen, R. F. Swier, I. Cortes-Ciriano, J. K. Wegner, J. P. Overington, A. P. IJzerman, H. W. T. van Vlijmen, A. Bender, *J. Cheminf.* **2013**, *5*, 42.
[25] A. Koutsoukas, R. Lowe, Y. Kalantar-Motamedi, H. Y. Mussa, W. Klaffke, J. B. O. Mitchell, R. C. Glen, A. Bender, *J. Chem. Inf. Model.* **2013**, *53*, 1957 – 1966.
[26] J. C. Costello, L. M. Heiser, E. Georgii, M. Gönen, M. P. Menden, N. J. Wang, M. Bansal, M. Ammad-ud-din, P. Hintsanen, S. A. Khan, J. P. Mpindi, O. Kallioniemi, A. Honkela, T. Aittokallio, K. Wennerberg, NCI DREAM Community, J. J. Collins, D. Gallahan, D. Singer, J. Saez-Rodriguez, S. Kaski, J.W. Gray, G. Stolovitzky, *Nat. Biotechnol.* **2014**, *32(12)*, 1202–1212
[27] H. P. Nguyen, A. Koutsoukas, M. F. Fauzi, G. Drakakis, M. Maciejewski, R. C. Glen, A. Bender, *Chem. Biol. Drug Des.* **2013**, *82*, 252 – 266.
[28] A. Koutsoukas, B. Simms, J. Kirchmair, P. J. Bond, A. V. Whitmore, S. Simmer, M. P. Young, J. L. Jenkins, M. Glick, R. C. Glen, A. Bender, *J. Proteomics* **2011**, *74*, 2554 – 2574.
[29] S. Liggi, A. Koutsoukas, Y. K. Motamedi, R. C. Glen, A. Bender, *J. Cheminf.* **2013**, 5, 15.
[30] S. Paricharak, T. Klenka, M. Augustin, U. A. Patel, A. Bender, *J. Cheminf.* **2013**, *5*, 49.
[31] A. C. Ravindranath, N. Perualila-Tan, A. Kasim, G. Drakakis, S. Liggi, S. C. Brewerton, D. Mason, M. J. Bodkin, D. A. Evans, A. Bhagwat, W. Talloen, H. W. Goehlmann, Qstar Consortium, Z. Shkedy, A. Bender, *Mol. Biosyst.* 2015, 11, 86 – 96.
[32] G. Drakakis, A. E. Hendry, K. Hanson, S C. Brewerton, M. J. Bodkin, D. A. Evans, G. N. Wheeler, A. Bender, *Med. Chem. Commun.* **2014**, *5*, 386 – 396.
[33] H. K. Keerthy, C. D. Mohan, K. S. Siveen, J. E. Fuchs, S. Rangappa, M. S. Sundaram, F. Li, K. S. Girish, G. Sethi, B. Basappa, A. Bender, K. S. Rangappa, *J. Biol. Chem.* **2014**, *289*, 31879 – 31890.
[34] C. D. Mohan, H. Bharathkumar, K. C. Bulusu, V. Pandey, S. Rangappa, J. E. Fuchs, M. K. Shanmugam, X. Dai, F. Li, A. Deivasigamani, K. M. Hui, A. P. Kumar, P. E. Lobie, A. Bender, B. Basappa, G. Sethi, K. S. Rangappa, *J. Biol. Chem.* **2014**, *289*, 34296 – 34307.
[35] A. P. Bento, A. Gaulton, A. Hersey, L. J. Bellis, J. Chambers, M. Davies, F. A. Kruger, Y. Light, L. Mak, S. McGlinchey, M. Nowotka, G. Papadatos, R. Santos, J. P. Overington, *Nucleic Acids Res.* **2014**, *42*, D1083 – D1090.
[36] A. G. Dossetter, G. Ecker, H. Laverty, J. Overington, *Future Med. Chem.* **2014**, *6*, 857 – 864.
[37] H. Bharathkumar, S. Paricharak, K. R. Dinesh, K. S. Siveen, J. E. Fuchs, S. Rangappa, C. D. Mohan, N. Mohandas, A. P. Kumar, G. Sethi, A. Bender, B. Basappa, K. S. Rangappa, *RSC Adv.* **2014**, *4*, 45143 – 45146.
