ICCS

International Conference
on Chemical Structures

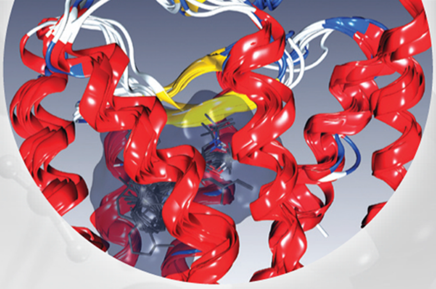
10th International Conference on Chemical Structures
10th German Conference on Chemoinformatics

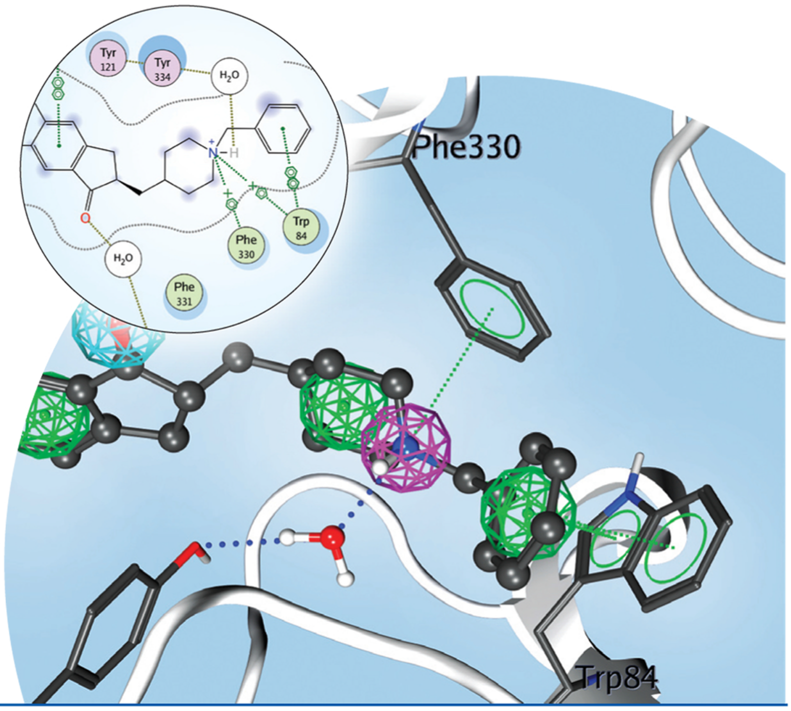
June 1–5 2014 // Noordwijkerhout // The Netherlands

Program & Abstracts

GCC

German Conference
on Chemoinformatics





Tyr 121
Tyr 334
H_2O
Phe 330
Trp 84
H_2O
Phe 331
Phe330
Trp84

Preface

Welcome to the 10th International Conference on Chemical Structures (ICCS) and the 10th German Conference on Chemoinformatics (GCC). For the 2014 event, the triennial ICCS and the annual GCC have decided to hold a joint conference covering cheminformatics, molecular modeling and all other aspects of the computer handling of chemical structures. This is to continue a long and successful history, which started for the ICCS with a NATO Advanced Study Workshop in 1973 and for the GCC with a CIC workshop in Hochfilzen, Austria, in 1986. Today, both conferences are among the most important events in this area of science and give an accurate picture of the state-of-the-art in the computer handling and manipulation of chemical structures.

We have received 144 abstract submissions from over 20 different countries. All submissions were subject to a review process carried out by a Scientific Advisory Board of 17 international reviewers from academia and industry. This allowed us to compile an outstanding scientific program of 34 plenary and 83 poster presentations. Additionally, the conference hosts an exhibition which allows a sizable number of scientific institutions and vendors to present their latest applications, content and software. And most important, sufficient time is provided for scientific exchange and discussion among the attending scientist both at the conference and also during the excursion to the beautiful city of Amsterdam on Wednesday afternoon.

Once again, the conference was chosen as the venue to present the triennial CSA Trust Mike Lynch Award. This year, it is granted to the “InChI Team” (Steve Heller, Alan McNaught, Igor Pletnev, Steve Stein, Dmitrii Tchekhovskoiin) in recognition of their outstanding accomplishments in the conception and development of the IUPAC International Chemical Identifier (InChI). Steve Heller will open the conference by receiving the award and delivering the keynote address titled The Status of the InChI Project on Sunday evening.

After the conference, you are encouraged to submit your presentation or poster for publication in a special ICCS/GCC issue of the Journal of Chemical Information and Modeling (JCIM) scheduled to appear in early 2015. Papers can be submitted at any date up to the 1st of September 2014, and authors should mention in their cover letter that the manuscript is intended to be included in the ICCS/GCC special issue. Of course, all manuscripts will be subject to a peer review following the JCIM guidelines.

This book of abstracts is intended to inform you about the scientific program of the conference and to help you to plan your attendance. Moreover, we also hope that the abstracts in this volume will serve you as a reminder of the presentations and posters as well as provide a snapshot of the current research in the area of cheminformatics and molecular modeling in 2014.

At this point, we would also like to thank the many sponsors for their financial support which helped us to provide bursaries to a considerable number of PhD-student attendants.

We hope that you enjoy the conference!

Markus Wagener, ICCS Chair
Frank Oellien, ICCS Co-Chair
Uli Fechner, GCC Chair

Contents

The Conference 5

Organizing Societies 7

Organizing Committee 7

Conference Office 7

Scientific Advisory Board 8

Sponsors 9

Exhibition 11

Workshops 13

Group Excursion 15

Scientific Program 17

Plenary Session 19

Poster Session RED 25

Poster Session BLUE 29

Plenary Session Abstracts 33

Keynote Address CSA Trust Mike Lynch Award 35

Session A: Cheminformatics 39

Session B: Structure-Based Drug Design and Virtual Screening 51

Session C: Analysis of Large Chemistry Spaces 61

Session D: Dealing with Biological Complexity 67

Session E: Integration of Chemical Information with other Resources 73

Session F: Structure-Activity and Structure-Property Prediction 79

Poster Session Abstracts RED 85

Poster Session Abstracts BLUE 123

The Conference

Organizing Societies

Chemical Information and Computer Applications Group of the Royal Society of Chemistry (RSC)	ROYAL SOCIETY OF CHEMISTRY
Chemical Structure Association Trust (CSA Trust)	CSA TRUST
Chemistry-Information-Computer Division of the German Chemical Society (GDCh)	GDCh
Division of Chemical Information of the American Chemical Society (ACS)	ACS
Division of Chemical Information and Computer Science of the Chemical Society of Japan (CSJ)	CSJ
Royal Netherlands Chemical Society (KNCV)	KNCV
Swiss Chemical Society (SCS)	

Organizing Committee

- Markus Wagener, ICCS Chair, Grünenthal
- Frank Oellien, ICCS Co-Chair, MSD
- Uli Fechner, GCC Chair, Beilstein-Institut

Conference Office

mcc Agentur für Kommunikation GmbH, Wölfertstraße 13, D-12101 Berlin
http://www.mcc-pr.de

Scientific Advisory Board

- Peter Ertl, Novartis Institute for Biomedical Research
- Kimito Funatsu, University of Tokyo
- Val Gillet, University of Sheffield
- Chris de Graaf, VU University Amsterdam
- Rajarshi Guha, NIH Chemical Genomics Center
- Michael Lajiness, Eli Lilly and Company
- Jordi Mestres, Hospital del Mar Research Institute(IMIM)
- Matthias Rarey, University of Hamburg
- Keith T. Taylor, Accelrys, Inc.
- Lothar Terfloth, Molecular Networks
- Markus Wagener, Grünenthal
- W. Patrick Walters, Vertex Pharmaceuticals, Inc.
- Egon Willighagen, Maastricht University
- Oliver Koch, TU Dortmund University
- Jan Kriegl, Boehringer-Ingelheim
- Wendy Warr, Wendy Warr & Associates
- Alex Tropsha, UNC Chapel Hill

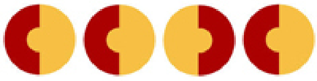
Sponsors

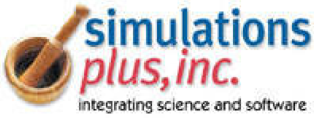
Premier Sponsor

Platinum Sponsor

Gold Sponsors

Silver Sponsors





Conference Bag Sponsor

evolvus

Poster Awards Sponsor

ChemInform®
RxnFinder

Publishing Sponsors

Journal of
Chemical Information
and Modeling

Journal of
Computational
Chemistry www.c-chem.org

molecular informatics
models – molecules – systems

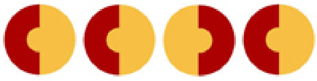
Other Sponsors

We would like to thank CCL.NET and Jan Labanowski for adding the conference to the CCL conferences webpage. We are also grateful to the Center of Bioinformatics of the University of Hamburg for hosting the conference webpage.

Exhibition

c.a.r.u.s®

cas®
A division of the American Chemical Society



CERTARA™
Implementing Translational Science

ChemAxon

CHEMICAL
COMPUTING
GROUP

™
cresset
smarter chemistry | smarter decisions

dotmatics
knowledge solutions

GVK BIO

KNIME

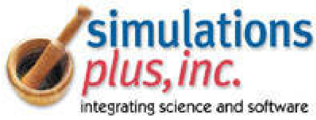
MEDIT
www.medit.fr

NextMove
SOFTWARE
t

OpenEye
Scientific Software

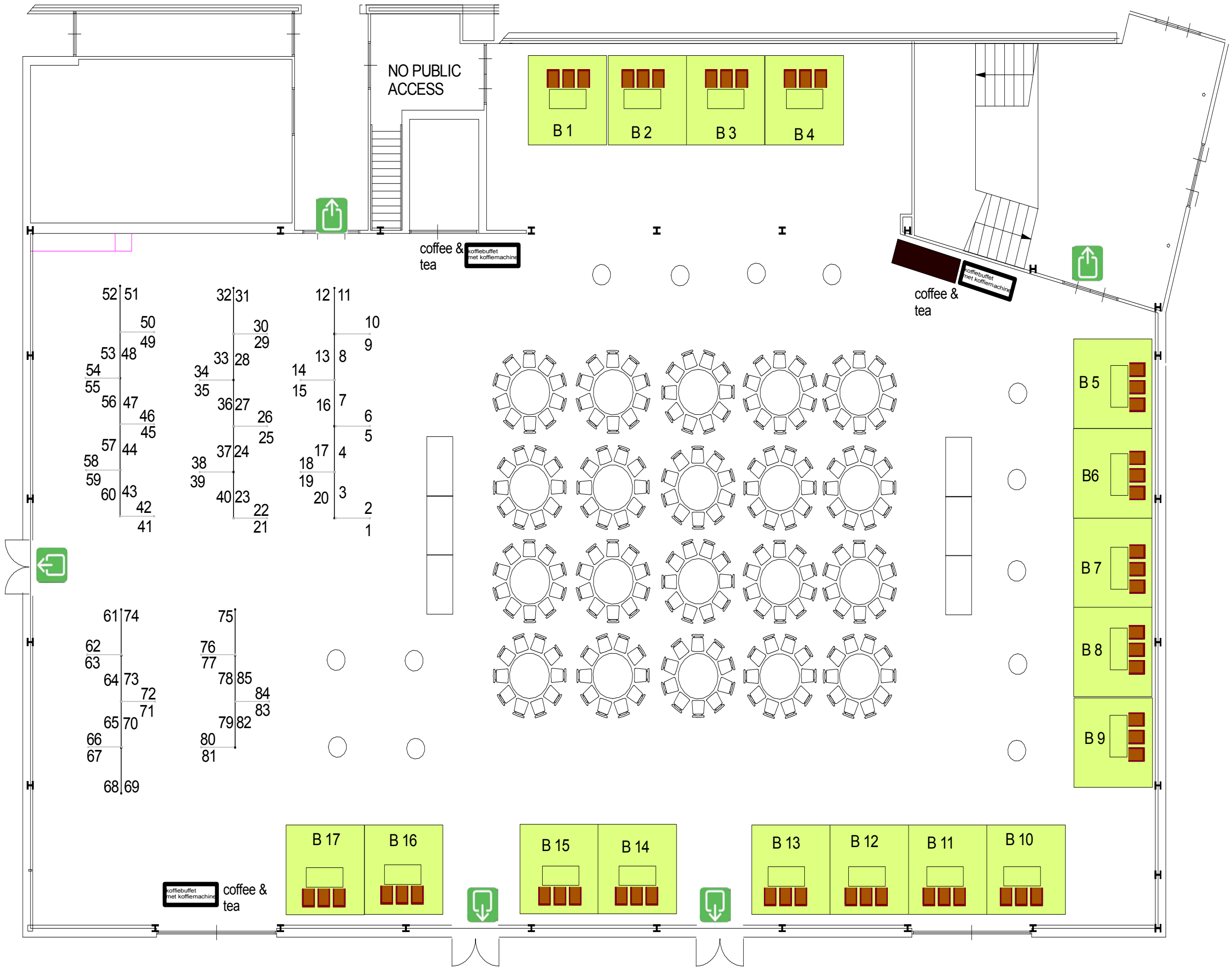
PerkinElmer®
For the Better

SCHRÖDINGER.


simulations
plus, inc.
integrating science and software

Xemistry chemoinformatics

Exhibition Layout



Exhibitors

Exhibitor	Booth	Exhibitor	Booth
Simulations Plus, Inc.	B 1	NextMove Software	B 10
Certara	B 2	Chemical Abstracts Service	B 11
Chemical Computing Group	B 3	Schrödinger	B 12
OpenEye Scientific Software	B 4	Dotmatics	B 13
Cambridge Crystallographic Data Centre	B 5	KNIME	B 14
Cresset	B 6	Xemistry	B 15
c.a.r.u.s. HMS GmbH	B 7	MEDIT	B 16
PerkinElmer	B 8	GVK BIO	B 17
ChemAxon	B 9		

Exhibition Hours

- Monday June 2, 2014 14:30 – 19:30
- Tuesday June 3, 2014 14:30 – 19:30

Workshops

Sunday June 1, 2014, 15:00 – 17:00

Certara

Muse Invention is a molecular design workflow to accelerate the identification and optimization of lead candidates, and enable rapid, multi-criteria drug design. The workshop will demonstrate how to use multiple parameters as criteria to optimize molecule design in ligand- and structure-based approaches.

A hands-on training is possible for participants using laptops and arriving one hour before the workshop starts for installation and setup (Windows 7, Linux Redhat or CentOS). Each participant of the workshop will be eligible to obtain a free license valid for one month.

The upcoming feature for reaction driven de novo design using chemical reactions will be presented and discussed.

Further information is available at www.certara.com/products/molmod/muse or by contacting us at eu_support@certara.com.

OpenEye

Prediction of which method or tool will work best for a particular task is a common but difficult problem in computational drug discovery. We will introduce best practice methods for comparison of virtual screening methods, exemplifying them with examples of

- Ligand-based screening with the extensively validated 3D shape-based tool ROCS
- Structure-based methods with the novel and effective OEDocking.

We will show how to select the single best performing method based on a set of retrospective experiments and how to assess the complementarity between methods based on measuring information recovery. We will also show how to reliably maximise the performance of ligand-based virtual screening tools by query optimization.

While the workshop will introduce the use of OpenEye's tools ROCS (shape-based virtual screening and lead-hopping) and OEDocking (structure-based pose prediction and virtual screening), the methodology to be presented is completely general.

Thursday June 5, 2014, 14:00 – 16:00

Cambrige Crystallographic Data Centre

The Cambridge Structural Database (CSD) now contains experimental crystal structure information for more than 700,000 organic and organometallic compounds. The CSD System further contains knowledge-based libraries providing easy access to extracted information about intermolecular interactions (IsoStar) and molecular geometries (Mogul). This workshop will concentrate on showing how to search the CSD itself to derive and analyse molecular geometrical information, and on the use of Mogul in applications relevant to drug discovery such as verifying conformational studies and explaining conformational trends in bound ligands. We will also discuss new uses of this molecular geometry information in constraining explored docking solutions, aiding the determination of ligand geometries in macromolecular crystallography, and the derivation of 3D conformations solely from crystallographic knowledge.

Chemical Computing Group

The course will focus on fragment-based drug design tools. Combinatorial fragment design and scaffold replacement in the receptor active site will be covered in detail, along with approaches for fragment linking and growing. A method for generating a series of closely related derivatives through medicinal chemistry transformations and the reaction based combinatorial builder will be presented. The use of pharmacophores and 2D/3D descriptors to guide drug design processes will also be discussed.

Group Excursion

Agenda

13:00	departure from the conference center, Noordwijkerhout
14:00	arrival at the Museumplein, Amsterdam possibility to visit the Rijksmuseum
18:00	boat cruise on the Amsterdam canals (Dutch: grachten) departure from the dock opposite of the Rijksmuseum
19:30	arrival by boat at the St. Olofskapel, NH Barbizon Palace, dinner
22:15	departure from the St. Olofskapel, NH Barbizon Palace, by bus
23:00	arrival at the conference center, Noordwijkerhout

Rijksmuseum

The Rijksmuseum – the Dutch national museum of art and history – possesses the largest and most important collection of classical Dutch art, e.g., works from Vermeer, Hals and Rembrandt (The Nightwatch). Additionally, the collection consists of a large variety of decorative art, e.g., Delftware or 17th century dollhouses, and an extensive collection of ship models including the original stern of the HMS Royal Charles. The museum has been under reconstruction for 10 years and has been open to the public again since 2013.

Scientific Program

Plenary Session

Sunday, June 1	
12:00 – 18:00	**Registration**
15:00 – 17:00	Workshop: Certara Meeting Room Boston 11
15:00 – 17:00	Workshop: OpenEye Meeting Room Boston 13
17:00 – 18:00	**Free Time**
18:00 – 19:00	**Welcome and Keynote Address** Rotonde
18:00 – 18:15	*Welcome and Introduction* ICCS/GCC Organizing Committee
18:15 – 19:00	Keynote Address, CSA Trust Mike Lynch Award *K-1: The status of the InChI project* Steve Heller, InChI Trust
19:00 – 20:00	**Welcoming Reception** Atrium
20:00 – 22:00	**Reception Dinner** Atrium

Monday, June 2	
08:30 – 14:30	**Session A – Cheminformatics** Matthias Rarey, Presiding Rotonde
08:30 – 09:00	*A-1: Enhancing the quality of HTS compound collections through multi-objective analysis* Britta Nisius, UCB Pharma, Belgium
09:00 – 09:30	*A-2: Compiling a medium-size screening library by polypharmacology directed virtual screening* Modest von Korff, Actelion Ltd., Switzerland
09:30 – 10:00	*A-3: Association Mining with Empirical Bayes and the Poisson-Gamma Model* Peter T. Corbett, Royal Society of Chemistry, United Kingdom
10:00 – 10:30	**Coffee Break** Atrium
10:30 – 11:00	*A-4: Computer assisted identification of green tea metabolites in human urine* Lars Ridder, Wageningen University, Netherlands
11:00 – 11:30	*A-5: A Novel Approach to De Novo Design using Reaction Networks* James E. A. Wallace, University of Sheffield, United Kingdom
11:30 – 12:00	*A-6: Improving substructure screens* Andrew Dalke, Dalke Scientific, Sweden
12:00 – 13:00	**Lunch** Atrium
13:00 – 13:30	*A-7: Generating Small Molecule Conformations from Structural Data* Patrick McCabe, Cambridge Crystallographic Data Centre, United Kingdom
13:30 – 14:00	*A-8: Physicochemical property trends in allosteric and orthosteric binders and their predicted presence in chemical databases* Gerard J. P. van Westen, EMBL-EBI, United Kingdom
14:00 – 14:30	*A-9: Multi-dimensional activity cliff analysis* Mark Mackey, Cresset, United Kingdom
14:30 – 15:00	**Coffee Break** Atrium
14:30 – 19:30	**Exhibition & Poster Session** Mic Lajiness, Presiding Atrium
15:00 – 17:00	**Poster Presentations RED** Atrium
18:30 – 19:30	**Reception** Atrium
19:30 – 21:30	**Dinner** Atrium

Tuesday, June 3	
08:30 – 14:30	**Session B – Structure-Based Drug Design and Virtual Screening** Pat Walters, Presiding Rotonde
08:30 – 09:00	*B-1: Nonadditivity in SAR analysis: It's not a bug – it's a feature!* Christian Kramer, University of Innsbruck, Austria
09:00 – 09:30	*B-2: Ligand structures in the Protein Data Bank – can we trust them?* David Sehnal, Masaryk University, Czech Republic
09:30 – 10:00	*B-3: An exploration of the 3D chemical space has highlighted a specific shape profile for the compounds intended to inhibit protein-protein interactions.* Mélaine A. Kuenemann, University Paris Diderot, France
10:00 – 10:30	**Coffee Break** Atrium
10:30 – 11:00	*B-4: The Motility of Water Molecules – A Statistical Evaluation of Water Molecules Based on Electron Density* Eva Nittinger, University of Hamburg, Germany
11:00 – 11:30	*B-5: How to treat waters in virtual screening? A case study on the Adenosine A2A Receptor* Bart Lenselink, Leiden Academic Centre for Drug Research, Netherlands
11:30 – 12:00	*B-6: WaterFLAP: Fast water prediction, scoring, and docking using GRID Molecular Interaction Fields* Simon Cross, Molecular Discovery Ltd, United Kingdom
12:00	**Group Photo**
12:00 – 13:00	**Lunch** Atrium
13:00 – 13:30	*B-7: Computational Methods Enabling Fragment Based Drug Discovery* Stephen D. Pickett, GlaxoSmithKline, United Kingdom
13:30 – 14:00	*B-8: Application of linear response MM-PBSA and QM/MM-GBSA rescoring for postprocessing of protein-ligand docking poses* Kanin Wichapong, Martin-Luther University Halle-Wittenberg, Germany
14:00 – 14:30	*B-9: Large Scale FEP on Congeneric Ligand Series – Have Practical Free Energy Calculations arrived at Last?* Thomas B. Steinbrecher, Schrodinger GmbH, Germany
14:30 – 15:00	**Coffee Break** Atrium
14:30 – 19:30	**Exhibition & Poster Session** Lothar Terfloth, Presiding Atrium
15:00 – 17:00	**Poster Presentations BLUE** Atrium
18:30 – 19:30	**Reception** Atrium
19:30 – 21:30	**Dinner** Atrium

Wednesday, June 4	
08:30 – 10:30	**Session C – Analysis of Large Chemistry Spaces** Thomas Engel, Presiding Rotonde
08:30 – 09:00	*C-1: The Proximal Lilly Collection Initiative: Design, Exploration and Application to Drug Discovery* Christos A. Nicolaou, Lilly Research Laboratories, United States
09:00 – 09:30	*C-2: Screening reaction-pathway driven very large chemical space: Discovery of potent mdm2-p53 antagonist* Alex Dömling, University of Groningen, Netherlands
09:30 – 10:00	*C-3: Large scale classification of chemical reactions from patent data* Gregory A. Landrum, Novartis Institutes for BioMedical Research, Switzerland
10:00 – 10:30	*C-4: Open Patent Data* Anna Gaulton, EMBL-EBI, United Kingdom
10:30 – 11:00	**Coffee Break** Atrium Lounge
11:00 – 13:00	**Session D – Dealing with Biological Complexity** Oliver Koch, Presiding Rotonde
11:00 – 11:30	*D-1: The impact of broad-scale genetic data on drug targets* Josef Scheiber, BioVariance GmbH, Germany
11:30 – 12:00	*D-2: Navigating chemical and biological spaces in the search of novel pharmaceuticals* Paula M. Petrone, Hoffmann-La Roche, Switzerland
12:00 – 12:30	*D-3: Using Information from Historical High-Throughput Screens to Predict Active Compounds* Sereina Riniker, Novartis Institutes for BioMedical Research, Switzerland
12:30 – 13:00	*D-4: PubChem Structure-Activity Relationship Clusters and the Difference between 2-D and 3-D Similarity* Volker D. Hähnke, National Center for Biotechnology Information, United States
13:00	**Box Lunch**
13:00 – 23:00	**Excursion** An afternoon in Amsterdam including a visit to one of the famous museums and a cruise on the Amsterdam canals. In the evening, the conference diner will take place in the St Olof Chapel.

Thursday, June 5	
07:30 – 08:30	**Hotel Check-out**
08:30 – 10:30	**Session E – Integration of Chemical Information with other Resources** Egon Willighagen, Presiding Rotonde
08:30 – 09:00	*E-1: Automatic completion of COSHH risk assessment forms using semantic representation of GHS and CLP regulation* Mark I. Borkum, University of Southampton, United Kingdom
09:00 – 09:30	*E-2: Scientific Lenses over Linked Chemistry Data using BridgeDB and the Open PHACTS Chemical Registration System* Colin R. Batchelor, Royal Society of Chemistry, United Kingdom
09:30 – 10:00	*E-3: Will the real drug targets please stand up?* Christopher Southan, University of Edinburgh, United Kingdom
10:00 – 10:30	*E-4: Large-scale integration of chemical and biological data for drug discovery* Jörg Degen, Hoffmann-La Roche Ltd., Switzerland
10:30 – 11:00	**Coffee Break & Hotel Check-out** Atrium Lounge
11:00 – 13:00	**Session F – Structure-Activity and Structure-Property Prediction** Peter Ertl, Presiding Rotonde
11:00 – 11:30	*F-1: Drug target prediction – comparison of methods* David Evans, Eli Lilly, United Kingdom
11:30 – 12:00	*F-2: Broadening the usefulness of pKa predictions by taking multiple protonation microstates into account* Robert D. Clark, Simulations Plus, Inc., United States
12:00 – 12:30	*F-3: Using Decision trees to rationalize the mechanism-of-action of hypnotics: Assessing predicted sedative effects of marketed drugs in vivo* Georgios Drakakis, Unilever Centre for Molecular Science Informatics, United Kingdom
12:30 – 13:00	*F-4: Existing and Developing Approaches for QSAR Modeling of Mixtures* Eugene Muratov, University of North Carolina, United States
13:00 – 13:15	**Closing Remarks** ICCS/GCC Organizing Committee
13:15 – 14:00	**Lunch**
13:30	**Shuttle busses leave for Schiphol Airport**
14:00 – 16:00	Workshop: Cambridge Crystallographic Data Centre Meeting Room Boston 11
14:00 – 16:00	Workshop: Chemical Computing Group Meeting Room Boston 13
14:30	**Shuttle busses leave for Schiphol Airport**
16:15	**Shuttle busses leave for Schiphol Airport**

Poster Session RED

Analysis of Large Chemistry Spaces

P-2: *Knowledge Discovery in Pharmaceutical Drug Transport using Emerging Graph Patterns*
Guillaume Poezevara, Université de Caen Basse-Normandie, France

P-4: *Comprehensive analysis of bioisosteric replacement in ligands of a serotonin receptors family*
Dawid Warszycki, Polish Academy of Sciences, Poland

Cheminformatics

P-6: *Discriminative Chemical Patterns: Automatic and Interactive Design*
Stefan Bietz, University of Hamburg, Germany

P-8: *Protein binding site comparison based on secondary structure elements*
Tobias Brinkjost, TU Dortmund, Germany

P-10: *fpserver: an in-process cheminformatics database system for Python*
Andrew Dalke, Dalke Scientific, Sweden

P-12: *Molecular Structure Similarity in the Context of Orphan Drug Legislation*
Pedro Franco, University of Sheffield, United Kingdom

P-14: *QSARINS: software for the development, analysis and validation of MLR models, and QSARINS-Chem: Insubria datasets and QSA(P)R models for environmental pollutants*
Paola Gramatica, University of Insubria – Varese, Italy

P-16: *A fragment-based computational approach to study the phase behavior of bio(polymers) and drug excipients*
Jan-Willem Handgraaf, Culgi B.V., Netherlands

P-18: *MONA – Intuitive, visual navigation through molecule collections*
Matthias Hilbig, University of Hamburg, Germany

P-20: *Ushering the Cactvs Toolkit into the Python Age (without breaking the legacy)*
Wolf D. Ihlenfeldt, Xemistry GmbH, Germany

P-22: *Combinatorial Library Optimization Process In The European Lead Factory Project*
Tuomo Kalliokoski, Lead Discovery Center GmbH, Germany

P-24: *Reaction Representation and Structure Transformation with Ambit-SMIRKS. Application in Metabolite Prediction*
Nikolay T. Kochev, University of Plovdiv, Bulgaria

P-26: *Multi-label Drug-like Compounds and Probabilistic Graphical Model Using Variational Mean Field Theory*
Hamse Y. Mussa, Unilever Center for Molecular Science Informatics, United Kingdom

P-28: *DataWarrior, a Free Tool for Chemistry Aware Data Visualization and Analysis*
Thomas Sander, Actelion Pharmaceuticals Ltd., Switzerland

P-30: *Expanding chemical space coverage in Matched Molecular Pairs Analysis*
Peter Schmidtke, Discngine, France

P-32: *Plexus – A Flexible Library Design Platform to Conduct Innovative Chemistry*
Iván Solt, ChemAxon, Hungary

Dealing with Biological Complexity

P-34: *Protein active site comparison with SiteHopper: phylogeny to polypharmacology*
Paul Hawkins, OpenEye Scientific, United States

P-36: *Assembly of Helical Chimeras with Natural Peptides: Simulating Molecular Interaction Surfaces*
Jeremie Mortier, FU Berlin, Germany

P-38: *Molecular simulations of peptides and proteins with Molecular Fragment Dynamics*
Andreas Truszkowski, Westphalian University of Applied Sciences, Germany

Integration of Chemical Information with other Resources

P-40: *KinHub: a Database to Enable Kinome-wide Analysis*
Sameh Eid, BioMed X Innovation Center, Germany

P-42: *Lessons learned from 30 years of developing Drug Discovery Informatics Systems*
Mic Lajiness, Eli Lilly, United States

P-44: *Bringing Desktop Solutions to the Mobile Space: the ChemDraw App Story*
Pierre Morieux, PerkinElmer, France

P-46: *Focus – a global communication platform for applied and theoretical medicinal chemists*
Nikolaus Stiefl, Novartis Institute for Biomedical Research, Switzerland

Structure-Activity and Structure-Property Prediction

P-48: *Identification of novel potential anti-cancer agents using network pharmacology based computational modelling*
Benjamin C. P. Allen, e-Therapeutics, United Kingdom

P-50: *Development of Medical Countermeasures against Organophosphorous Intoxication; From High Throughput Screening to Quantitative Structure-Activity Relationships*
Susanne Johansson, FOI - Swedish Defence Research Agency, Sweden

P-52: *Surflex-QMOD: Protein Pocket Modeling for Affinity Prediction*
Alexander Steudle, Certara, Germany

P-54: *Metabolism simulation and toxicity prediction in the evaluation of food*
Lothar Terfloth, Molecular Networks GmbH, Germany

P-56: *Single Target SAR and Multiple Target Polypharmacology Analysis Using inSARa*
Knut Baumann, Technische Universität Braunschweig, Germany

Structure-Based Drug Design and Virtual Screening

P-58: *Repurposing of known kinase inhibitors for inhibition of trypanothione synthetase*
Christiane Ehrt, TU Dortmund, Germany

P-60: *Retrieving 'hits' through in silico screening and expert assessment*
Renate Griffith, University of New South Wales, Australia

P-62: *Structural requirements of drug candidates to cause cholestatic side effects*
Susanne Hermans, Radboud Institute for Molecular Life Sciences (RIMLS), Netherlands

P-64: *MotiveQuery: Language and Web Service for Fast Identification of Protein Structural Motifs in the Entire Protein Data Bank*
Lukáš Pravda, Masaryk University, Czech Republic

P-66: *Exploiting Solvent Effects in Drug Design and Optimization*
Guido Kirsten, Chemical Computing Group, Montreal, Canada

P-68: *In silico prediction of antitumor cytotoxicity of pharmacologically active substances for human breast cancer and normal cell lines*
Varvara Konova, Russian Academy of Medical Sciences, Russian Federation

P-70: *Modular Interactive Structure-Based Pharmacophore Searching*
Jens Kunze, ETH Zurich, Switzerland

P-72: *Identification of potential protein-protein binding sites by using a 3 phased self-contained in silico workflow*
Christian Jäger, Fraunhofer Institute for Cell Therapy and Immunology, Germany

P-74: *NUCLEO.QUERY: A free web-based virtual screening platform targeting nucleotide cofactor proteins*
Constantinos Neochoritis, University of Groningen, Netherlands

P-76: *Small molecular turn mimetics as protein-protein interaction inhibitor building blocks*
Anna Rudo, TU Dortmund, Germany

P-78: *Development of a Protein-ligand Interaction Database for Structure-based Drug Design*
Richard Sherhod, Vernalis R&D, United Kingdom

P-80: *The Pocketome of Human Kinases: Prioritizing (yet) Untapped Protein Kinases for Drug Development*
Andrea Volkamer, BioMed X Innovation Center, Germany

Poster Session BLUE

Analysis of Large Chemistry Spaces

P-1: *Chemical Space Analysis of Ion Channel Ligands Identified by High-Throughput Screening*
Alexander Böcker, Evotec AG, Germany

P-3: *Drug safety assessment through automatic extraction of structure-activity relationships*
Alexander Steudle, Certara, Germany

Cheminformatics

P-5: *ChemTrove: Enabling a generic ELN to support Chemistry by integrating ChemSpider widgets and templates*
Colin R. Batchelor, Royal Society of Chemistry, United Kingdom

P-7: *A computationally efficient structure key for large proteins*
Gerd Blanke, StructurePendium Technologies GmbH, Germany

P-9: *Estimating Classification Uncertainty for Unbalanced Ensemble Models*
Robert D. Clark, Simulations Plus, Inc., United States

P-11: *Maximum Common Substructure-based Data Fusion in Similarity Searching and Virtual Screening*
Edmund V. Duesbury, University of Sheffield, United Kingdom

P-13: *Web server for the rapid calculation of empirical atomic charges with QM accuracy*
Stanislav Geidl, Masaryk University, Czech Republic

P-15: *Towards a new chemical standardization pipeline in PubChem*
Volker D. Hähnke, National Center for Biotechnology Information, United States

P-17: *Recent Advancements in Chemoinformatics for Porous Materials*
Maciej Haranczyk, Lawrence Berkeley National Laboratory, United States

P-19: *Chemogenomics analysis of small molecule bioactivity data: Privileged scaffolds and conserved structural elements in proteins*
Lina Humbeck, TU Dortmund, Germany

P-21: *Identification of functionally active residues in* α*1-AR by computational approaches*
Kapil Jain, The University of Queensland, Australia

P-23: *HELM: an open standard for handling large Biologics*
Iván Solt, ChemAxon, Hungary

P-25: *Bioisosteres in accessible chemistry space*
Mark Mackey, Cresset, United Kingdom

P-27: *A Neural Gas based Approach towards Pharmacophore Model Elucidation*
Daniel Moser, Goethe University Frankfurt, Germany

P-29: *Vectorizing hydrophobicity: From in silico models to membrane-active peptides*
Max Pillong , ETH Zurich, Switzerland

P-31: *An Upper Bound to the Effectiveness of Substructural Analysis Methods*
Nor S. Sani, University of Sheffield, United Kingdom

P-33: *A machine learning-based protocol for docking results analysis*
Sabina Smusz, Polish Academy of Sciences, Poland

P-35: *Molecular Fragment Dynamics Study of the Interaction between Zinc Ricinoleate and the Complexing Agent Methylglycinediacetic Acid as a new System for Enzyme Purification*
Karina van den Broek, Westfälische Hochschule, Germany

Dealing with Biological Complexity

P-37: *Heteromeric assembly of voltage gated ion channels*
Guido Humpert, Grünenthal, Germany

P-39: *HCS-fingerprints opening new routes to target identification*
Paul Selzer, Novartis Institutes for Biomedical Research, Switzerland

P-41: *The lipophilicity mirage: logD as an endpoint in drug discovery*
Christian Tyrchan, AstraZeneca, Sweden

Integration of Chemical Information with other Resources

P-43: *A new platform to solve the Computational Chemistry's BigData problem*
Carles Bo, Institute of Chemical Research of Catalonia (ICIQ), Spain

P-45: *Liberating Laboratory Data*
Jeremy G. Frey, University of Southampton, United Kingdom

P-47: *Paper is not dead, and PDFs feel very well, thank you*
Wolf D. Ihlenfeldt, Xemistry GmbH, Germany

P-49: *Peptide Line Notations for Biologics Registration and Patent filings*
Roger Sayle, NextMove Software Limited, United Kingdom

P-51: *An automated document classifier to retrieve ChEMBL-like papers*
Gerard J. P. van Westen, EMBL-EBI, United Kingdom

P-53: *Open PHACTS: Solutions and the Foundation*
Egon Willighagen, Maastricht University, Netherlands

Structure-Activity and Structure-Property Prediction

P-55: *The role of negative evidence in fingerprint-based Naïve Bayes models*
Nikolas H. Fechner, Novartis Institutes for BioMedical Research, Switzerland

P-57: *A Novel Mechanistic Approach to Free-Wilson SAR Analysis Enabling the Use of Potency Data in Computational Lead Optimization*
Kiril Lanevskij, ACD/Labs, Inc., Canada

P-59: *First-principles search for molecular structures with a genetic algorithm*
Adriana Supady, Fritz-Haber-Institut der Max-Planck-Gesellschaft, Germany

P-61: *Exploratory Data Analysis & Visualization Applied to Structure-Activity Relationships*
Pierre Morieux, PerkinElmer, France

P-63: *VAMMPIRE-LORD: an open access web server for targeted lead optimization based on Matched Molecular Pairs*
Julia Weber, Goethe University Frankfurt, Germany

Structure-Based Drug Design and Virtual Screening

P-65: *A-WOL Ligand Based Screening combined with HTS*
Jaclyn Bibby, University of Liverpool, United Kingdom

P-67: *MOARF: A novel workflow for the multiobjective optimisation of drug-like molecules*
Nicholas Firth, Institute of Cancer Research, United Kingdom

P-69: *Interactions of Strongly and Weakly Bound Water Molecules to Protein Binding Sites and Propensity for Replacements*
Stefan Güssregen, Sanofi, Germany

P-71: *Maximising recovery in virtual screening using information entropy*
Paul Hawkins, OpenEye Scientific, United States

P-73: *Knowledge-Based Potentials in Protein-Protein Docking*
Dennis M. Krüger, Chemical Genomics Centre of the Max Planck Society, Germany

P-75: *Identification of novel tubulin inhibitors by parallel virtual screening protocol of reaction-based combinatorial library of combretastatin CA-4 derivatives*
Rafal Kurczab, Polish Academy of Sciences, Poland

P-77: *PTP1B – Combining Inhibitory Activity with Selectivity*
Alexandra Naß, FU Berlin, Germany

P-79: *Rational Design Supported by Ligand-Based NMR Data*
Ionut Onila, University Konstanz, Germany

P-81: *Pharmacophore based Virtual-Screening for the Identification of Covalent Coxsackievirus 3C Protease Inhibitors*
Robert Schulz, FU Berlin, Germany

P-83: *Use of Back-Calculated Protein Chemical Shift Perturbations in Fragment Docking*
Tim ten Brink, Universite Lyon 1, France

P-85: *Small molecule inhibitors of CD40-TRAF6 interaction reduce atherosclerosis by targeting its inflammatory natur*e
Barbara A. Zarzycka, Maastricht University, Netherlands

Plenary Session Abstracts

Keynote Address CSA Trust Mike Lynch Award

K-1: The status of the InChI project

Steve Heller[1], Stephen Stein[2], Dmitrii Tchekhovskoi[2], Alan McNaught[1], Igor Pletnev[3]

[1]InChI Trust, [2]The National Institute of Standards and Technology (NIST), Gaithersburg, MD, United States, [3]Lomonosov Moscow State University, Moscow, Russia

The IUPAC International Chemical Identifier (InChI), an Open Source algorithm for uniquely representing a defined chemical structure standard, has found acceptance both quickly and widely due primarily to three factors, two of which are technical. The first is that, with the availability of vast amounts of information and data on the Internet, the chemical community needs a way to link this information and data. The second is that being Open Source and free it is easy to obtain and use. The third, and the reason I am giving this lecture is the unusual "perfect storm" of having a team of five people who work together so smoothly that the project has progressed with no personality or political conflicts – an unusual feat for a group of high quality scientists from three countries, a US Government lab, and an international chemical standards organization.

Developing a standard is easy. The hard thing is getting a standard adopted and used. As the newly appointed Associate Director for Data Science at NIH has said – "Standards are like toothbrushes. Everyone has one, but no one wants to use someone else's." The InChI algorithm was developed and made available in short period of time (about 5 years) and quickly filled a clear and large void. The policies of IUPAC (the international chemicals standards organization, NIST (where the algorithm was initially programmed), and the InChI Trust (the non-profit organization that funds the ongoing project costs), all have the same policies and goals – to actually deliver and maintain a real, viable, and valuable standard.

This presentation will describe what chemistry the InChI algorithm currently covers and the status of the many areas of chemistry and biochemistry that are under development. Further information on the project can be found at http://www.inchi-trust.org.

Session A: Cheminformatics

A-1: Enhancing the quality of HTS compound collections through multi-objective analysis

Britta Nisius

UCB Pharma, UCB New Medicines, Braine-l'Alleud, Belgium

High-throughput screening (HTS) is one of the most widely used approaches to identify starting points for drug discovery [1]. Even though compound screening collections typically contain a few thousands to millions of compounds, they are still very small in numbers compared to the vastness of chemical space and the diversity of biological space.

To positively impact the hit-rate of a HTS run and the quality of the obtained hits, it is imperative to carefully design a HTS compound collection. Consequently, there are several key requirements one should consider in order to generate a good screening collection:

- the compounds need to possess desirable physicochemical properties,
- they should be chemically & structurally diverse and
- they should cover a broad biological space.

Therefore, designing HTS collections is a multi-objective task, which necessitates effective combination of varied computational approaches.

In this presentation, we will present how diverse chemoinformatics approaches, e.g., Murcko scaffolds [2], drug-likeness scores, structural fingerprints [3] and Bayesian affinity fingerprints [4], were combined in a multi-objective way by using Pareto ranking and Derringer functions [5] to expand the UCB HTS collection.

Multiple external compound collections were analyzed and a few were selected, from which a limited number of high-quality compounds complementing the existing HTS collection were purchased.

1. Bleicher K. H.; Böhm, H. J.; Müller, K.; Alanine, A. I. Hit and lead generation: Beyond high-throughput screening. *Nat. Rev. Drug Discov.* **2003***,* 2, 369–378.
2. Bemis, G.W.; Murcko, M. A. The properties of known drugs: 1. Molecular frameworks. *J. Med. Chem.* **1996**, 39, 2887–2893.
3. Rogers, D.; Hahn, M. Extended connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, 50, 742–754.
4. Bender, A.; Jenkins, J. L.; Glick, M.; Deng, Z.; Nettles,J. H.; Davies, J. W. "Bayes Affinity Fingerprints" improve retrieval rates in virtual screening and define orthogonal bioactivity space: When are multitarget drugs a feasible concept? *J. Chem. Inf. Model.* **2006**, 46, 2445–2456.
5. Segall, M. D. Multi-parameter optimization: identifying high quality compounds with a balance of properties. *Curr. Pharm. Des.* **2012**, 18, 1292-1310.

A-2: Compiling a medium-size screening library by polypharmacology directed virtual screening

Modest von Korff, Thomas Sander, Urs Lüthi, Oliver Peter, Romain Siegrist, Thierry Kimmerlin

Actelion Ltd., Gewerbestrasse 16, CH-4123 Allschwil, Switzerland

This study describes how to compile a diverse library of 30,000 small molecules for medium thoughput screening. A medium sized library was needed because of the implementation of phenotypic assays in drug discovery. The limited throughput of these assays restricted the number of compounds to be tested. Therefore, it was desired to select molecules having a high probability of being active on different targets in drug discovery. The compounds, available for the selection, were restricted to 440,000 in-house molecules. To ensure a high bioactivity rate of the chosen molecules it was decided to go for an indirect polypharmacology approach [1]. Polypharmacology refers to the observation that a molecule that is biologically active on one target has a certain probability of being active on another target too [2]. Consequently, a subset of biologically highly active molecules ($\text{ChEMBL}_{\text{active}}$) from the ChEMBL database was used as query for the in-house library in a virtual screening experiment. In order to compare the molecules a three dimensional pharmacophore descriptor, the Flexophore, was used [3]. This descriptor was already shown to describe the bio-chemical space in a meaningful way. By applying the Flexophore, a diverse set of 30,000 compounds (PocketLibrary) was selected from the virtually screened in-house compounds. All selected compounds had a high pharmacophore similarity to at least one of the molecules in $\text{ChEMBL}_{\text{active}}$. A detailed analysis of the matching molecules from $\text{ChEMBL}_{\text{active}}$ revealed that many of these molecules showed polypharmacology. Subsequently, a phenotypic screening was performed with the PocketLibrary. A high number of confirmed hits in this first screening demonstrated the validity of the approach described here.

1. Hu, Y.; Bajorath, J. Polypharmacology directed compound data mining: identification of promiscuous chemotypes with different activity profiles and comparison to approved drugs. *J. Chem. Inf. Model.* **2010**, *50*, 2112–2118.
2. Hopkins, A. L. Drug discovery: Predicting promiscuity. *Nature* **2009**, *462*, 167–168.
3. von Korff, M.; Freyss, J.; Sander, T. Flexophore, a new versatile 3D pharmacophore descriptor that considers molecular flexibility. *J. Chem. Inf. Model.* **2008**, *48*, 797–810.

A-3: Association Mining with Empirical Bayes and the Poisson-Gamma Model

Peter T. Corbett

Royal Society of Chemistry, Cambridge, UK

Large amounts of information are available in the chemical literature and text mining represents one way to gather this. The use of named entity recognition, name-to-structure conversion, chemical databases, controlled vocabularies and ontologies allows mentions of entities of interest to chemists to be detected and to be linked to chemical structures, databases and other forms of structured data. The statistical analysis of co-occurrences – entity-entity, entity-word and word-word – in large corpora, provides one way of making use of this information. In this presentation we will be demonstrating techniques for analysing co-occurrences in the corpus of more than 100,000 research papers published by the Royal Society of Chemistry in the 21st century, and visualising the results.

Very simple approaches to co-occurrence mining have problems. One approach is to score each co-occurrence and to sort the list of observed co-occurrences by score. Scoring by crude number of co-occurrences leads to a bias towards trivial co-occurrences between common terms. Scoring by some statistical significance measure, such as the result of a chi-squared test, leads to similar issues. Scoring by pointwise mutual information (PMI), a function of the ratio of the observed number of co-occurrences to the number expected under independence assumptions, gives high scores to strong, non-trivial associations, but also produces large amounts of noise – the associations found are often not reproducible, in that high-PMI associations found in one sample are often not found in another. *Ad hoc* combinations of these measures, involving thresholds on one measure and sorting by another, can be helpful but require careful tuning of parameters. What is needed is a measure that combines association strength and reliability into a single statistic. Using Empirical Bayes methods and the Poisson-gamma model, we have developed a measure of association strength – PGPMI - which estimates the likely PMI in a large population based on counts in a small sample. This measure finds reproducible, non-trivial co-occurrences.

Another analysis and visualisation technique is the use of Self-Organizing Maps. As well as providing a large-scale overview of a corpus, or the section of a corpus containing papers/paragraphs with a given term, these provide another useful way of detecting nontrivial co-occurrences. Given a term of interest, we can generate an SOM of paragraphs containing that term. We can then examine the distribution of paragraphs containing other terms across the SOM, and quantify how much that distribution varies from the SOM as a whole. This provides another useful score – one that allows the identification of terms strongly associated with some particular aspect of the original term.

A-4: Computer assisted identification of green tea metabolites in human urine

Lars Ridder[1,2], **J. J. J. van der Hooft**[1,3], **S. Verhoeven**[2], **R. C. H. de Vos**[3,4], **J. Vervoort**[1], **R. J. Bino**[1]

[1]Laboratory of Biochemistry, Wageningen University, The Netherlands, [2]Netherlands eScience Center, Amsterdam, The Netherlands, [3]Netherlands Metabolomics Centre, Leiden, The Netherlands, [4]Plant Research International, Wageningen, The Netherlands

The intestinal degradation and human biotransformation of drugs and small molecules present in food can give rise to a large variety of potentially bioactive metabolites in the human body. However, the absence of reference data for many of these components limits their identification, e.g., in plasma and urine. We present an *in silico* workflow to enhance the extraction of chemical information from metabolic profiling, based on liquid chromatography coupled to multistage accurate mass spectrometry (LC-MS^n), in these complex biological samples.

Our computational workflow reverses the conventional metabolite profiling process: First, database searches are combined with *in silico* metabolite predictions to generate a virtual library of both known and novel metabolites that could potentially be present in a sample. Subsequently, metabolite identification is assisted by matching the predicted structures to measured accurate masses, and by automatic annotation of MS^n fragment ions with substructures of these candidate metabolites. The substructure assignment algorithm makes use of the hierarchical information in the multistage fragmentation data. An associated scoring function is used to rank alternative chemical structures in order to direct the identification efforts towards the most likely candidates [1].

We evaluated the workflow by applying it to LC-MS^n datasets of green tea and of urine after tea consumption. For 85 known components in the green tea sample, correct structures were highly ranked (with a median of 3.5) among hundreds of alternative candidates automatically retrieved from the PubChem database and 24 novel putative assignments were made [2]. Subsequently, the 109 compounds found in green tea were systematically converted by *in silico* biotransformation rules defining possible modifications in the human gut and liver [3] before their excretion in urine. With this virtual library of 27245 potential metabolites, 74 previously identified components [4] were automatically annotated in the urine LC-MS^n datasets and 26 additional urinary metabolites originating from green tea consumption were putatively identified (Figure 1). 77% of the annotated metabolites were not present in the Pubchem database, indicating the importance of using *in silico* biotransformation to discover novel metabolites. The *in silico* metabolite network provides hypothetical pathways that link the annotated metabolites to known molecules in the sample. Further applications will be highlighted to illustrate the potential of our approach to support annotation of unknown metabolites in LC-MS^n metabolite profiling data.

A-5: A Novel Approach to De Novo Design using Reaction Networks

James E. A. Wallace[1,2], Beining Chen[2], Michael J. Bodkin[3], Val J. Gillet[1]

[1]Information School, University of Sheffield, Sheffield, UK, [2]Department of Chemistry, University of Sheffield, Sheffield, UK, [3]Eli Lilly, Erl Wood, UK

One of the main issues with existing *de novo* design methods is that many produce large numbers of hypothetical molecules that ultimately prove impossible to synthesise in real world conditions. To retain practicability, the transformations used to build molecules should be driven by knowledge of chemical reactions such as the reaction vector approach as previously developed at Sheffield [1]. At its core, a reaction vector is effectively a list of the molecular changes that occur within a chemical reaction. The vectors can be applied to previously unseen molecules containing similar structural features, thereby generating potentially novel molecules as a result. As the applied changes are based on real chemical reactions, the virtual products are more likely to be synthetically accessible. However, in its original form, the reaction vector is based on single step reactions, which can lead to difficulties when incorporated within a multi-step *de novo* design tool. For example, it is often the case that the intermediate structures in a reaction sequence do not resemble the ultimate endpoint in terms of molecule shape, dimensions or functionality. This is problematic for scoring functions for *de novo* design that assume a smooth progression from the starting molecule to potential products, where such discontinuities can lead to many useful routes being inappropriately scored, and therefore discarded. This is particularly an issue in reaction sequences that involve the use of protecting groups.

In this project, we propose a solution to the above issue by encoding the changes that occur in an entire reaction sequences into a single vector enabling the bypassing of the intermediates. In order to achieve this, it is first necessary to create a reaction network from a database of single step reactions [2]. A reaction network can be considered as a graph in which molecules form the nodes of the graph, linked by edges that represent reactions that transform one molecule to another. By connecting molecules in this manner, every path within the network will represent a reaction sequence. Reaction sequence vectors are then generated by tracing paths in the network. A reaction sequence vector can then be generated by recoding the difference between the end point (product) of the sequence and the start molecule. A reaction sequence vector can then be applied to an unseen molecule, in a similar manner to reaction vectors, to generate a product molecule in a single step, thereby passing over potentially poor scoring intermediates. The reaction sequence vector system and associated structure generation methods have been built into an open source *de novo* tool using the KNIME data mining framework [3].

1. Patel, H.; Bodkin, M. J.; Chen, B.; Gillet, V. J. Knowledge-Based Approach to de Novo Design Using Reaction Vectors. *J. Chem. Inf. Model.* **2009,** *49*, 1163–1184.
2. Soh, S.; Wei, Y.; Kowalczyk, B.; Gothard, C. M.; Baytekin, B.; Gothard, N.; Grzybowski, B. A. Estimating Chemical Reactivity and Cross-influence from Collective Chemical Knowledge. *Chem. Sci.* **2012,** *3*, 1497–1502.
3. Berthold, M.; Cebron, N.; Dill, F.; Gabriel, T.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: The Konstanz Information Miner. In *Data Analysis, Machine Learning and Applications*, Preisach, C.; Burkhardt, H.; Schmidt-Thieme, L.; Decker, R. Eds. Springer Berlin Heidelberg: 2008; pp 319–326.

A-6: Improving substructure screens

Andrew Dalke

Dalke Scientific, Trollhättan, Sweden

This presentation covers my ongoing research in improving substructure screening methods. This topic is not widely explored in the literature, in part because it's mostly seen as an performance issue rather than a scientifically interesting problem, and in part because there is no good benchmark data. I will attempt to convince people that there are interesting and unexplored areas, and present an open challenge for others to work on improved screening methods.

In some respect, substructure search is a solved problem in small molecule chemistry. The core algorithm, subgraph isomorphism, is NP-complete but N is usually small and an efficient VF2 algorithm can easily process 10-100 thousand records per second.

Fast searches enable new ways to work with data. Similarity searches of multi-million compound data sets like ChEMBL take less than 0.1 seconds, which is fast enough that humans consider it to be "instant". Results this fast can be displayed dynamically during structure input instead of waiting for a "submit" action, or done incidentally, to provide extra helper information in a display.

Brute-force substructure isn't fast enough. Instead, nearly every search system precomputes target characteristics that can reject obvious mismatches, then use an isomorphism test to identify the true positives. A good screen can reduce search time by several orders of magnitude, and easily achieve sub-second average performance against a 1 million record database, though with a wide variance that makes it difficult to use interactively.

Structure screens generally fall into two camps. The MACCS-166 keys, CACTVS/PubChem fingerprints, and OrChem fingerprints use specific substructure patterns based on human observations of which queries have poor performance against a specific database. The Daylight fingerprints, Open Babel's FP2, Bingo, RDKit's 'Pattern' fingerprints, and ABCD use subgraph enumeration and a hash function to define a more target independent screening method.

One of the few published screening comparisons [1] shows that the 166-bit MACCS keys are clearly worse than the CACTVS 881-bit keys. CACTVS is sometimes noticeably better than the 1021-bit FP2 keys and sometimes worse but overall they appear to be equally matched. However, CACTVS's smaller fingerprints imply that its bits are more selective and thus more RAM friendly. ABCD's screening ability is better, though it uses a sparse count-based inverted index which is hard to retrofit in an existing binary fingerprint-based cheminformatics system.

Screening efficiency is not the only concern, as otherwise the best solution would combine multiple screens. Additional filters have reduced sensitivity due to correlations with previous filters. At some point it's a better use of time to do the substructure search than to continue to screen.

It's reasonable to believe that judiciously chosen substructure keys are more selective than hash-based ones. The CACTVS/FP2 comparison supports that belief. But pattern-based screens involves a lot of human involvement, which isn't worthwhile when hash keys may be almost as good.

I have developed a method to automate substructure key selection. It enumerates canonical subgraph patterns for sample queries and records, then heuristically identifies those patterns which best improves the worst-case time. It is similar to decision tree methods, but since the results must fit into a standard fingerprint there's the unusual condition that all levels ask the same question.

Preliminary tests show that the first few patterns found are a useful augment to RDKit's substructure screen fingerprint, though not a replacement. I believe there are better algorithms, but the problem is surprisingly complex and there's very little literature on how to improve substructure performance, I want to encourage others to work in this area. The biggest problem is the almost complete lack of real-world queries for evaluating improvements. Companies have the data but do not want to chance revealing proprietary structures. Commercial database vendors have little reason to share their knowledge with competitors. Commercial database providers have non-disclosure policies and signed confidentiality agreements because they want companies to trust them with proprietary data. Even some non-commercial database providers like PubChem and EMBL have non-disclosure policies that prevent outsiders from access to use query data.

BindingDB, however, has contributed past query data for exact search, similarity, and substructure search. I have started the Structure Query Collection [2] with the BindingDB dataset and a few SMARTS data sets, and I am searching for additional data sets. I will discuss also some of the legal and ethical considerations which have come up during the process of releasing this data.

1. Dimitris, A. K.; Lobanov, V. S.; Shemanarev, M.; Rassokhin, D. N.; Izrailev, S.; Jaeger, E. P.; Alex, S.; Farnum, M. Efficient Substructure Searching of Large Chemical Libraries: The ABCD Chemical Cartridge *J. Chem. Inf. Model.* **2011**, *51*, 3113–3130.
2. Structure Query Collection - https://bitbucket.org/dalke/sqc

A-7: Generating Small Molecule Conformations from Structural Data

Patrick McCabe[1], **Oliver Korb**[1], **Jason Cole**[1], **Robin Taylor**[2]

[1]Cambridge Crystallographic Data Centre, Cambridge, UK ,[2]Taylor Cheminformatics Software, Rickmansworth, UK

Conformer generation is important in computer aided drug design and discovery and many programs have been developed that attempt to reproduce biologically relevant conformations from initial basic chemical models. We present the scientific methods and results, based on a test set of 3291 Cambridge Structural Database (CSD) derived structures, for a new knowledge based conformer generator based on CSD data.

Given an input molecule, which is optionally minimised, pre-determined CSD rotamer and CSD ring distributions are incrementally applied to a fragmented view of the molecule and the generated conformers assigned scores based on sample relative frequencies (i.e. approximate probabilities) of the geometric parameters assigned. Conformers are monitored for clashes during the incremental build up procedure allowing early rejection of clashing conformations. In the final stage a diverse subset of conformations is derived by clustering.

A-8: Physicochemical property trends in allosteric and orthosteric binders and their predicted presence in chemical databases

Gerard J. P. van Westen, Simone Trubian, John P. Overington

ChEMBL / Chemogenomics Group, European Molecular Biology Laboratory European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom

Introduction

Generating drug-like lead and candidate molecules against a specific molecular target remains a major challenge in drug discovery. These challenges partially fall into two basic themes – a) the diversity and size of the set of compounds used in the initial screen, and b) the physicochemical properties of the binding site of the target, which may contain obligate features that are incompatible with binding molecules with drug-like properties [1]. One strategy to consider in some cases is the use of allosteric modulators. Previous work shows that for allosteric modulators trends in chemistry, bioactivity, and targets hit can be distinguished and here this work is followed up [2,3].

In parallel, chemogenomic (and proteochemometric) methods that were predicted to link targets from a ligands (chemical) perspective have not always lived up to their expected performance [4]. A strategy to improve the performance of these computational approaches is considering their binding mode when known (be it allosteric or orthosteric).

Aims

The aim of this work was to characterize the properties of allosteric and orthosteric binders both as classes of molecules, but also on a per target class basis. Separation of known ligands for a target class based on competitive and non-competitive mechanisms should increase performance of chemogenomic methods.

Methods

Physicochemical and structural features of both a large set of allosteric ligands and orthosteric ligands, extracted from the ChEMBL database of bioactive molecules, were analyzed and compared to the background of ChEMBL. Complementary to this ligand based approach the targets represented in both classes were analyzed and bioactivity was characterized in all sets.

Results

From our dataset a series of classifier models were created that could distinguish between the three classes (allosteric, orthosteric or unknown, accuracy 80%, MCC 0.50). Furthermore, using the target hierarchy of the ChEMBL database target class dependent models were constructed (accuracy 86%, MCC 0.63). Subsequently a number of chemical databases were screened (DrugBank, PDB, eMolecules, Maybridge, etc). From these results we could classify the allosteric likeness and orthosteric likeness of these databases.

1. Overington, J. P.; Al-Lazikani, B.; Hopkins, A. L. How many drug targets are there? *Nat. Rev. Drug Discov*. **2006**, *5*, 993–996.
2. Lindsley, J. E.; Rutter, J. Whence cometh the allosterome? *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 10533–10535.
3. van Westen, G. J. P.; A Gaulton, and J.P. Overington Chemical, Target, and Bioactive Properties of Allosteric Modulation. *PLoS Comp. Biol.* **2014** *Revised version submitted*
4. van Westen, G. J. P.; Overington, J. P. A ligand's-eye view of protein similarity. *Nat. Methods* **2013**, *10,* 116–117.

A-9: Multi-dimensional activity cliff analysis

Mark Mackey, Tim Cheeseright, Rae Lawrence, Martin Slater

Cresset, Cambridge, UK

During lead optimization the stepwise progression of compound activity is often disrupted by compounds that cause a disproportionately large (positive or negative) change in the biological response. These activity cliffs have long been recognized as an important source of information about the requirements of the protein for the series of interest [1]. However detection and analysis of these critical regions of SAR has been generally limited to the use of 2D similarity methods such as fingerprints.

Here, we present a method for expanding the dimensionality of activity cliff detection to include the 3D shape and electrostatic character of the ligands. In contrast to fingerprint similarity methods, accurate 3D similarity methods treat bioisosteres correctly which allows the identification of cliffs which the 2D methods fail to find.

The use of 3D similarity also often provides intuitive explanation of the fundamental reasons for these cliffs, leading to a much better understanding of why the structural change(s) led to such an unexpected change in binding. In particular, examining the differences in electrostatic potential between a pair of molecules can illuminate why structurally small changes (such as different aromatic substitution patterns) can lead to significant activity cliffs. Doing this requires accurate representations of molecular electrostatic potentials (Figure 1).

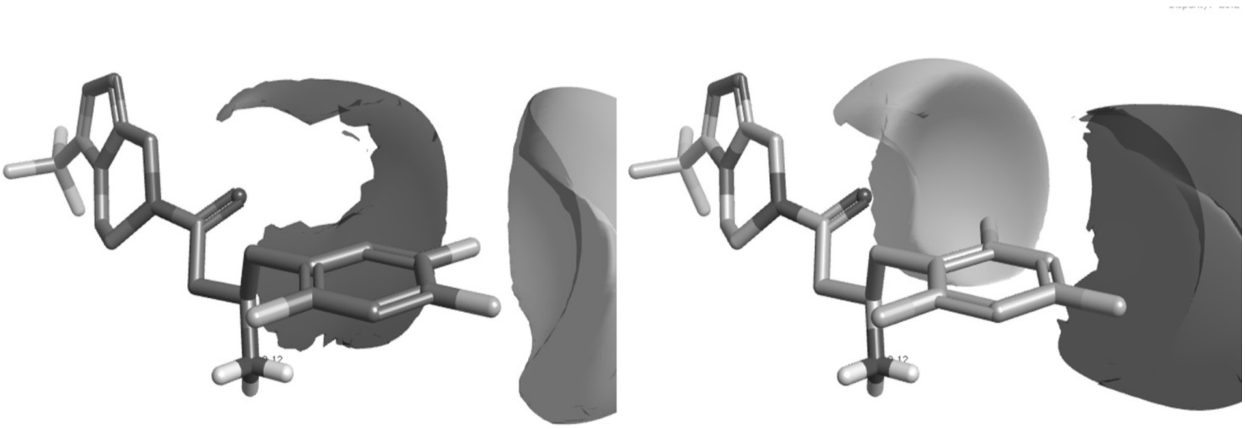


Figure 1: Difference maps showing the effect of moving a fluorine atom on electrostatic potential for two DPP IV inhibitors.

3D alignment and similarity methods can also be used to enable the rationalization of activity cliffs that are detected using 2D similarity methods. Small structural changes giving rise to large conformational differences are an obvious example. The combination of 2D and 3D activity cliff detection can provide powerful insights into the SAR around a lead series.

The detection of activity cliffs for the primary activity end point is a valuable addition to the arsenal of drug discovery scientists. However, modern drug discovery rarely proceeds through the optimization of a single end point. More often project teams are tasked with optimizing the primary activity while minimizing the effect on a secondary, selectivity target or on a critical ADMET parameter. We have therefore studied the application of the 3D activity cliff analysis to multiple activity endpoints. These “selectivity cliffs” highlight where molecular changes have a large effect on the activity against one target but not another. Visualization of this data is a challenging task. If there are 500 molecules in the dataset then there around 250,000 data points to be analyzed for two activity end points. We will discuss the challenges and present some novel visualization techniques to deal with this data (Figure 2).

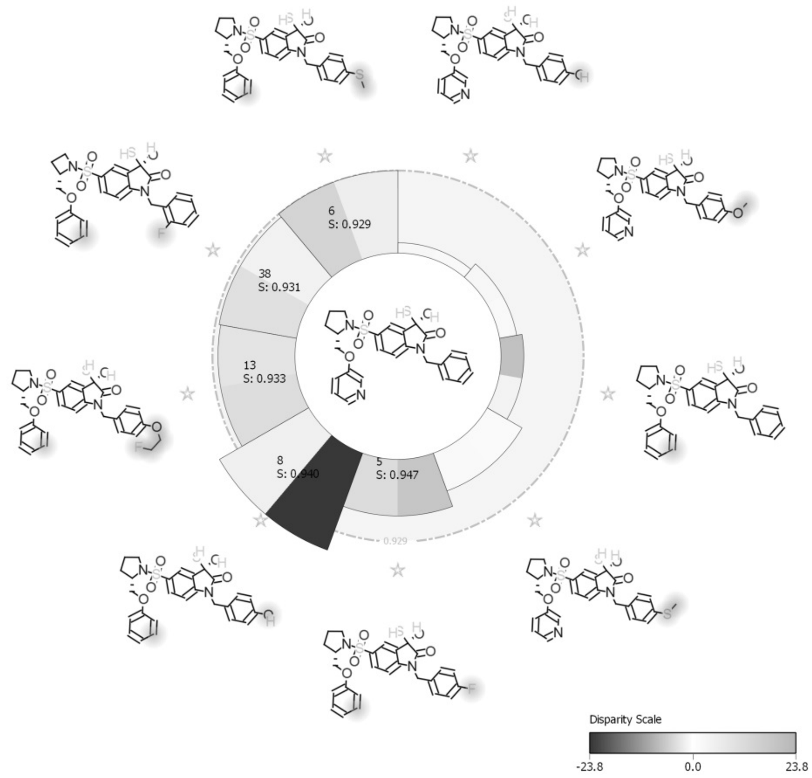


Figure 2: Visualizing selectivity cliffs.

1. Stumpfe , D.; Hu , Y.; Dimova , D.; Bajorath , J. Recent Progress in Understanding Activity Cliffs and Their Utility in Medicinal Chemistry. *J. Med. Chem.* **2014**, *57*, 18–28.

Session B: Structure-Based Drug Design and Virtual Screening

B-1: Nonadditivity in SAR analysis: It's not a bug – it's a feature!

Christian Kramer, Julian E. Fuchs, Klaus R. Liedl

Department of General, Inorganic and Theoretical Chemistry, University of Innsbruck, 6020 Innsbruck, Austria

Nonadditivity is probably perceived as most annoying feature of ligand datasets: Most scoring functions and QSAR models, especially Free Wilson analysis depend on additive contributions of fragments. If there is a substantial degree of nonadditivity in datasets, many types of data analysis and predictions cannot be expected to work any more. In reality it turns out that standard compound libraries indeed show a substantial degree of nonadditivity on different targets [1]. A lot of pioneering work in structural nonadditivity analysis has been carried out by the Klebe group. They found that nonadditivity can be structurally rationalized in the context of differing water networks and residual mobility [2,3].

We have used a big data chemoinformatics approach to analyze large ligand datasets for the nonadditivity content: Using an automatic double substitution cycle extraction based on the Matched Molecular Pair formalism, we are able to quickly profile large ligand datasets for many different targets. Structural analysis of the nonadditivity cycles reveals that all cases of strong nonadditivity can be traced back to geometric changes in ligand or protein. Each single case yields important insights for structural SAR analysis and can directly be translated into system specific physical restraints that govern binding.

In this talk we want to show how nonadditivity can be turned into something useful for SAR analysis. The systematics of nonadditivity analysis will be presented with special emphasis on the consideration of experimental uncertainty. If ignored, experimental uncertainty [4,5] will lead to false-positive identification of nonadditive features. Using a number of structural examples, we will illustrate which structural features lead to nonadditivity and what we can learn out of these.

1. Patel, Y.; Gillet, V. J.; Howe, T.; Pastor, J.; Oyarzabal, J.; Willett, P. Assessment of Additive/Nonadditive Effects in Structure–Activity Relationships: Implications for Iterative Drug Design. *J. Med. Chem.* **2008**, *51*, 7552–7562.
2. Baum, B.; Muley, L.; Smolinski, M.; Heine, A.; Hangauer, D.; Klebe, G. Non-additivity of Functional Group Contributions in Protein–Ligand Binding: A Comprehensive Study by Crystallography and Isothermal Titration Calorimetry. *J. Mol. Biol.* **2010**, *397*, 1042–1054.
3. Biela, A.; Betz, M.; Heine, A.; Klebe, G. Water Makes the Difference: Rearrangement of Water Solvation Layer Triggers Non-additivity of Functional Group Contributions in Protein–Ligand Binding. *ChemMedChem* **2012**, *7*, 1423–1434.
4. Kramer, C.; Kalliokoski, T.; Gedeck, P.; Vulpetti, A. The Experimental Uncertainty of Heterogeneous Public Ki Data. *J. Med. Chem.* **2012**, *55*, 5165–5173.
5. Kalliokoski, T.; Kramer, C.; Vulpetti, A.; Gedeck, P. Comparability of Mixed IC50 Data – A Statistical Analysis. *PLoS ONE* **2013**, *8*, e61007.

B-2: Ligand structures in the Protein Data Bank – can we trust them?

David Sehnal[1,2,3], **Radka Svobodová Vařeková**[1,2], **Crina-Maria Ionescu**[1], **Stanislav Geidl**[1,2], **Lukáš Pravda**[1,2], **Deepti Jaiswal**[1], **Michaela Wimmerová**[1,2], **Jaroslav Koča**[1,2]

[1]CEITEC - Central European Institute of Technology, Masaryk University, Brno, Czech Republic, [2]National Centre for Biomolecular Research, Masaryk University, Brno, Czech Republic, [3]Faculty of Informatics, Masaryk University, Brno, Czech Republic

Validation of biomolecular structures arose as a major issue in the structural biology community when it became apparent that some published structures contained serious errors. An essential step in the validation process is checking the ligand structure, because ligands play a key role in protein function, and also because they are the main source of errors in structures. Nonetheless, ligand validation is a very challenging task, because of the high diversity and nontriviality of their structure, and the general lack of information about correct structures. Therefore, software tools focused on ligand validation were developed only recently (PDB-care [1], ValLigURL [2]), or are still under development (e.g., MotiveValidator [3]).

These tools are able to validate one or more structures (even thousands of structures), but they are not able to do a wider analysis of all the ligands.

In our work, we focus on this challenge. Specifically, we would like to answer this question: How reliable are ligand structures in the PDB? We employed advanced superimposition approaches [4] and and the web application MotiveValidator to collect and validate ligand structures from the PDB. Subsequently, we put together the new database MotiveValidatorDB, which contains statistical data about the validation of all ligands in the PDB (i.e., number of structures with correct/wrong chirality, missing atoms or rings etc.). The user may also inspect the details of the validation results for each ligand or chemical component (e.g., all NAG instances in the PDB).

1. Lütteke, T.; von der Lieth, C.-W. pdb-care (PDB carbohydrate residue check): a program to support annotation of complex carbohydrate structures in PDB files. *BMC Bioinf.* **2004**, *5*, 69.
2. Kleywegt, G. J.; Harris, M. R. ValLigURL: a server for ligand-structure comparison and validation. *Acta Crystallogr. D. Biol. Crystallogr.* **2007**, *63*, 935–938.
3. MotiveValidator. http://ncbr.muni.cz/MotiveValidator (accessed Feb 1, 2014)
4. Sehnal, D.; Svobodová Vařeková, R.; Huber, H. J.; Geidl, S.; Ionescu, C.-M.; Wimmerová, M.; Koča, J. SiteBinder: an improved approach for comparing multiple protein structural motifs. *J. Chem. Inf. Model.* **2012**, *52*, 343-359.

B-3: An exploration of the 3D chemical space has highlighted a specific shape profile for the compounds intended to inhibit protein-protein interactions.

Mélaine A. Kuenemann[1,2], **Laura M. L. Bourbon**[1,2], **Céline M. Labbé**[1,2] , **Bruno O. Villoutreix**[1,2], **Olivier Sperandio**[1,2]

[1]University Paris Diderot, Paris, France, [2]Inserm UMR-S 973, Paris, France

The vital role of Protein-Protein Interactions (PPI), articulated through numerous cellular processes, makes them the subject of a growing number drug discovery projects. Yet, the specific properties of PPI (often described as flat, large and hydrophobic) require a dramatic paradigm shift in our way to design the small compounds meant to modulate them for therapeutic interventions. To this end, successful inhibitors of PPI targets (iPPI) may be used to discover what singular properties make this type of inhibitors capable of binding such intricate surfaces. Among the properties from which lessons could be learnt, the 3D characteristics of iPPI have been pinpointed as essential. Understanding the putative shape profile of iPPI could help the design of a new generation of inhibitors with improved ligand efficiencies.

In an attempt to identify such putative 3D specificities, we have collected the bioactive conformations of 60 orthosteric iPPI and compared them to those of 1623 inhibitors of regular targets from different databases (2P2I, PDBind, PDB). The selection of discriminative 3D molecular descriptors has highlighted new characteristics of iPPI. Because the heavier and more hydrophobic character of iPPI could have impeded the identification of genuine specificities, we have imposed that none of the identified descriptors could correlate with the hydrophobicity or the size of the compound. The newly identified properties were further confirmed as specific to iPPI using the data of much larger datasets including our iPPI-DB, eDrug and a representative subset of the bindingDB. Most noticeably, the essential property revealed by this study illustrates how iPPI manage to bind to hydrophobic patches. Interestingly, the absence of correlation of such property with the hydrophobicity and the size of the compounds, that can be a liability for further developments, opens new ways to design potent iPPI with a better balance for some of the pharmacokinetic features.

B-4: The Motility of Water Molecules – A Statistical Evaluation of Water Molecules Based on Electron Density

Eva Nittinger[1], **Nadine Schneider**[1], **Gudrun Lange**[2], **Matthias Rarey**[1]

[1]University of Hamburg, Center for Bioinformatics, Bundesstraße 43, 20146 Hamburg, Germany, [2]Bayer CropScience AG, Industriepark Hoechst, G836, 65926 Frankfurt am Main, Germany

Water molecules are essential for biological processes in many different ways. They serve as media for biological macromolecules, are directly involved in enzymatic reactions such as hydrolysis or stabilize different biological complexes by mediating protein-protein or protein-ligand interactions.

Several studies exist that analyze water molecules in crystal complexes with respect to their local environment and structural properties [1,2]. Moreover, sophisticated methods like WaterMap [3] and Szmap [4] aim to predict strongly bound and replaceable water molecules independent of crystallographically identified water molecules. None of these predictive methods use the experimentally available electron density to characterize water molecules in crystal structures and to distinguish between strongly bound water molecules and those which may be too flexible to be resolved in the electron density map.

First results of our statistical analysis have shown the necessity of differentiating water molecules into actually observed ones in the electron density or those placed by automatic structural refinement tools [5] without underlying electron density. Neglecting this separation, noise would unintentionally be integrated into any result based on the statistical analysis. Different quality criteria already exist for evaluating electron density e.g. real space R-factor or real space correlation coefficient [6]. However, all existing measurements show their drawbacks on single water molecules. Due to this, we have developed a new estimate for the distribution of electron density around an individual atom taking the surrounding density values in its van-der-Waals radius into account. Based on this value we performed a statistical analysis, wherein every single water molecule was analyzed for a diverse set of structural properties, e.g. the number of hydrogen bonds, the hydrogen bond partners and potential preferences for specific hydrogen bond partners such as backbone or side chain amino acids.

Using these structural properties we are able to distinguish between confined water molecules in biological complexes and artificial ones without electron density placed by automatic refinement programs. These results will gain further significance in structure-based drug design methods such as docking, wherein especially docking poses would benefit from correctly placed water molecules.

1. Park, S.; Saven, J. G. Statistical and Molecular Dynamics Studies of Buried Waters in Globular Proteins. *Proteins* **2005**, *60*, 450–463.
2. Renthal, R. Buried Water Molecules in Helical Transmembrane Proteins. *Protein Sci.* **2008**, *17*, 293–298.
3. Young, T.; Abel, R.; Kim, B.; Berne, B. J.; Friesner, R. A. Motifs for Molecular Recognition Exploiting Hydrophobic Enclosure in Protein–ligand Binding. *Proc. Natl. Acad. Sci.* **2007**, *104*, 808–813.
4. Grant, J. A.; Pickup, B. T.; Nicholls, A. A Smooth Permittivity Function for Poisson-Boltzmann Solvation Methods. *J. Comput. Chem.* **2001**, *22*, 608–640.
5. Afonine, P. V; Grosse-Kunstleve, R. W.; Echols, N.; Headd, J. J.; Moriarty, N. W.; Mustyakimov, M.; Terwilliger, T. C.; Urzhumtsev, A.; Zwart, P. H.; Adams, P. D. Towards Automated Crystallographic Structure Refinement with Phenix.refine. *Acta Crystallogr. D. Biol. Crystallogr.* **2012**, *68*, 352–367.
6. Warren, G. L.; Do, T. D.; Kelley, B. P.; Nicholls, A.; Warren, S. D. Essential Considerations for Using Protein–ligand Structures in Drug Discovery. *Drug Discov. Today* **2012**, *17*, 1270–1281.

B-5: How to treat waters in virtual screening? A case study on the Adenosine A_{2A} Receptor

<u>Eelke B. Lenselink</u>[1], Thijs Beuming[2], Woody Sherman[2], Herman W. T. van Vlijmen[1], Adriaan P. IJzerman[1]

[1]Leiden Academic Centre for Drug Research, Leiden, Netherlands [2]Schrödinger Inc., New York, US

One of the major challenges in virtual screening (VS) is the incorporation of explicit water molecules during docking. As waters are frequently involved in ligand-protein interactions, it is clear that docking results could benefit from a correct prediction of bridging waters. It is however not straightforward to determine which waters should be selected, and in which orientations, especially when waters are not present in the crystal structure. In addition, since water bridges are often non–conserved, it is likely necessary to use an ensemble of water containing structures. We have investigated these issues here in the context of the Adenosine A_{2A} receptor, where water molecules have been shown to be vital for VS 1[,2].

To generate appropriate water containing A_{2A} models, we followed two approaches. First, different X-ray waters and their (multiple) conformations were generated using the interactive optimizer (PROPKA) in Maestro, using a recently published high-resolution structure of the A_{2A} receptor. However, due to the relatively low resolution of typical GPCR structures, water molecules are often not even present. Therefore, we also tested WaterMap as a way to predict water positions and orientations in the absence of such information from crystallography. The effect of these water orientations on VS enrichment was assessed using a set of high affinity antagonists and an appropriate set of decoys. In many cases enrichments were improved significantly using water molecule orientations derived from MD simulations, suggesting it is important to evaluate multiple water positions and orientations during docking. In addition, waters placed using the MD based approach proved to be as accurate as those derived from crystallography. Finally, we applied a machine learning method to idenitfy a small ensemble of water molecules with higher enrichment than the single best water.

In general, this water optimization strategy could be applied to any target where waters mediate protein-ligand interactions.

1. Roumen, L.; Sanders, M. P.; Vroling, B.; de Esch; I. J. P.; de Vlieg, J.; Leurs, R.; de Graaf, C. In Silico Veritas: The Pitfalls and Challenges of Predicting GPCR-Ligand Interactions. *Pharmaceuticals* **2011**, *4*, 1196–1215.
2. Katritch, V.; Jaakola, V.-P.; Lane, J. R.; Lin, J.; IJzerman, A. P.; Yeager, M.; Abagyan, R. Structure-based discovery of novel chemotypes for adenosine A(2A) receptor antagonists. *Journal of Medicinal Chemistry* **2010**, *53*, 1799–1809.