[38] T. A. Hopf, L. J. Colwell, R. Sheridan, B. Rost, C. Sander, D. S. Marks, *Cell* **2012**, *149*, 1607 – 1621.
[39] R. S. Dwyer, D. P. Ricci, L. J. Colwell, T. J. Silhavy, N. S. Wilgreen, *Genetics* **2013**, *195*, 443 – 455.
[40] R. C. Glen, *J. Comput. Aided Mol. Des.* **2012**, *26*, 47 – 49.
[41] P. Murray-Rust, H. S. Rzepa, J. J. Stewart, Y. Zhang, *J. Mol. Model.* **2005**, *11*, 532 – 541.
[42] D. M. Lowe, P. T. Corbett, P. Murray-Rust, R. C. Glen, *J. Chem. Inf. Model.* **2011**, *51*, 739 – 753.
[43] D. M. Jessop, S. E. Adams, E. L. Willighagen, L. Hawizy, P. Murray-Rust, *J. Cheminf.* **2011**, *3*, 41.
[44] G. Grethe, J. M. Goodman, C. H. Allen, *J. Cheminform.* **2013**, *5*, 45.
[45] P. Murray-Rust, J. A. Townsend, S. E. Adams, W. Phandungsukanan, J. Thomas, *J. Cheminf.* **2011**, *3*, 43.
[46] P. Murray-Rust, S. E. Adams, J. Downing, J. A. Townsend, Y. Zhang, *J. Cheminf.* **2011**, *3*, 42.
[47] P. Murray-Rust, *Nature* **2008**, *451*, 648 – 651.
[48] N. M. O'Boyle, R. Guha, E. L. Willighagen, S. E. Adams, J. Alvarsson, J. C. Bradley, I. V. Filippov, R. M. Hanson, M. D. Hanwell, G. R. Hutchison, C. A. James, N. Jeliazkova, A. S. Lang, K. M. Langner, D. C. Lonie, D. M. Lowe, J. Pansanel, D. Pavlov, O. Spjuth, C. Steinbeck, A. L. Tenderholt, K. J. Theisen, P. Murray-Rust, *J. Cheminf.* **2011**, *3*, 37.
[49] S. Orchard, B. Al-Lazikani, S. Bryant, D. Clark, E. Calder, I. Dix, O. Engkvist, M. Forster, A. Gaulton, M. Gilson, R. Glen, M. Grigorov, K. Hammond-Kosack, L. Harland, A. Hopkins, C. Larminie, N. Lynch, R. K. Mann, P. Murray-Rust, E. Lo Piparo, C. Southan, C. Steinbeck, D. Wishart, H. Hermjakob, J. Overington, J. Thornton, *Nat. Rev. Drug Discov.* **2011**, *10*, 661 – 669.

[50] L. Hawizy, D. M. Jessop, N. Adams, P. Murray-Rust, *J. Cheminf.* **2011**, *3*, 17.
[51] H. van de Waterbeemd, E. Gifford, *Nat. Rev. Drug Discov.* **2003**, *2*, 192 – 204.
[52] J. Kirchmair, A. Howlett, J. E. Peironcely, D. S. Murrell, M. J. Williamson, S. E. Adams, T. Hankemeier, L. van Buren, G. Duchateau, W. Klaffke, R. C. Glen, *J. Chem. Inf. Model.* **2013**, *53*, 354 – 367.
[53] S. Boyer, C. Hasselgren Arnby, L. Carlsson, J. Smith, V. Stein, R. C. Glen, *J. Chem. Inf. Model.* **2007**, *47*, 583 – 590.
[54] L. Carlsson, O. Spjuth, S. Adams, R. C. Glen, S. Boyer, *BMC Bioinformatics* **2010**, *11*, 362.
[55] Z. Wang, Y. Chen, H. Liang, A. Bender, R. C. Glen, A. Yan, *J. Chem. Inf. Model.* **2011**, *51*, 1447 – 1456.
[56] J. Comer, S. Judge, D. Matthews, L. Towes, B. Falcone, J. Goodman, J. Dearden, *ADMET & DMPK* **2014**, *2*, 18 – 32.