B-6: WaterFLAP: Fast water prediction, scoring, and docking using GRID Molecular Interaction Fields

Simon Cross[1], Massimo Baroni[1], Gabriele Cruciani[2]

[1]Molecular Discovery Ltd, UK, [2]Department of Chemistry, University of Perugia, Italy

For 30 years GRID Molecular Interaction fields (MIFs) [1] have been at the forefront of structure-based drug design; the neuraminidase inhibitor Zanamivir (Relenza^{TM}) was the first marketed drug that was identified directly using structure-based design [2]. Since the early focus on structure-based characterisation of targets using GRID, ligand-based approaches have also been applied and developed to aid other critical aspects of drug discovery such as the optimisation of pharmacokinetic properties [3] and metabolic stability [4]; the general applicability of GRID illustrates the importance of characterising molecules through their molecular interactions as opposed to their two-dimensional representation. The FLAP approach [5] combines this GRID MIF characterisation along with a pharmacophoric fingerprint representation that can be used to align molecules using a *common reference framework* that includes small molecules and macromolecules and can therefore be used for ligand-based and structure-based approaches, with applications including virtual screening [6–8], pharmacophore elucidation [9,10], 3D-QSAR, and docking.

In the first publication of the GRID method [1], the OH2 water probe was used to highlight the importance of structural water in dihydrofolate reductase, and in recent years the role of water molecules in ligand binding has become the subject of an increasing number of publications.

In this work we present a significant update to the original approach, called WaterFLAP, which automatically uses GRID OH2 Molecular Interaction Fields to predict binding site water molecule positions, and subsequently scores the waters according to their 'happiness', using their predicted enthalpy of interaction, hydrophobicity, and a newly developed entropic score. The new approach is described along with examples of use for design, and integration within the FLAP software for improved ligand-docking.

1. Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28*, 849–857.
2. von Itzstein, M.; Wu, W. Y.; Kok, G. B.; Pegg, M. S.; Dyason, J. C.; Jin, B.; van Phan, T.; Smythe, M. L.; White, H. F.; Oliver, S. W.; et al. Rational design of potent sialidase-based inhibitors of influenza virus replication. *Nature* **1993**, *363*, 418–423.
3. Crivori, P.; Cruciani, G.; Carrupt, P. A.; Testa, B. Predicting blood-brain barrier permeation from three-dimensional molecular structure. *J. Med. Chem.* **2000**, *43*, 2204–2216.

4. Cruciani, G.; Carosati, E.; De Boeck, B.; Ethirajulu, K.; Mackie, C.; Howe, T.; Vianello, R. MetaSite: understanding metabolism in human cytochromes from the perspectiva of the chemist. *J. Med. Chem.* **2005**, *48*, 6970–6979.
5. Baroni, M.; Cruciani, G.; Sciabola, S.; Perruccio, F.; Mason, J. S. A common reference framework for analyzing/comparing proteins and ligands. Fingerprints for Ligands and Proteins (FLAP): theory and application. *J. Chem. Inf. Model.* **2007**, *47*, 279–294.
6. Carosati, E.; Mannhold, R.; Wahl, P.; Hansen, J. B.; Fremming, T.; Zamora, I.; Cianchetta, G.; Baroni, M. Virtual screening for novel openers of pancreatic K(ATP) channels. *J. Med. Chem.* **2007**, *50*, 2117–2126.
7. Cross, S.; Baroni, M.; Carosati, E.; Benedetti, P.; Clementi, S. FLAP: GRID Molecular Interaction Fields in Virtual Screening. Validation using the DUD Data Set. *J. Chem. Inf. Model.* **2010**, *50*, 1442–1450.
8. Sirci, F.; Goracci, L.; Rodriguez, D.; van Muijlwijk-Koezen, J.; Gutierrez-de-Teran, H.; Mannhold, R. Ligand-, structure- and pharmacophore-based molecular fingerprints; a case study on adenosine A1, A2A, A2B, and A3 receptor antagonists. *J. Comput. Aided. Mol. Des.* **2012**.
9. Cross, S.; Baroni, M.; Goracci, L.; Cruciani, G. GRID-Based Three-Dimensional Pharmacophores I: FLAPpharm, a Novel Approach for Pharmacophore Elucidation. *J. Chem. Inf. Model.* **2012**, *52*, 2587–2598.
10. Cross, S.; Ortuso, F.; Baroni, M.; Costa, G.; Distinto, S.; Moraca, F.; Alcaro, S.; Cruciani, G. GRID-Based Three-Dimensional Pharmacophores II: PharmBench, a Benchmark Data Set for Evaluating Pharmacophore Elucidation Methods. *J. Chem. Inf. Model.* **2012**, *52*, 2599–2608.

B-7: Computational Methods Enabling Fragment Based Drug Discovery

Stephen D. Pickett, **Ian Wall, Martin Saunders, Ceara Rea**

GlaxoSmithKline, Stevenage, UK

Fragment based drug discovery has evolved over the past decade into an important method for lead discovery. However, successful implementation requires the integration, presentation and utilisation of data from a variety of sources including protein ligand structural information, biochemical and biophysical screening data, physicochemical properties and complementary high throughput screening data.

We describe the development of the FBDD computational platform at GSK. The methods and tools presented will be illustrated by case studies from in-house fragment programs and address key issues encountered within the course of a fragment program. The development of a specific fragment screening set is described, including a novel approach for assessing the likely chemical tractability of any fragment hit. A common method for following up fragment hits is the use of similarity searching to identify readily available analogues (SARbyCatalog). We describe the FindAnalogues method that combines multiple descriptors and similarity searching methods and has proved successful in both SARbyCatalog and remining high throughput screening data.

FBDD programs can generate a large number of protein ligand structures and ready access to these structures at the desktop, coupled with computational analyses such as GRID maps greatly facilitates interactive discussion within teams. We describe tools developed to integrate screening, physicochemical, DMPK and computational data that can be presented alongside protein ligand structural information. These tools have enabled a comparison of factors that influence success in different biophysical screening methods and subsequent structure elucidation. Results of the analysis are presented that provide guidelines for the design of fragment sets.

The application of these methods is described in the context of a successful fragment based drug discovery program and the presentation includes a successful application of structure based drug design in fragment optimization.

B-8: Application of linear response MM-PBSA and QM/MM-GBSA rescoring for postprocessing of protein-ligand docking poses

Kanin Wichapong[1,2] , Gerry A. F. Nicolaes[2], Wolfgang Sippl[1]

[1]Department of Pharmaceutical Chemistry, Martin-Luther University Halle-Wittenberg, Halle(Saale), Germany, [2]The Cardiovascular Research Institute Maastricht (CARIM), Department of Biochemistry, Maastricht University, Maastricht, The Netherlands

Accurate prediction of the binding strength of small molecules to a specific target protein is a nontrivial task in structure-based drug design. Molecular docking and subsequent scoring of docked poses are commonly used to rapidly screen compounds from large databases. However, it is frequently found that docking scores show no correlation to experimentally determined binding affinities or cannot be used to discriminate between active and inactive compounds. Therefore, more accurate approaches, such as binding free energy calculation, are nowadays widely used to rescore docking results. Traditionally, methods for the calculation of binding free energies of protein-ligand complexes, MM-PB(GB)SA or QM/MM-GBSA, are calculated by processing an ensemble of snapshots derived from equilibrated molecular dynamics simulations. However, in our previous work [1], we have shown that even a single protein-ligand complex obtained after a short energy minimization can be used for this purpose. Moreover, we have successfully generated a linear response MM-PBSA model for prediction of binding free energies of Wee1 kinase inhibitors. The derived model showed a high correlation between calculated binding free energies and biological binding affinities ($r^2 = 0.67$) and performed very well in the discrimination between active inhibitors and decoys. We subsequently extended the scope of our study by application of our methodology to the related kinase Myt1, for which a few inhibitors have been reported [2], such as to identify novel and potent Myt1 kinase inhibitors. In a recent study [3], we have applied several rescoring approaches (MM-PBSA, MM-GBSA and QM/MM-GBSA) to rescore docking results and we have discovered that the QM/MM-GBSA rescoring method performed better than the other approaches to distinguish active compounds from inactive ones. Finally, we have applied QM/MM-GBSA rescoring to estimate binding free energies of compounds derived from virtual screening of Myt1 kinase. By use of our novel QM/MM-GBSA rescoring protocol we were able to identify novel and potent Myt1 kinase inhibitors with corresponding K_i values from low micromolar level to as low as 8.1 nM. The developed protocol is a fast and simple approach, thus it can be used to filter and prioritize compounds during virtual screening experiments.

1. Wichapong, K.; Lawson, M.; Pianwanit, S.; Kokpol, S.; Sippl, W. Postprocessing of protein-ligand docking poses using linear response MM-PB/SA: application to Wee1 kinase inhibitors. *J. Chem. Inf. Model.* **2010**, *50*, 1574–1588.
2. Rohe, A.; Göllner, C.; Wichapong, K.; Erdmann, F.; Al-Mazaideh, G. M.; Sippl, W.; Schmidt, M. Evaluation of potential Myt1 kinase inhibitors by TR-FRET based binding assay. *Eur. J. Med. Chem.* **2013**, *61*, 41–48.

3. Wichapong, K.; Rohe, A.; Platzer, C.; Slynko, I.; Erdmann, F.; Schmidt, M.; Sippl, W. Application of Docking and QM/MM-GBSA Rescoring to Screen for Novel Myt1 Kinase Inhibitors. *J. Chem. Inf. Model.* **2014**, DOI: 10.1021/ci4007326.

B-9: Large Scale FEP on Congeneric Ligand Series – Have Practical Free Energy Calculations arrived at Last?

Thomas B. Steinbrecher, Teng Lin, Lingle Wang, Goran Krilov, Robert Abel, Woody Sherman, Richard Friesner

Schrodinger GmbH, Dynamostr. 13, 68165 Mannheim, Germany

The holy grail of computational structure based ligand design has long been the accurate prediction of binding free energies for novel compounds. Molecular Dynamics based free energy calculations (FEC) have been proposed as one of the most suitable methods to reach this goal, which would significantly impact the modern drug design process. However, despite many successful studies, FEC have for more than 20 years failed to fulfill this promise. Possible reasons for this include force field deficiencies, insufficient sampling and difficulties in assessing the quality of simulation results. One of the main obstacles in addressing these issues has been the lack of large scale validation studies on diverse series of ligands, due to the lack of computational resources and the time consuming process of simulation setup and analysis.

Here, we will present results from FEC conducted on several protein-ligand systems of pharmaceutical interest. Covering more than 10 targets and more than 200 compounds, the results offer more than an order of magnitude more data than typical FEC studies and allow statistically valid conclusion about their efficacy. We show that relative binding free energies can be calculated with good accuracy in most cases, typically with R^2 values in the range of 0.5-0.8 and mean unsigned errors (MUE) of less than 1 kcal/mol on average when comparing to experimental data. We show that FEC consistently outperforms other binding energy estimation methods. Statistical error estimates from individual calculations are much smaller than observed deviations from experimental results, but improved error estimates can be obtained from constructing redundant graphs of ligand transformations.

Session C: Analysis of Large Chemistry Spaces

C-1: The Proximal Lilly Collection Initiative: Design, Exploration and Application to Drug Discovery

Christos A. Nicolaou, Ian Watson, Jibo Wang

Lilly Research Laboratories, Eli Lilly & Company, Indianapolis, IN, USA

Modern drug discovery typically involves the exploration of a large set of compounds to identify promising hits. The initial set of compounds may be real, i.e. found in in-house molecular libraries or vendor collections, or virtual, found only in electronic form in virtual libraries. The initial exploration of these compound sets takes the form of experimental or virtual screening, with the latter being the only option in the case of virtual libraries. Promising structures are evaluated and upon confirmation of activity may serve as a starting point for further research either via an additional, more focused exploration of the starting set of compounds to retrieve near neighbors or via synthesis of analog compounds. In such process, the importance of the size, diversity and quality of the initial set of compounds cannot be overlooked. Given the practical limitations of expanding real compound libraries (cost, maintenance, logistics) and the sheer number of theoretically feasible compounds, the drug discovery community has invested heavily in virtual library design and exploitation. Moreover, numerous efforts have been reported focusing on mapping the chemical space and in designing large virtual libraries with diverse structures reaching out to less explored, but potentially promising regions for pharmaceutical development. A common practical concern hampering such efforts has been the synthesizability of the chemical structures proposed which is often questionable.

This presentation provides a description of the Proximal Lilly Collection (PLC), a large virtual compound library consisting of structures readily synthesizable using Eli Lilly knowhow, technology and starting materials. PLC has been designed to bridge the chemical synthesis knowhow and potential at Eli Lilly with the needs of ongoing discovery chemistry projects and thereby enable the exploitation of a larger chemical space in every day efforts. We also describe the PLC Initiative, a new drug discovery paradigm founded on the tight integration of PLC-based virtual hit identification with automated synthesis, purification and testing in order to expedite lead discovery.

At the core of PLC is the Lilly Annotated Reaction Repository (LARR) which contains reactions validated on Lilly automated synthesis systems. Each reaction is encoded using an annotation scheme which captures a wealth of information including the detailed profile of the reagents that may (or may not) be used. A flexible virtual synthesis engine (VSE) supported by a high- performance computing system enables the generation of chemical structure designs with high automated synthesizability confidence. The VSE is tightly coupled with the Lilly chemical sample management system to ensure that the reagent sets used in the process can readily be accessible for chemical synthesis. Users interact with the system through a collection of search and retrieve utilities which enable them to effectively query the PLC space. We present these utilities and describe our approach to efficiently explore the PLC space operating on LARR reaction and reagent similarity. In the final part of our presentation we describe a number of potential usage scenarios. A discussion on lessons learned, issues to be resolved, and future development directions will conclude the presentation.

C-2: Screening reaction pathway-driven very large chemical space: Discovery of potent mdm2-p53 antagonist

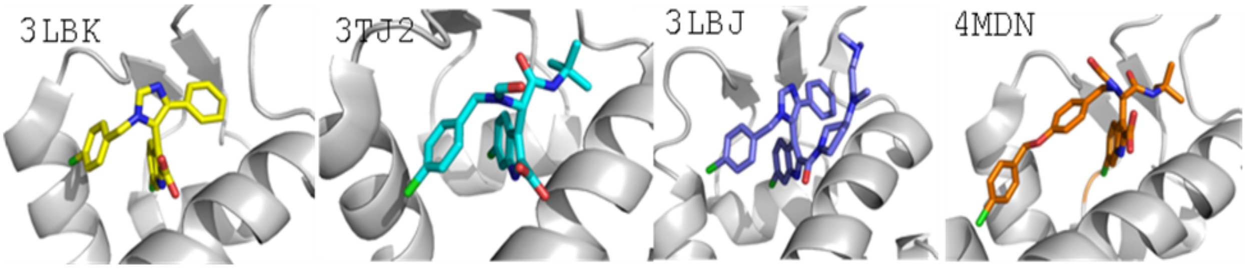
Alex Dömling[1], Carlos Camacho[2], Tad Holak[3], Dave Koes[2], Constantinos Neochoritis[1], Kareem Khoury[4]

[1]Dept. Drug Design, RUG, Netherlands, [2]University of Pittsburgh, USA, [3]Jagiellonian University, Krakow, Poland, [4]Carmolex Inc, Pittsburgh, USA.

Current industrial screening paradigm is HTS: Most medicinal chemistry program start with hits from high throughput screening. Virtual screening of very large chemistry space combined with structure-based drug discovery offers a valuable alternative. However, until recently, most virtual screening exercises used rather small libraries of limited chemical diversity. And, although, identified compounds are often commercially available for validation, such as in ZINC, PUBCHEM or ChEMBL databases, follow-up SAR is often challenging and expensive. For instance, GB-11 (what is GB11) is very large and hard to test due to the lack of efficient synthetic access to hit compounds for validation.

We have introduced ANCHOR.QUERY, a web-based and google-like technology for the structure-based mining of billions of small molecules (anchorquery.csb.pitt.edu/) [1]. The compound database is based on >20 diverse scaffolds with a defined reaction pathway based on efficient and rapid multicomponent reaction chemistry (MCR) [2]. The virtual compound library is design for a high level of confidence in synthetic feasibility and speed: Every hit compound can be rapidly accessed from commercial starting materials in less than four chemical steps. Protocols for the synthesis are provided online. To increase virtual screening hit rates the compound library is biased towards deeply buried anchor residues which play a key role in molecular recognition, e.g. deeply buried amino acid side chains in PPIs.

We have validated ANCHOR.QUERY with the discovery of more than 10 different compound classes able to antagonize the protein protein interaction p53-Mdm2-Mdm4 [3–6]. Several of the predicted compounds could be validated by cocrystal structure analysis and compared with the predicted binding poses (Figure). Some scaffolds were optimized towards low nM compounds highly active in cancer cells and a xenograft.



1. Koes, D. et al. Enabling large-scale design, synthesis and validation of small molecule protein-protein antagonists. *PLoS One* **2012**, 7: e32839. doi:10.1371/journal.pone.0032839.
2. Dömling, A.; Wang, K.; Wang, K. Chemistry and Biology of Multicomponent Reactions *Chem. Rev.* **2012**, *112*, 3083–3135.
3. Czarna, E. et al. Robust generation of lead compounds for protein protein interactions by computational and MCR chemistry: p53/mdm2 antagonists. *Angew. Chem. Intl. Ed.* **2010**, *48*, 5352–5356.
4. Bista, M. et al. Transient protein states in designing inhibitors of the mdm2-p53 interaction. *Structure* **2013**, *21*, 2143–2151.
5. Huang, Y. et al. Discovery of highly potent p53-mdm2 antagonists and structural basis for anti-acute myeloid leukemia activities. *ACS Chem. Biol.* **2014**, doi: 10.1021/cb400728e.

C-3: Large scale classification of chemical reactions from patent data

Gregory A. Landrum[1], Daniel M. Lowe[2], Roger A. Sayle[2]

[1]Novartis Institutes for Biomedical Research, Basel, Switzerland, [2]NextMove Software Limited, Cambridge, U.K.

There are numerous public data sources available to scientists who wish to carry out chemogenomic analyses of the interaction of molecules with many proteins (ChEMBL), learn about the SAR information available from a particular high-throughput screen (PubChem), or study the 3D details of how a molecule binds to a protein (PDB). Researchers who are interested in studying how those molecules were made, on the other hand, are somewhat out of luck. With the exception of the KEGG collection of information about enzymatic reactions, there are no large, public, machine-readable sources of information about chemical reactions.

Here we present a first attempt at remedying this situation: a collection of >1 million chemical reactions gathered by applying text mining to patent data [1]. We will also describe a workflow for assigning these reactions to classes in the RSC's RXNO ontology [2] that combines expert inspection and a clustering scheme based on a fingerprint for chemical reactions implemented as part of the open-source RDKit [3]. Finally we'll show results from some analyses made possible by this large data set: an analysis of reaction yield by reaction class and an application of machine learning to classify reactions.

1. Lowe, D. M. Extraction of chemical structures and reactions from the literature. PhD thesis. University of Cambridge: Cambridge, UK; 2012.
2. http://www.rsc.org/ontologies/RXNO/index.asp
3. RDKit: open source cheminformatics. http://www.rdkit.org

C-4: Open Patent Data

Anna Gaulton[1], Jon Chambers[1], Mark Davies[1], Lee Harland[2], George Papadatos[1], John P. Overington[1]

[1]European Molecular Biology Laboratory European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, United Kingdom, [2]SciBite Limited, Office 8, 10 Buckhurst Road, Bexhill-On-Sea, East Sussex, United Kingdom

While the availability of resources such as ChEMBL [1] and PubChem [2] have, in recent years, dramatically increased access to chemistry and bioactivity data in the public domain, much drug discovery-relevant information still remains buried in patent documents. In general, peer-reviewed literature lags behind patent documents, and some valuable information may never be published in journals at all. However, knowledge of the chemical space contained in patent documents is a crucial component of drug discovery that has thus far largely required access to commercial databases.

In an attempt to bridge this gap, the European Bioinformatics Institute has acquired the SureChem system, previously developed by Digital Science (http://www.digital-science.com), and is making this resource open to all users as SureChEMBL. The SureChEMBL system uses a number of chemical name recognition algorithms to identify ~15 million distinct chemical structures within 20 million US, EP and WO full text patent documents and Japanese patent abstracts, and is updated on a daily basis. This structural annotation allows users to search for patents containing a compound or substructure of interest, and retrieve other compounds from those patents.

Access to the resource is available via a variety of methods – through a dedicated web interface, RESTful web services, workflow tools and through integration with other resources. The UniChem database

[3], also based at EMBL-EBI, has been used to cross-reference the patent structures with a large number of other chemical resources, providing users with the ability to rapidly address overlap between various datasets, or provide dynamic cross-references to patent documents containing a structure of interest.

In seeking to answer complex drug-discovery questions, it is vital to be able to combine data from multiple different domains (e.g., chemistry, pharmacology, genomics, disease biology). In this respect, semantic web technologies can provide a significant advantage over traditional data integration methods, providing a more flexible and scalable platform for the integration of multiple diverse datasets. Open PHACTS aims to provide such a platform for drug discovery [4], and as part of this project, the ChEMBL database has already been made available in Resource Description Framework (RDF) format [5], providing access to pharmacology data through the system. The SureChEMBL dataset will also be made available via RDF, and integrated into the Open PHACTS platform. Alongside the extraction of chemical structures, text-mining will also be used to recognize biological entities within the patent documents, allowing the data set to be interrogated from both a chemical and biological viewpoint within the platform.

Funding Statement: The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement n° 115191 and 284209, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in kind contribution. Funding has also been provided by a Strategic Award for Chemogenomics from the Wellcome Trust (086151/Z/08/Z).

1. Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington J. P. The ChEMBL bioactivity database: an update. *Nucleic Acids Res*. **2014**, *42*, D1083–D1090.
2. 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*, D1075–D1082.
3. Chambers, J.; Davies, M.; Gaulton, A.; Hersey, A.; Velankar, S.; Petryszak, R.; Hastings, J.; Bellis, L.; McGlinchey, S.; Overington, J. P. UniChem: A Unified Chemical Structure Cross-Referencing and Identifier Tracking System. *J. Cheminform*. **2013**, *5*, 3.
4. Williams, A. J.; Harland, L.; Groth, P.; Pettifer, S.; Chichester, C.; Willighagen, E. L.; Evelo, C.T.; Blomberg, N.; Ecker, G.; Goble, C.; Mons, B. Open PHACTS: Semantic interoperability for drug discovery. *Drug Discov. Today*. **2012**, *17*, 1188–1198.
5. Jupp, S.; Malone, J.; Bolleman, J.; Brandizi, M.; Davies, M.; Garcia, L.; Gaulton, A.; Gehant, S.; Laibe, C.; Redaschi, N.; Wimalaratne, S. M.; Martin, M.; Le Novère, N.; Parkinson, H.; Birney, E.; Jenkinson, A. M. The EBI RDF Platform: Linked Open Data for the Life Sciences. *Bioinformatics*. **2014**.

Session D: Dealing with Biological Complexity

D-1: The impact of broad-scale genetic data on drug targets

Josef Scheiber, Görkem Sahin, Dennis Schwartz, Ariane Böhm

BioVariance GmbH, Munich, Germany

The recently announced 1000$-genome will inevitably have significant impact on the way Drug Discovery and Development is performed. Having human sequences along with model organism sequences available on a broad scale will enable insights that have not been possible thus far. Namely, we will be able to better design drugs for specific population groups or dial out population-specific off-targets, which will of course open up new routes for cheminformatics applications as well.

Within this study, we have analyzed the genomic data made available from the 1000 genomes project [1]. More specifically, we have been studying drug response rates on a molecular level. From each of the 1092 available human genomes we identified the sequences of all known drug targets covered in DrugBank [2]. We then extracted the sequences and annotated respective variations (mutations, deletions, insertions, duplications, inversions, translocations) for each drug target for each individual genome. Then we compared the sequences on a target-level and quantified the variations in order to correlate them with drug response rates. Respective results will be shown in this contribution.

Furthermore we identified a set of targets where the variations lead to immediate impact on binding pocket composition in certain populations or population groups. For these we built homology models based on existing 3D structures from the PDB database where a structure with a bound ligands was available [3]. It turns out that in a few cases drug responses can be explained on a molecular level. Specific examples will be shown.

1. The 1000 Genomes Project Consortium, An integrated map of genetic variation from 1,092 human genomes, *Nature* **2012**, *491*, 56–65.
2. Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V.; Djoumbou, Y.; Eisner, R.; Guo, A. C.; Wishart, D. S. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.* **2011**, *39* (Database issue), D1035–D1041.
3. http://www.pdb.org/pdb/general_information/news_publications/annual_reports/annual_report_year_2013.pdf

D-2: Navigating chemical and biological spaces in the search of novel pharmaceuticals

Paula Petrone

Hoffmann-La Roche, Basel, Switzerland

Since the advent of high-throughput screening, there has been an urgent need for methods that facilitate the interrogation of large-scale chemical biology data to build a mode of action (MoA) hypothesis for compounds. Despite the broad availability of biological data and the opportunities for data integration, the industrial -and often the academic- standard practice is to limit virtual screening techniques and compound searches to chemical structure considerations. This work raises awareness about the benefits of integrating bioinformatics considerations and pharmacogenomics data into the standard cheminformatics approaches.

Combining machine learning techniques and biological data, we offer a method to navigate the biologically relevant chemical space in search of new chemotypes: the "high-throughput screening fingerprint" (HTS-FP). This tool uses chemical biological descriptors which compare compounds solely on the basis of their bioactivity. In the current embodiment, data are aggregated from 195 biochemical and cell-based assays and can be used to identify bioactivity relationships among a collection comprising ~1.5 million compounds. We demonstrate the value of the HTS-FP for virtual screening and in particular scaffold hopping based on relationships of biological similarity. We also use these descriptors to define a metric of biological diversity which we apply to the design of high-throughput (HTS) screening libraries. We show how diversity in biological space is an essential requirement of high-throughput screening libraries as opposed to chemical space coverage. Efforts of data integration at the interface between chemistry and biology are gaining momentum in industry as data mining turns retrospective data into actionable asset. Roche has given a definite step to integrate these principles into its drug discovery workflows.

1. Fliri, A. F., Loging, W. T., Thadeio, P. F., and Volkmann, R. A. Biological spectra analysis: Linking biological activity profiles to molecular structure *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 261–266.
2. Petrone, P. M., Simms, B., Nigsch, F., Lounkine, E., Kutchukian, P., Cornett, A., Deng, Z., Davies, J.W., Jenkins,J. and Glick, M. Rethinking molecular similarity: comparing compounds on the basis of biological activity. *ACS Chem. Biol.* **2012**, *7,* 1399–1409.
3. Petrone, P. M., Wassermann, A. M., F., Lounkine, E., Kutchukian, P., Simms, B., Jenkins, J., Selzer, P. and Glick, M. Biodiversity of small molecules – a new perspective in screening set selection. *Drug Discov. Today* **2013**, *18*, 674–680.

D-3: Using Information from Historical High-Throughput Screens to Predict Active Compounds

Sereina Riniker[1], Yuan Wang[2], Jeremy L. Jenkins[2], Gregory A. Landrum[1]

[1]Novartis Institutes for BioMedical Research, Novartis Pharma AG, Novartis Campus, 4056 Basel, Switzerland, [2]Novartis Institutes for BioMedical Research Inc., 250 Massachusetts Avenue, Cambridge, MA 02139, USA

High-throughput screening (HTS), which enables the testing of a huge number of molecules within a relatively short time frame, is a well-established approach for hit finding in drug discovery. Conventional HTS routinely employed in pharmaceutical industry allows screening of 1-5 million compounds within a few weeks [1]. However, as the industry shifts to more disease-relevant but more complex phenotypic screens, the focus has moved to piloting smaller but smarter chemically/biologically diverse subsets followed by an expansion around hit compounds in the entire library [2]. One standard method for doing this is to train a machine-learning (ML) model, such as a Naïve Bayes (NB) classifier, with the chemical fingerprints of the tested subset of molecules and then select the next compounds based on the predictions of this model. An alternative approach would be to take advantage of the wealth of information contained in older (full-deck) screens using HTS fingerprints (HTS-fps) [3], where each element of the fingerprint corresponds to the activity outcome of a particular assay (Figure 1). These HTS-fps can be used like chemical fingerprints to train a ML model and generate predictions for untested compounds.

Here, we constructed HTS-fps using two collections of HTS data: (i) ~100 full-deck assays from Novartis with ~1.7 million compounds, and (ii) ~100 assays from PubChem with ~430K compounds. For each source, an additional set of 60 assays was used for testing. In these experiments, a subset of molecules was randomly selected for training of a random forest (RF) classifier and predictions were generated for the remaining compounds. The performance was evaluated using the area under the

ROC curve (AUC) and enrichment factors at different percentages, and compared to that of a NB model trained with the chemical fingerprint ECFP4. The RF(HTS-fps) was found to outperform NB(ECFP4) for the large majority of test assays in both collections, especially in the early recognition metrics. In addition, the use of classifier fusion [4], which allows the combination of the chemical fingerprints and HTS-fps, increased the predictive power even further.

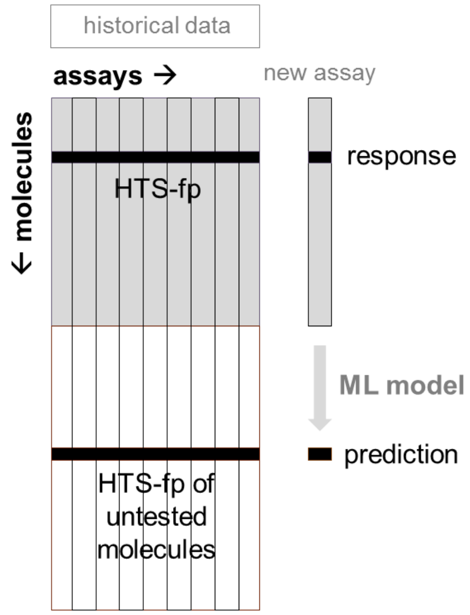


Figure 1: Schematic representation of the HTS fingerprints (HTS-fps) approach. HTS-fps are constructed from historical assays. A machine-learning (ML) model is trained using the HTS-fps and the response (i.e. activity outcome) of a new assay. For an untested molecule, a prediction of the activity outcome is obtained from the ML model based on its HTS-fp.

1. Battersby, B. J.; Tau, M. Novel Miniaturized Systems in High-Throughput Screening. *Trends Biotechnol.* **2002**, *20*, 167–173.
2. Petrone, P. M.; Wassermann, A. M.; Lounkine, E.; Kutchukian, P.; Simms, B.; Jenkins, J. L.; Selzer, P.; Glick, M. Biodiversity of Small Molecules – A New Perspective in Screening Set Selection. *Drug Discov. Today* **2013**, *18*, 674–680.
3. 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*, 1399–1409.
4. Riniker, S.; Fechner, N.; Landrum, G. A. Heterogeneous Classifier Fusion for Ligand-Based Virtual Screening: Or, How Decision Making by Committee Can Be a Good Thing. *J. Chem. Inf. Model.* **2013**, *53*, 2829–2836.

D-4: PubChem Structure-Activity Relationship Clusters and the Difference between 2-D and 3-D Similarity

Volker D. Hähnke, Lianyi Han, Sunghwan Kim, Evan Bolton, Stephen Bryant

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Department of Health and Human Services, Bethesda, Maryland, U.S.A.

A structure-activity relationship (SAR) connects the structural differences between molecules to changes in their shared bioactivity. Consequently, deducing a SAR is a crucial step in the hit exploration stage of a drug discovery campaign. PubChem contains more than 215 million biological activity results from 720 thousand biological assay experiments testing nearly two million unique chemical structures against more than six thousand defined protein targets [1]. Given the diversity of the data, establishing a meaningful SAR from this wealth of information can be a challenge. In order to facilitate easy access to groups of similar structures with common bioactivity in PubChem we performed a cluster analysis of the available data. The obtained clustering is accessible as a public resource [2].

For this purpose, 803,268 bioactive compounds were selected from PubChem and grouped into bioactivity-centered clusters based on their shared activity and their computed structural similarities. Activity was considered in three different contexts: i) active in the same assay, ii) active for the same target, and iii) active in the same pathway. Similarity as assessed with five different similarity measures. Four of those quantify the similarity of three-dimensional (3-D) structures based on the Rapid Overlay of Chemical Structures [3] using conformations available from PubChem3D [4]. PubChem structural keys [5] were used to quantify two-dimensional (2-D) structural similarity. That way we obtained 18 million clusters with molecules similar in structure and activity that can be used to infer a structure-activity relationship.

In this presentation, we will present the methodology we used to obtain clusters of bioactive structures, including the selection criteria for bioactive compounds, employed similarity measures and clustering approach. The obtained clusters will be used to illustrate the differences between 2-D and 3-D similarity. Furthermore, we will present examples of structure-activity relationship clusters.

1. Bolton, E.; Wang, Y.; Thiessen, P. A.; Bryant, S.H. PubChem: Integrated Platform of Small Molecules and Biological Activities. In *Annual Reports in Computational Chemistry. Volume 4*; Wheeler, R. A.; Spellmeyer, D. C., Eds., Elsevier: Amsterdam, 2008, pp217.
2. PubChem Structure-Activity Relationship Clusters, http://pubchem.ncbi.nlm.nih.gov/sar (accessed Feb 14, 2014).
3. ROCS – Rapid Overlay if Chemical Structures; OpenEye Scientific Software, Inc.: Santa Fe, NM, 2009, http://www.eyesopen.com/rocs (accessed Feb 14, 2014).
4. Kim, S.; Bolton, E. E.; Bryant, S. H. PubChem3D: Biologically relevant 3-D similarity. *J. Cheminf.* **2011**, *3*, 26.
5. PubChem Substructure Fingerprint, ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.pdf (accessed Feb 14, 2014).

Session E: Integration of Chemical Information with other Resources

E-1: Automatic completion of COSHH risk assessment forms using semantic representation of GHS and CLP regulation

Mark I. Borkum, **Jeremy G. Frey**

University of Southampton, U.K.

The completion of a risk assessment form is a sequence of three events: chemical information discovery and acquisition; chemical information manipulation; and, assessment form template interpolation. For organizations, the costs associated with manual risk assessment are very significant; ranging from the wages of those who complete and review the forms, to the damages paid to those who are affected by any errors. We present a software system that uses Semantic Web technologies to assist with and automate the completion of COSHH [1] risk assessment forms. We use an RDF [2] representation of the GHS [3] and CLP [4] regulation to describe the hazards and precautionary statements that are associated with a library of chemical substances. Aggregating the RDF data for a collection of chemical substances enables the automatic generation of a ready-to-use, human- and machine-readable risk assessment form. The software system is backed by an RDF triple-store and SPARQL [5] endpoint, and fronted by a lightweight, mobile-friendly Web application. We find that the introduction of software systems for automated risk assessment is, on the whole, beneficial to organizations. However, we also encounter a number of new risks, including: the risk that users obtain and utilise the "correct" answer to an "incorrect" question; the risk that users implicitly trust and/or rely upon the software system; and, the risk that the data gathered by the software system could be used to infer the research agendas of both individual users and organizations as a whole. Finally, we report on recent work to implement our software system using the Sigma-Aldrich catalogue as the primary data source.

1. Health and Safety Executive (HSE). A *Step by Step Guide to COSHH Assessment*; HSE Books, U.K., 2004.
2. Manola, F.; Miller, E.; McBride, B. *RDF Primer* [Online]; W3C Recommendation. Published online: February 10, 2004. http://www.w3.org/TR/2004/REC-rdf-primer-20040210/ (accessed January 31, 2014).
3. United Nations (UN). *Globally Harmonized System of Classification and Labelling of Chemicals*, 5th rev. ed.; UN, New York, U.S.A., 2013.
4. European Union (EU). European Regulation (EC) No 1272/2008; Official Journal of the European Union. **2008**, *353*, EU, Brussels, Belgium, 2008.
5. Prud'hommeaux, E.; Seaborne, A. *SPARQL Query Language for RDF* [Online]; W3C Recommendation. Published online: January 15, 2008. http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/ (accessed January 31, 2014).

E-2: Scientific Lenses over Linked Chemistry Data using BridgeDb and the Open PHACTS Chemical Registration System

Colin R. Batchelor[1], Christian Brenninkmeijer[2], Chris Evelo[3], Carole Goble[2], Alasdair J. G. Gray[4], Ken Karapetyan[1], Valery Tkachenko[1], Egon Willighagen[3]

[1]Royal Society of Chemistry, Cambridge, UK, [2]University of Manchester, Manchester, UK, [3]Maastricht University, Maastricht, The Netherlands, [4] Heriot-Watt University, Edinburgh, UK

Fluvastatin is a compound that exhibits stereochemistry. Three public chemical datasets – ChemSpider, ChEMBL, and DrugBank – list different stereo-forms. The question for a linked data platform is, "should these be treated as being the same?"

In linked data applications, the links between data sets are (generally) defined by the data providers. Very often, the meaning of the link is not provided at all and the two records may not be truly equivalent, but they may be treated as such under certain circumstances. Typically datasets will capture alternative views of the world at different levels of granularity which do not necessarily match the view of the world being presented in an application using the data. In our presentation we argue that the application using the linked data should decide upon the operational equivalence to apply between instances in different data sets, depending on the research question, using a *scientific lens* [1]. This is a set of rules that modifies the links between data sets according to the research question. For the fluvastatin example, we would have two different lenses, the first ensuring that the records would only be linked if they represented the same stereoisomer of the drug, the second allowing all stereoisomers to be treated as equivalent.

On the Open PHACTS Discovery Platform [2] we implement these lenses using the BridgeDb coreference service [3,4] and actively curate the links paying attention to the scientific meaning of the relationships claimed by the underlying data sets. The chemical linksets are provided by the Open PHACTS Chemical Registration Service (CRS) [5] which canonicalizes structures in various public chemical datasets, using a public ruleset based loosely on FDA guidelines and InChI normalization. It then generates links using different chemical equivalence conditions, disregarding for example stereochemistry, isotopic substitution or different salt forms. These links are defined in the open CHEBI and CHEMINF ontologies, providing human-readable explanations for their justification.

Using this approach, applications built upon the Open PHACTS Discovery Platform can vary the connections between chemistry datasets by applying different lenses and define the scope of the matching and therefore the amount and meaning of the search hits returned with little effect on the performance of the system.

1. Brenninkmeijer, C. Y. A.; et al Scientific Lenses over Linked Data: An approach to support task specific views of the data. A vision. In *Proceedings of 2nd International Workshop on Linked Science 2012 (LISC2012)*, 2012.
2. Gray, A. J. G.; et al Applying Linked Data Approaches to Pharmacology: Architectural Decisions and Implementation. *Semantic Web Journal*, **2014**, *5*, 101–113.
3. Brenninkmeijer, C. Y. A.; et al. Including Co-referent URIs in a SPARQL Query. In *4th International Workshop on Consuming Linked Data*. 2013.
4. Van Iersel, M. P.; Pico, A. R.; Kelder, T.; Gao, J.; Ho, I.; Hanspers, K.; Conklin, B. R.; Evelo, C. T. The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. *BMC Bioinformatics*, **2010**, 11.
5. Karapetyan, K.; Tkachenko, V.; Batchelor, C.; Sharpe, D.; Williams, A. RSC chemical validation and Standardization platform: A potential path to quality-conscious databases. In *Abstracts of Papers of the American Chemical Society* 2013, (Vol. 245), USA.

E-3: Will the real drug targets please stand up?