[57] J. Kirchmair, M. J. Williamson, A. M. Afzal, J. D. Tyzack, A. P. Choy, A. Howlett, P. Rydberg, R. C. Glen, *J. Chem. Inf. Model.* **2013**, *53*, 2896 – 2907.
[58] J. D. Tyzack, M. J. Williamson, R. Torella, R. C. Glen, *J. Chem. Inf. Model.* **2013**, *53*, 1294 – 1305.
[59] J. D. Tyzack, H. Y. Mussa, M. J. Williamson, J. Kirchmair, R. C. Glen, *J. Cheminf.* **2014**, *6*, 29.
[60] A. Llinas, R. C. Glen, J. M. Goodman, *J. Chem. Inf. Model.* **2008**, *48*, 1289 – 1303.
[61] A. J. Hopfinger, E. X. Esposito, A. Llinas, R. C. Glen, J. M. Goodman, *J. Chem. Inf. Model.* **2009**, *49*, 1 – 5.
[62] D. S. Palmer, N. M. O'Boyle, R. C. Glen, J. B. Mitchell, *J. Chem. Inf. Model.* **2007**, *47*, 150 – 158.
[63] N. J. Macaluso, S. L. Pitkin, J. J. Maguire, A. P. Davenport, R. C. Glen, *ChemMedChem* **2011**, *6*, 1017 – 1023.
[64] G. P. Moloney, A. Garavelas, G. R. Martin, M. Maxwell, R. C. Glen, *Eur. J. Med. Chem.* **2004**, *39*, 305 – 321.
[65] B. C. S. Cross, P. J. Bond, P. G. Sadowski, B. K. Jha, J. Zak, J. M. Goodman, R. H. Silverman, T. A. Neubert, I. R. Baxendale, D. Ron, H. P. Harding, *Proc. Natl. Acad. Sci.* **2012**, *109*, 5559 – 5560.
[66] B. C. S. Cross, P. J. Bond, P. G. Sadowski, B. K. Jha, J. Zak, J. M. Goodman, R. H. Silverman, T. A. Neubert, I. R. Baxendale, D. Ron, H. P. Harding, *Proc. Natl. Acad. Sci.* **2012**, *109*, E869 – E878.
[67] S. M. Tomasio, H. P. Harding, D. Ron, B C. S. Cross, P. J. Bond, *Mol. BioSyst.* **2013**, *9*, 2408 – 2416.
[68] S. Fergus, A. Bender, D. R. Spring, *Curr. Opin. Chem. Biol.* **2005**, *9*, 304 – 309.
[69] A. Isidro-Llobet, T. Murillo, P. Bello, A. Cillibrizzi, J. T. Hodgkinson, W. R. Galloway, A. Bender, M. Welch, D. R. Spring, *Proc. Natl. Acad. Sci.* **2011**, *108*, 6793 – 6798.
[70] A. Koutsoukas, S. Paricharak, W. R. Galloway, D. R. Spring, A. P. IJzerman, R. C. Glen, D. Marcus, A. Bender, *J. Chem. Inf. Model.* **2014**, *54*, 230 – 242.
[71] S. Khalid, P. J. Bond, *Methods Mol. Biol.* **2012**, *924*, 635 – 657.
[72] T. Paramo, A. East, D. Garzon, M. B. Ulmschneider, P. J. Bond, *J. Chem. Theory Comput.* **2014**, *10*, 2151 – 2164.
[73] R. H. Currie, J. M. Goodman, *Angew. Chem. Int. Ed. Engl.* **2012**, *51*, 4695 – 4697.
[74] J. M. Goodman, I. M. Socorro, *J. Comput. Aided Mol. Des.* **2007**, *21*, 351 – 357.
[75] R. C. Glen, *Chem. Commun.* **2002**, *23*, 2745 – 2747.
[76] Y. Wang, T. Suzek, J. Zhang, J. Wang, S. He, T. Cheng, B. A. Shoemaker, A. Gindulyte, S. H. Bryant, *Nucleic Acids Res.* **2014**, *42*, D1075 – D1082.