Christopher Southan

IUPHAR Database and Guide to PHARMACOLOGY Web Portal Group, Centre for Integrative Physiology, School of Biomedical Sciences, University of Edinburgh, Edinburgh, EH8 9XD, UK

Discerning the molecular mechanisms of action (mmoa) for drugs treating human diseases is crucially important. This presentation will provide an overview of target numbers in IUPHAR/BPS Guide to PHARMACOLOGY [1], the curatorial challenges and compare these to other sources and consider the wider implications for drug discovery. We have developed stringent mapping criteria for primary targets (i.e., identifying those direct protein interactions mechanistically causative for therapeutic efficacy). This includes inter-source corroboration by intersecting multiple drug sources inside PubChem to produce consensus structure sets. An analogous approach is used to intersect published target lists and database subsets at the UniProtKB/Swiss-Prot identity level (a selection of drug and target lists is now hosted on our website http://www.guidetopharmacology.org/lists.jsp). Our cumulative curation results reveal that structure representation differences, data provenance and variability of assay results, are major issues for experimental pharmacology and global database quality. While our activity mappings encompass some polypharmacolgy (e.g., dual inhibitors and kinase panel screens) our strategic choice is to annotate minimal, rather than maximal target sets. The consequent increased precision gives our database high utility for data mining, linking and cross-referencing. Our own database figures are currently converging to ~200 human protein primary targets for ~900 consensus chemical structures of approved small-molecule drugs. Target lists from other sources are typically larger. Comparative analysis of these lists by their UniProt ID content and Gene Ontology distributions suggests curatorial differences are the main cause of divergence [2]. The global target landscape thus shows paradoxical trends. On the one hand, cumulative drug research output and recent expansions (e.g., epigenetic targets and orphan diseases) have pushed bioactive compounds from papers or patents to above 2 million and chemically modulatable human proteins above 1500 [3]. On the other hand, reports of Phase II clinical efficacy failure, with implicit target de-validation, are frequent. In addition, our assessment of drug approval data from 2009 to 2013 indicates new targets (i.e., first-in-class mmoas) are so low as to threaten the sustainability of the pharmaceutical industry. Causes and consequences of these paradoxes, along with utilities for minimal and maximal druggable genomes, will be discussed.

1. Pawson, A. J.; Sharman, J. L.; Benson, H. E.; Faccenda, E.; Alexander, S. P.; Buneman, O. P.; Davenport, A. P.; McGrath, J. C.; Peters, J. A.; Southan, C.; Spedding, M.; Yu, W.; Harmar, A. J. NC-IUPHAR. The IUPHAR/BPS Guide to PHARMACOLOGY: an expert-driven knowledgebase of drug targets and their ligands. *Nuc. Acids Res*. **2014**, *42*.
2. Southan, C.; Sitzmann; Muresan, S. Comparing the Chemical Structure and Protein Content of ChEMBL, DrugBank, Human Metabolome Database and the Therapeutic Target Database. *Mol. Inf.* **2013**, *32*, 881–897.
3. Southan, C.; Varkonyi, P.; Boppana, K.; Jagarlapudi, S. A.; Muresan, S. Tracking 20 years of compound-to-target output from literature and patents. *PLoS One* **2013**, *8*, e77142.

E-4: Large-scale integration of chemical and biological data for drug discovery

Jörg Degen[1], Paula Petrone[1], David Herzig[1], Lisa Sach-Peltason[1], Nina Jeliazkova[2], Franziska Romrig[3], Olivier Roche[1], Daniel Stoffler[1]

[1]*F. Hoffmann-La Roche Ltd., Basel, Switzerland,* [2]*DEAconsult Ltd., Sofia, Bulgaria, Arcondis GmbH, Munich, Germany*

In a typical drug discovery project many decisions need to be taken on the basis of heterogeneous data from a variety of sources such as activities from high-throughput and low-throughput screening assays, measured and predicted physicochemical and pharmacological properties as well as information from patents, literature, and around availability of substances. One of the big challenges that arise in this context is the effort required for integration of the corresponding data as it is often fragmented across multiple systems. While this circumstance is not new, it becomes more apparent and problematic as data volume and complexity of questions in research are ever increasing.

Efforts of data integration at the interface between chemistry and biology are gaining momentum in industry and academia as data mining turns retrospective data into actionable assets. But even though pharmaceutical companies and academic institutions have taken significant steps to harmonize and streamline data capture and storage to better facilitate access and increase data quality, many organizations are still faced with the complexity of historically grown data landscapes where it is oftentimes impractical to fundamentally restructure existing solutions within reasonable effort. In this context the authors particularly appreciate the undertaking of the Innovative Medicines Initiative OpenPHACTS consortium [1] and the providers of the integrated data sources.

In an effort to strengthen our ability to answer essential questions in a drug discovery program [2] and by capitalizing on the OpenPHACTS initiative, we have established a new dedicated data warehouse and query framework that enables us to aggregate and interpret data from different domains. For example, computational predictions that rely on historical activity data such as target identification in phenotypic screens or virtual screenings based on biological profiles, compound repurposing or pathway identification can now be addressed in a straight-forward manner. By integrating publicly available data through OpenPHACTS with our internal data and mapping the corresponding concepts, we now have the possibility to query both data sources in a unified manner, thus enabling more comprehensive data analysis and more efficient decision making.

With our new framework we have taken a significant step forward to integrate these principles into our drug discovery workflows and to utilize the value of both internally and externally generated data more thoroughly and efficiently. In this presentation we will briefly discuss architectural and performance considerations during the design process as well as showcase the usability with application examples.

1. Williams, A. J.; Harland, L.; Groth, P.; Pettifer, S.; Chichester, C.; Willighagen, E. L.; Evelo, C. T.; Blomberg, N.; Ecker, G.; Goble, C.; Mons, B. Open PHACTS: Semantic interoperability for drug discovery. *Drug Discovery Today* **2012**, *17*, 1188–1198.
2. Azzaoui, K.; Jacoby, E.; Senger, S.; Rodríguez, E. C.; Loza, M.; Zdrazil, B.; Pinto, M.; Williams, A. J.; de la Torre, V.; Mestres, J.; Pastor, M.; Taboureau, O.; Rarey, M.; Chichester, C.; Pettifer, S.; Blomberg, N.; Harland, L.; Williams-Jones, B.; Ecker, G. F. Scientific competency questions as the basis for semantically enriched open pharmacological space development. *Drug Discovery Today* **2013**, *18*, 843–852.

Session F: Structure-Activity and Structure-Property Prediction

F-1: Drug target prediction – comparison of methods

David Evans[1],Thibault Varin[2], Jeffrey Sutherland[2], Michael Lajiness[2], Suzanne Brewerton[1], Georgios Drakakis[3], Andreas Bender[3], Michal Vieth [2]

[1]Eli Lilly and Company, Windlesham, UK[2]Eli Lilly and Company, Indianapolis, US, [3]Unilever Centre, University of Cambridge, UK

Phenotypic screens where the primary interaction site of active compounds may not be not known are of increasing importance, creating opportunities for in silico target prediction methods. Target prediction is also highly complementary to knowledge-based chemogenomics databases such as Metacore and ChEMBL, allowing novel compounds to be viewed in the context of annotated ligand-protein and protein-protein interactions. The technology also allows for the prediction of off-target activities at an early stage in hit-to-lead scaffold selection, but in all these cases a realistic assessment of the likely accuracy of the predictions is important to minimize experimental work in their confirmation.

The work presented here compares the effectiveness of three well-known methodologies (support vector machines, multiclass naïve Bayes and SEA) at predicting the known targets of 811 marketed drugs, based on training sets from ChEMBL15 and in house assay data. Results indicate that correct targets are consistently ranked in to top 2-3 predictions but that different algorithms are more effective for different types of data source. In addition we examine the ability of models trained on external data sources to predict those of in house assays and vice-versa.

1. MetaCoreTM http://thomsonreuters.com/metacore/
2. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2011**, *40* D1100–D1107.
3. In Silico Target Predictions: Defining a Benchmarking Data Set and Comparison of Performance of the Multiclass Naïve Bayes and Parzen-Rosenblatt Window. *J. Chem. Inf. Model.* **2013**, *53*, 1957–1966.
4. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* **2007**, *25*, 197–206.

F-2: Broadening the usefulness of pK_a predictions by taking multiple protonation microstates into account

Robert D. Clark[1], Robert Fraczkiewicz[1], Marvin Waldman[1], Mario Lobell[2], Andreas Göller[2], A. Hillisch[2], Ursula Krenz[2], Rolf Schoenneis[2]

[1]Simulations Plus, Lancaster, CA, USA, [2]Bayer Pharma AG, Global Drug Discovery, Wuppertal, Germany

The ultimate measure of value for a quantitative structure-property relationship (QSPR) is how much difference it makes to the productivity of those who use it to make predictions. This depends in part on how broadly applicable the predictions are but also in part on the context within which the predictions are presented. The first facet relates to how much structural variation there is in the chemical space covered by the model, which is commonly referred to as its "applicability domain." The contextual aspect involves how immediately relevant the predictions made are to the task at hand and how effectively they are presented to those who need to make use of them.

Quantitative estimates of the protonation and deprotonation behavior of organic molecules in aqueous solution is critically important to those doing pharmaceutical research, but getting good predictions for molecules that contain two or more ionizable centers can be very difficult. The most rigorous and accurate way to predict the pK_a's measurable by titration – macroscopic pK_a's – in such cases is to take the interactions between all accessible ionization states of the molecule – all of its microstates – into account [1]; however, adequate parameterization of such models requires very large, well-curated, and structurally diverse training sets. Simulations Plus and Bayer Pharma AG recently joined forces to develop a new *in silico* pK_a prediction tool by applying Simulation Plus's advanced microstates analysis and artificial neural net technology to data for ~27000 compounds having a total of ~33500 associated pK_a values: ~14000 values from the literature and ~19500 values determined experimentally at Bayer Pharma.

Model validation was performed with >16000 pK_a measurements taken on a completely different set of compounds analyzed after model training was complete. The model achieved a mean absolute error (MAE) of 0.50, a root mean square error (RMSE) of 0.67 and a coefficient of determination (R^2) of 0.93 on this external test set. This represents a substantial improvement over the corresponding performance statistics for the pK_a model in version 6.5 of ADMET Predictor, which was trained on the literature data alone: MAE = 0.72, RMSE = 0.94, and R^2 = 0.87. The new model is being released as part of ADMET Predictor™ 7.0.

The collaboration went beyond simply correctly predicting macroscopic pK_a predictions. It also involved putting those predictions into a context useful for medicinal chemists and other scientists at Bayer Pharma by integrating them into Pipeline Pilot and Bayer's PharmacophorInformatics (PIx) platform. In addition to predicting macroscopic pK_a's, key insights derived from the associated high-dimensional microstate information for multiprotic compounds are conveyed graphically via plots of averaged single proton acidity (ASPA), averaged site protonation (ASP) and pK50s for individual sites – i.e., pH values at which each site is half-protonated across all microstates. The fast calculation speed (>100,000 compounds/h [CPU: Intel Xeon L5420, 2.5 GHz]) allows interactive as well as batch processing.

1. Fraczkiewicz, R. In silico Prediction of Ionization. In *Reference Module in Chemistry, Molecular Sciences, and Chemical Engineering*; Reedijk, J. Ed.; Elsevier, 2013: http://www.sciencedirect.com/science/article/pii/B978012409547202610X.

F-3: Using Decision trees to rationalize the mechanism-of-action of hypnotics: Assessing predicted sedative effects of marketed drugs in vivo

Georgios Drakakis[1], Keith Wafford[2], Suzanne Brewerton[2], Michael Bodkin[2], David Evans[2], Andreas Bender[1]

[1] Unilever Centre for Molecular Science Informatics, Cambridge, UK, [2] Eli Lilly and Company, Windlesham, UK

The increase of processing power and bioactivity data has led to the augmented development and usage of *in silico* bioactivity prediction methods [1]. Protein target predictions can help establish the link between compound and phenotypic effect on a biological system. Such mechanism-of-action rationalization of a particular compound class has been recently carried out in our group on a *Xenopus laevis* dataset via polypharmacological bioactivity profile analysis [2,3]. In particular, protein targets were predicted using a multiclass Naïve Bayes bioactivity prediction method [4] and decision trees were then employed to derive the mode of action by separating active and inactive compounds based on the observed phenotype. In this work, we applied the described methodology to a subset of the Eli Lilly SCORE™ dataset of hypnotics, comprising 845 data instances (491 unique compounds). The resulting model (68.4% accuracy using 10-fold cross-validation) was used to predict potential sedative effects of marketed and experimental Drugbank [5] compounds. More specifically, five polypharmacological bioactivity profiles were derived for explaining and predicting hypnotic effects, which can be described as a sequence of nodes/protein targets in the decision tree, from root to leaf. For example, if a query compound is predicted as active on Dopamine D2, Histamine H1 and 5-HT2A receptors, it should promote a sedative effect. Those predicted as sedative but had no -to our best knowledge- prior literature annotation were tested *in vivo* for confirmation in the SCORE^{TM} rat model. Five out of the seven tested Drugbank[5] molecules showed sleep activity in the SCORE^{TM} experiment.

1. Koutsoukas, A.; Simms, B.; Kirchmair, J.; Bond, P. J.; Whitmore, A. V; Zimmer, S.; Young, M. P.; Jenkins, J. L.; Glick, M.; Glen, R. C.; et al. From in Silico Target Prediction to Multi-Target Drug Design: Current Databases, Methods and Applications. *J. Proteomics* **2011**, *74*, 2554–2574.
2. Liggi, S.; Drakakis, G.; Hendry, A. E.; Hanson, K. M.; Brewerton, S. C.; Wheeler, G. N.; Bodkin, M. J.; Evans, D. A.; Bender, A. Extensions to In Silico Bioactivity Predictions Using Pathway Annotations and Differential Pharmacology Analysis: Application to Xenopus Laevis Phenotypic Readouts. *Mol. Inf.* **2013**, *32*, 1009–1024.
3. Drakakis, G.; Hendry, A. E.; Hanson, K. M.; Brewerton, S. C.; Bodkin, M. J.; Evans, D. A.; Bender, A. Comparative Mode-of-Action Analysis Following Manual and Automated Phenotype Detection in Xenopus Laevis. *Med. Chem. Commun. [Online early access] DOI:10.1039/C3MD00313B.*
4. Koutsoukas, A.; Lowe, R.; Kalantar-Motamedi, Y.; Mussa, H. Y.; Mitchell, J. B. O.; Glen, R.; Bender, A. In Silico Target Predictions: Comparing Multiclass Naïve Bayes and Parzen-Rosenblatt Window and the Definition of a Benchmarking Dataset for Target Prediction. *J. Chem. Inf. Model.* **2013**, *53*, 1957–1966.
5. Knox, C.; Law V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V.; Djoumbou, Y.; Eisner, R.; Guo, A. C.; Wishart, D. S. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.* **2011**, *39 (Database issue)*, D1035–1041.

F-4: Existing and Developing Approaches for QSAR Modeling of Mixtures

Eugene Muratov[1,2], **Ekaterina Varlamova**[2], **Anatoly Artemenko**[2], **Denis Fourches**[1], **Victor Kuz'min**[2], **Alexander Tropsha**[1]

[1]Laboratory for Molecular Modeling, UNC Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599, USA, [2]Laboratory of Theoretical Chemistry, Department of Molecular Structure, A.V. Bogatsky Physical-Chemical Institute NAS of Ukraine, Odessa, 65080, Ukraine

Most of (if not all) the time chemicals act on humans, animals, or environment in combination with other compounds. However, Quantitative Structure-Activity (Property) Relationship (QSAR or QSPR) modeling of mixtures is only beginning to emerge because of the scarcity of reliable experimental data on biological and/or physical-chemical effects of mixtures [1]. In this presentation, we will discuss most recent theoretical developments and applications in this new area of QSAR.

The QSAR modeling of organic mixtures requires the use of specific descriptors to characterize the different chemicals involved, taking into account their stoichiometry. All studies published to date on the subject can be divided into several groups depending on the descriptor types used: *(i)* descriptors based on the mixture partition coefficients or biological descriptors; *(ii)* additive molecular descriptors (weighted sum of descriptors of individual components); *(iii)* integral non-additive descriptors of mixtures (mixture components are taken into account in a different manner from the additive scheme); and *(iv)* fragment non-additive descriptors (structural parts of different mixture components simultaneously taken into account by the same descriptor; see Figure 1) [1].

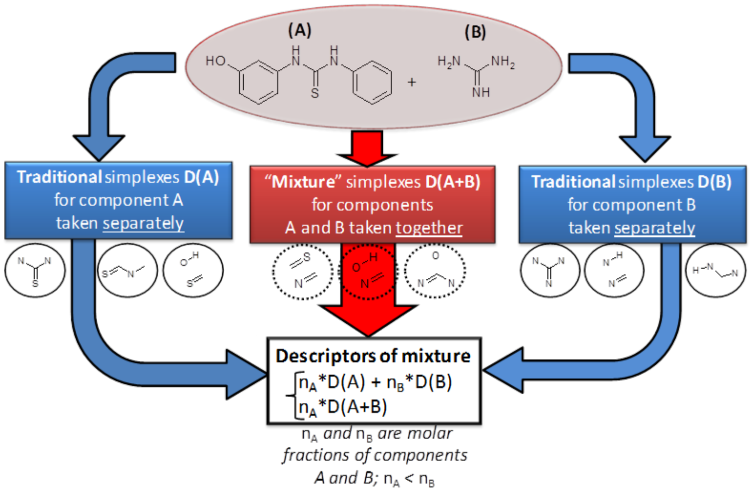

(A)
(B)
HO
H N
H N
S
+
H_2N
NH_2
NH
Traditional simplexes D(A) for component A taken separately
"Mixture" simplexes D(A+B) for components A and B taken together
Traditional simplexes D(B) for component B taken separately
Descriptors of mixture
n_A*D(A) + n_B*D(B)
n_A*D(A+B)
n_A and n_B are molar fractions of components A and B; n_A < n_B

Figure 1: Generation of Simplex descriptors for mixtures.

Depending on the dataset and potential application(s) of the models, three different strategies of external validation could be used: *(i)* "points out" – prediction of the investigated property for any composition of mixture from the modeling set, *(ii)* "mixtures out" – filling of missing cells in the initial mixture data matrix, *i.e.,* prediction of the investigated property for mixtures with unknown activity created by combining pure compounds from the modeling set, and *(iii)* "compounds out" – prediction of the investigated property for mixtures formed by novel pure compound(s) absent in the modeling set, which is the most rigorous method of external validation in QSAR modeling of mixtures.

We will present several case studies including QSAR modeling of antipoliovirus activity of binary combinations of antivirals and QSPR modeling of $T_{boiling}$ of mixtures of organic solvents. Given the importance and the growing need for such models in drug discovery and chemical hazard assessment, we expect the development of innovative modeling workflows and the improvement of existing QSAR/QSPR approaches for mixtures in the near future. Specifically, the accumulation of additional data and its thorough curation as well as rigorous internal and external validation can significantly improve the quality of QSAR models of mixtures and enable their application for virtual screening of large databases of actual or uncharacterized mixtures.

1. Muratov, E. N.; Varlamova, E. V.; Artemenko, A. G.; Polishchuk, P. G.; Kuz'min, V. E. Existing and Developing Approaches for QSAR Analysis of Mixtures. *Mol. Inform.* **2012**, *31*, 202–221.

POSTER SESSION ABSTRACTS RED

P-2: Knowledge Discovery in Pharmaceutical Drug Transport using Emerging Graph Patterns

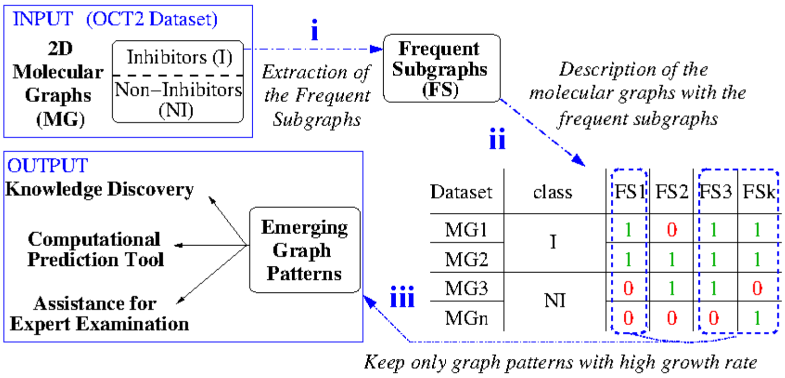
Guillaume Poezevara[1,2], **Alban Lepailleur**[2], **Sylvain Lozano**[3], **Bertrand Cuissart**[1], **Bruno Crémilleux**[1], **Ronan Bureau**[2], **Philippe Vayer**[3]

[1]*Groupe de Recherche en Informatique, Image, Automatique et Instrumentation de Caen (GREYC CNRS UMR 6072), Campus Côte de Nacre, Université de Caen Basse-Normandie, France,* [2]*Centre d'Etudes et de Recherche sur le Médicament de Normandie (CERMN UPRES EA-4258 FR CNRS INC3M), UFR des Sciences Pharmaceutiques Université de Caen Basse-Normandie, France,* [3]*Technologie Servier, 27 rue Eugène Vignat, 45000 Orléans, France*

The clinical importance of efflux and uptake transporters in drug disposition is widely acknowledged, and membrane transporter anomalies are the basis for certain clinical disorders. Particularly, the Organic Cation Transporter 2 (OCT2) is a renal transporter that plays a key role in disposition and renal clearance of most currently prescribed drugs. Computational models could predict undesirable effects that are based on drug transporter interactions, and statistical models like quantitative structure-activity relationships and pharmacophores have been proposed [1,2].

Due to the evolution of the modern information methods and technology, collecting, combining, storing, and mining huge amounts of data can now be done very efficiently. Several works have been conducted in chemoinformatics to extract the frequent substructures from a dataset of graphs and to link them to a biological activity. However, in the recent years studies have given more attention to the discovery of patterns that are significant, like the emerging patterns [3], than simply frequent.

In this study, a dataset [2] containing 910 compounds has been considered, each compound being represented by its 2D molecular graph together with its ability to inhibit the OCT2 transporter. Based on these information, the dataset has been partitioned into two classes: inhibitors vs. non-inhibitors. As shown on the figure, our method [4,5] finds representative subsets of frequent patterns which occur more frequently in the inhibitors than in the non-inhibitors, or inversely. The extracted patterns are said to be "emerging" [5] and are characterized by their contrast (growth rate) between the two classes.


INPUT (OCT2 Dataset)
2D Molecular Graphs (MG)
Inhibitors (I)
Non–Inhibitors (NI)
i
Extraction of the Frequent Subgraphs
Frequent Subgraphs (FS)
Description of the molecular graphs with the frequent subgraphs
ii
OUTPUT
Knowledge Discovery
Computational Prediction Tool
Assistance for Expert Examination
Emerging Graph Patterns
iii
Keep only graph patterns with high growth rate

Dataset	class	FS1	FS2	FS3	FSk
MG1	I	1	0	1	1
MG2		1	1	1	1
MG3	NI	0	1	1	0
MGn		0	0	0	1

As the method is not based on expert knowledge, it allows automatic knowledge discovery. Then, the extracted patterns have been implemented in a computational prediction tool and used as assistance for expert examination.

1. Suhre, W. M.; Ekins, S.; Chang, C.; Swaan, P. W.; Wright, S. H. Molecular Determinants of Substrate/inhibitor Binding to the Human and Rabbit Renal Organic Cation Transporters hOCT2 and rbOCT2. *Mol. Pharmacol.* **2005**, *67*, 1067–1077.
2. Kido, Y.; Matsson, P.; Giacomini, K. M. Profiling of a Prescription Drug Library for Potential Renal Drug-Drug Interactions Mediated by the Organic Cation Transporter 2. *J. Med. Chem.* **2011**, *54*, 4548–4558.

3. Sherhod, R.; Gillet, V. J.; Judson, P. N.; Vessey, J. D. Automating Knowledge Discovery for Toxicity Prediction Using Jumping Emerging Pattern Mining. *J Chem Inf Model* **2012**, *52*, 3074–3087.
4. Cuissart, B.; Poezevara, G.; Crémilleux, B.; Lepailleur, A.; Bureau, R. Emerging Patterns as Structural Alerts for Computational Toxicology in *Contrast Data Mining: Concepts, Algorithms and Applications;* Dong, G.; Bayley, J.; Taylor & Francis Group; **2012**; 259–270.
5. Lozano, S.; Poezevara, G.; Halm-Lemeille, M.P.; Lescot-Fontaine, E.; Lepailleur, A.; Bissell-Siders, R.; Crémilleux, B.; Rault, S.; Cuissart, B.; Bureau, R. Introduction of Jumping Fragments in Combination with QSARs for the Assessment of Classification in Ecotoxicology; *J. Chem. Inf. Model.* **2010,** *50*, 1330–1339.
6. Dong, G.; Li, J. Efficient Mining of Emerging Patterns: Discovering Trends and Differences. *KDD* **1999**; 43–52.

P-4: Comprehensive analysis of bioisosteric replacement in ligands of a serotonin receptors family

Dawid Warszycki, Jakub Staroń, Rafał Kafel, Andrzej J. Bojarski

Department of Medicinal Chemistry, Institute of Pharmacology Polish Academy of Sciences, Cracow, Poland

A bioisosteric replacement transforms an active compound into another one by exchanging a group of atoms with broadly similar (in physicochemical properties) groups. Implementations of this technique are aimed on increase of affinity, improvement of pharmacokinetic properties or exploration of new, unknown scaffolds.

For compounds with determined affinity for any serotonin receptor stored in the ChEMBL [1] database (version 16 May 2013) all possible bioisosteres were generated in Pipeline Pilot [2]. Analysis of this collection, consisting of more than 1 million structures, showed that in average 31% of known ligands of a particular target are mutual bioisosteres. Further data exploration revealed the most frequent and the most efficient replacements in modulating ligands activity for different subtypes of serotonin receptors.

As regards, for example, 5-HT$_6$ receptor ligands, the most frequently explored modifications were halogen substitution and ring modification (contracting, expanding, changing linear fragments to rings, etc). Analysis showed that the most appropriate fragments for increasing ligands affinity for 5-HT$_6$R are phenyl and sulfonamide. Moreover, it was found that ring modifications in ligands of other targets may result in more potent compounds acting on 5-HT$_6$ receptor. On the other hand, substitution to nitrile group or introduction of any pyridines instead of other aromatic ring, caused decrease of ligands activity.

Similar observations and selectivity analysis, are presented and discussed for each of serotonin receptor subtype.

1. Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nuc. Acids Res.* **2011**, *40*, D1100–D1107.
2. Pipeline Pilot, version 6.0, Accelrys, Inc., San Diego, CA, USA.

P-6: Discriminative Chemical Patterns: Automatic and Interactive Design

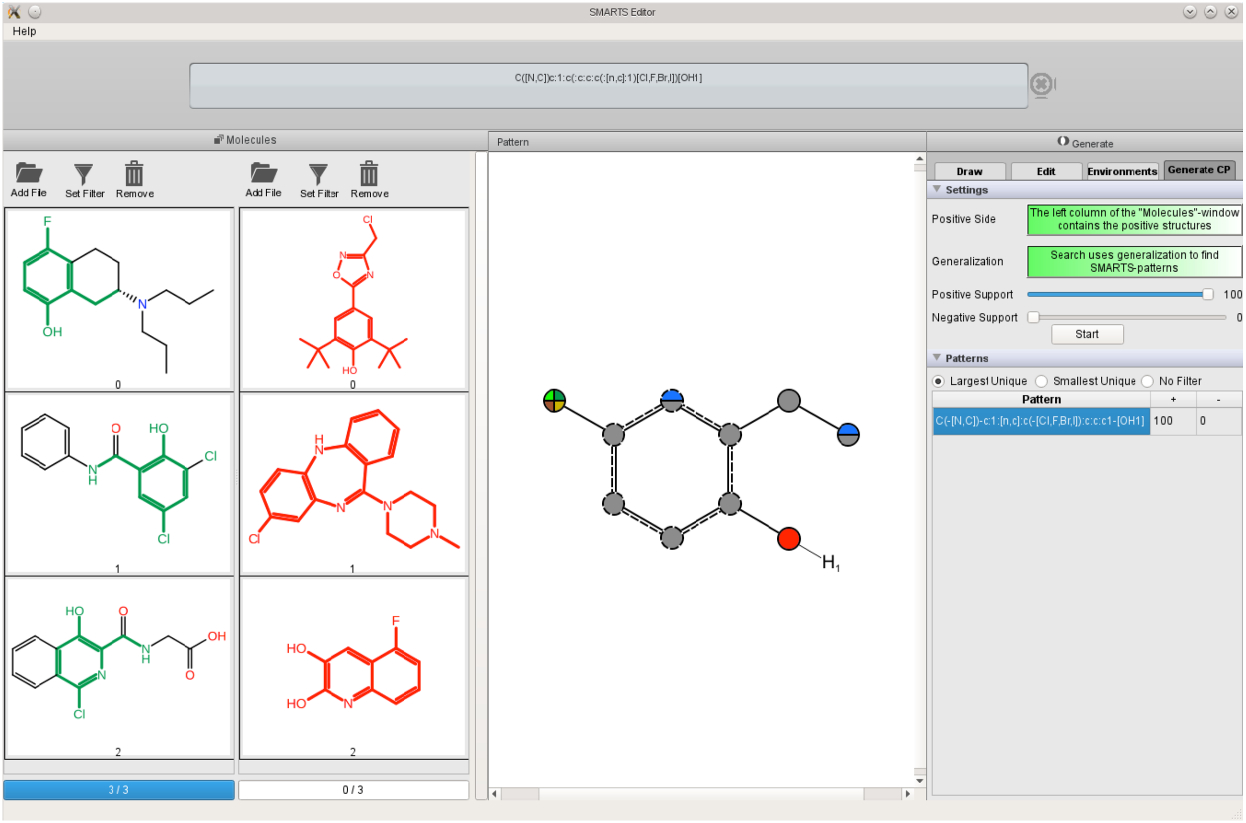
Stefan Bietz, Karen T. Schomburg, David Seier, Matthias Rarey

University of Hamburg, Center for Bioinformatics, Bundesstrasse 43, 20146 Hamburg, Germany

The concept of chemical patterns is a commonly used instrument in Cheminformatics. Its applications range from the definition of chemical moieties and the specification of chemical search queries to the description of whole molecular classes. A special field of application is the differentiation of molecules by so-called discriminative chemical patterns. These can for instance describe common structural properties of one molecule class which are not present in another. While such patterns are obviously helpful in classification tasks like toxicology prediction or the discovery of active drugs, finding the discriminative features is not trivial. Computational methods proved to be helpful for the automatic generation of discriminative molecular patterns on the basis of predefined molecule sets [1,2].

A challenge that occurs when working with chemical pattern is their abstract representation. Generally, chemical patterns are annotated in string-based pattern languages like SMARTS [3] , which indeed offer powerful opportunities to describe chemical properties, although this also complicates the interpretation and manual generation of these patterns. The SMARTSeditor [4] is an editor for an interactive design of SMARTS patterns. Based on the intuitive SMARTSviewer visualization concept [5] and a comprehensive support of the SMARTS language, even highly complicated patterns can easily be created in a stepwise construction procedure where the user can freely combine predefined scaffold templates and specify molecular properties.

We present a new extension of the SMARTSeditor for an automatic detection, visualization, and modification of discriminative chemical patterns. The underlying subgraph mining approach supports the discovery of SMARTS patterns that separate two given molecule sets. In general, our algorithm can create discriminative patterns with any desired atomic SMARTS expressions of both specializing and generalizing manner. A prioritizing search strategy further enables a rapid detection of quasi-optimal solutions. This particularly allows the user to immediately work with sensible preliminary results while the search runs in background. The seamless integration of the automatic pattern detection method with the editor enables quick manual adaptions of the pattern with direct feedback on the matching behavior on the molecule sets. Several practical examples of semi-automatic discriminant pattern design will be presented.



C([N,C])c:1:c(:c:c:c(:[n,c]:1)[Cl,F,Br,I])[OH1]

Settings

Positive Side: The left column of the "Molecules"-window contains the positive structures

Generalization: Search uses generalization to find SMARTS-patterns

Positive Support: 100

Negative Support: 0

Patterns

● Largest Unique ○ Smallest Unique ○ No Filter

Pattern	+	-
C(-[N,C])-c:1:[n,c]:c(-[Cl,F,Br,I]):c:c:c1-[OH1]	100	0

Figure 1: Screenshot of the SMARTSeditor: On the left side, two molecule sets are shown from which a discriminative pattern can be derived automatically. The matchings of the pattern are high-lighted for each molecule. The center panel shows the pattern itself enabling interactive manipulation of the pattern with several editing tools shown on the right.

1. Borgelt, C.; Meinl, T.; Berthold, M. Moss: a program for molecular substructure mining. In *Proceedings of the 1st international workshop on open source data mining: Frequent Pattern Mining Implementations*; ACM Press: New York, 2005; pp 6–15.
2. Kazius, J.; Nijssen, S.; Kok, J.; Bäck , T.; Ijzerman, A. P. Substructure mining using elaborate chemical representation. *J. Chem. Inf. Model.* **2006**, *46*, 597—605.
3. James, C. A., Weininger D., Daylight Theory Manual. Daylight Chemical Information Systems, Inc. of Aliso Viejo, CA, available at www.daylight.com
4. Schomburg, K. T.; Wetzer, L.; Rarey, M. Interactive design of generic chemical patterns. *Drug Discov. Today* **2013**, *18*, 651—658.
5. Schomburg, K. T.; Ehrlich, H.-C.; Stierand, K.; Rarey, M. From structure diagrams to visual chemical patterns. *J. Chem. Inf. Model.* **2010**, *50*, 1529—1535.

P-8: Protein binding site comparison based on secondary structure elements

Tobias Brinkjost[1,2], **Christiane Ehrt**[1], **Petra Mutzel**[2], **Oliver Koch**[1]

[1]*Department of Chemistry and Chemical Biology, TU Dortmund, Germany,* [2]*Department of Computer Science, TU Dortmund, Germany*

The investigation of protein-ligand interactions is one of the prerequisites for structure-based design of small molecule modulators of protein function. These interactions can be regarded based on structural similarity of secondary structure elements with impact on rational drug design [1]. The basic idea of the presented approach is the fact that a similar spatial arrangement of secondary structure elements around the binding site ('ligand-sensing cores') can recognize similar scaffolds independent of the overall fold [2]. The discovery of Namoline as a lysine-specific demethylase 1 (LSD1) inhibitor, which impairs the growth of prostate cancer cells, by Willmann et al. demonstrated the pharmaceutical relevance of this concept [3]. However, to date there is no automated procedure available to compare 'ligand-sensing cores' of various proteins.

We will present the results of our ongoing progress to develop an automated computational method to identify 'ligand-sensing cores' in binding pockets of otherwise unrelated proteins for all known protein structures and possibilities. Our current approach is based on detecting maximum common subgraphs (*MCS*) of labeled graphs determined by variants of the Bron Kerbosch [4] maximum clique detection algorithm in appropriately defined product graphs. Since our goal is to process all known protein structures and the maximum common sub-graph isomorphism problem is well known to be *NP-complete*, we will introduce a pre-filtering procedure to limit the number of graph-comparisons to promising ones.

In the end, the complete information about all similar ligand-sensing cores within all known protein structures would provide access to previously unused data to predict polypharmacology and to identify new lead structures. Therefore, this development should lead to a valuable tool for rational drug design.

1. Koch, O. The Use of Secondary Structure Element Information in Drug Design: Polypharmacology and Conserved Motifs in Protein-Ligand Binding and Protein-Protein Interfaces. *Fut. Med. Chem.*, **2011,** *3,* 699–708.
2. Koch, M. A., Waldmann, H.; Protein structure similarity clustering and natural product structure as guiding principles in drug discovery. *Drug Discov. Today*, **2005**; *10*, 471–483.
3. Willmann, D., Lim, S., Wetzel, S., Metzger, E., Jandausch, A., Wilk, W., Jung, M., Forne, I., Imhof, A., Janzer, A., Kirfel, J., Waldmann, H., Schüle, R., Buettner, R.; Impairment of prostate cancer cell growth by a selective and reversible lysine-specific demethylase 1 inhibitor. *Internat. J. Canc.* **2012**, *131*, 2704–2709.
4. Bron, C., Kerbosch, J. Finding all Cliques of an Undirected Graph. *Communications of the ACM*, **1973**, *16*, 575–579.

P-10: fpserver: an in-process cheminformatics database system for Python

Andrew Dalke

Dalke Scientific, Trollhättan, Sweden

fpserver is an in-process cheminformatics data management system for Python. It implements standard cheminformatics methods including record lookup, Tanimoto search, substructure search, as well as an off-line build system to precompute data and index records for faster searches. The methods can be called from the command-line or through a Python API. fpserver includes a built-in Django app to publish the methods through a JSON-based web services interface.

These standard methods are easy to implement with any of the nearly dozen widely used chemistry toolkits [1], though with substantial linear-time overhead. For better performance, most turn towards off-the-shelf database servers with chemistry-aware extensions [2]. They improve performance through a combination of precomputation, caching, improved data structures, and query planners. A DBMS is especially appropriate in enterprise environments where the underlying data changes frequently.

Server solutions have some negatives. A database server may introduce excessive complexity when a research group wants fast access to a relatively static dataset but doesn't have people with the specialized database management skills to set up and maintain the server. The popular database servers scale well across multiple cores, but can't easily be scaled across a distributed memory cluster, or scaled down to a single machine, or used in one-off command-line searches.

An alternative to a server is an in-process database, where database functionality is through function calls to a library rather than network calls to a stand-alone server. These can be simpler to maintain because the database follows the same rules as any other library. SQLite [3] is the most widely used in-process SQL database.

fpserver is an in-process database management system. It uses SQLite to store and retrieve structure record data and chemfp [4] for fingerprints generation and fast Tanimoto searches and substructure screens. Worst-case similarity search timings for the nearest 100 neighbors of a 1 million record set is under 0.1 seconds, which is fast enough that it can be used in an interactive sketcher.

The fpserver design supports multiple toolkits, though only RDKit support is currently implemented. Substructure search uses RDKit's "pattern" fingerprints for screening, followed by the full subgraph isomorphism validation. Search performance depends on the query, database size, and number of matches requested. A benchmark based on the Structure Query Collection's BindingDB queries [5] averages 0.9 seconds per query against ChEMBL. With a warm OS disk cache, RDKit can validate about 1,000 structures per second, or 6,000 per second once fpserver's molecule cache is warm.

The fpserver API is used the the command-line tool and by the built-in Django app. The app makes it easy to add cheminformatics web services to any Django server.

fpserver is commercial open source under the BSD license.

1. A partial list of toolkits or toolkit providers include: Open Babel, RDKit, Indigo, CDK, Daylight, OpenEye, CACTVS, and ChemAxon.
2. A partial list of database extensions or extension providers include: Bingo, OrChem, MyChem, ChemAxon, Daylight, OpenEye, and Accelrys.
3. SQLite - http://sqlite.org/ (accessed Feb 14, 2014).
4. chemfp - http://chemfp.com/ (accessed Feb 14, 2014).
5. Structure Query Collection - https://bitbucket.org/dalke/sqc (accessed Feb 14, 2014).

P-12: Molecular Structure Similarity in the Context of Orphan Drug Legislation

<u>Pedro Franco</u>[1], N. Porta[2], John Holliday[1], Peter Willett[1]

[1]University of Sheffield, Sheffield, UK, [2]European Medicines Agency, London, UK

An orphan drug is a medicinal product that is intended for the treatment of a rare disease that affects only a small number of patients, e.g., five in ten-thousand. According to the current European orphan drug legislation [1], the Community and the Member States shall not, for a period of 10 years, accept another orphan medicinal product, or accept an application to extend an existing marketing authorisation, for the same therapeutic indication, in respect of a similar medicinal product. Thus far, the European Medicines Agency, the regulatory authority, has used human judgments of similarity when assessing new medicines for rare diseases. The project reported here seeks to develop quantitative methods for this purpose.

The project began with an analysis of the correlation between human and computed judgments of similarity for 100 pairs of molecules chosen from the Drug Bank 3.0 database [2,3]. The human similarity assessments for these pairs of molecules were obtained from a total of 143 experts from Asia, Europe and the USA, with the experts being asked to state whether each pair was, or was not, similar. The fraction of the experts judging a pair to be similar was then compared with the Tanimoto coefficient computed using a range of different types of 2D fingerprints, with the aim of identifying those fingerprints that correlated most closely with the human judgments.

The following types of fingerprint were studied: ECFP4, ECFC4, Daylight, Unity, BCI, MDL as implemented in the Pipeline Pilot system; and Extended, Standard, Estate, PubChem, MACCS, Morgan, Feat Morgan, Atom Pair, Torsion, RDKit, Avelon and Layers, as implemented in the KNIME system. 1D molecular property descriptors were also studied but these proved to be of only limited effectiveness for this application. Logistic regression models were developed for each type of fingerprint, relating the Tanimoto similarity for a pair of molecules computed using that fingerprint with the probability that a majority of the human experts would regard that pair as being similar. The resulting regression models were then validated using a separate test-set containing 100 pairs of molecules that had previously been considered for the assignment of orphan-drug status by the European Medicines Agency. The best models (those based on the BCI, Daylight, Unity, ECFP4, Extended, Standard, Morgan, Feat Morgan, Torsion, Avalon and Layers fingerprints) were able to reproduce over 95% of the human judgments. This success rate was increased to 98% using a simple data fusion approach in which a pair of molecules is classified as similar (or non-similar) when three or more of the individual fingerprints are in agreement.

The results obtained here suggest that computed Tanimoto values could provide a useful source of information when assessing new active substances that are being proposed for the treatment of rare diseases.

1. Regulation (EC) No. 141/2000 of the European Parliament and Council of 16 December 1999.
2. Drug Bank, Drug Bank 3.0, 2011, available online at http://www.drugbank.ca/ (accessed Jan 7th, 2014).
3. Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Branco, K.; Mak, C.; Neveu, V.; Djoumbou, Y.; Eisner, R.; Guo, A. C.; Wishart, D. S. *Nuc. Acids Res.* **2011**, 39, 1035–1041.

P-14: QSARINS: software for the development, analysis and validation of MLR models, and QSARINS-Chem: Insubria datasets and QSA(P)R models for environmental pollutants

Paola Gramatica, Nicola Chirico, Alessandro Sangion, Stefano Cassani

QSAR Research Unit in Environmental Chemistry and Ecotoxicology, Department of Theoretical and Applied Sciences (DiSTA), University of Insubria, Varese, Italy.

QSA(P)R models, when correctly developed and rigorously validated, are highly useful for screening and prioritizing chemicals without experimental data, even before their synthesis: this can be done in the ”benign by design” approach in green chemistry. Their use in regulation is strongly suggested by the European legislation of chemicals REACH, in particular to reduce tests on animals and experimental costs. Recently, particular attention has been devoted to the validation of QSAR models, and the “OECD principles for the validation of QSARs models for their application in regulation” [1] have been established to increase the reliability of data predicted by QSAR models. We here propose the new software QSARINS (QSAR-INSUBRIA) [2], for the development of Multiple Linear Regression (MLR) models, by Ordinary Least Squares (OLS) and Genetic Algorithm (GA) for variable selection. This program is mainly focused on internal and external validation of models by different statistical parameters, and is a user-friendly platform for QSAR modeling in agreement with the OECD Principles and for the analysis of reliability of the predictions. Additional features implemented include tools for explorative analysis of the datasets by Principal Component Analysis (PCA), dataset splitting, Applicability Domain analysis (e.g., detection of outliers and interpolated or extrapolated predictions), consensus modelling, selection of the best model by Multi-Criteria Decision Making (MCDM) and various informative and useful plots. QSARINS-Chem, a specific module of QSARINS, includes several datasets of environmental pollutants with the chemical structures (Hyperchem and MDL MOL formats) and the corresponding end-points (physico-chemical properties and biological activities), modeled by Insubria group during the last fifteen years. The chemicals with the related available data can be accessed in different ways (by CAS RN, SMILES, names, etc.) and their 3D structure can be visualized. Additionally, some QSAR models based on molecular 0-2D descriptors calculated by the free open source software PaDEL-Descriptor [3] are implemented in QSARINS-Chem. Among them, there is the Insubria Persistent Bioaccumulative and Toxic (PBT) Index model for the prediction of the cumulative behavior of new chemicals as PBTs. The new PaDEL-Descriptor models can be easily applied for future predictions on chemicals without experimental data, checking the Applicability Domain. The QMRF of all these PaDEL-Descriptor models is available. QSARINS-Chem can be also used as a management tool of personal datasets and models and additional chemometric analyses can be done by PCA and MCDM for screening and ranking chemicals in order to prioritize the most dangerous.

1. OECD Principles, **2004**, Available online at: http://www.oecd.org/dataoecd/33/37/37849783.pdf (accessed Jan 28, 2014).
2. Gramatica, P.; Chirico, N.; Papa, E.; Cassani, S.; Kovarich, S. QSARINS: A new software for the development, analysis, and validation of QSAR MLR models. *J. Comput. Chem. (Software News and Updates)*. **2013**, *34*, 2121–2132.
3. Yap, C. W. PaDEL-Descriptor: An open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* **2011**, *32*, 1466–1474. Available online at: http://padel.nus.edu.sg/software/padeldescriptor/index.html (accessed Jan 28, 2014).

P-16: A fragment-based computational approach to study the phase behavior of bio(polymers) and drug excipients

Jan-Willem Handgraaf[1], Monica Bulacu[1], Rubèn Serral Gracià[1], Joanne Klein Wolterink[1], Johannes G. E. M. Fraaije[1,2]

[1]Culgi B.V., Leiden, The Netherlands, [2]Leiden University, Leiden, The Netherlands

The computational study of molecular processes in nature, such as polymer dynamics or aggregation, are for the most part still out of reach for conventional simulation methods such as Molecular Dynamics or Monte Carlo. Here we present a fragmentation engine designed to handle in principle any molecular architecture. The methodology is based on the well-known notion that atomic groups in molecules can be lumped into single particles, or “beads” [1]. This so-called “coarse-graining” allows for (much) larger length and time scales that can typically be attained by conventional simulation methods.

The problem of fragmenting an atomistally detailed molecule is transformed to a global minimization problem. A scoring function determines how good or bad a given fragmentation is. Then the algorithm is reduced to finding the global minimum of the scoring function. The global minimization is performed by a Monte Carlo evolution where the lowest scoring fragmentation is stored [2]. This typically succeeds finding the global minimum for a given molecular architecture within a reasonable amount of time.

From a fundamental chemical informatics point of view, it is interesting to note that that we can represent a collection of different molecules by a much smaller set of distinct fragments, and that the frequency distribution of fragments follows Heap's power law. Extrapolation indicates that one could fragment the entire PubChem database into a mere few thousands distinct fragments. The great practical advantage is that in this way, by pre-parameterization, one can in principle speed up molecular affinity calculations of thermodynamic accuracy by many orders of magnitude, while still maintaining chemical specificity. Thus we eliminate the need for Quantitative Structure–Activity Relationship models or other ad hoc scoring function approaches.

So far we have successfully applied the fragment-based computational approach to molecular systems and architectures found in the chemical, oil and pharmaceutical industry. The method has been refined and validated against industrially relevant data. Here we will demonstrate the versatility of the methodology by applying it to polymer adhesives [3], drug-excipient phase diagrams, drug-peptide interactions, the biopolymer lignin, and a collection of drugs for cancer therapies.

1. Carbone, P.; Karimi-Varzaneh, H. A.; Müller-Plathe, F. *Faraday Discuss*. **2010**, *144*, 25–42. And references therein.
2. Fraaije, J. G. E. M., Nath, S. K., van Male, J., Becherer, P., Klein Wolterink, J., Handgraaf, J.-W., Case F., Tanase,, C. Serral Gracià, R. Culgi Manual version 8.0, www.culgi.com (2013) (ISBN: 978-90-817846-0-3).
3. Handgraaf, J.-W., Serral Gracia, R., Nath, S K., Chen Z., Chou, S.-H., Ross, R. B., Schultz, N. E., Fraaije, J. G. M. E., *Macromolecules* **2011**, *44*, 1053-1061.

P-18: MONA – Intuitive, visual navigation through molecule collections

Matthias Hilbig, Therese Inhester, Matthias Rarey

Zentrum für Bioinformatik, Hamburg, Germany

The visual inspection, filtering and preparation of organic small-molecules is at the core of many cheminformatics applications. Pipelining tools like KNIME [1] or Pipeline Pilot [2] are the methods of choice when all steps are known in advance and no intervention is necessary. MONA [3] on the other hand is a GUI Tool which supports intuition-driven processing of molecule collections. Molecules are organized in sets which can be visualized as tables of depictions or histogram plots of single properties as well as easily combined with classical set operations like union, intersection, or complement.

To better facilitate the visual analysis of sets, we present a new view for sets in MONA. While recent visualization methods like molecule clouds [4] or scaffold trees [5] calculate a large static arrangement of molecules in advance, MONA allows to interactively create clusterings by molecular similarity and to navigate them individually. A two-dimensional arrangement enables fast browsing through cluster representatives as well as through cluster members. In order to easily detect small structural differences between all molecules in one cluster, a local search optimization algorithm was developed which aligns two-dimensional depictions on the fly.

Another often requested feature is the ability to handle arbitrary numerical SDF properties which are of importance for postprocessing screening data. The already present filtering and sorting routines were therefore extended to handle any numerical value annotated in the SDF, e.g. bioaffinity data, measured experimental physical-chemical properties, price informations or in-house IDs. Additional SDF properties can be imported on a case by case basis and used in MONA to filter sets and to sort clusters or sets.

Great care was taken in integrating this new visualization method as seamless as possible without disturbing the intuitively understandable user interface of MONA. Several use cases of MONA ranging from library preprocessing, comparing large datasets to screening postprocessing will be presented.

1. Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: The Konstanz Information Miner. In *Data Analysis, Machine Learning and Applications*; Preisach, C.; Burkhardt, P. D. H.; Schmidt-Thieme, P. D. L.; Decker, P. D. R., Eds.; Studies in Classification, Data Analysis, and Knowledge Organization; Springer Berlin Heidelberg, 2008; pp. 319–326.
2. Accelrys Software Inc: Pipeline Pilot 9.0. 2013
3. Hilbig, M.; Urbaczek, S.; Groth, I.; Heuser, S.; Rarey, M. MONA – Interactive Manipulation of Molecule Collections. *Journal of Cheminformatics* **2013**, *5*, 38.
4. Ertl, P.; Rohde, B. The Molecule Cloud - Compact Visualization of Large Collections of Molecules. *J. Cheminf.* **2012**, *4*, 12.
5. Wetzel, S.; Klein, K.; Renner, S.; Rauh, D.; Oprea, T. I.; Mutzel, P.; Waldmann, H. Interactive Exploration of Chemical Space with Scaffold Hunter. *Nat Chem Biol* **2009**, *5*, 581–583.

P-20: Ushering the Cactvs Toolkit into the Python Age (without breaking the legacy)

Wolf D. Ihlenfeldt

Xemistry GmbH, Königstein; Germany

The Cactvs Chemoinformatics Toolkit is probably the most powerful general-purpose chemical information processing toolkit on the market. Since its inception about twenty years ago, its main language for rapid script development has been Tcl – at that time a language at the forefront of lazily-typed rapid prototyping and interface programming language design.

While Tcl is still actively maintained, and does provide features not matched by many of the nowadays more popular competitors – most notably impressive multi-threading capabilities, which are fully accessible from within the toolkit – history has passed on. Tcl has undeniably fallen out of the public eye, and there is an understandable reluctance by users to learn new languages which are effectively only used by one of their tools.

This problem has finally been addressed. The Cactvs toolkit is now available with Python as a second alternative (or parallel) interface language. The new interface closely follows the established Tcl command patterns to support easy migration by experienced users, while still providing true "pythonesque" constructs. Since significant functionality of the toolkit is implemented as external Tcl script function snippets, and future enhancements will probably preferably be coded in Python without providing also a Tcl port, providing automatic and fully transparent access to language-mismatched components has been an important and rather peculiar design goal.

Examples of the new toolkit scripting capabilities shall be presented, as well as a documentation of the challenges involved in the design of a parallel multi-language interface to a large software system.

Acknowledgment: We gratefully acknowledge support by Vertex Pharmaceuticals, Inc. for this project.

P-22: Combinatorial Library Optimization Process In The European Lead Factory Project

Tuomo Kalliokoski

Medicinal Chemistry, Lead Discovery Center GmbH, Dortmund, Germany

The European Lead Factory (ELF) is a 60 month, 196 million EUR project which has started on January 2013 [1]. There are 30 partners from both academia and the pharmaceutical industry. The aim of the project is to combine large combinatorial library with High-Throughput Screening (HTS)-facility. The 500,000-membered compound library is named as Joint European Compound Collection (JECC). JECC consists of 300,000 compounds contributed the pharmaceutical companies Bayer, AstraZeneca, Lundbeck, Janssen, Merck, Sanofi and UCB. The additional 200,000 compounds are to be synthesized during the project by European academics and small- and medium-sized enterprises (SMEs). The proposals for the new compound libraries are scored by a committee against calculated molecular properties, structural features, novelty, diversity, synthetic tractability and innovation.

After library proposal has been accepted by the committee, the initial library designs from the academic partners and SMEs are optimized before the production phase at Lead Discovery Center so that resulting combinatorial libraries are chemically diverse as possible and still have suitable physicochemical properties. This presentation gives an introduction to the ELF project and a detailed description of the library optimization process employed.

1. Mullard, A. European Lead Factory opens for business. *Nat. Rev. Drug Discov.* **2013**, *12*, 173–175.

P-24: Reaction Representation and Structure Transformation with Ambit-SMIRKS. Application in Metabolite Prediction

Nikolay T. Kochev[1], Svetlana L. Avramova[1], Nina G. Jeliazkova[2]

[1] University of Plovdiv, Department of Analytical Chemistry and Computer Chemistry, 24 Tsar Assen St. Plovdiv, Bulgaria,
[2] Ideaconsult Ltd, 4 A. Kanchev str., Sofia 1000, Bulgaria

Ambit-SMIRKS is an extension of the Ambit-SMARTS Java library [1], both part of the Ambit2 project [2]. The modules are implemented on top of Chemistry Development Kit (CDK) [3]. Ambit-SMIRKS performs two main tasks: (1) parsing of SMIRKS linear notations into internal reaction (transformation) representations based on CDK objects and (2) application of the stored reactions against target molecules for actual transformation of the target chemical objects. The transformations can be applied on various sites of the target molecule in several modes: single, non-overlapping, non-identical, non-homomorphic or externally specified list of sites. Ambit-SMARTS implements the entire SMARTS language specification as defined by the Daylight, plus additional syntax extensions to make software compliant with SMARTS modifications made by third party software packages such as OpenEye, MOE and OpenBabel. These extensions can be switched on/off by a customer request. The SMIRKS library utilizes the Ambit-SMARTS parser and the efficient substructure searching algorithm implemented within Ambit-SMARTS package [1]. Typically most SMIRKS implementations support SMILES plus simple SMARTS syntax features. However, Ambit-SMIRKS module supports full SMARTS syntax for reactions specification.

The SMIRKS module is used to enable metabolite predictions in Toxtree (since version 2.5.0) [4], once these sites of metabolisms are predicted by SMARTCyp [5]. Toxtree is a flexible and user-friendly open-source application that predicts various kinds of toxic effects, mostly by applying structural alerts, arranged in a decision tree fashion. SMARTCyp (Cytochrome P450-Mediated Drug Metabolism) model is originally developed by Patrik Rydberg et al. [5] and was included as Toxtree module since Toxtree 2.1.0.

Ambit-SMIRKS is available as a Java library and as OpenTox Algorithm API compatible Web service. A web page facilitating SMIRKS testing is available at http://apps.ideaconsult.net:8080/ambit2/depict/reaction and an online version of Toxtree (including site of metabolism and metabolite predictions) is available at http://toxtree.sourceforge.net/predict.

1. Jeliazkova, N.; Kochev, N. AMBIT-SMARTS: Efficient Searching of Chemical Structures and Fragments. *Mol. Inf.* **2011**, *30*, 707–720.
2. Jeliazkova N., Jeliazkov V. AMBIT RESTful web services: an implementation of the OpenTox application programming interface. *J. Cheminf.* **2011**, *3*, 18.
3. Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493–500.
4. http://toxtree.sourceforge.net last accessed February 10, 2014.
5. Rydberg, P.; Gloriam, D.; Zaretzki, J.; Breneman, C.; Olsen, L. SMARTCyp: A 2D Method for Prediction of Cytochrome P450-Mediated Drug Metabolism. *ACS Med. Chem. Lett.* **2010**, *1*, 96-100.

P-26: Multi-label Drug-like Compounds and Probabilistic Graphical Model Using Variational Mean Field Theory

Hamse Y. Mussa, Robert C. Glen

Unilever Center for Molecular Sciences Informatics, Cambridge, UK

Employing ligand-based screening approaches to devise multi-target protein prediction (multi-class classification) models is ubiquitous in current cheminformatics. However, these classification approaches are based on an unrealistic and somewhat simplistic assumption: each ligand (drug-like compound) can interact with only one specific protein out of the multitude of proteins in a living cell. This means that in our cheminformatics classification context each compound belongs only to a single class, i.e., a ligand has a single-label [1].

In nature, however, a ligand can typically simultaneously interact with multiple proteins in a living cell, irrespective of whether or not these proteins are the desired target protein(s). Furthermore, the ligand promiscuity – its multi-label aspect – can be beneficial to human health; it can also trigger lethal side effects.

Thus, it is extremely important that the multi-label nature of a ligand is taken into consideration when developing *in silico* multi-target protein prediction (multi-class classification) models. In recent years, there has been a surge in the application of the multi-label concept in numerous pattern recognition disciplines, such as text mining, scene analysis, bioinformatics, etc [1]. However, the same cannot be said about ligand-based screening approaches in cheminformatics.

In this work we present a multi-label multi-class classification scheme for classifying multi-label ligands. The presented model is based on the powerful (probabilistic) graphical method [2] trained using a variational mean field theory [3]. In order to assess the predictive power of the proposed classifier, it has been built and tested on a large chemical dataset comprising over 120,000 unique compounds distributed over 100 target proteins, whereby each compound has been annotated against more than one target protein. The generalization ability of the new algorithm is compared with the performance of the more "traditional" single-label multi-class classifiers currently employed in cheminformatics.

1. Tsoumakas, G.; Katakis, I. Multi-label Classification: An Overview. *International Journal of Data Warehousing and Mining* **2007**, *3*, 1–17.
2. Koller, D.; Friedman, N. *Probabilistic Graphical Models: Principles and Techniques,* 1st ed.; Springer-Verlag: New York, 2006.
3. Yedidia, J.S. An Idiosyncratic Journey Beyond Mean Field Theory. In *Advanced Mean Field Methods: Theory and Practice*; Opper, M.; Saad, D., Ed.; MIT Press: MA, 2001; pp 37.

P-28: DataWarrior, a Free Tool for Chemistry Aware Data Visualization and Analysis

Thomas Sander, Modest von Korff, Joel Freyss

Actelion Ltd., Gewerbestrasse 16, CH-4123 Allschwil, Switzerland

In the pharmaceutical industry as well as in science in general we notice a paradigm shift [1] from hypothesis driven reasoning to data driven approaches. Data mining, machine learning [2] and data visualization [3,4] are increasingly used to extract the knowledge hidden in ever growing data sets. The availability of a multitude of software tools is both, evidence and driver of this evolution. And yet, while existing software covers multidimensional alphanumerical data rather well, it is still not an easy task using software to correlate biological effects with the chemical structures causing the effects. In an attempt to contribute filling the gap, we release our in-house developed chemistry aware data visualization and analysis program DataWarrior for free public use.

DataWarrior was developed as part of Actelion's drug discovery software platform OSIRIS [5]. It combines dynamic graphical views and interactive row filtering with chemical intelligence. Scatter plots, box plots, bar charts and pie charts not only visualize numerical or category data, but also show trends of multiple scaffolds or compound substitution patterns. Chemical descriptors encode various aspects of chemical structures, e.g. the chemical graph, chemical functionality from a synthetic chemist's point of view or 3-dimensional pharmacophore features [6]. These allow for fundamentally different types of molecular similarity measures, which can be applied for many purposes including row filtering and the customization of graphical views. DataWarrior supports the enumeration of combinatorial libraries as the creation of evolutionary libraries. Compounds can be clustered and diverse subsets can be picked. Calculated compound similarities can be used for multidimensional scaling methods, e.g. Kohonen nets. Physicochemical properties can be calculated, structure activity relationship tables can be created and activity cliffs be visualized.

In this contribution some of DataWarrior's visualization and analysis features will be demonstrated using chemical and biological sample data.

1. The Forth Paradigm: Data-Intensive Scientific Discovery; Hey, T.; Tansley, S.; Tolle, K., Eds.; Microsoft Research: 2009; ISBN-13: 9780982544204.
2. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H. The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl.* **2009**, *11*, 10–18.
3. TIBCO Spotfire; TIBCO Software Inc.; 3303 Hillview Avenue, Palo Alto, CA 94304, USA.
4. Vortex; Dotmatics; Bishops Stortford, Herts, CM23 2ND, UK.
5. Sander, T., Freyss, J.; Korff, M. v.; Reich; J.R.; Rufener, C. OSIRIS, an entirely in-house developed drug discovery informatics system. *J. Chem. Inf. Model.* **2009**, *49*, 232–246.
6. Korff, M. v.; Freyss, J.; Sander, T. Flexophore, a new versatile 3D pharmacophore descriptor that considers molecular flexibility. *J. Chem. Inf. Model.* **2008**, *48*, 797–810.

P-30: Expanding chemical space coverage in Matched Molecular Pairs Analysis

Peter Schmidtke, Vincent Le Guilloux

Discngine, Paris, France

Already since the 1980's Matched Molecular Pairs Analysis (MMPA) has been used to analyze structure/property data in a medicinal chemistry and later chemoinformatics context. Despite the age of the first MMPAs, rather few conceptual novelties have been introduced since that time. As a consequence, current use of MMPA is still restricted to molecular property prediction and usually unsuitable for structure activity prediction.

Furthermore, rather than using MMPA as a pure "catalog of already completed chemical transformations" few attempts have been made to use MMPA to provide a more general predictive model for property as well as activity prediction.

Here our ongoing developments on MMPA are shown. Using the previously presented fcsMMP approach [1], we are able to create fuzzy context specific transformation rules on medium sized datasets based on reduced graphs [2]. Integrating fuzziness to represent the common core of a matched molecular pair (MMP) allows for more stringent and accurate transformation rules using the full common core (or context), rather than considering only a local context around the attachment point. Furthermore, evidence is shown that such fuzziness can be used to expand the chemical space represented by a classical MMPA without losing the precision gained by considering the full context. This allows us to build more accurate predictive models that strictly follow the principles of basic MMPA regarding chemical interpretability.

Last, results obtained by using fuzzy fragments in MMPs are shown. Here again, probabilistic prediction results are shown on the outcome of chemical transformations not contained in the original data, allowing thus for a more accurate prospective solution to the "what to do next?" challenge.

1. Schmidtke, P. & Le Guilloux, V. Fuzzy Context Specific Matched Molecular Pairs. *J. Cheminf.* **2014**, *6 (Suppl. 1)*, P44.
2. Birchall, K.; Gillet, V. J. Reduced graphs and their applications in chemoinformatics. *Methods Mol. Biol.* **2011**, *672*, 197–212.

P-32: Plexus – A Flexible Library Design Platform to Conduct Innovative Chemistry

Krisztian Niesz, Ivan Solt, Andras Stracz

ChemAxon Ltd, Budapest, Hungary

Here we describe Plexus, ChemAxon's new discovery platform providing easy to use cheminformatics tools for the bench researchers. Plexus allows chemists to easily get on the most common virtual library design processes, including compound enumeration, property based filtering, pharmacophore similarity searching and clustering as simple and straightforward as possible.

ChemAxon's Plexus aims to give chemists powerful cheminformatics capabilities, advanced visualization and analytics tools within a series of intuitive interfaces which simplify and speed up their everyday research. Under the hood it is wiring all of ChemAxon's extensive and industry leading cheminformatics solutions and presenting a series of workflows to let researches more quickly and easily design, evaluate and take action on new chemical entities.

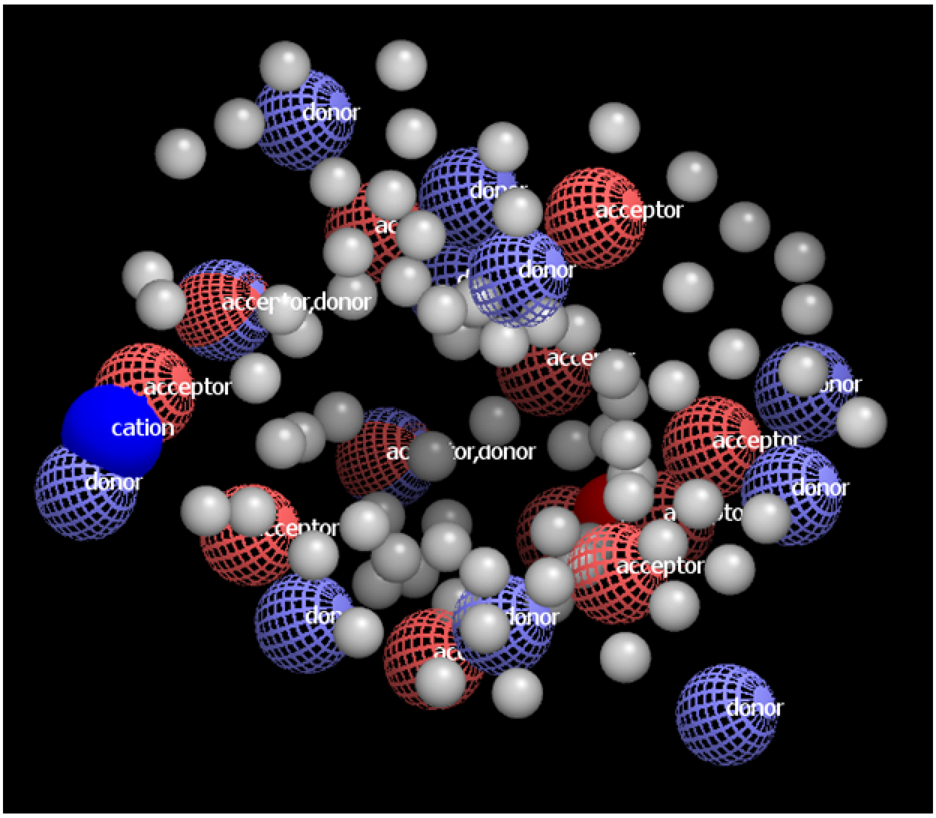
As a desktop/web application, Plexus can be installed for individual users on all major devices and operating systems, but can also be centrally deployed as a collaborative research environment.

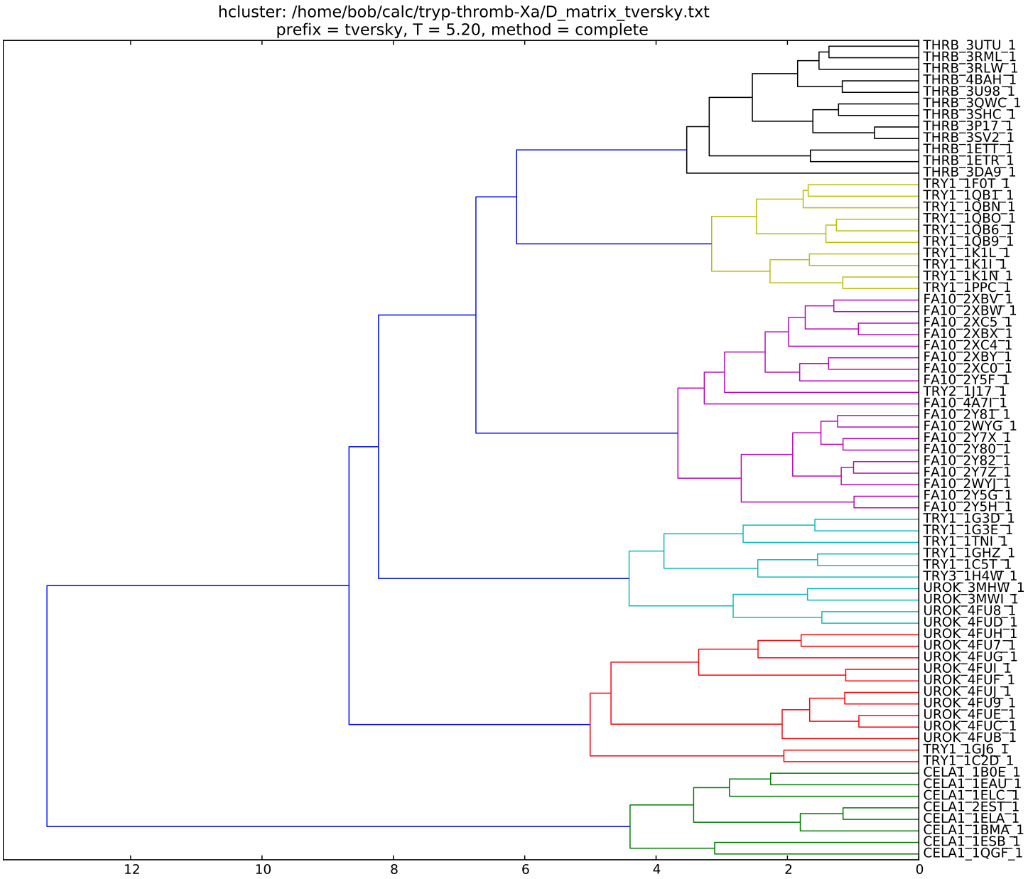
P-34: Protein active site comparison with SiteHopper: phylogeny to polypharmacology

Paul Hawkins, Gregory Warren

OpenEye Scientific, Santa Fe, USA

There is a long history of using sequence alignment data to understand evolutionary relationships [1]. More recently attempts to use sequence alignment and comparison to predict cross-reactivity and polypharmacology have been made with varying degrees of success [2]. We present a new method, SiteHopper, which rapidly aligns and compares a three-diminsional representation of protein active or binding sites. This method is expected to show superior performance to sequence comparison in compound cross-reactivity/polypharmacology versus sequence because it directly compares the shape and underlying chemistry of different protein binding sites. Case studies will be presented to show that SiteHopper is able to find similarity between binding sites for targets with very different sequences.





1. Lotytynoja, A.; Goldman, N. Phylogeny-Aware Gap Placement Prevents Errors in Sequence Alignment and Evolutionary Analysis. *Science*, **2008**, *320*, 1632–1635.
2. Kalinina, O.V. *et al.* SDPpred: a tool for prediction of amino acid residues that determine differences in functional specificity of homologous proteins. *Nucleic Acids Res.*, **2004**, *32 (Web Server issue)*, W424–W428.

P-36: Assembly of Helical Chimeras with Natural Peptides: Simulating Molecular Interaction Surfaces

Jeremie Mortier[1], **Elisabeth K. Nyakatura**[1], **Raheleh Rezaei Araghi**[1], **Sebastian Wieczorek**[1], **Carsten Baldauf**[2], **Gerhard Wolber**[1], **Beate Koksch**[1]

[1]*Department of Biology, Chemistry and Pharmacy, Freie Universität Berlin, Germany,* [2]*Fritz-Haber-Institute der MPG, Berlin, Germany*

Peptide foldamers with specific biological activities are of substantial interest as modulators of protein-protein interactions. Such structures can be designed substituting α-amino acids by β- and γ-amino acids in a peptide sequence (Figure 1A and 1B) [1], although this generally affects its overall structure and topology. In this work, ideal core packing between natural α-helical peptides and an αβγ-chimera peptide that forms coiled coil was identified by phage display. Selected peptides assembled with the chimeric sequence in tetrameric complexes, showing thermal stabilities that are comparable to that of the parent bundle consisting purely of α-amino acids.

3D-models of the selected heteromers were investigated by molecular dynamics simulations and compared to the parent complex made of natural α-residues (Figure 1C). This study led to the identification of atypical interhelical interactions between the selected residues and the β/γ-segment, providing insight into the outstanding stability of the new heterodimers [2]. Gained results are of high interest for the design of unnatural oligomers that resemble protein structures and can interfere with protein-protein interactions.

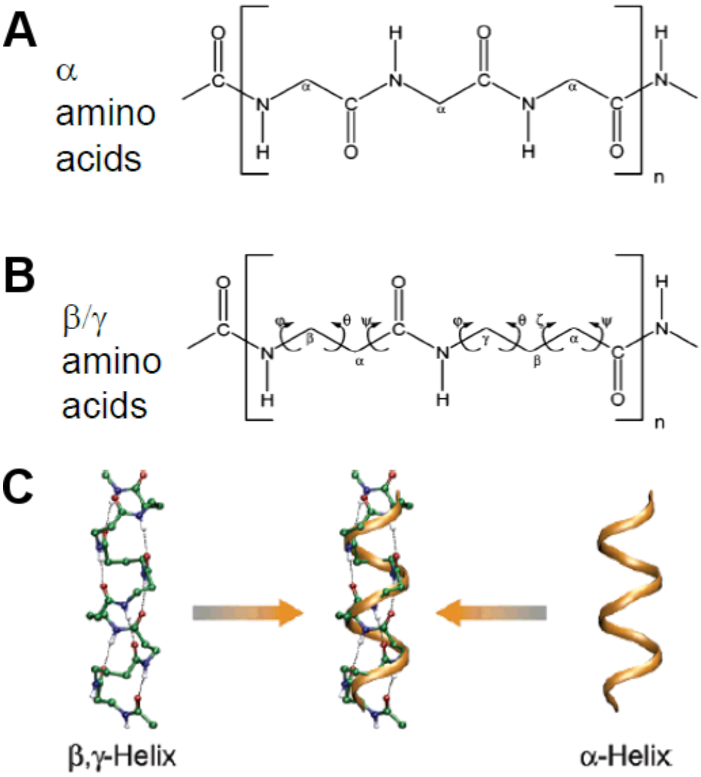


Figure 1: **A**. 2D-structure of α- and β/γ-residues. **B**. Superposition of a β/γ-helix and an α-helix in 3D. **C**. Heterotetramer coiled coil with two αβγ-chimera peptides (β- and γ-residues shown as sticks).

1. Baldauf, C.; Gunther, R.; Hofmann, H. J. Helix formation in α,γ- and β,γ-hybrid peptides: theoretical insights into mimicry of α- and β-peptides. *J. Org. Chem.* **2006**, *71*, 1200–1208.
2. Nyakatura, E. K.; Rezaei Araghi, R.; Mortier, J.; Wieczorek, S.; Baldauf, C.; Wolber, G.; Koksch, B. An Unusual Interstrand H-Bond Stabilizes the Heteroassembly of Helical αβγ-Chimeras with Natural Peptides. *ACS Chem. Biol.* **2014**.

P-38: Molecular simulations of peptides and proteins with Molecular Fragment Dynamics

Andreas Truszkowski[1], Annamaria Fiethen[2],Hubert Kuhn[2], Achim Zielesny[3], Matthias Epple[1]

[1]Inorganic Chemistry and Center for Nanointegration, University of Duisburg-Essen, Essen, Germany, [2]CAM-D Technologies, Essen, Germany, [3]Institute for Bioinformatics and Chemoinformatics, Westphalian University of Applied Sciences, Recklinghausen, Germany

Molecular Fragment Dynamics (MFD) is a mesoscopic simulation technique based on Dissipative Particle Dynamics (DPD). Whereas DPD beads in general may not necessarily be identified with chemical compounds at all the MFD variant uses specific molecules or molecular fragments as its basic interacting entities. MFD has been successfully applied for studying formulations of surfactants and polymers[1-3] but has yet not been utilized for systems containing biopolymers like peptides or proteins. Therefore it is the aim of this study to extend the MFD technique to the biomolecular realm.

A basic requirement for the setup of biomolecular MFD simulations is an editor for the construction of peptides and proteins from molecular fragments. It includes the fragment building blocks for all 20 proteinogenic amino acids, their charged species and disulfide bonds. A flexible input of one-letter and three-letter amino acid codes is supported. In addition amino acid sequences can be obtained from the Protein Data Bank (PDB)[4]: PDB files are comfortably evaluated and analyzed with the open-source library BioJava[5] and graphically displayed with the chemical open-source visualizer Jmol[6]. Since a PDB file contains atomic 3D structure coordinates this spatial information may be used for a corresponding spatial arrangement of the backbone fragments. Last but not least the editor has the capability to set charges of amino acid side chains in a manual or automatic manner.

The 3D structures of peptides and proteins are stabilized by anisotropic interactions between polar parts of their molecular structures like hydrogen bonds. However, the MFD technique describes fragment interactions by isotropic repulsion parameters only. Thus specific potentials between backbone fragments are introduced which are able to define spatial orientation and stiffness - ranging from full flexibility over partial spatial restrictions up to a fully conserved backbone structure.

1. Ryjkina, Ekaterina, R.; Kuhn, H.; Rehage, H.; Müller, F.; Peggau, J. Molecular dynamic computer simulations of phase behavior of non-ionic surfactants. *Angew. Chem. Int. Ed.* **2002**, *41*, 983–986.
2. Schulz, S. G.; Kuhn, H.; Schmid, G.; Mund, C.; Venzmer, J. Phase behavior of amphiphilic polymers: A dissipative particles dynamics study. *Colloid. Polym. Sci.* **2004**, *283*, 284–290.
3. Truszkowski, A.; Epple, M.; Fiethen, A.; Zielesny, A.; Hubert, K. Molecular fragment dynamics study on the water-air interface behavior of non-ionic polyoxyethylene alkyl ether surfactants. *J. Colloid. Interface. Sci.* **2013**, *410*, 140–145.
4. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
5. Prlic, A.; Yates, A.; Bliven, S.E.; Rose, P.W.; Jacobsen, J.; Troshin, P.V.; Chapman, M.; Gao, J.; Koh, C.H.; Foisy, S.; Holland, R.; Rimsa, G.; Heuer, M.L.; Brandstatter-Muller, H.; Bourne, P.E.; Willis, S. BioJava: an open-source framework for bioinformatics in 2012. *Bioinformatics* **2012**, *28*, 2693–2695.
6. Jmol: an open-source Java viewer for chemical structures in 3D. http://www.jmol.org/ (accessed Jan 27, 2014).

P-40: KinHub: a Database to Enable Kinome-wide Analysis

Sameh Eid[1], Samo Turk[1], Andrea Volkamer[1], Friedrich Rippmann[2], Simone Fulle[1]

[1]BioMed X Innovation Center, Im Neuenheimer Feld 583, 69120 Heidelberg, Germany, [2]Merck KGaA, Merck Serono, Global Computational Chemistry, Frankfurter Str. 250, 64293 Darmstadt, Germany

Human kinases are among the most studied drug targets for a variety of diseases including cancer, inflammation, and auto-immune disorders. Accordingly, a few databases focusing on kinase data [1–3] are available which provide the user with useful information regarding e.g. the structural determinants of kinase binding. However, each of these databases has its strengths but also some limitations concerning availability, searching criteria or crosslinks between data points. In particular, there is lack of a database which integrates 3D structures of the entire kinome, compounds and screening data in one knowledge base.

To fill this gap, we are developing KinHub, a centralized knowledge base comprising a regularly updated and curated database of all human kinase structures in a unified naming scheme. KinHub brings together kinase-relevant information from KinBase [4], PDB, UniProt, and various other sources, in a relational database. Furthermore, KinHub provides a user-friendly interface so that researchers interested in protein kinases can easily search the database for structural information such as available 3D structures, conformational state of the kinase (e.g. DFG loop and αC helix), annotation of active sites, chemical nature of the bound ligands, and relationships between protein kinase structures. Yet unknown human kinase structures are provided as homology models; attempts along this line to model the entire kinome will be presented together with the outline of the database.

Overall, KinHub is suitable to put the pool of structural information to practical use and will be continuously growing with new structural data or classification published.

1. van Linden, O. P.; Kooistra, A. J.; Leurs, R.; de Esch, I. J.; de Graaf, C. KLIFS: A knowledge-based structural database to navigate kinase-ligand interaction space. *J. Med. Chem.* **2013**. doi:10.1021/jm400378w.
2. Kinase SARfari available at https://www.ebi.ac.uk/chembl/sarfari/kinasesarfari.
3. Molecular Operating Environment (MOE), 2013.08, Chemical Computing Group Inc., 1010 Sherbrooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2013.
4. KinBase available at http://kinase.com/kinbase/.

P-42: Lessons learned from 30 years of developing Drug Discovery Informatics Systems

Mic Lajiness, Tom Hagadone

Eli Lilly & Company, Indianapolis, Indiana, USA

One of the most sought after goals in the pharmaceutical industry is to have an integrated cheminformatics database system (CIDBS) that provides scientists with the ability to access and exploit a wide variety of corporate research data to facilitate the discovery and development of new therapeutic agents. In spite of this, it is quite surprising that there are few examples of commercial or proprietary solutions that are well regarded, have stood the test of time and are considered to be successful by their users. Some notable exceptions in this challenging landscape are a series of related in-house developed systems, Cousin, ChemLink, and Mobius that were first put into production in 1981 (Cousin) and continue to the present day (Mobius). The authors are the principal scientists involved in all of these systems and will, in this chapter, describe aspects of their work and highlight many of the lessons learned in the past 30+ years that led to these successful systems.