[77] V. Law, C. Knox, Y. Djoumbou, T. Jewison, A. C. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson, V. Neveu, A. Tang, G. Gabriel, C. Ly, S. Adamjee, Z. T. Dame, B. Han, Y. Zhou, D. S. Wishart, *Nucleic Acids Res.* **2014**, *42*, D478 – D484.
[78] UniProt Consortium. *Nucleic Acids Res.* **2014**, *42*, D191 – D198.
[79] P. W. Rose, A. Prlic, C. Bi, W. F. Bluhm, C. H. Christie, S. Dutta, R. K. Green, D. S. Goodsell, J. D. Westbrook, J. Woo, J. Young, C. Zardecki, H. M. Berman, P. E. Bourne, S. K. Burley, *Nucleic Acids Res.* **2015**, *43*, D345 – D356.
[80] R. D. Finn, A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. Sonnhammer, J. Tate, M. Punta, *Nucleic Acids Res.* **2014**, *42*, D222 – D230.
[81] M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, M. Tanabe, *Nucleic Acids Res.* **2014**, *42*, D199 – D205.
[82] Q. Duan, C. Flynn, M. Niepel, M. Hafner, J. L. Muhlich, N. F. Fernandez, A. D. Rouillard, C. M. Tan, E. Y. Chen, T. R. Golub, P. K. Sorger, A. Subramanaian, A. Ma'ayan, *Nucleic Acids Res.* **2014**, *42*, W449 – W460.
[83] X. M. Fernandez-Suarez, D. J. Rigden, M. Y. Galperin, *Nucleic Acids Res.* **2014**, *42*, D1 – D6.
[84] R. Salomon-Ferrer, A. W. Goetz, D. Poole, S. Le Grand, R. C. Walker, *J. Chem. Theory Comput.* **2013**, *9*, 3878 – 3888.
[85] L. C. T. Pierce, R. Salomon-Ferrer, C. A. F. de Oliveira, J. A. McCammon, R. C. Walker, *J. Chem. Theory Comput.* **2012**, *8*, 2997 – 3002.
[86] D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. Shan, W. Wriggers, *Science* **2010**, *330*, 341 – 346.
[87] P. Bisignano, S. Doerr, M. J. Harvey, A. D. Favia, A. Cavalli, G. De Fabritiis, *J. Chem. Inf. Model.* **2014**, *54*, 362 – 366.
[88] I. Buch, T. Giorgino, G. De Fabritiis, *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 10184 – 10189.
[89] J. E. Fuchs, S. von Grafenstein, R. G. Huber, H. G. Wallnoefer, K. R. Liedl, *Proteins* **2014**, *82*, 546 – 555.
[90] A. C. Pan, D. W. Borhani, R. O. Dror, D. E. Shaw, *Drug Discov. Today* **2013**, *18*, 667 – 673.
[91] R. O. Dror, H. F. Green, C. Valant, D. W. Borhani, J. R. Valcourt, A. C. Pan, D. H. Arlow, M. Canals, J. R. Lane, R. Rahmani, J. B. Baell, P. M. Sexton, A. Christopoulos, D. E. Shaw, *Nature* **2013**, *503*, 295 – 299.
[92] J. E. Fuchs, S. von Grafenstein, R. G. Huber, M. A. Margreiter, G. M. Spitzer, H. G. Wallnoefer, K. R. Liedl, *PLoS Comput. Biol.* **2013**, *9*, e1003007.
[93] K. Haug, R. M. Salek, P. Conesa, J. Hastings, P. de Matos, M. Rijnbeek, T. Mahendraker, M. Williams, S. Neumann, P. Rocca-Serra, E. Maguire, A. Gonzalez-Beltran, S. A. Sansone, J. L. Griffin, C. Steinbeck, *Nucleic Acids Res.* **2013**, *41*, 781 – 796.
[94] J. M. Foster, P. Moreno, A. Fabregat, H. Hermjakob, C. Steinbeck, R. Apweiler, M. J. Wakelam, J. A. Vizcaino, *PLoS One* **2013**, *8*, e61951.

Received: November 10, 2014
Accepted: December 16, 2014
Published online: ■■ ▮▮, 0000

*J. E. Fuchs, A. Bender, R. C. Glen**

■■–■■

Cheminformatics Research at the Unilever Centre for Molecular Science Informatics Cambridge