P-44: Bringing Desktop Solutions to the Mobile Space: the ChemDraw App Story

Pierre Morieux[1], **Philip McHale**[2]

[1]PerkinElmer SAS, Villebon-sur-Yvette, France, [2]PerkinElmer, Waltham, MA, U.S.A.

Over the last decade, mobile technologies and mobile applications have drastically changed the way we live, communicate and interact with one another, enabling us to instantly share pictures with people, book a flight or manage our bank account from virtually anywhere [1]. Mobile devices have also started infiltrating laboratories and universities worldwide [2,3], creating a new generation of online socially active scientists, and the rapid adoption rate of tablets as a mobile platform is certainly a key factor in this development [4]. It is therefore important to better understand and capitalize on what these changes mean for scientific research. Indeed, today's online video streaming and blogging platforms are changing the way scientists are searching for information, exchanging, or even publishing their results [5].

At PerkinElmer we have acquired solid experience in this area in the past year as we have started to actively engage with our customers in the social media space. More importantly we have taken advantage of the new mobile and virtual platforms to create and promote novel, trend-setting applications in the field of chemical information. The release in 2013 of the tablet mobile version of the widely used software application ChemDraw was an important milestone in the era of scientific mobile applications [6–8]. The mobile app was designed to incorporate seamless sharing functionalities to various social media and cloud computing platforms, as well as a novel device-to-device content sharing capability called Flick-to-Share. Specifically, these new ways to capture and share chemical content via tablets have been tested during proof-of-concept pilot studies in American universities and have yielded very interesting results [9]. These results suggest that the Flick-to-Share functionality can revolutionize the way Chemistry is taught in the classroom. In this paper we present the rationale and methodology behind the development of the PerkinElmer mobile applications, and describe how they can improve the way scientists learn, exchange, and work in the laboratory environment.

1. Gibbs, N. Your Life is Fully Mobile, Time Tech [Online] **2012**, http://techland.time.com/2012/08/16/your-life-is-fully-mobile/ (accessed Jan 06, 2014).
2. Pennock, R. Going mobile. *Nature* **2009**, *461*, 1157.
3. Looi, C. K., Seow, P., Zhang, B., So, H. J., Chen, W.; Wong, L. H. Leveraging mobile technology for sustainable seamless learning: a research agenda. *Brit. J. Educ. Technol.* **2010**, *41*, 154–169.
4. Cantrill, S. All you can tweet. *Nat. Chem.* **2013**, *5*, 247.
5. Miah, A. Twitter journal: would you share your original research on social media? The Guardian [Online] **2014**, http://www.theguardian.com/higher-education-network/blog/2014/jan/27/twitter-only-journal-academic-research (accessed Jan 29, 2014).
6. Karlin, M. ChemDraw – Draw & Share Chemical Structures. The Ed Tech Round Up [Online] **2013**, http://www.edtechroundup.org/6/post/2013/07/chemdraw-draw-share-chemical-structures.html (accessed Aug 15, 2013).
7. Lowe, D. A New Version of ChemDraw Mobile. In the Pipeline [Online] **2013**, http://pipeline.corante.com/archives/2013/09/11/a_new_version_of_chemdraw_mobile.php (accessed Sep 12, 2013).

8. Top 10 Innovations: Honorable Mentions. The Scientist [Online] **2013**, http://www.the-scientist.com/?articles.view/articleNo/38571/title/Top-10-Innovations--Honorable-Mentions/ (accessed Dec 02, 2013).
9. Morsch, L. Using ChemDraw for iPad and Flick-to-Share to Increase Engagement in Organic Chemistry. DivCHED Committee on Computers in Chemical Education [Online] **2013**, http://www.ccce.divched.org/P8Fall2013CCCENL (accessed Nov 23, 2013).

P-46: Focus – a global communication platform for applied and theoretical medicinal chemists

Nikolaus Stiefl[1], Donovan Chin[2], Peter Hunt[3], Richard Lewis[1], Mika Lindvall[4], Katrin Spiegel[5], Clayton Springer[2], Yongjin Xu[4], Peter Gedeck[6]

[1]Novartis Institute for Biomedical Research, Basel, Switzerland, [2]Novartis Institute for Biomedical Research, Cambridge, US, [3]Novartis Institute for Biomedical Research, Horsham, UK, [4]Novartis Institute for Biomedical Research, Emeryville, US, [5]AB Science, Paris, France, [6]Novartis Institute for Tropical Diseases, Singapore, Singapore

With *in silico* data and data analysis becoming more and more important for both the applied as well as theoretical medicinal chemists creating a work environment that can be used by all team members is a crucial part of local and global collaboration. However, often users are overwhelmed by the amount of new software being introduced and hence cannot afford to spent too much time on learning complicated tools.

Here we present a software tool developed and optimized at NIBR since 2006 based on the ICM software suite [1] called FOCUS. It is built around tasks that can be assigned to specific modules like Data Handling, Design, Analysis, etc. The modules are built in a workflow type style so users can easily mix and match depending on their specific project needs without too much training.

FOCUS is tightly integrated with internal services that comprise – among others - data retrieval systems, *in silico* models, as well as simple automatized modelling procedures such as pharmacophore searches, R-group analysis, and similarity searches. In addition, an interactive 3D editor was implemented in the underlying software for simple compound analoguing. From a technical perspective, FOCUS is based on a mix of the ICM scripting language and standard technologies such as javascript, HTML, and web services, respectively. Hence, FOCUS sessions can be highly optimized for specific project needs by experts from the computational community.

1. http://www.molsoft.com.

P-48: Identification of novel potential anti-cancer agents using network pharmacology based computational modelling

Benjamin C. P. Allen, Victoria Flores

e-Therapeutics plc, Oxford, UK

Network pharmacology models cells as networks of interacting proteins. Within this paradigm, a disease state is identified as a disorder of large-scale sub-networks, and the target for pharmacological intervention becomes a set of proteins. Network modelling can be used to identify a set of key interactions which can be targeted to correct the network disorder. The chemoinformatic problem is then to identify compounds with desirable interactions with the set of proteins; multi-target drugs.

At e-Therapeutics we have developed a workflow based on this model, using a combination of open-source and proprietary databases and in-house proprietary chemoinformatics tools. Our current lead anti-cancer asset, dexanabinol, was identified from its network interactions, and is now completing Phase I clinical trials. Using dexanabinol's properties and its interactions with the network of cancer related proteins in the cell as a template, this approach has been used to identify a further set of novel anti-cancer compounds. 85 compounds have been tested against a panel of 3 cancer cell lines, and 51 show some activity at the 100µm level, with 15 active at below 15µm. These results demonstrate that the network pharmacology approach is capable of generating significantly enriched sets of compounds, and is a valuable tool for early stage drug discovery.

P-50: Development of Medical Countermeasures against Organophosphorous Intoxication; From High Throughput Screening to Quantitative Structure-Activity Relationships

Susanne Johansson[1],Christine Akfur[1], David Andersson[2], Cecilia Lindgren[2], Lotta Berg[2], Weixing Qian[2], Anna Linusson[2], Fredrik Ekström[1]

[1]Swedish Defence Research Agency, CBRN Defence and Security, SE-901 82 Umeå, Sweden, [2]Department of Chemistry, Umeå University, SE-901 87 Umeå, Sweden

Organophosphorous compounds e.g. insecticides, pesticides and nerve agents inhibit acetylcholinesterase (AChE) by covalently binding to the catalytic serine residue of the enzyme, and causes acute toxicity or death of the intoxicated individual. Medical countermeasures such as reactivators restore the enzymatic activity by cleaving the bond between the phosphorous conjugate and the enzyme. The efficacy of reactivators depends on several factors including the properties of the inhibitor, nucleophilicity of the reactivator, steric and electronic factors that influence approach of the phosphorous atom by the nucleophile, and the affinity for the inhibited enzyme. In our research we combine biochemical-, structural- and computational techniques to identify new chemical leads and to develop rational methods for their conversion to reactivators. The work presented herein is based on an *in-vitro* high throughput screening (HTS) of 17 500 drug-like compounds, where 124 novel inhibitor "hits" of AChE were identified. The hits display a large diversity in physicochemical properties, e.g., size, flexibility, polarity and acidity, and they span a new and significantly larger chemical space than previously described AChE reactivators. Based on the HTS-hits 18 structurally analogous compounds, statistically balanced in terms of physicochemical properties, were designed and synthesized to establish a relationship between the chemical structures and affinity to AChE. Using calculated physicochemical properties to describe the structural features of the 18 compounds, and an activity based assay to measure their affinity to AChE, a quantitative-structure activity relationship (QSAR) for ligand binding to *apo*-AChE has been developed. By applying an activity-independent time correlated single photon counting (TCSPC) based assay, QSAR models for binding to nerve agent inhibited AChE have also been established. Noteworthy is that the QSAR models predict different physicochemical properties to be important for affinity to *apo*-AChE and nerve agent inhibited AChE. The successful design and synthesis of a potential broad-spectrum reactivator verify our hypothesis that the HTS-hits, together with detailed knowledge of their molecular interactions and SAR, serves as chemical starting points for the development of new reactivators of organophosphorous inhibited AChE.

P-52: Surflex-QMOD: Protein Pocket Modeling for Affinity Prediction

Rocco Varela[1], Ajay N. Jain[2], Alexander Steudle[3]

[1]Certara, St. Louis, MO, United States, [2]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA, United States, [3]Certara, Martin-Kollar-Str. 17, 81829 München, Germany

Computational approaches for binding affinity prediction are most frequently demonstrated through cross-validation within a series of molecules or through performance shown on a blind test set. Previous reports of the Surflex-QMOD approach demonstrated its ability to produce accurate and scaffold-independent predictions of binding affinity by constructing an interpretable physical model of a binding site based solely on the structures and activities of ligands [1]. We now demonstrate how such a system performs in an iterative, temporal lead optimization exercise [2]. A series of gyrase inhibitors with known synthetic order formed the set of molecules that could be selected for "synthesis" [3]. Beginning with a small number of molecules, based only on structures and activities, a model was constructed. Compound selection was done computationally, each time making selections based on confident predictions of high potency and selections based on quantitative measures of three-dimensional structural novelty. Compound selection was followed by model refinement using the new data. Iterative computational candidate selection produced rapid improvements in selected compound activity, and explicit incorporation of novel compounds uncovered more structurally diverse potent inhibitors than strategies lacking active novelty selection.

We also present a new hybrid structure-guided strategy that incorporates protein structures to inform models of structure–activity relationships [4]. Many QSAR methods have utility in making predictions within a highly related chemical series, but cannot generally be fruitfully applied to novel compounds due to limited domains of applicability. A new structure-guided Surflex-QMOD method demonstrates the ability to use protein structures as well as ligand structure-activity data to construct more robust physical models. These models can accurately predict binding affinities over a broad class of compounds while more accurately representing physical protein pockets and ligand binding modes. The structure-guided method was applied to CDK2 [4], with detailed comparisons to standard QSAR approaches and docking-based predictions. Additional comparisons included structure activity data for urokinase, Chk1, and PTP1b [5]. Results will be presented establishing a new integrated modeling approach that leverages molecular similarity, docking, and multiple-instance learning to produce broadly applicable accurate predictions in cases with limited protein structures but with high ligand diversity.

1. Langham, J. J.; Cleves, A. E.; Spitzer, R.; Kirshner, D.; Jain, A. N. Physical binding pocket induction for affinity prediction. *J. Med. Chem.* **2009**, *52*, 6107–6125.
2. Varela, R., Walters, W. P., Goldman, B. B., Jain, A. N. Iterative refinement of a binding pocket model: active computational steering of lead optimization. *J. Med. Chem.* **2012**, *55*, 8926–8942.
3. Charifson, P.; Grillot, A.; Grossman, T.; Parsons, J.; Badia, M.; Bellon, S.; Deininger, D.; Drumm, J.; Gross, C.; LeTiran, A. Novel dual-targeting benzimidazole urea inhibitors of DNA gyrase and topoisomerase IV possessing potent antibacterial activity: intelligent design and evolution through the judicious use of structure-guided design and structure- activity relationships. *J. Med. Chem.* **2008**, *51*, 5243–5263.
4. Varela, R.; Cleves, A. C.; Spitzer, R.; Jain, A. N. A Structure-Guided Approach for Protein Pocket Modeling and Affinity Prediction. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 917–934.
5. Brown, S. P.; Muchmore, S. W. Large-scale application of high-throughput molecular mechanics with Poisson–Boltzmann surface area for routine physics-based scoring of protein–ligand complexes. *J. Med. Chem.* **2009**, *52*, 3159–3165.

P-54: Metabolism Simulation and Toxicity Prediction in the Evaluation of Food Ingredient/Contaminant Safety

Lothar Terfloth[1], Kirk Arvidson[2], Kristi Muldoon-Jacobs[2], Patra Volarath[2], Aleksey Tarkhov[1], Tomasz Magdziarz[1], James Rathman[3,4], Dimitar Hristozov[4], Chihae Yang[1,4]

[1]Molecular Networks GmbH, Erlangen, Germany, [2]Center for Food Safety and Applied Nutrition, Office of Food Additive Safety, U.S. Food and Drug Administration, College Park, MD, USA, [3]Chemical and Biomolecular Engineering, The Ohio State University, Columbus, OH,USA, [4]Altamira LLC, Columbus, OH,USA

During the safety assessment of food additives and their impurities, metabolic knowledge becomes critical when *in vivo* data are unavailable for the specific compound. Inclusion of metabolism information into the *in silico* workflow is therefore a pre-requisite for the US FDA's Chemical Evaluation and Risk Estimation System (CERES). The chemical space of food additives was profiled using public ToxPrint and metabolic (human liver and S9-fraction) chemotypes. Chemotypes are structural fragments encoded with physicochemical properties, and are considered as alerts when associated with a specific endpoint. The metabolic potential was predicted by applying a diverse set of over 100 phase I as well as more than 20 phase II conjugation reactions of human liver metabolic chemotypes. As an example, the effect of S9 metabolic chemotypes on Ames mutagenicity was analysed against the food additives.

Although nearly 20% of the compounds were perceived as genotoxic carcinogens by chemotype alerts, only 4% of such compounds were actually predicted to be mutagenic under CERES mechanistic QSAR paradigm. Nearly 50% of the food additives predicted to be mutagenic were matched with chemotype alerts for genotoxic carcinogens. When repeating the analysis after removing the compounds that may be detoxified by S9, the reliability of the genotoxic alerts is increased almost 3 times.

This study demonstrates the value of the use of metabolic rules in conjunction with known chemotype alerts to reduce the false positive rate of structural rules. Implementation of the metabolic rulebase in CERES is also presented.

Funding Statement: The research was funded by EU FP7 and Cosmetics Europe (Grant n° 266835).

P-56: Single Target SAR and Multiple Target Polypharmacology Analysis Using inSARa

Sabrina Wollenhaupt, Knut Baumann

Institut für Medizinische und Pharmazeutische Chemie, Technische Universität Braunschweig, Braunschweig, Germany

The versatile inSARa method (**i**ntuitive **n**etworks for **S**tructure-**A**ctivity **R**elationship **a**nalysis) was primarily developed with the objective to support medicinal chemists in tackling large-scale SAR analysis and visualization in an intuitive, straightforward way. The main features of the approach are hierarchical networks of clearly-defined substructure relationships based on common pharmacophoric features. The method takes advantage of the synergy resulting from the combination of reduced graphs (RG) and the intuitive concept of the maximum common substructure (MCS) [1].

One key concept in SAR analysis is the similar property principle which states that similar molecules tend to have similar properties including their biological activity [2]. For this assumption to hold and consequently successful SAR interpretation and (on-/off-)target prediction, it is of utmost impor-

tance to encode molecular similarity in a meaningful way. Various analyses (e.g. comparisons based on the prediction of bioactivities using *k*-Nearest-Neighbor (*k*NN) regression) show that the way of molecular representation and perception of similarity used in the inSARa approach is superior to the commonly used concept of fingerprint-based similarity analysis. Due to the better choice of nearest neighbors in chemical space (demonstrated e.g. by smaller prediction error in *k*NN regression), benefits are expected by using the inSARa concept in SAR analysis. The inSARa Hybrid approach, which combines inSARa with fingerprint-based similarity networks in different ways, highlights the advantages resulting from the combination of both concepts.

Using inSARa networks, common molecular or pharmacophoric features crucial for bioactivity modification are easily identified in data sets of different size (up to thousands of molecules) and heterogeneity [3]. When focusing on a set of active molecules at one single target, the resulting inSARa networks are shown to be valuable for various essential tasks in SAR analysis. Based on simple rules not only common pharmacophoric patterns but also bioisosteric exchanges, activity cliffs or 'SAR hotspots' and 'activity switches' are easily identified. These different types of SAR information are either identified by interactive navigation of the hierarchical networks or automated network analysis (inSARa^{auto}). In Analogy to the fingerprint-based SAR-Index, the SARdisco Score which is based on inSARa^{auto} globally characterize the portion of SAR (dis)continuity in inSARa networks.

Additionally, inSARa networks of a large number of different targets were pairwisely compared on the basis of the portion of common RG-MCSs. The results indicate that inSARa networks which were primarily developed for SAR interpretation are also valuable for gaining insights in polypharmacology. The promising results of the analysis show that the RG-MCS based concept can complement published chemogenomic approaches for ligand-based analysis of targets similarities and the identification of cross-reactivities/off-target-relationships. The advantage of the developed RG-MCS approach is the easy interpretability and the fact that molecular features involved in protein-ligand binding are represented.

1. Gardiner, E. J.; Gillet, V. J.; Willett, P.; Cosgrove, D. A. Representing Clusters Using a Maximum Common Edge Substructure Algorithm Applied to Reduced Graphs and Molecular Graphs. *J. Chem. Inf. Model.* **2007,** *47*, 354–366.
2. Johnson M. A., Maggiora G. M., Eds. *Concepts and Applications of Molecular Similarity;* John Wiley & Sons: New York, 1990.
3. Wollenhaupt, S.; Baumann, K. inSARa: Intuitive and Interactive SAR Interpretation by Reduced Graphs and Hierarchical MCS-based Network Navigation. *J. Chem. Inf. Model.* (in revision).

P-58: Repurposing of known kinase inhibitors for inhibition of trypanothione synthetase

Christiane Ehrt[1], Nana Boateng[1], Dennis M. Krüger[2], Oliver Koch[1]

[1]Department of Chemistry and Chemical Biology, TU Dortmund, Germany, [2]Chemical Genomics Centre of the Max Planck Society, Dortmund, Germany

Trypanothione synthetase (TryS) catalyzes the two-step biosynthesis from spermidine, glutathione (GSH) and ATP to trypanothione, which is a key intermediate in trypanosomatid parasites of the species *Trypanosoma* and *Leishmania*. Thus, TryS is an attractive new drug target to cope with diseases like Chagas disease, leishmaniasis, or African trypanosomiasis that affect approximately 15 to 20 million people worldwide and cause 100.000 to 150.000 deaths per annum. Interestingly, paullones, a chemical class of potent cyclin-dependent kinase (CDK) and glycogen synthase kinase 3 (GSK-3) inhibitors, were shown to inhibit TryS [1]. Based on a comparison of available X-ray structures, the binding of paullones to TryS seems to be in accordance with the principle of 'ligand-sensing cores' [2]. The spatial arrangement of secondary structure elements around the ATP binding sites of TryS and kinases is quite similar, independent of the overall fold, which indicates similar ligand binding.

Unfortunately, the available TryS X-ray structure (pdb-id 2vps) was determined without substrates and an important loop region of the ATP grasp fold is missing. Preliminary modeling and molecular dynamics simulation studies revealed the binding modes of ATP, GSH and glutathionylspermidine (GSP). However, the bound state of the loop region that closes after substrate binding remained unexplained [3]. A considerably more detailed analysis that utilizes the X-ray structure information of the related GSP synthetase (pdb-id 2io7) led to a complete model of TryS, containing all substrates and the closed ATP grasp fold loop. Exhaustive molecular dynamics simulations have confirmed this model as reasonable and revealed that the presence of all substrates leads to a partial closure of an associated β-sheet over the bound triphosphate.

Here we present the results of the performed molecular dynamics simulations and show how the obtained model can be used to identify further (kinase) inhibitors that own a similar molecular scaffold. These inhibitors can then be diversified to obtain selective TryS inhibitors. The approach includes a detailed comparison of the full ATP binding pocket of TryS to known kinase X-ray structures as well as docking studies of kinase inhibitors received from the kinase SARfari database (www.ebi.ac.uk/chembl/sarfari/kinasesarfari). Biochemical testing of newly identified kinase inhibitors for inhibition of TryS will be carried out.

1. Koch, O.; Jäger, T.; Flohé, L.; Selzer, P. M. Inhibition of Trypanothione Synthetase as a Therapeutic Concept. In *Drug Discovery in Infectious Diseases*; Selzer, P. M., Ed.; Wiley: Weinheim, **2013**; Vol. 4; p 429–443.
2. Koch, O. The Use of Secondary Structure Element Information in Drug Design: Polypharmacology and Conserved Motifs in Protein-Ligand Binding and Protein-Protein Interfaces. *Future Med. Chem.* **2011,** *3*, 699–708.
3. Koch, O.; Cappel, D.; Nocker, M.; Jäger, T.; Flohé, L.; Sotriffer, C. A.; Selzer, P. M. Molecular Dynamics Reveal Binding Mode of Glutathionylspermidine by Trypanothione Synthetase. *PLOS ONE* **2013**, *8*, 1–10.

P-60: Retrieving 'hits' through in silico screening and expert assessment

Renate Griffith[1], **Malgorzata Drwal**[1,2]

[1] UNSW Sydney, Australia, [2] Charité Berlin, Germany

A strategy will be presented involving combination of ligand- and structure-based drug design methods to retrieve structurally novel compounds through virtual screening of large chemical databases. In particular, various pharmacophore methods (ligand-based, complex-based and structure-based) form an integral part of this strategy and will be discussed.

It will be shown that the combination of pharmacophores, docking, and expert assessment can be successfully applied in database screening to retrieve structurally novel compounds as topoisomerase I and II inhibitors [1–3].

1. Drwal, M. N.; Agama, K.; Pommier, Y.; Griffith, R. Development of purely structure-based pharmacophores for the topoisomerase I-DNA- ligand binding pocket. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 1037–1049.

2. Drwal, M. N.; Griffith, R. Combination of Ligand- and Structure-Based Methods in Virtual Screening. *Drug Discov Today: Tech* **2013**, *10*, 395–401.
3. Drwal, M. N.; Agama, K.; Wakelin, L. P. G.; Pommier, Y.; Griffith, R. Exploring DNA Topoisomerase I Ligand Space in Search of Novel Anticancer Agents. *PLoS ONE* **2011**, *6*, e25150.

P-62: Structural requirements of drug candidates to cause cholestatic side effects

Tina Ritschel[1], Susanne M. A. Hermans[1,2], Rick Greupink[2], Frans G. M. Russel[2]

[1]Computational Discovery and Design (CDD) Group, Centre for Molecular and Biomolecular Informatics (CMBI), Radboud University Medical Centre, PO Box 9101, 6500 HB Nijmegen, NL [2]Department of Pharmacology and Toxicology, Radboud Institute for Molecular Life Sciences (RIMLS), Radboud University Medical Centre, PO Box 9101, 6500 HB Nijmegen, NL

Drug-induced cholestasis is a frequently observed side effect and is often caused by an unexpected interaction of the drug with the bile salt export pump (BSEP/ABCB11). BSEP is the key membrane transporter responsible for the active excretion of bile acids from hepatocytes into bile.

In our study we shed further light on potential inhibitors of BSEP activity and the molecular structure of such compounds. Current *in silico* models for the prediction of BSEP inhibitors are only based on physicochemical parameters or 2D structural data of interacting small molecules [1]. We describe a pharmacophore model that takes into account the 3D configuration of the interacting small molecules [2]. To generate input and validation data sets, isolated membrane vesicles over-expressing human BSEP were used to assess the effect of compounds on BSEP-mediated ^{3}H-taurocholic acid transport.

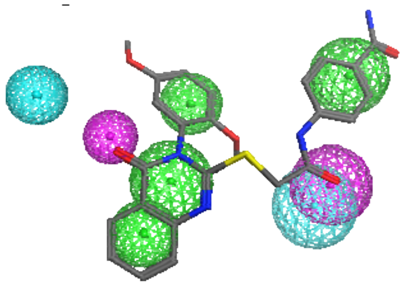


Figure 1. BSEP pharmacophore model with one inhibitor. Hydrophobic/aromatic features (green), H-bond acceptor / anionic features (pink),H-bond acceptor projection vectors (cyan).

The pharmacophore model contains eight features (Figure 1) and exclusion volumes (not shown). The pharmacophore was validated against a set of 59 compounds, including registered drugs. The model recognized 9 out of 12 inhibitors (true positive rate: 75%), but only 8 of the 47 non-inhibitors (false positive rate: 17%). We demonstrate that this alternative approach is not only successful in differentiating inhibitors from non-inhibitors of BSEP, but also to improve current *in silico* approaches for the identification of BSEP inhibitors by 37%.

In addition, we performed a virtual screening to detect potential BSEP inhibitors. A structurally diverse selection of the hits was purchased and tested *in vitro*. All compounds displayed statistically significant BSEP inhibition, ranging from 13±1% to 67±7% of control (P<0.05).

Finally, the method we propose could be applied in drug discovery research to flag compounds with potential adverse cholestatic effects in an early stage of drug development.

1. Warner, D. J.; Chen, H.; Cantin, L. D.; Kenna, J. G.; Stahl, S.; Walker, C. L.; Noeske, T., Mitigating the Inhibition of Human Bile Salt Export Pump by Drugs: Opportunities Provided by

Physicochemical Property Modulation, In-silico Modeling and Structural Modification. *Drug Metab. Dispos.* **2012**, *40*, 2332–2341.
2. Molecular Operating Environment (MOE), 2013.08. *Chemical Computing Group Inc., 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7,* **2011**.

P-64: MotiveQuery: Language and Web Service for Fast Identification of Protein Structural Motifs in the Entire Protein Data Bank

Lukáš Pravda[1], Radka Svobodová Vařeková[1,2], David Sehnal[1,2], Crina-Maria Ionescu[1], Jaroslav Koča[1,2]

[1]CEITEC - Central European Institute of Technology, Masaryk University, Brno, Czech Republic, [2]National Centre for Biomolecular Research, Masaryk University, Brno, Czech Republic

Only recently, the number of experimentally determined structures stored in the Protein Data Bank (PDB) has reached 100 000. Both their quality and quantity has been increasing rapidly in recent years. However, a significant number of structures still lack proper characterization of their biological function in literature. The detection and the analysis of various structural motifs may shed light on a protein's previously unknown function. Furthermore, such an analyses can prove of great use for areas like drug design, signal transduction, protein-ligand interactions, or structure validation [1–3].

Here, we present MotiveQuery (MQ), a user friendly chemical language and a web service primarily designed for the definition and extraction of structural motifs from the Protein Data Bank. MQ combines the clarity and brevity of programming languages with the versatility of natural language, aiming for an efficient inclusion of chemical and biochemical knowledge into the definition of structural motifs. MQ allows definitions based on the chemical connectivity and the three-dimensional structure at the same time. Additionally, in the case of molecules based on residue chains (such as proteins, nucleic acids, saccharides, etc.), MQ allows the user to include any amount of information regarding the residue level structure directly into the definition of the motifs. On top of that, MQ allows the user to fine-tune the search in a manner similar to the criteria used by the PDB, such as method of structure determination, resolution, enzymatic class, organism of origin etc. The MQ service is available free of charge at http://webchem.ncbr.muni.cz/Platform/MotiveQuery.

1. Konc, J.; Janežič, D. Binding site comparison for function prediction and pharmaceutical discovery. *Curr. Opin. Struct. Biol.* **2014**, *25*, 34–39.
2. Nakashima, R.; Sakurai, K.; Yamasaki, S.; Hayashi, K.; Nagata, C.; Hoshino, K.; Onodera, Y; Nishino, K.; Yamaguchi, A. Structural basis for the inhibition of bacterial multidrug exporters. *Nature* **2013**, *500*, 102–106.
3. Sehnal, D.; Vařeková, R. S.; Huber, H. J.; Geidl, S.; Ionescu, C.-M.; Wimmerová, M.; Koča, J. SiteBinder: an improved approach for comparing multiple protein structural motifs. *J. Chem. Inf. Model.* **2012**, *52*, 343–359.

P-66: Exploiting Solvent Effects in Drug Design and Optimization

Guido Kirsten, Jean-Francois Truchon, Al Ajamian, Chris Williams, Paul Labute

Chemical Computing Group, Montreal, Canada

There is significant interest in understanding the behavior of water molecules as it relates to ligand-receptor interactions. In specific cases, ambiguous and counterintuitive SAR seems to be linked to solvent effects. Ligand affinity and specificity appear to be influenced by the action of water molecules on the solvated ligand-receptor complex As such, a deeper analysis of solvent effects would expose potential ligand design opportunities that were previously not conceivable. Here, we report the application of the 3D Reference Interaction Site Model as a potential method to account for such solvent effects.

1. Luchko, T.; Gusarov, S.; Roe, D. R.; Simmerling, C.; Case, D. A.; Tuszynski, J.; Kovalenko, A. Three-dimensional molecular theory of solvation coupled with molecular dynamics in Amber. *J. Chem. Theory Comput.* **2010**, *6*, 607–624.
2. Kovalenko, A.; Hirata, F. Self-consistent description of a metal-water interface by the Kohn-Sham density functional theory and the three-dimensional reference interaction site model. *J. Chem. Phys.* **1999**, *110*, 10095–10112.

P-68: In silico prediction of antitumor cytotoxicity of pharmacologically active substances for human breast cancer and normal cell lines

Varvara Konova[1], Alexey Lagunin[1], Pavel Pogodin[1,2], Anastassia Rudik[1], Sergey Ivanov[1], Dmitry Druzhilovsky[1], Dmitry Filimonov[1], Vladimir Poroikov[1,2]

[1]Orekhovich Institute of Biomedical Chemistry of Rus. Acad. Med. Sci., Moscow, Russia, [2]National Research Medical University, Department of Biochemistry of Medical-Biological Faculty, Moscow, Russia

Modeling of pathologic processes for the analysis of causes and development of disease on the molecular genetic level contributes to better understanding of neoplastic disease pathogenesis and opens new perspectives for early disease detection [1,2].

Therefore, the systems biology approaches are required to study complex diseases at the proteins' and genes' level and the interactions between them [3]. These approaches allow the obtained genomic, transcriptomic, proteomic and metabolomic data to be integrated for the analysis of complex disorders occurring in the neoplastic disease development [4]. Currently, bio- and chemoinformatics methods, and mathematical modeling are widely used for the detection of pathogenic mechanisms of cancer and quest for potential drug-targets and their ligands. Regulatory network analysis based on signaling pathways and cell cycle regulation data combines in single system the genomics and proteomics data. The use of differential gene expression tumor data results in obtaining information on new therapeutic targets and treatment choices for individual patient [5].

The present work was focused on the development of an approach for the *in silico* prediction of the cytotoxic effect of chemical compounds in normal and breast cancer cell lines based on the prediction of their cytotoxicity and action on human proteins, regulatory networks modeling and gene expression.

We have created the cell cycle regulatory networks models including 2141 regulatory interactions between 1265 breast cancer genes/proteins and hypo-and hyperexpressed genes for 4 breast cancer cell lines (BT-20, MCF7, SK-BR-3, T47D) and 2 normal cell lines (HaCaT, WI-38). In the breast cancer cell cycle regulation modeling, several known and new pharmacological targets were found that further should be validated experimentally.

The computer program PASS [6,7] (Prediction of Activity Spectra for Substances) was used to construct the structure-activity relationships models. The appropriate training sets were created based on the information from ChEMBLdb 16 database (www.ebi.ac.uk/chembldb/) about cytotoxicity and interactions of chemicals with 661 proteins involved in human cell cycle regulation. The average prediction accuracy calculated by a leave-one-out cross-validation procedure was approximately 96% for cytotoxicity prediction for 24 breast cancer cell lines and 31 normal cell lines and 97% for protein-ligand interactions.

Libraries of commercially available samples of chemical compounds (Asinex, ChemBlock, ChemBridge, InterBioScreen) containing more than million structures, were used for *in silico* screening of promising antitumor ligands.

While screening, we have selected few dozen promising compounds for which the interactions with identified targets, the cytotoxicity for breast cancer cell lines, and the absence of cytotoxicity for 31 normal human cell lines were predicted. Several selected compounds were experimentally tested for cytotoxicity to some breast cancer cell lines. Selection of possible antitumor compounds is shown schematically in Figure 1.

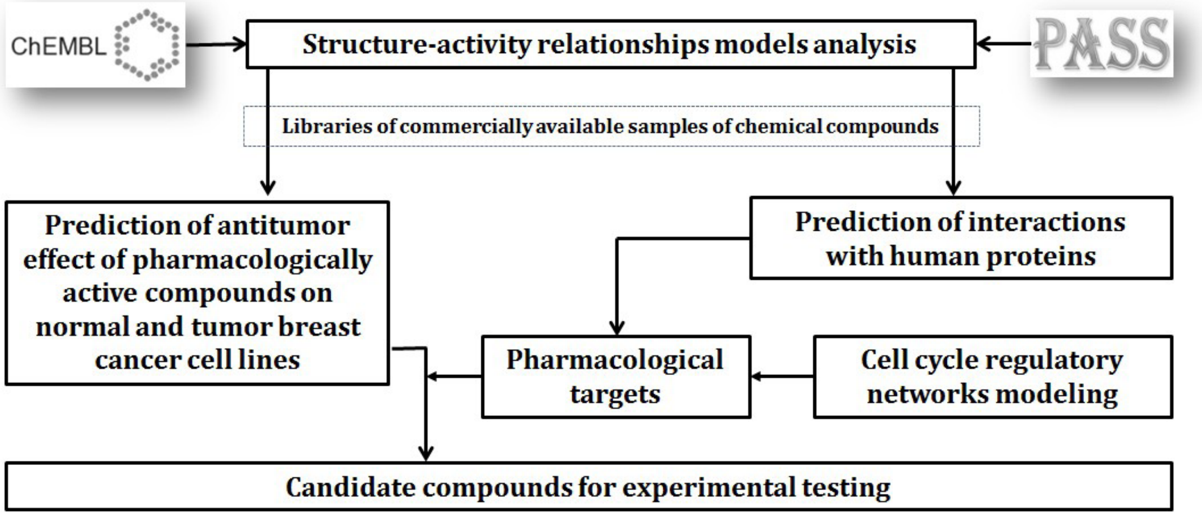

ChEMBL
Structure-activity relationships models analysis
PASS
Libraries of commercially available samples of chemical compounds
Prediction of antitumor effect of pharmacologically active compounds on normal and tumor breast cancer cell lines
Prediction of interactions with human proteins
Pharmacological targets
Cell cycle regulatory networks modeling
Candidate compounds for experimental testing

Figure 1: Scheme of selection of possible antitumor compounds.

Thus, the developed approach allows one to reveal compounds possessing antitumor activity for breast cancer cell lines and action on proteins responsible for the cell cycle arrest and apoptosis.

1. Vera-Licona, P.; Bonnet, E.; Barillot, E.; Zinovyev, A. OCSANA: optimal combinations of interventions from network analysis. *Bioinformatics* **2013**, *29*, 1571–1573.
2. Menden, M. P.; Iorio, F.; Garnett, M.; McDermott, U.; Benes, C. H.; Ballester, P. J.; Saez-Rodriguez J. Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLoS One* **2013**, *8*, 613–618.
3. Lamb, J.; Crawford, E. D.; Peck, D.; Modell, J. W.; Blat, I. C.; Wrobel, M. J.; Lerner J.; Brunet, J. P.; Subramanian, A.; Ross, K. N.; Reich, M.; Hieronymus, H.; Wei, G.; Armstrong, S. A.; Haggarty, S. J.; Clemons, P. A.; Wei, R.; Carr, S. A.; Lander, E. S.; Golub, T. R. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **2006**, *313*, 1929–1935.

4. Iorio, F.; Rittman, T.; Ge, H.; Menden, M.; Saez-Rodriguez J. Transcriptional data: a new gateway to drug repositioning? *Drug. Discov. Today* **2013**, *18*, 350–357.
5. Koborova, O. N.; Filimonov, D .A.; Zakharov, A. V.; Lagunin, A. A.; Ivanov, S. M.; Kel, A.; Poroikov, V. V. In silico method for identification of promising anticancer drug targets. *SAR and QSAR Environ. Res.* **2009**, *20*, 755–766.
6. Poroikov, V. V.; Filimonov, D. A.; Borodina, Yu. V.; Lagunin, A. A.; Kos, A. Robustness of biological activity spectra predicting by computer program PASS for non-congeneric sets of chemical compounds. *J. Chem. Inform. Comput. Sci.* **2000**, *40*, 1349-1355.
7. Filimonov, D. A.; Poroikov, V. V. In *Chemoinformatics Approaches to Virtual Screening*; Varnek, A.; Tropsha, A. Ed.; RSC Publishing: Cambridge (UK), 2008, 182–216.

P-70: Modular Interactive Structure-Based Pharmacophore Searching

Jens Kunze, Gisbert Schneider

Department of Chemistry and Applied Biosciences, Swiss Federal Institute of Technology (ETH), Zürich, Switzerland

We present an interactive tool for structure-based virtual screening (SBVS). The modular software is suited for searching transient ligand binding sites and is able to explicitly take inaccuracies of atom positions into account for pharmacophore modeling and matching. The tool allows the users to integrate their own implementations of pocket detection, potential pharmacophore point (PPP) generation and pharmacophore descriptor calculation. In the present version, a combination of our own in-house tools is incorporated for SBVS. Pocket detection is achieved with PocketPicker [1], and receptor-derived pharmacophore models are generated with VirtualLigand [2]. PPPs are formed according to the geometric rules supposed by Bissantz et al. [3] and translated to a cross-correlation vector representing idealized "virtual ligand" pharmacophores. All steps are interactive and tailored to enable focused supervised SBVS campaigns. The target pocket can be rigged and unwanted PPPs may be deleted. It is also possible to consider an error estimate for the protein atom positions, i.e., for single structures by the crystallographic B-factor, for structure ensembles a global alignment or a local pocket residue alignment can be performed during the process. The method has been evaluated in retrospective and prospective applications. These results will be presented.

1. Weisel, M.; Proschak, E.; Schneider, G. PocketPicker: analysis of ligand binding-sites with shape descriptors. *Chem. Cent. J.* **2007**, *1*, 7.
2. Löwer, M.; Geppert, T.; Schneider, P.; Hoy, B.; Wessler, S.; Schneider, G. Inhibitors of *Helicobacter pylori* Protease HtrA Found by 'Virtual Ligand' Screening Combat Bacterial Invasion of Epithelia. *PLoS One* **2011**, *6*, 3.
3. Bissantz, C.; Kuhn, B.; Stahl, M. A Medicinal Chemist's Guide to Molecular Interactions. *J. Med. Chem.* **2010**, *53*, 5061–5084.

P-72: Identification of potential protein-protein binding sites by using a 3 phased self-contained in silico workflow

Christian Jäger[1], Anett Stephan[2], Stephan Schilling[1], Mirko Buchholz[1]

[1]Fraunhofer Institute for Cell Therapy and Immunology, Department of Drug Design and Target Validation (IZI-MWT), 06120 Halle (Saale), Germany, [2]NOMAD Bioscience, 06120 Halle (Saale), Germany

Many proteins of the posttranslational processing machinery, such as glycosyltransferases and the recently discovered isoenzyme of glutaminyl cyclase (QC), are localized in the Golgi apparatus. Thereby the retention of proteins in the Golgi is a common sorting signal for the cell. For yeast it is well known, that the cytosolic protein Vps74 causes the retention by binding to type II transmembrane proteins at the consensus sequence (F/L)-(L/I/V)-X-X-(R/K) as a retention signal in the protein [1,2].

Recently, it was shown, that the same retention signal also exists in a subset of human Golgi-resident proteins such as for isoQC and GnT-IX. Furthermore the retention in the Golgi in mammalian cells is mediated by GPP34 (also named GOLPH3, GmX33α, MIDAS), an ortholog of VPS74 [3], both sharing sequence identity of 62%. Additionally, GPP34 binds to phosphatidylinositol-4-phosphate (PtdIns(4)P), a prevalent part of the Golgi membrane.

For a detailed analysis of the protein-protein binding sites, we developed a multiphase workflow, combining different *in silico* methods within 3 steps.

In Step 1, a model of GPP34 was generated, taking as much as possible experimental knowledge into account, such as information on dimerization, PtdIns(4)P binding and sequence identities.

Step 2 consists of a fragment based approach for the identification of reasonable initial positions of the isoQC- as well as the GnT-IX- cytosolic N-terminus for the binding to GPP34. Based on the retention motifs fragment libraries were created, which contain short overlapping peptide snippets with a length of 3 amino acids. Using these snippets, multi-fragment searches (MFS) were initiated by the random placement of 5000 copies of each fragment on GPP34. The resulting fragment positions were scored according to their interaction potentials.

Step 3 includes a docking of the whole peptides of isoQC and Gnt-IX into the model complex of GPP34. Therefore, we selected the top three scored fragments as templates for the initial positioning of the whole peptide, followed by a peptide-protein docking approach using the FlexPepDock Server [4,5].

Several analysis steps led to the identification of at least 4 meaningful hot spots for a possible binding of isoQC and GnT-IX to the GPP34-homodimer, which are presented here.

1. Tu, L.; Tai, W. C. S.; Chen, L.; Banfield, D. K. Signal-mediated dynamic retention of glycosyltransferases in the Golgi. *Science* **2008**, *321*, 404–407.
2. Bell, A. W.; Ward, M. A.; Blackstock, W. P.; Freeman, H. N.; Choudhary, J. S.; Lewis, A. P.; Chotai, D.; Fazel, A.; Gushue, J. N.; Paiement, J.; Palcy, S.; Chevet, E.; Lafrenière-Roula, M.; Solari, R.; Thomas, D. Y.; Rowley, A.; Bergeron, J. J. Proteomics characterization of abundant Golgi membrane proteins. *J. Biol. Chem.* **2001**, *276*, 5152–5165.
3. Dippold, H. C.; Ng, M. M.; Farber-Katz, S. E.; Lee, S.-K.; Kerr, M. L.; Peterman, M. C.; Sim, R.; Wiharto, P. A.; Galbraith, K. A.; Madhavarapu, S.; Fuchs, G. J.; Meerloo, T.; Farquhar, M. G.; Zhou, H.; Field, S. J. nihms144236 // GOLPH3 Bridges Phosphatidylinositol-4- Phosphate and Actomyosin to Stretch and Shape the Golgi to Promote Budding. *Cell* **2009**, *139*, 337–351.
4. London, N.; Raveh, B.; Cohen, E.; Fathi, G.; Schueler-Furman, O. FlexPepDock Server // Rosetta FlexPepDock web server – high resolution modeling of peptide-protein interactions; 2009. http://flexpepdock.furmanlab.cs.huji.ac.il/cite.php.

5. Raveh, B.; London, N.; Schueler-Furman, O. Sub-angstrom modeling of complexes between flexible peptides and globular proteins. *Proteins* **2010**, *78*, 2029–2040.

P-74: NUCLEO.QUERY: A free web-based virtual screening platform targeting nucleotide cofactor proteins

Constantinos Neochoritis[1], Alexander Dömling[1], Tryfon Zarganes-Tzitzikas[1], Carlos Camacho[2], Dave Koes[2], Kareem Khoury[3]

[1]Dept. Drug Design, RUG, Netherlands, [2]PITT, Pittsburgh, USA, [3]Carmolex BV, Groningen, Netherlands

Today's industrial screening paradigm is high throughput (HTS). The majority of all medicinal chemistry projects start with the expensive, time consuming screening of millions of library compounds. Nevertheless, industry largely fails to serve the vast number of genomics derived non-traditional targets (e.g. protein protein interactions) and bringing breakthrough medications to the patients. Instead historical rather limited chemical libraries continue to be the mainstay and low numbers of drugs showing only incremental benefits for the patients enter the market every year. At the same time a frightening and general strategy-switch from small molecules to biotechnology drugs is observed in most big pharma companies.

Recently virtual screening has evolved as an alternative to HTS with impressive results. Most virtual screening campaigns, however, rely either on rather small libraries of commercially accessible traditional compounds (e.g. ZINC) or very large exhaustively enumerated libraries of mostly difficult to synthesize compounds (e.g. GDB). We have recently introduced ANCHOR.QUERY (http://anchorquery.csb.pitt.edu/) a virtual pharmacophore-based screening platform where screening hits can be instantaneously and with high precision and confidence resynthesized [1]. The screening library of the public available ANCHOR.QUERY comprise a ~2 billion-sized chemical space of multicomponent reaction (MCR) derived compounds, while a proprietary version consists of >500 unique scaffolds based on MCR and traditional chemistry. The value of ANCHOR.QUERY was recently shown by the discovery of multiple novel antagonists of the protein protein interaction p53/MDM2 [2].

We now introduce the free web-based screening platform NUCLEO.QUERY (http://nucleoquery.csb.pitt.edu/). NUCLEO.QUERY leverages the concept of *anchors*, nucleic acid residues that bury a large amount of solvent accessible surface area at the protein-nucleic acid interface. The interactive nature of the software allows querying the library for different pharmacophore features. Each query yields a diverse set of small molecules that match different properties of the receptor beyond the nucleotide. It is the first time that software attempts to rationally design nucleo-base type inhibitors. NUCLEO.QUERY is a specialized pharmacophore search technology that brings interactive virtual screening of novel protein-nucleic acid inhibitors to the drug hunters desktop.

Here we demonstrate the powerful usage of NUCLEO.QUERY for the rapid discovery of potent cell active anti tuberculosis agents by targeting *Mycobacterium tuberculosis* thymidylate kinase (TMK). Thymidylate kinase (TMK) has emerged as an attractive therapeutic target because inhibiting TMK functions blocks DNA synthesis in replicating organisms, such as *Mycobacterium tuberculosis* and no shunt-pathway is known.

1. Koes, D. et al. Enabling large-scale design, synthesis and validation of small molecule protein-protein antagonists. *PLoS One* **2012**, *7*, e32839.
2. Huang, Y. et al. Discovery of highly potent p53-mdm2 antagonists and structural basis for anti-acute myeloid leukemia activities. *ACS Chem. Biol.* **2014**, doi: 10.1021/cb400728e.

P-76: Small molecular turn mimetics as protein-protein interaction inhibitor building blocks

Anna Rudo, **Oliver Koch**

TU Dortmund, Department of Chemistry and Chemical Biology, Otto-Hahn-Str. 6, 44227 Dortmund, Germany

Protein-protein interactions (PPIs) are ubiquitous in nature and essential to almost all biological processes including signal transduction, gene-expression and pathogenicity. Therefore, modulation offers attractive therapeutic opportunities. Interfaces, intrinsically planar and large, were deemed to be hardly accessible targets for small molecular modulation due to the lack of well-defined binding-pockets [1]. For this reason, an approach for the design of PPI inhibitors is to elucidate and mimic the important, limited-sized elements that act as actual protein recognition motifs [2].

Up to now, only strand and helix mimetics have been used as interaction inhibitors [3], although just in weak and transient heterodimer interfaces of greater pharmacological interest, irregular turn structures prevail as regions of high affinity binding [4]. The turn backbones provide valuable information for the design of new drugs, since they act as scaffolds for presenting the relevant side chains in the correct and specific orientation [5].

We aim for a uniquely rational way for the computational elucidation of small molecules as PPI inhibitor turn mimetics by making use of the uniform and complete classification of turns. It implies 157 turn types showing different backbone conformations [5]. With computer-based methods small-molecular scaffold mimetics are to be found for each of the turn types. These compounds should be further functionalized with appropriate side-chains to reproduce the bioactive structure, leading to a library of turn mimicking inhibitor building blocks.

As a first example, our analysis of a bacterial GTPase-activating PPI being responsible for correct formation of flagella leads us to a crucial interaction turn entity of type n(4)I. These kinds of β-turns are already known to be excellently replaced by the benzodiazepines scaffold. The structure was further *in silico* functionalized by fragment growing and linking steps to also address neighboring anchor points within the surface to increase binding and specificity.

1. Che, Y.; Brooks, B. R.; Marshall, G. R. Development of small molecules designed to modulate protein-protein interactions. *J. Comput. Aided Mol. Des.* **2006**, *20*, 109–130.
2. Labbé, C. M.; Laconde, G.; Kuenemann, M. A.; Villoutreix, B. O.; Sperandio, O. iPPI-DB: a manually curated and interactive database of small non-peptide inhibitors of protein-protein interactions. *Drug Discov. Today* **2013**, *18*, 958–968.
3. Yin, H.; Hamilton, A. D. Strategies for targeting protein-protein interactions with synthetic agents. *Angew. Chem. Int. Ed.* **2005**, *44*, 4130–4163.
4. Guharoy, M; Chakrabarti, P. Secondary structure based analysis and classification of biological interfaces: identification of binding motifs in protein–protein interactions. *Bioinformatics* **2007**, *23*, 1909–1918.
5. Koch, O.; Klebe, G. Turns revisited: a uniform and comprehensive classification of normal, open, and reverse turn families minimizing unassigned random chain portions. *Proteins* **2009**, *74*, 353–367.

P-78: Development of a Protein-ligand Interaction Database for Structure-based Drug Design

Richard Sherhod[1], **Adrian Schreyer**[2], **James Davidson**[1]

[1] Vernalis R&D Ltd., Cambridge, UK, [2] University of Cambridge, Cambridge, UK

There is a wealth of macromolecular structural data available in the public domain, most notably *via* the Protein Data Bank (PDB) [1], which represents an invaluable resource for facilitating the structure-based drug design process. However there are a number of challenges facing medicinal chemists who wish to use this information to better understand their drug targets; for example, the PDB does not store structures for complete biological assemblies, visualisation of protein-ligand interactions may require the use of expensive software tools that may be of limited availability within an organisation, unpublished in-house macromolecular data may be stored in quite different formats to those in the PDB, and unpublished structures may not be searchable alongside public structures. In order to address these issues, we have applied the experience of developing CREDO [2] to construct a database of biological assemblies and their protein-ligand interactions for both public and in-house macromolecular structures. The database schema has been specifically designed to accommodate the inconsistencies observed in PDB records and the differences between public and in-house structural data. To provide access to this database a set of RESTful web services have been deployed and two complementary data access and visualisation tools are currently being developed.

The database schema has been designed to include a range of descriptors for small molecules and macromolecules, while the web services provide a number of methods for searching different aspects of the data. The USRCAT [3] method has been implemented to provide alignment-independent 3D similarity searching across ligands, both in an ensemble of conformations and in their protein-bound conformation. This is supplemented by some commonly used techniques for 2D similarity searching across small molecules. The FuzCav [4] descriptor has been incorporated to permit 3D similarity searching across protein binding sites. Structural Interaction Fingerprint [5] similarity searching has been implemented to allow similar and diverse protein-ligand interaction patterns and ligand binding modes to be identified. Finally, all protein residues are labelled with their appropriate annotations, obtained from EMBL-EBI's SIFTS database [6], to highlight important protein sequence and functional information and to allow searching by protein families, domains, and functional sites.

Two software tools are being developed to provide data access: a PyMOL [7] plugin, and a web portal interface. The PyMOL plugin allows the database and RESTful services to be integrated with PyMOL's advanced visualisation features, while providing tools to make PyMOL more user friendly for chemists. The web portal is designed to provide a more flexible working environment, with more searching options and WebGL-based visualisation tools. Furthermore, the web portal can readily be made accessible over the internet to allow sharing of structural information with project collaborators.

1. Protein Data Bank http://www.rcsb.org/pdb/ (accessed Jan 5, 2014).
2. Schreyer, A.; Blundell, T. CREDO: a protein-ligand interaction database for drug discovery. *Chem. Biol. Drug Des.* **2009**, *73*, 157–167.
3. Schreyer, A. M.; Blundell, T. USRCAT: real-time ultrafast shape recognition with pharmacophoric constraints. *J. Cheminform.* **2012**, *4*, 27.
4. Weill, N.; Rognan, D. Alignment-free ultra-high-throughput comparison of druggable protein-ligand binding sites. *J. Chem. Inf. Model.* **2010**, *50*, 123–135.
5. Deng, Z.; Chuaqui, C.; Singh, J. Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions. *J. Med. Chem.* **2004**, *47*, 337–344.

6. Velankar, S.; Dana, J. M.; Jacobsen, J.; van Ginkel, G.; Gane, P. J.; Luo, J.; Oldfield, T. J.; O'Donovan, C.; Martin, M.-J.; Kleywegt, G. J. SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Res.* **2013**, *41*, D483–D489.
7. Schrödinger LLC. The PyMOL Molecular Graphics System, version 1.5.0.3.

P-80: The Pocketome of Human Kinases: Prioritizing (yet) Untapped Protein Kinases for Drug Development

Andrea Volkamer[1], Sameh Eid[1], Samo Turk[1], Friedrich Rippmann[2], Simone Fulle[1]

[1]BioMed X Innovation Center, Im Neuenheimer Feld 583, 69120 Heidelberg, Germany, [2]Merck KGaA, Merck Serono, Global Computational Chemistry, Frankfurter Str. 250, 64293 Darmstadt, Germany

Protein kinases are involved in a variety of diseases including cancer, inflammation, and auto-immune disorders. Thus, the development of new kinase inhibitors is of utmost importance in pharmaceutical research.

Due to the high amount of investigations into this drug target class, more than 3700 experimentally determined structures are currently available in the Protein Data Bank (PDB), covering 245 distinct kinases out of the known 518 kinases representing the human kinome [1]. Additionally, on average 10 to 15 new kinase structures are deposited every month to the PDB. This large data pool provides crucial insights into the binding sites of protein kinases and should facilitate efforts to develop kinase inhibitors with improved selectivity.

Here, we present a comprehensive analysis of the ATP-binding pockets of representative structures of the entire human kinome. Statistics over the properties of all pockets and their distribution over the individual kinase groups are curated, including structural and physicochemical descriptors, e.g., pocket volume, buriedness or hydrophobicity. Furthermore, the ATP-binding pockets are clustered with respect to their structural and physicochemical similarity [2] and druggability scores [3] of the individual pockets are generated.

The presented analysis provides new insights into the binding pockets of the entire kinome and will help to prioritize so far untapped protein kinase structures for drug discovery efforts.

1. Manning, G.; Whyte, D. B.; Martinez, R.; et al. The Protein Kinase Complement of the Human Genome, *Science* **2002**, *298*, 1912–1934.
2. von Behren, M. M.; Volkamer, A.; Henzler, A. M.; *et al.* Fast protein binding site comparison via an index-based screening technology. *J. Chem. Inf. Model.* **2013**, *53*, 411–422.
3. Volkamer, A.; Kuhn, D.; Grombacher, T.; Rippmann, F.; Rarey, M. Combining global and local measures for structure-based druggability predictions. *J. Chem. Inf. Model.* **2012**, *52*, 360–372.

POSTER SESSION ABSTRACTS BLUE

P-1: Chemical Space Analysis of Ion Channel Ligands Identified by High-Throughput Screening

Alexander Böcker, **Pierre Ilouga, Stephen Hess, Annett Müller**

Evotec AG, Essener Bogen 7, 22419 Hamburg, Germany

At Evotec a carefully-selected maximum diversity library of more than 250,000 lead-like compounds is readily available for high throughput screening and has been applied in screening projects of several ion channel targets. Highly robust fluorescence-based assays of ion flux or membrane potential have been developed as indicated by excellent Z' values above 0.7 [1]. Automated analysis scripts were developed in the statistical programming language R to fully capitalize on the measured kinetic traces, discriminate potential agonists from antagonists, identify auto-fluorescent compounds or correct for potential edge effects. Our approach was able to successfully identify sufficiently large sets of potential agonists and antagonists. This outcome is supported by confirmation of the hits using manual and automated patch-clamp measurements.

In this study we have also systematically investigated the chemical space of the screening collection using principal component analysis (PCA), clustering [2] and selforganizing maps [3] in combination with different 2D and 3D descriptors or chemical fingerprints 4–6]. The identified ion channel ligands have been compared to compounds detected as hits by cell-based HTS assays for Gprotein coupled receptor ligands. Our analyses show that ion channel HTS hits are highly diverse and broadly cover 2D, 3D and fingerprint spaces. Various key physicochemical and 3D properties were further studied individually. As an example a paucity of low molecular weight ion channel HTS hits with low lipophilicity was observed. This observation is not obvious for GPCR HTS hits.

In conclusion we would like to emphasize that hit discovery for ion channel targets requires detailed knowledge of the biophysical and pharmacological properties of the channel pharmacology. This allows the required assay development & statistical hit analyses necessary for successful identification of ion channel ligands for subsequent hit-to-lead programs. Chemical space analyses demonstrate that highly diverse sets of potential ligands can be expected both in 2D, 3D and fingerprint space. Simple 2D and 3D descriptors like logP/logD [6] or molecular weight may be applied as triaging filters for HTS hit sets.

1. Zhang, J. H., Chung, T. D. Y., Oldenburg, K. R. J. A Simple Statistical Parameter for Use in Evaluation and Validation of High Throughput Screening Assays. *J. Biomol. Screen.* **1999**, *4*, 67–73.
2. Weizhong, L. A Fast Clustering Algorithm for Analyzing Highly Similar Compounds of Very Large Libraries. *J. Chem. Inf. Mod.* **2006**, *46*, 1919–1923.
3. Ultsch, A.; Moerchen, F. ESOM-Maps: Tools for Clustering, Visualization and Classification with Emergent SOM. In *Technical Report Dept. of Mathematics and Computer Science*, University of Marburg, Germany, 2005, 46.
4. Koutsoukas, A.; Paricharak, S.; Galloway, W. R. J. D.; Spring, D. R.; Ijzerman, A. P.; Glen, R. C.; Marcus, D.; Bender, A. *J. Chem. Inf. Model.* **2014**, *54*, 230–242.
5. SYBYLX Version 1.2; Tripos International: St Louis, USA, 2009.
6. JChem, Version 5.9; ChemAxon Ltd.: Budapest, Hungary, 2012.

P-3: Drug safety assessment through automatic extraction of structure-activity relationships

Bernd Wendt, Alexander Steudle

Certara, Martin-Kollar-Str. 17, 81829 München, Germany

An improved workflow for automatic building of 3D-QSAR models from chemical biology databases such as ChEMBL [1] will be presented. Starting with a chemical structure of choice a 3D-smilarity search identifies neighborhood compounds in chemical biology space. These compounds form the basis for an iterative procedure to produce significant and robust 3D-QSAR models [2]. The resulting models provide indication of potential drug safety threats but also enable a mechanistic understanding of potential toxicity effects by relating the model back to the structure and highlighting those parts of the structure that renders it toxic. Such a drug safety assessment goes beyond similarity and physicochemical property-based computational models and can help the medicinal chemist to make better compounds.

On the basis of prospective [3] and retrospective examples the presentation will showcase simple and straightforward visual analysis of structure-activity relationships that yield insights that would not be revealed on the basis of chemical similarity alone.

1. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2001**, *40*, D1100–D1107.
2. Wendt, B.; Uhrig, U.; Bös, F. Capturing Structure– Activity Relationships from Chemogenomic Spaces. *J. Chem. Inf. Model.* **2011**, *51*, 843–851.
3. Wendt, B.; et al. Toluidinesulfonamide Hypoxia-Induced Factor 1 Inhibitors: Alleviating Drug–Drug Interactions through Use of PubChem Data and Comparative Molecular Field Analysis Guided Synthesis. *J. Med. Chem.* **2011**, *54*, 3982–3986.

P-5: ChemTrove: Enabling a generic ELN to support Chemistry by integrating ChemSpider widgets and templates

Simon J. Coles[1], Aileen E. Day[2], Jeremy G. Frey[1], Richard J. Whitby[1], Colin R. Batchelor[2]

[1]Chemistry, University of Southampton, UK, [2]Royal Society of Chemistry, Cambridge, UK

LabTrove [1] has been developed by the Southampton University since 2005 as a multidisciplinary, open-source electronic (laboratory) notebook for researchers to plan experiments and save and share their results. 'Out of the box' it contains generic tools for researchers from any discipline to upload, display and share data of any file type (either within LabTrove itself or via links), however specific chemistry file types would need to be created and edited outside LabTrove.

Over the last year the Royal Society of Chemistry's ChemSpider [2] has been working with Southampton University in a collaboration guided by the aims of the Dial-a-Molecule Grand Challenge [3] to enhance LabTrove with chemistry-specific functionality and retrieve chemical information from ChemSpider when creating LabTrove entries. We are also working towards a future vision of publishing experimental compound and reaction data from LabTrove to ChemSpider.

So far we have a working prototype of ChemTrove with the added features when editing an entry to:

- Search ChemSpider by name for a compound and retrieve a structure image, name, molecular formula, molecular weight and/or ChemSpider link to add to the entry
- Create or edit a stoichiometry table of the chemicals used and produced in a reaction (with: retrieval of compound properties from ChemSpider; inter-conversion of substance amounts; calculation of product yields; ability to record both planned and actual amounts)
- Draw structures within the LabTrove edit page using JSDraw and render them in a LabTrove post using JSmol lite

This functionality has been added by developing a set of jQuery ChemSpider widgets and TinyMCE editor plugins which add buttons to the "edit" page of LabTrove and load these widgets when clicked. We have demonstrated these widgets in LabTrove but they could in theory be included into any web-page (so could be incorporated into any web-based ELN).

By allowing compound and reaction data to be published from ELNs directly to repositories such as ChemSpider, ChemSpider Synthetic Pages and ChemSpider Reactions a major enabling step towards making accessible the type of detailed reaction data, of both successful and less successful reactions , required for Dial-a-Molecule will have been realized. Allowing users of ELNs to easily search on and retrieve reactions and characterisations that have previously been performed when planning a new reaction will also aid progress. Towards this aim LabTrove templates have been written to structure entries that contain compound and reaction data so that they will be more understandable when deposited to the ChemSpider repositories than a free-format block of text and files. There is a compound template for entries that contain compound structures, properties and spectra to be published to ChemSpider, and two templates for structuring reaction entries – one more detailed version for submission to ChemSpider SyntheticPages and one more simple one for submission to ChemSpider Reactions (under development). These templates have been tested during a student intern project to digitise thesis data and publish it into 1035 new LabTrove entries (which use these templates) and 208 new compounds and over 600 spectra in ChemSpider.

ChemTrove is being hosted on a cloud server by the company Liberata to undergo a usability trial by selected academics to guide its future development.

1. LabTrove. http://www.labtrove.org/ (accessed Feb 11, 2014).
2. ChemSpider. http://www.chemspider.com (accessed Feb 11, 2014).
3. Dial a Molecule Grand Challenge. http://www.dial-a-molecule.org (accessed Feb 11, 2014).

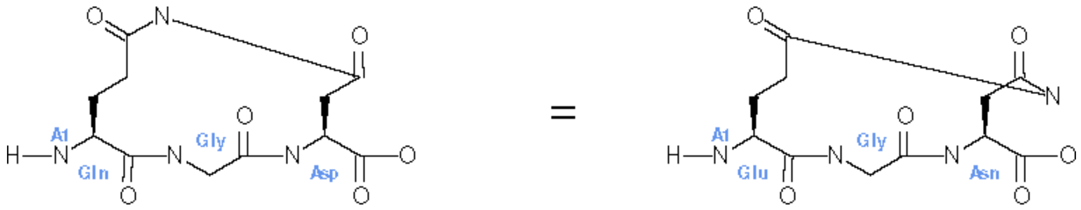
P-7: A computationally efficient structure key for large proteins

Gerd Blanke[1], **Jan H. Jensen**[2]

[1]StructurePendium Technologies GmbH, Essen, Germany, [2]Biochemfusion ApS, Copenhagen, Denmark

Work in progress: The poster will look further into the issues of structural representation of peptides and proteins.

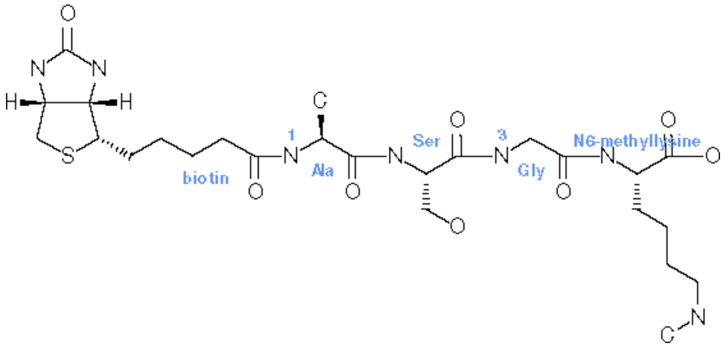
Protein structures that have chemical modifications or crosslinks are not uniquely described by their plain amino acid sequence. Obviously, the plain sequence does not capture sidechain and terminal modifications. Crosslinks complicate matters even further; e.g. lactam cyclization between Gln and Asp yields the same chemical structure as a lactam cyclization between Glu and Asn. Thus the plain sequences Gln-Gly-Asp and Glu-Gly-Asn will represent the same chemical structure if the first and third residues are lactam cyclizised.



One solution is to calculate the InChI key for the corresponding whole chemical structure but that is not feasible (nor efficient) for structures having more than 1024 atoms, roughly corresponding to a protein with > 130 residues. The approach pursued here is to keep to the protein plain sequence to the extent possible, but replacing chemically modified parts by the InChI key of those chemical fragments. This is very efficient when the majority of the protein structure is described by its plain sequence.

The structure key generation is trivial for sidechain- or terminal-modified residues. An example of this where we have a terminally-modified residue and another with a sidechain modification:

[biotin]-Ala-Ser-Gly-[N6-methyllysine]-OH (*)



Key = [JQGJUMKGKLLWLT-ZFLHDRRCSA-N]-Ser-Gly-[KLXPPJKQDSBAIA-FVMUQQIWSA-N]

The first InChI key above is the InChI key of the "[biotin]-Ala" fragment and the second is for the "N6-methyllysine" fragment. The generated InChI keys contain information about the fragment attachment points although InChI does not support attachment points. We have found that a careful selection of short-lived isotope-complexes can be used to represent attachment points reliably, and that the probability that these complexes occur by accident in reality can be kept to a minimum.

(*) Sequence is written in Biochemfusion PLN (Protein Line Notation) – see http://www.biochemfusion.com/doc/#Specifications

Special care is however required for crosslinked residues. Consider a three-residue peptide with residues one and three crosslinked. The structure key is then of the form

[fragment-key](1,a)-Gly-[fragment-key](1,b)

The number identifies the crosslink. The a/b. suffix is used to reference attachment point (pairs) within the crosslink.

The fragment key representing crosslinked residues has to take the following into account:

- The directionality of the crosslink fragment matters: The InChI key of a binary crosslink fragment as seen from the first residue to the second is different from the structure seen from second residue to first (since we encode attachment points).

- Chemical symmetry must be detected so attachment points do not carry “false information”. In the example peptide above, a fully symmetrical crosslink fragment should result in a structure key of “[key](1,a)-Gly-[key](1,a)”.
- The fragment key must be normalized so a direction-invariant version is recorded in the structure key.

P-9: Estimating Classification Uncertainty for Unbalanced Ensemble Models

Robert D. Clark, Wenkel Liang, Robert Fraczkiewicz, Marvin Waldman

Simulations Plus, Inc., Lancaster, CA, USA

Quantitative structure-activity relationships (QSARs) are poised to play an important role in regulatory decision-making [1], but for that to happen there has to be an increased focus on quantifying predictive uncertainty for individual compounds. Most research in the area has involved work on regression models (e.g., [2]) or aggregate predictivity rather than on assessment of the confidence one can have in individual predictive classifications.

Our group has recently shown that the distribution of errors for an artificial neural net ensemble (ANNE) classification model follows, to a good approximation, a beta binomial distribution with respect to the degree of classification consensus among the networks that make up the model [3]. The associated predictions follow a beta binomial distribution as well, albeit a different one. The ratio of the two distributions at any given tally can then be used to provide a good estimate of the likelihood that a predictive classification getting that number of positive votes is incorrect – i.e., its uncertainty. Moreover, we were able to show that the distributions of predictions and error in the training pool provide a good indication of the predictive uncertainties for large external validation sets.

The only models described in our earlier report [3] were those in which classification was determined by majority vote. In addition, the data sets analyzed were all more or less balanced, in that the numbers of positive and negative examples were similar. Here we describe modifications that extend the scope of the technique to include unbalanced data sets of the sort commonly encountered for properties involved in absorption, distribution, metabolism, excretion, and toxicity – i.e., ADMET properties. Finally, we demonstrate that the technique is not limited in applicability to one particular type of ensemble model: it works equally well for ensemble models in which classification is based on averaged network output as well as those which are based on voting.

1. Worth, A.P. The role of QSAR methodology in the regulatory assessment of chemicals. In *Advances in Computational Chemistry and Physics Volume 8: Recent Advances in QSAR Studies.* Puzyn, T.; Leszczynski, J.; Cronin, M.T., Eds. Springer: the Netherlands, 2010; pp 367.
2. Beck, B.; Breindl, A.; Clark, T. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1046–1051.
3. Clark, R.D.; Liang, W.; Fraczkiewicz, R.; Waldman, M. Estimating Classification Uncertainty for Ensemble Models. *6th Joint Sheffield Conference on Chemoinformatics*, http://cisrg.shef.ac.uk/shef2013/talks/31.pdf

P-11: Maximum Common Substructure-based Data Fusion in Similarity Searching and Virtual Screening

Edmund V. Duesbury, John Holliday, Peter Willett

University of Sheffield, Sheffield, UK

Two techniques exist in data fusion which have been proven to work in various forms in chemoinformatics: similarity fusion, where different similarity measures are combined; and group fusion, where similarities are combined from multiple reference molecules. Data fusion has been shown to work very well when applied to fingerprint-based similarity searching [1] , yet little is known on its application to Maximum Common Substructure (MCS)-based similarity searching. With recent reductions in the time complexity of MCS-finding algorithms, it is worth investigating MCS similarity methods on large numbers of compounds to better assess its feasibility.

Two applications of the MCS have been mentioned in the literature. The first method uses a similarity coefficient for similarity searching, relying on the MCS to determine the proportion of overlap between two molecules [2]. The other application is that of the chemical hyperstructure. The hyperstructure concept is an alternative form of data fusion, being a hypothetical molecule that is constructed from the overlap of a set of existing molecules. Initially proposed to reduce the time of database searching, it has also been used directly for virtual screening on two occasions since its inception [3,4], the latter of which showed it to be useful in QSAR studies. The concept's performance however in 2D similarity searching has to date not been evaluated thoroughly on large sets of compounds.

The work being carried out in this project aims to evaluate hyperstructures as a complementary method for fusion-based similarity-searching, with an emphasis on virtual screening. The poster will describe how hyperstructures are constructed, evaluated for virtual screening and compared with existing search methods. The main focus here is the comparison between fingerprint group fusion (ECFP_6 fingerprints), MCS similarity group fusion, and hyperstructure similarity, and data fusion of combinations of these three techniques. The BEDROC score [5] with an α of 32.2 is used to measure virtual screening performance – a higher value indicating a better recall of actives. To obtain a measure of diversity, the top 5% of active compounds obtained from each search method were converted to Bemis-Murcko frameworks, with bond and atom label information removed from the frameworks. The number of these active compounds per unique framework represents the diversity.

Results in this work show that the hyperstructure concept is consistently and significantly less effective for virtual screening, compared to the MCS and fingerprint group fusion techniques in terms of numbers of actives retrieved (mean BEDROC scores of 0.854, 0.651 and 0.460 for fingerprints, MCS and hyperstructures respectively). Hyperstructure searches however retrieve a greater diversity of active molecules than fingerprints, where the former obtained 1.565 active molecules per framework, compared to 1.934 for fingerprints.

The MAX rule fusion of the three types of search in pairwise combinations generally do not outright outperform pure fingerprint searches, although of note is that fingerprint and MCS fusion yields a comparable recall to fingerprints, and fingerprint and hyperstructure fusion, whilst somewhat less accurate than fingerprints – yields a greater diversity of frameworks (mean molecules per framework of 1.767). In addition, hyperstructures are shown to be quicker to search with than MCS group fusion, requiring about a quarter of the time to perform a similarity search yet with little loss in recall and improved active framework retrieval.

1. Willett, P. Combination of Similarity Rankings Using Data Fusion. *J. Chem. Inf. Model.* **2013**, *53*, 1–10.

2. Raymond, J. W.; Willett, P. Effectiveness of Graph-Based and Fingerprint-Based Similarity Measures for Virtual Screening of 2D Chemical Structure Databases. *J. Comput. Aided Mol. Des.* **2002**, *16*, 59–71.
3. Brown, N. Generation and Application of Activity-Weighted Chemical Hyperstructures. PhD Thesis, University of Sheffield: Sheffield, 2002.
4. Palyulin; Radchenko; Zefirov. Molecular Field Topology Analysis Method in QSAR Studies of Organic Compounds. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 659–667.
5. Truchon, J.-F.; Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the “Early Recognition” Problem. *J. Chem. Inf. Model.* **2007**, *47*, 488–508.

P-13: Web server for the rapid calculation of empirical atomic charges with QM accuracy

Stanislav Geidl[1,2], David Sehnal[1,2,3], Crina-Maria Ionescu[1], Radka Svobodová Vařeková[1,2], Purbaj Pant[1], Jaroslav Koča[1,2]

[1]CEITEC - Central European Institute of Technology, Masaryk University, Brno, Czech Republic, [2]National Centre for Biomolecular Research, Masaryk University, Brno, Czech Republic, [3]Faculty of Informatics, Masaryk University, Brno, Czech Republic

Partial atomic charges describe the distribution of electron density in a molecule, and therefore they provide clues regarding the chemical behaviour of molecules. Atomic charges are frequently used in molecular modelling applications such as molecular dynamics, docking, conformational searches, binding site prediction, etc. Recently, partial atomic charges have also become popular chemoinformatics descriptors. Charge calculation methods can be divided into two main groups, namely quantum mechanical (QM) approaches and empirical approaches. QM approaches provide accurate charges, but they are very slow and therefore not feasible for large sets of molecules. Empirical charges can be calculated quickly and their accuracy is similar to QM, making empirical charges more appropriate for chemoinformatics applications. A very useful empirical charge calculation method is EEM (Electronegativity Equalization Method) [1,2]. This method provides charges comparable to the QM approach for which the given EEM model was parameterized [3,4]. Many EEM parameterizations were published and new parameterizations are still in development [5].

We present a web server able to calculate EEM atomic charges for one or more molecules of any size. This application provides all EEM parameters sets published to date, and additionally allows custom configurations. The calculated charges can be visualized and compared via several statistical descriptors directly inside the web browser, or downloaded for further processing. The web server is able to process small drug-like molecules, as well as extremely large biomolecules (e.g., proteins and their complexes), owing to a novel approach that allows solving the EEM equation in a sophisticated, yet very efficient manner.

1. Mortier, W. J.; Ghosh, S. K.; Shankar, S. Electronegativity Equalization Method for the Calculation of Atomic Charges in Molecules. *J. Am. Chem. Soc.* **1986**, *108*, 4315–4320.
2. Svobodová Vařeková, R.; Koča, J. Optimized and Parallelized Implementation of the Electronegativity Equalization Method and the Atom-Bond Electronegativity Equalization Method. *J. Comput. Chem.* **2006**, 27, *3*, 396–405.
3. Bultinck, P.; Langenaeker, W.; Lahorte, P.; De Proft, F.; Geerlings, P.; Van Alsenoy, C.; Tollenaere, J. P. The Electronegativity Equalization Method II: Applicability of Different Charge Schemes. *J. Phys. Chem. A.* **2002**, *106*, 7895–7901.

4. Svobodová Vařeková, R.; Jiroušková, Z.; Vaněk, J.; Suchomel, S.; Koča, J. Electronegativity Equalization Method: Parameterization and Validation for Large Sets of Organic, Organohalogene and Organometal Molecules. *Int. J. Mol. Sci.* **2007**, *8*, 572–582.
5. Ionescu, C. M.; Geidl, S.; Svobodová Vařeková, R.; Koča, J. Rapid Calculation of Accurate Atomic Charges for Proteins via the Electronegativity Equalization Method. *J. Chem. Inf. Model.* **2013**, *53*, 2548–2558.

P-15: Towards a new chemical standardization pipeline in PubChem

Volker D. Hähnke, Evan Bolton, Stephen Bryant

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Department of Health and Human Services, Bethesda, Maryland, U.S.A.

PubChem is an open repository for molecular structures and their biological activities maintained by the National Center for Biotechnology Information (NCBI) [1]. PubChem includes two major databases containing chemical structures: Substance and Compound. PubChem Substance contains descriptions of chemical substances from individual contributors. PubChem Compound is derived from Substance through automated structure standardization protocols, identifying equivalent chemicals between depositors and generating a preferred chemical representation. This allows the chemical structure to be used to aggregate information between contributors.

Since the creation of PubChem in 2004, the number of contributed substance records is steadily increasing, reaching 127 million in early 2014. There are 48 million corresponding entries in PubChem Compound. Aggregation of information using chemical structure is not trivial. The entries in PubChem Substance stem from more than 250 contributing organizations, each with their own preference on how to represent a given chemical. Tautomerism, mesomerism and ionization additionally increase the number of valid – but not identical – representations of what is essentially the same structure.

Enforced standards for chemical representation do not exist. The IUPAC guidelines [2] and the FDA Substance Registration Manual [3] provide suggestions; however, even guidelines such as these are not generally followed and are predated by a large corpus of chemical structures. To correct this, it is common practice for software algorithms to be used to 'clean up' chemical structure representation. Sadly, automated processing software is not full-proof and can, at times, modify chemical representation in bizarre ways or corrupt the original information (e,g., by making it ambiguous). In addition, how a chemical is represented is important as it has direct ramifications on the predictive power of derived models (such as fragment-based LogP prediction) or concepts like molecular similarity (since some equivalent structure representations may be considered dissimilar), adding to the burden of structure standardization software when generating a preferred representation.

In the case of PubChem, each new contributor potentially adds previously unknown or unanticipated chemical representation. Modifying existing SMARTS-based pattern matching to account for an ever-increasing variation of representations adds to the risk of unknown, unwanted, or unpredictable behavior (such as fixing too much or too little, or conflicting rules). For this reason, PubChem is transitioning away from SMARTS-based pattern matching to a more flexible rule-based standardization system that not only increases the efficiency of the standardization process but also allows for easy modification and validation. Generally speaking, the new standardization pipeline is based on spherical atom environments of appropriate size. These have the advantage of being easy and fast to generate, while covering most known functional groups. The new system is parameterized to reproduce the results of the previous PubChem standardization protocols, and then curated to improve upon these results.

In this presentation, the general concept of atom-environment based structure standardization will be outlined. This will cover the presentation of basic statistics for fragments derived from the PubChem Substance and PubChem Compound databases. Prominent examples of functional groups particularly affected by the aforementioned effects are highlighted along with examples of erroneous structures beyond hope deposited in PubChem. Furthermore, cases of added standardization rules will be shown that widen the applicability of the new approach, while presenting circumstances still subject to research for a meaningful standardized representation. Finally, potential future directions using statistic-based approaches to detect and remove valid but improbable structures are considered.

1. Bolton, E.; Wang, Y.; Thiessen, P. A.; Bryant, S.H. PubChem: Integrated Platform of Small Molecules and Biological Activities. In *Annual Reports in Computational Chemistry. Volume 4*; Wheeler, R. A.; Spellmeyer, D. C., Eds., Elsevier: Amsterdam, 2008, pp 217.
2. Brecher, J. Graphical representation standards for chemical structure diagrams (IUPAC Recommendations 2008). *Pure. Appl. Chem.* **2008**, *80*, 277–410.
3. Food and Drug Administration Substance Registration System Standard Operating Procedure Substance Definition Manual, http://www.fda.gov/downloads/ForIndustry/DataStandards/SubstanceRegistrationSystem-UniqueIngredientIdentifierUNII/ucm127743.pdf (accessed Feb 14, 2014).

P-17: Recent Advancements in Chemoinformatics for Porous Materials

Maciej Haranczyk, Richard L. Martin

Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

We present recently developed techniques for high-throughput analysis and discovery of porous materials, inspired by established chemoinformatics approaches for small molecules. We describe how concepts from similarity searching and screening of small molecule databases can be extended to overcome the unique challenges presented by periodic, porous crystal structures.

Crystalline porous materials have been exploited in industrial applications for many years, for instance as catalysts for oil refinement, water softeners, and membranes for separations. The scale of their use is enormous, and their commercial impact reaches hundreds of billions dollars globally. Moreover, they are also promising candidates for emerging technologies such as economically competitive vehicular natural gas storage and carbon dioxide capture. Porous materials exhibit complex networks of internal space, which permit the diffusion of guest chemical species. The size, shape and connectivity of a material's pore structure determine the interactions which occur between the guest and the host material.

Identifying materials with the specific pore structure advantageous for a particular application is a great challenge. For example, zeolites, the most well-known class of porous materials, consist solely (in their simplest, all-siliceous form) of tetrahedral arrangements of silicon and oxygen; nevertheless, approximately 200 unique zeolite topologies are known to exist, and over two million hypothetically achievable zeolite topologies have been computationally enumerated. With such a vast search space of possible materials, it is clear that exhaustive synthesis is not a feasible strategy for materials discovery, and so there is a pressing need for computational methods for high-throughput analysis of porous materials.

In this contribution we present recently developed techniques for the representation, comparison and screening of very large sets of porous materials. We describe Voronoi holograms, pore descriptors based on the Voronoi decomposition, and a similarity function tailored for their comparison. We illustrate similarity searching and dissimilarity-based selection for porous materials using this approach, which has enabled the automated discovery of similar structures and construction of diverse training sets of materials. Finally, we describe how these advancements have facilitated a high-throughput database screening technique which has identified thousands of previously unexplored promising candidate materials for carbon dioxide capture.

1. Haranczyk, M.; Sethian, J.A. Automatic structure analysis in high-throughput characterization of porous materials. *J. Chem. Theory Comput.* **2010**, *6*, 3472–3480.
2. Martin, R. L.; Willems, T. F.; Lin, L.-C.; Kim, J.; Swisher, J.; Smit, B.; Haranczyk, M. Similarity-driven Discovery of Porous Materials for Adsorption-based Separations. *ChemPhysChem*, **2012**, *13*, 3595–3597.
3. Martin, R. L.; Smit, B.; Haranczyk, M. Addressing challenges of identifying geometrically diverse sets of crystalline porous materials. *J. Chem. Inf. Model.* **2012**, *52*, 308–318.

P-19: Chemogenomics analysis of small molecule bioactivity data: Privileged scaffolds and conserved structural elements in proteins

Lina Humbeck[1], Till Schäfer[2], Nils Kriege[2], Petra Mutzel[2], Oliver Koch[1]

[1]Department of Chemistry and Chemical Biology, [2]Department of Computer Science, TU Dortmund, Dortmund, Germany

The term “privileged scaffolds” is often used for multiple molecules that show bioactivity on different targets but consist of the same scaffold [1]. The question, whether the interaction of such scaffolds is also represented by conserved structural elements in the respective protein target, still remains open, although the protein folding space is limited and occasional examples indicate that this assumption is reasonable[2]. Similar structural elements can be found in different proteins, ranging from conserved motifs that interact with specific functional groups to proteins with similar folding but different function that still can bind similar ligands [2].

Unfortunately, software that allows identifying similar structural elements in protein structures independent from the overall fold is still under development and is not yet available. Therefore, we decided to analyze molecule databases that also contain bioactivity data (like the ChEMBL database - www.ebi.ac.uk/chembldb/) in combination with structures of the protein targets to identify possible structural similarities.

For this analysis the tool Scaffold Hunter is used [3], which is a tool for the visual analysis of chemical compound databases. It provides integrated visualization as well as analysis of biological activity data and fosters the interactive exploration of data imported from a variety of sources. Based on the underlying scaffold tree algorithm, each molecule is associated with its unique scaffold, which is stepwise simplified by removing single rings. This allows identifying groups of molecules sharing similar scaffolds even in huge databases.

We will present the result of this exhaustive chemogenomics analysis of small molecule bioactivity data and we will show if privileged scaffolds also interact with conserved structural elements in proteins.

1. Welsch, M.E.; Snyder, S.A.; Stockwell, B.R. Privileged scaffolds for library design and drug discovery. *Curr. Opin. Chem. Biol.* **2010**, *14*, 347–361.

2. Koch, O. The Use of Secondary Structure Element Information in Drug Design: Polypharmacology and Conserved Motifs in Protein-Ligand Binding and Protein-Protein Interfaces. *Fut. Med. Chem.* **2011**, *3*, 699–708.
3. Klein, K.; Koch, O.; Kriege, N.; Mutzel, P.; Schäfer, T. Visual Analysis of Biological Activity Data with Scaffold Hunter. *Mol. Inf.* **2013**, *32*, 964–975.

P-21: Identification of functionally active residues in α_1-AR by computational approaches

Kapil Jain, Lotten Ragnarsson, Åsa Andersson, Richard J. Lewis

The University of Queensland, Australia

Allosteric binding sites of G protein-coupled receptors (GPCRs) have emerged as important targets for drugs with improved selectivity profiles for the treatment of a broad range of diseases. To identify functionally important residues specific to α_1-adrenergic receptor (α_1-AR), which regulates vasoconstriction, hypertension and anxiety disorders, we built a homology model of α_{1B}-AR based on the crystal structure of the turkey β_1-AR. We then performed steered and accelerated molecular dynamics to identify residues that could regulate conformational changes during the egress of NE and receptor activation respectively. Residues predicted to be functionally significant were mutated to alanine and their effect on agonist binding and efficacy were determined. Of the fourteen residues predicted by steered dynamics, mutations on Tyr-110, Trp-121, Tyr-203 and Cys-195 caused significant drops in affinity of NE by 1000 and 500 folds respectively. An additional 16 residues identified as being involved in receptor activation were also experimentally tested and the consequence of these mutations assessed using accelerated molecular dynamics. Of these, mutations at Cys-195 and Asp-190 caused 1000- and 250-fold drops in NE affinity, respectively, demonstrating their role in receptor activation. Of the five functionally significant residues identified, two residues, Tyr-110 and Asp-190 are unique to the α_{1B}-AR, providing an opportunity for the design of subtype specific drugs. This study demonstrates the utility of computational approaches in structure-activity relationships to guide the rational design of drugs with improved therapeutic potential.

P-23: HELM: an open standard for handling large Biologics

Roland Knispel[1], Tianhong Zhang[2], Claire Bellamy[3], Sergio Rotstein[4], HELM Project Team[5], Iván Solt[1]

[1]Business Analyst Biologics, ChemAxon, Budapest, Hungary, [2]Manager Research Informatics, Pfizer, Cambridge, MA, USA, [3]HELM Project Manager, Pistoia Alliance, Wakefield, MA, USA, [4]Director Research Business Technologies, Pfizer, Cambridge, MA, USA, [5]Pistoia Alliance, Wakefield, MA, USA

The Pharmaceutical Industry faces many challenges these days. While the annual number of novel medical entities (NMEs) is dwindling, so do the revenues from many blockbuster-selling drugs as they reach the end of their patent protection. At the same time increasingly strict safety requirements and assessments of the extra benefit over existing therapeutics, which are demanded by regulatory authorities, are posing challenging hurdles to overcome for NME approval.

This situation has led many organizations within this industry to dramatically reshape their R&D processes. An increasing number of chunks of the R&D cycle are "externalized" to numerous highly specialized partner organizations, thus creating a highly interconnected ecosystem, in which innovation is accelerated and overall development costs are lowered and shared between these partners. New strategies and technologies helping to nurture these ecosystems need to be established, i.e. enabling barrier-free and efficient data exchange between partners; efforts, which are best solved in the pre-competitive domain. Actively engaged in supporting this new R&D paradigm is the Pistoia Alliance, a unique, formally-incorporated, cross-company, open-innovation organization with over 80 members, established to lowering the barriers to innovation by business process optimization in the pre-competitive domain.

By initiating the HELM project the Pistoia Alliance acknowledged the lack of tools for properly representing the molecular structure of biological macromolecules, which substantially contribute to the current R&D portfolios of the pharmaceutical industry. Existing tools were primarily designed to work with either small molecules or unmodified, unconjugated amino acid and nucleotide sequences. The Hierarchical Editing Language for Macromolecules (HELM), previously developed by Pfizer researchers (see J. Chem. Inf. Model 2012, 52, 2796-2806), enables the representation of a wide range of biomolecules (e.g. proteins, nucleotides, antibody drug conjugates) through a hierarchical notation that represents complex macromolecules as polymeric structures with support for unnatural components (e.g. unnatural amino acids) and chemical modifications.

In early 2013 the Pistoia Alliance formalized the HELM notation as an open standard and prepared the existing proprietary editor and toolkit for public release as reference implementation of the HELM standard. After scoring this first project milestone, facilitating adoption of HELM as an industry standard for the manipulation and exchange of complex biomolecule data, gradual evolution of the standard according to industry needs, and setting up an organizational infrastructure to govern such changes are on the agenda.

The first part of this talk will briefly summarize current key activities of the Pistoia Alliance followed by insights into the HELM project – sharing challenges faced, and a few lessons learned while pursuing to establish HELM as an open standard. The second part will elaborate on leveraging the HELM standard for transforming and consolidating structure management of large bio-molecules with novel tools, currently in the making.

P-25: Bioisosteres in accessible chemistry space

Mark Mackey, Tim Cheeseright, Rae Lawrence, Martin Slater

Cresset, Cambridge, UK

Searching for bioisosteric replacements is a valuable part of a medicinal chemist's toolbox. A bioisosteric core replacement can solve an ADMET or IP issue and move development into a new lead series, while bioisosteric replacements for leaf groups enable fine tuning of molecular properties without affecting the fundamental activity.

A fundamental limitation of most current bioisosteric replacement search tools is the synthesisability of their suggestions. Existing software solutions either leave it up to the user to triage the results according to synthetic accessibility, or provide a complexity or synthesisability score to the user. However, these scores generally correlate poorly with medicinal chemists' assessments and struggle further when trying to distinguish the relative synthesisability of a set of structurally related compounds.

We present an alternative approach in which the chemist is able to define the accessible synthetic space around the core of their lead molecule. The search for bioisosteres is confined to this space, so that the results are all known to be synthesisable using accessible reagents. In order to do this, multiple databases of fragments are created from the reagent sets and classified according to the synthons present and the desired chemistries.

Despite the very large number of chemical reactions in the modern medicinal chemistry tool set, we have found that a very limited number of synthetic transforms are needed to fully represent this space. By focusing on the structural transformation rather than the chemical reaction, many different chemistries can be summarized into a small set of rules.

We implement these rules in a special-purpose chemical transformation language, ATPAT. The ATPAT language combines a molecular regular expression syntax that is simpler, more extensible and more powerful than SMARTS with a set of simple transformation procedures that can be automatically applied on a successful match. The ATPAT engine allows new chemical transformation rules to be generated effortlessly.

The result is an integrated system that allows the chemist to easily process his or her available reagents into a list of potential molecules to make. Wrapping this system in a KNIME or Pipeline Pilot wrapper allows automation and simple integration into existing cheminformatics systems.

P-27: A Neural Gas based Approach towards Pharmacophore Model Elucidation

Daniel Moser[1,2], Sandra Wittmann[1], Ewgenij Proschak[1,2]

[1]Institute of Pharmaceutical Chemistry, Goethe University, Frankfurt, Germany, [2]German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany

The pharmacophore concept is commonly employed in virtual screening for hit identification. A pharmacophore is generally defined as the three-dimensional arrangement of the structural and physicochemical features of a compound responsible for its affinity to a pharmacological target [1,2]. Given a number of active ligands binding to a particular target in the same manner, it can reasonably be assumed that they share some common features, a common pharmacophore. Although there are a number of tools available for the elucidation of such common pharmacophore models, development is still ongoing [3]. The improvement of computational time for pharmacophore elucidation from large ligand data sets is in the focus of current research efforts. We present a Growing Neural Gas (GNG)-based approach for the extraction of the relevant features. The GNG, developed by Fritzke [4], can be seen as an advancement of the general Neural Gas (NG) algorithm presented by Martinetz and Schulten [5] in the early Nineties. A (G)NG is an artificial neural network which is used in different application fields, e.g. image processing [6], cluster analysis [7], and pattern recognition, but is also employed in chemoinformatics-related applications [8]. Our approach consists of four general steps: alignment of ligand conformations based on their pharmacophore annotations, feature extraction using a GNG, refinement of pharmacophore feature coordinates and radii, and scoring of the generated pharmacophore. Results of retrospective validation indicate an acceptable quality of the generated models along with shorter computational times. Additionally a prospective virtual screening for leukotriene A4 hydrolase (LTA4H) inhibitors was performed. LTA4H is a bifunctional zinc metalloprotease which displays both epoxide hydrolase and aminopeptidase activity [9]. It catalyses the stereo selective hydrolysis of leukotriene A4 to the inflammatory lipid mediator leukotriene B4 in a two-step reaction [10], as a part of the arachidonic acid cascade.

1. Ehrlich, P. Über Den Jetzigen Stand Der Chemotherapie. *Ber. Dtsch. Chem. Gesell.* **1909**, *42*, 17–47.
2. Kier, L. B. Molecular Orbital Calculation of Preferred Conformations of Acetylcholine, Muscarine, and Muscarone. *Mol. Pharmacol.* **1967**, *3*, 487–494.
3. Leach, A. R.; Gillet, V. J.; Lewis, R. A.; Taylor, R. Three-Dimensional Pharmacophore Methods in Drug Discovery. *J. Med. Chem.* **2010**, *53*, 539–558.
4. Fritzke, B. A Growing Neural Gas Network Learns Topologies. *Adv. Neural Inf. Process. Syst.* **1995**, *7*, 625–632.
5. Martinetz, T.; Schulten, K. A "Neural-Gas" Network Learns Topologies. *Artif. Neural Networks* **1991**, *1*, 397–402.
6. Angelopoulou, A.; Psarrou, A.; Rodríguez, J.; Revett, K. Automatic Landmarking of 2D Medical Shapes Using the Growing Neural Gas Network. In *Computer Vision for Biomedical Image Applications*; Liu, Y.; Jiang, T.; Zhang, C., Eds.; Springer: Berlin / Heidelberg, 2005; Vol. 3765, pp. 210–219.
7. Canales, F.; Chacón, M. Modification of the Growing Neural Gas Algorithm for Cluster Analysis. In *Progress in Pattern Recognition, Image Analysis and Applications*; Rueda, L.; Mery, D.; Kittler, J., Eds.; Springer: Berlin / Heidelberg, 2007; Vol. 4756, pp. 684–693.
8. Weisel, M.; Kriegl, J.; Schneider, G. PocketGraph: Graph Representation of Binding Site Volumes. *Chem. Cent. J.* **2009**, *3*, P66.
9. Haeggström, J. Z.; Kull, F.; Rudberg, P. C.; Tholander, F.; Thunnissen, M. M. G. M. Leukotriene A4 Hydrolase. *Prostaglandins Other Lipid. Mediat.* **2002**, *68-69*, 495–510.
10. Paige, Y. M. S. Inflammatory Diseases - Immunopathology, Clinical and Pharmacological Bases; Khatami, M., Ed.; 2012.

P-29: Vectorizing hydrophobicity: From in silico models to membrane-active peptides

Max Pillong, Yen-Chu Lin, Katharina Stutz, Sarah Haller, Petra Schneider, Jan Hiss, Gisbert Schneider

Institute of Pharmaceutical Sciences, ETH Zürich, Switzerland

With dramatically increasing pathogen resistances to classic antibiotics, antimicrobial peptides (AMPs) represent a promising class of potential future antimicrobial agents[1]. The modes of action of AMPs range from the inhibition of metabolic pathways, over membrane rupture by peptide aggregation, to the formation of pore-like multimeric complexes spanning bacterial lipid membranes.

Early studies aimed to classify AMPs based on their amino acid sequence[2]. Only few attempts are based on their structural aspects and even less attempts have targeted the design of an AMP exhibiting a specific mode of action. This is due to lacking structural data for AMPs and uncertainty in the required biophysical properties of their different modes of action.

In this study, we address these issues by the compilation of a computationally motivated *in silico* test set of peptide structures and the implementation of a novel, hydrophobicity-based peptide representation aiming at the identification of synthetic pore-forming AMPs. The descriptor is targeted at global, as well as local physicochemical profiles[3] in three-dimensional space in correlation to the peptide's sequence. The resulting descriptor allows for rapid comparison of peptides with regard to these properties.

We compared known pore-forming AMPs with trans-membrane domains of integral membrane proteins and the *in silico* generated peptide structures. We succeeded in the prospective application and identified an innovative, membrane-active designer peptide, exhibiting similar behavior to a known pore-forming AMP. Membrane activity was investigated in several biophysical assays. The conducted experiments include vesicle rupture assays, isothermal titration calorimetry (ITC) measurements and atomic force microscopy (AFM) imaging.

1. Fjell, C. D.; Hiss, J. A.; Hancock R. E. W.; Schneider, G. Designing Antimicrobial Peptides: Form Follows Function. *Nat. Rev. Drug. Discov.* **2011**, *11*, 37–51.
2. Jenssen, H. Descriptors for Antimicrobial Peptides. *Expert Opin. Drug. Discov.* **2011**, *2*, 171–184.
3. Eisenberg, D. Three-dimensional Structure of Membrane and Surface Proteins. *Annu. Rev. Biochem.* **1984**, *53*, 595–623.

P-31: An Upper Bound to the Effectiveness of Substructural Analysis Methods

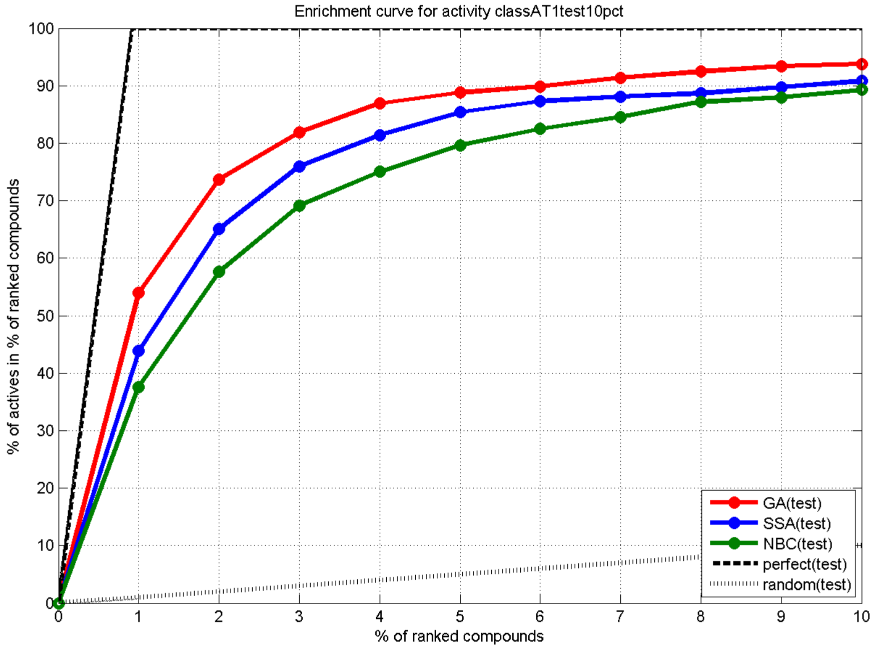
Nor S. Sani[1,2], John Holliday[1], Peter Willett[1]

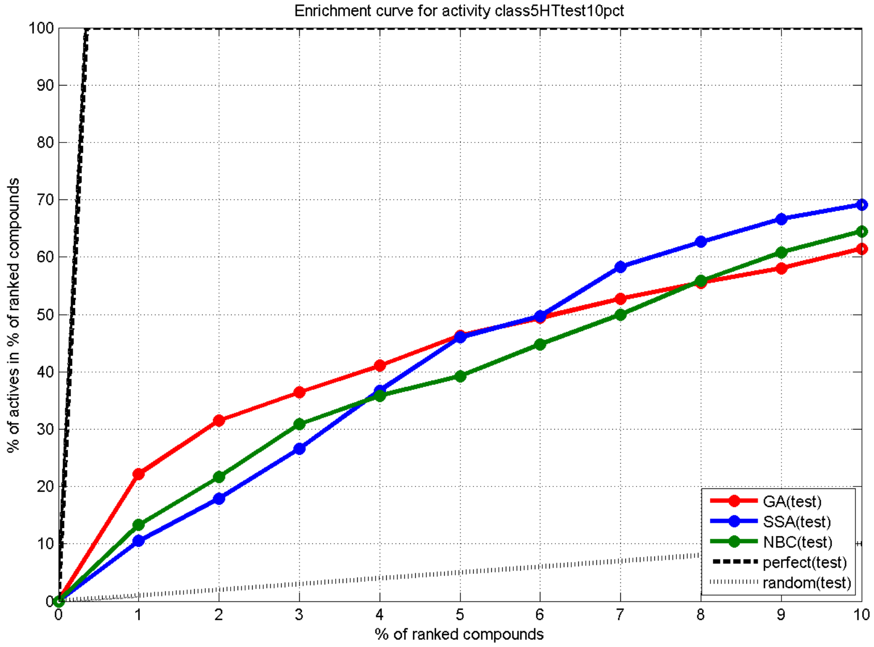
[1]University of Sheffield, Sheffield, UK, [2]National University of Malaysia, Bangi Selangor, Malaysia

Substructural analysis (SSA) was one of the very first machine learning techniques to be applied to chemoinformatics. Given a set of compounds that have been tested in some together with fragment occurrence data (typically in the form of 2D fingerprints) for those compounds, biological screen, SSA computes weights for each of the fragments that denote its contribution to the activity (or inactivity) of compounds containing that fragment [1]. A simple fragment weight, for example, might be the fraction of the active compounds in a training-set that contain that fragment. The overall probability of activity for a compound is then computed by summing (or otherwise combining) the weights for the fragments present in that compound. A variety of weighting schemes are available for this purpose [2]. The approach, which was first described some 40 years ago [1], is very closely related to a naïve Bayesian classifier (NBC) [3], a machine learning method that has become very popular in the last few years with its availability in the Pipeline Pilot system [4].

This poster reports work that seeks to identify an upper-bound to the effectiveness of SSA methods using a genetic algorithm (GA) that we have developed for the calculation of fragment weights based on 2D fingerprints. The chromosome for the GA is a vector containing N real numbers, where the i-th element is the fragment weight for the i-th bit in the fingerprint; and the fitness function for the GA is the number of active molecules that occur in the top-1% of a ranking of a training-set of active and inactive molecules when the molecules are ranked using the set of N weights encoded in a chromosome. The GA, which uses single-point crossover and single-bit mutation, is run for a pre-set number of generations or until the weights have stabilized, thus providing an estimate of the best possible SSA weights that can be obtained using that training-set. The resulting weights can then be applied to a separate test-set. Our experiments used the MDL fingerprints that can be generated using the Pipeline Pilot NBC software and an MDDR dataset containing 102,540 compounds and comprising eleven different bioactivity classes. The training-set contained 10% of the actives for a class and 10% of the inactives, with the test-set comprising the remaining 90% of the dataset.

Building on previous work [2,3], we compared a total of nine published SSA weighting schemes, and found that the most effective across the eleven bioactivity classes was the R4 weight (which is based on work originally carried out by Robertson and Sparck Jones in the context of text search engines). This was found to perform at broadly the same level of effectiveness as the naïve Bayesian classifier in the Bayesian modeling routine in the Pipeline Pilot; however, both of these were slightly less effective at the very top of the ranked molecules than the GA-based weighting scheme, though the differences were not large in some cases. Examples of the enrichment curves for two of the bioactivity classes are shown below, where red, blue and green denote the GA, the R4 SSA weight and the Pipeline Pilot NBC, respectively. We draw two conclusions from these experiments. First, that a GA provides a possible non-deterministic method for generating the fragment weights for use in SSA-based virtual screening. Second, and more importantly, that the results obtained are only slightly more effective than those obtained from existing, deterministic methods for generating such weights. Given the second finding, it would appear that the existing methods perform at a level not far from the best possible level that can be achieved using this approach to ligand-based virtual screening.





1. Cramer, R. D.; Redl, G.; Berkoff, C. E. *J. Med. Chem.* **1974**, *17*, 533–535.
2. Ormerod, A.; Willett, P.; Bawden, D. *Quant. Struct.-Act. Rel.* **1989**, *8*, 115–129.
3. Hert J.; Willet,t P.; Wilton, D.J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. *J. Chem. Inf. Model.* **2006**, *46*, 462–470.
4. Rogers, D.; Brown, R. D.; Hahn, M. *J. Biomol. Screen.* **2005**, *10*, 682–686.

P-33: A machine learning-based protocol for docking results analysis

Sabina Smusz[1,2], **Stefan Mordalski**[1], **Jagna Witek**[1], **Krzysztof Rataj**[1], **Andrzej J. Bojarski**[1]

[1]Department of Medicinal Chemistry, Institute of Pharmacology Polish Academy of Sciences, Kraków, Poland, [2]Faculty of Chemistry, Jagiellonian University, Kraków, Poland

Docking is an important part of virtual screening campaigns and belongs to the group of the most popular chem- and bioinformatics procedures [1]. Its aim is to predict the conformation of the ligand and the receptor in their complex. A major challenge, however, is still connected with the analysis of the results. Although advanced scoring schemes were developed for the prediction of interaction energies between ligand and the protein, they still do not fully solve the problem, nor the visual inspection, which is very time-consuming and subjective, especially in case of large number of diverse compounds.

In this study, a novel protocol for automatic evaluation of massive docking results is proposed. It is a combination of the description of docking results in the form of a string with machine learning approach [2].

The docking results are described by means of Structural Interaction Fingerprints (SIFts) [3] and Spectrophores [4]. SIFts provide information about the interactions between ligand and each of the amino acids of the receptors, whereas Spectrophores are the source of information about the conformation of the docked compound, as they consist of atomic properties values calculated in a way that is dependent on the actual spatial orientation of a molecule. Such prepared representation of ligand-receptor complexes constitutes an input for machine learning experiments, with the use of 5 different classification algorithms, followed by multi-step results analysis taking into account the quality of the model of the receptor structure, various conformations of the docked compound and the performance of particular docking algorithm.

The pilot studies were performed for the serotonin receptors 5-HT$_6$ and 5-HT$_7$, however the tool is constructed in a way enabling its application for any target.

Acknowledgments: The study is supported by a grant PRELUDIUM 2013/09/N/NZ2/01917 financed by the National Science Centre.

1. Breda, A.; Basso, L. A; Santos, D. S.; De Azevdo, W. F., Jr; Virtual Screening of Drugs: Score Functions, Docking and Drug Design. *Curr. Comput. Aided Drug. Des.* **2008**, *4*, 265–272.
2. Kotsiantis, S. B. Supervised Machine Learning: A Review of Classification Techniques. *Informatica* **2007**, *31*, 249–268.
3. Deng, Z.; Chuaqui, C.; Singh, J. Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions. *J. Med. Chem.* **2004**, *47*, 337–344.
4. Bultinck, P.; Langenaeker, W.; Lahorte, P.; De Proft, F.; Geerlings, P.; Van Alsenoy, C.; Tollenaere, J. P. The Electronegativity Equalization Method II: Applicability of Different Atomic Charge Schemes. *J. Phys. Chem.* **2002**, *106*, 7895–7901.

P-35: Molecular Fragment Dynamics Study of the Interaction between Zinc Ricinoleate and the Complexing Agent Methylglycinediacetic Acid as a new System for Enzyme Purification

Karina van den Broek[1], **Annamaria Fiethen**[2], **Andreas Truszkowski**[3], **Achim Zielesny**[1], **Hubert Kuhn**[2]

[1]Institute for Bioinformatics and Chemoinformatics, Westphalian University of Applied Sciences, Recklinghausen, 45665, Germany, [2]CAM-D Technologies GmbH, Essen, 45127, Germany, [3]Inorganic Chemistry and Center for Nanointegration, University of Duisburg-Essen, Essen, 45141, Germany

Large-sized mesoscopic systems have been analyzed effectively with appropriate simulation techniques like atomistic Molecular Dynamics Simulations (MD) [1] and the recently developed coarse-grained Molecular Fragment Dynamics method (MFD) which is based on the Dissipative Particle Dynamics technique (DPD) [2,3]$^{)}$.

In MD the simulated molecules are calculated by interactions of single atom types. The intra and intermolecular forces in MFD are no longer described in an atomistic manner. In MFD atoms are grouped together to molecular fragments which represent their chemical properties [4]. Consequently, in MFD molecules are constructed as chains of connected beads. Due to the reduction of the size of the model the MFD method is useful for simulations of specific dynamical processes at the microsecond and micrometer scale. MFD can be generally applied in Material Engineering and Life Science simulation studies.

Zinc ricinoleate ($Zn(Ri)_2$) is a zinc salt of ricinoleic acid. The chemical structure consists of one Zn^{2+} ion stabilized by two molecules of ricinoleic acid. The metal soap became known as a substance which has the ability to work as a pollutant or odor absorber [5]. Since $Zn(Ri)_2$ was proposed to interact with substances containing sulphuric- and nitrogen-containing functional groups it was in addition successfully developed as a separation media in the process of protein and enzyme purification [6,7]. Experimental data showed that the activity of $Zn(Ri)_2$ can be significantly increased by activating the molecule with a complexing agents like methylglycinediacetic acid (MDGA) [7–9].

In perfect agreement with experimental results, MD and MFD simulations revealed the mechanism of activation of $Zn(Ri)_2$ by MGDA. In an aqueous solution this activated $Zn(Ri)_2$ changes to a conformational state in which the zinc atom turns out to be more available for electron donor atoms. Therefore, this activated molecular conformation supports a nucleophilic attack from the exterior.

The simulation technique of MFD was applied to study the interaction of $Zn(Ri)_2$ with proteins and amino acid in detail. With the results of these simulation studies a detailed insight into this complex mechanism could be obtained.

1. van Gunsteren, W. F.; Berendsen, H. J. C. Computer Simulation of Molecular Dynamics: Methodology, Applications, and Perspectives in Chemistry. *Angew. Chem. Int. Ed.* **1990**, *9*, 992–1023.
2. Hoogerbrugge, P. J.; Koelman, J. M. V. A. Simulating Microscopic Hydrodynamic Phenomena with Dissipative Particle Dynamics. *Europhys. Lett.* **1992**, *3*, 155–160.
3. Koelman, J. M. V. A; Hoogerbrugge, P. J. Dynamic Simulations of Hard-Sphere Suspensions Under Steady Shear. *Europhys. Lett.* **1993**, *3*, 363–368.
4. Truszkowski, A.; Epple, M.; Fiethen, A.; Zielesny, A.; Kuhn, H. Molecular fragment dynamics study on the water–air interface behavior of non-ionic polyoxyethylene alkyl ether surfactants. *J. Coll. Interface Science* **2013**, *410*, 140–145.
5. Lowicki, N.; Sidillo, M.; Neunhoeffer, O. Toxikologische und hautspezifische Eigenschaften des GRILLOCIN. Fette, Seifen, *Anstrichm.* **1974**, *3*, 136–140.

6. Beutel, S.; Kasper, C.; Kuhn, H.; Ralla, K.; Ruf, F.; Sander, F.; Prof. Dr. Scheper, T.; Sohling, U.; Thie, G. A process for the preparation of separating media for purification and / or isolation of enzymes and / or proteins. Süd-Chemie AG; publication number: DE102011101880 A1, **2012.**
7. Kuhn, H.; Müller, F.; Peggau, J.; Zekorn, R. Mechanism of the odor-adsorption effect of zinc ricinoleate. A molecular dynamics computer simulation. *J. Surfact. Deter.* **2000**, *3*, 335–343.
8. Kuhn, H.; Thie, G. Surface-active metal complexes for adsorbing noxious substances and method for producing the same. Cam-d Technologies Gmbh; publication number: EP2222684 A2, **2008**.
9. Kuhn, H.; Ruf, F.; Sohling, U.; Thie, G. Solid odour adsorbent based on zinc ricinoleates and related compounds. Süd-Chemie Ag; publishing number: EP2108446 A1, **2009**.

P-37: Heteromeric assembly of voltage gated ion channels

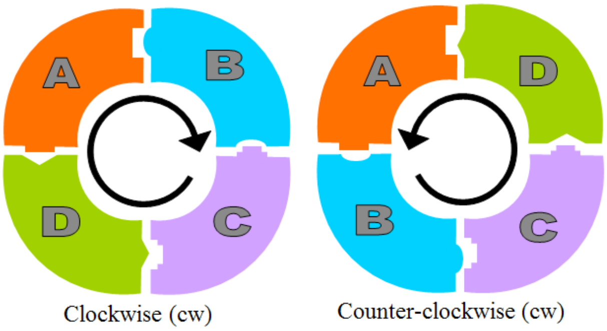
Guido Humpert[1,2], **Daniel Hoffmann**[2], **Achim Kless**[1]

[1]Grünenthal, Drug Discovery, Aachen, Germany, [2]University Duisburg-Essen, Bioinformatics, Germany

In this paper we focus on the analysis of the formation of heteromeric voltage gated ion channels. In contrast to bacterial homomeric ion channels that have been investigated in the last decade by crystallographic methods the assembly of heteromeric human ion channels has yet to be determined. For that reason we have built several homology models of human calcium and sodium channels with up to 2000 AA to study their assembly by computational methods. For all models we have built the principle possible orientations like clockwise, counter-clockwise and cross-linked. Subsequently we have optimized these decoys by energy minimization and molecular dynamics simulations.

The protein interfaces between the domains as well as the formed patches of these models have been analyzed by various energetic measures (e.g., Rosetta scores, forcefield energies) at different time points.

As a result the key drivers for the domain associations will be summarized with respect to the applied methods.



1. *Molecular Operating Environment (MOE)*, 2013.08; Chemical Computing Group Inc., 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, **2013**.
2. Ronald, A.; Li, Irene L.; Ennis, Robert J. French; Samuel C. Dudley, Jr.; Tomaselli , Gordon F. and Marbán, Eduardo Clockwise Domain Arrangement of the Sodium Channel Revealed by μ-Conotoxin (GIIIA) Docking Orientation. *J. Biol. Chem.* **2001**, *276*, 11072–11077.
3. Dudley, S. C., Jr.; Chang, N.; Hall, J.; Lipkind, G.; Fozzard, H. A.; French, R. J. μ-Conotoxin Giiia Interactions with the Voltage-Gated Na+ Channel Predict a Clockwise Arrangement of the Domains. *J. Gen. Physiol.* **2000**, *116*, 679–690.
4. Leaver-Fay, A.; Tyka, M.; Lewis, S. M.; Lange, O. F.; Thompson, J.; et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules *Methods Enzymol.* **2011**, *487*, 545–574.

P-39: HCS-fingerprints opening new routes to target identification

Felix Reisen, Amelie Sauty de Chalon, Martin Pfeifer, Xian Zhang, Daniela Gabriel, Paul Selzer

Novartis Institutes for BioMedical Research, Novartis Campus Forum 1, Basel CH-4056, Switzerland

High content screening (HCS) is a powerful tool for drug discovery being capable of measuring cellular responses to chemical disturbance in a high throughput manner. It provides an image based highly multiplexed and quantitative readout of cellular phenotypes including features such as shape, intensity or texture. The corresponding feature vectors can be used to characterize cellular phenotypes and are thus defined as HCS-fingerprints [1]. HCS-fingerprints can be employed for multivariate hit-calling and for image-based profiling to support target hypothesis generation for uncharacterized compounds [2,3].

We present the analysis and exploration of a novel dataset consisting of marketed drugs which were tested for their phenotypic effects on 6 different cellular compartments (nucleus, cytoplasm, Golgi apparatus, endoplasmic reticulum, mitochondria, and cytoskeleton). By analyzing phenotypic consistencies among treatments we were able to link a broad range of targets to characteristic phenotypes. We further investigated phenotypic similarities among modulation of targets that are known to be part of the same pathways. This revealed several biological processes for which distinctive cellular phenotypes could be observed. In the context of pharmaceutical research, image-based profiling can thus be employed to support target identification for compounds originating from phenotypic drug discovery campaigns.

1. Reisen, F.; Zhang, X.; Gabriel, D.; Selzer, P. Benchmarking of Multivariate Similarity Measures for High-Content Screening Fingerprints in Phenotypic Drug Discovery. *J. Biomol. Screen.* **2013**, *18*, 1284–1297.
2. Gustafsdottir, S. M.; Ljosa, V.; Sokolnicki, K. L.; Anthony, W. J.; Walpita, D.; Kemp, M. M.; Petri Seiler, K.; Carrel, H. A.; Golub, T. R.; Schreiber, S. L.; Clemons, P. A.; Carpenter A. E.; Shamji, A. F. Multiplex cytological profiling assay to measure diverse cellular states. *PLoS One* **2013**, *8*, e80999.
3. Young, D. W.; Bender, A.; Hoyt, J.; McWhinnie, E.; Chirn, G.-W.; Tao, C. Y.; Tallarico, J. A.; Labow, M.; Jenkins, J. L.; Mitchison, T. J.; Feng, Y. Integrating high-content screening and ligand-target prediction to identify mechanism of action. *Nat. Chem. Biol.* **2008**, *4*, 59–68.

P-41: The lipophilicity mirage: logD as an endpoint in drug discovery

Christian Tyrchan, Fredrik Bergström, Johan Ulander

AstraZeneca, Pepparedsleden 1,431 83 Mölndal, Sweden

Several metrics and rules such as LLE or Lipinski rules incorporate lipophilicity and are often seen as valid endpoints in drug discovery, which is in fact a Fata Morgana (eng. Mirage). This italian phrase is derived from Latin meaning “fairy”, from the belief that these mirages were fairy castles in the air to lure sailors to their death. These mirages distort the object which they are based on significantly, often such that the object is completely unrecognizable [1]. The use of distribution and partition coefficients to explain biological potency ranges back into the first half of 1900 [2–4]. The relationship between the potency of compounds and their organic solvent- aqueous buffer distribution coefficients was further developed by Hansch [5]. The octanol water partion coefficient ($\log D_{\text{octanol/water}}$) is defined as the logarithm of the ratio of the concentration of all ionized and un-ionized forms at a specific pH in the two phases.

The conceptual simplicity and appealing correlations between $\log D_{\text{octanol/water}}$ and e.g. potency and pharmacokinetic parameters has put it in focus for many discussions with tremendous impact in drug discovery.

To identify specific interactions or highlight possible sweet spots for optimisation, a reference point needs to be defined. In biological systems it is natural to use as reference the chemical potential in aqueous phase of pharmacological active compounds. Differences in logD have become a surrogate measure for the difference in the chemical potential between compounds, but the choice of $\log D_{octanol/water}$ is arbitrary. Lipophilicity can be strongly correlated to physico-chemical parameters like solubility or permeability which are related to bioavailability. The same is true for the most important complex endpoint in medicinal chemistry the dose to man. logD may influence simultaneously in a non-linear way all dose dependent parameters such as potency, clearance (CL), fraction unbound (fu) and volume of distribution (Vss). Any optimisation strategy based on bulk properties such as logD does not capture the underlying complexity [6,7]. This is exemplified by using published human PK data combined with internal measured $\log D_{octanol/water}$ data [8].

In fact it should be expected not to find linear correlations to complex phenomena. Therefore the contribution of lipophilicity to *in-vivo* endpoints can be considerably smaller than often assumed.

1. Wikipedia, http://en.wikipedia.org/wiki/Fata_Morgana_(mirage), Feb. 2014.
2. Overton E. Studien ueber die Narkose; Jena: Gustav Fischer Verlag, Germany, 1901.
3. Meyer H. Zur Theorie der Alkoholnarkose. *Arch Exp. Pathol. Pharmakol.* **1899**, *42*, 109–118.
4. Fieser, L.; Ettlinger, M.; Fawaz, G. Naphthoquinone Antimalarials; Distribution between Organic Solvents and Aqueous Buffers. *J. Am. Chem. Soc.* **1948**, *70*, 3228–3232.
5. Hansch, C.; Maloney, P.; Fujita, T.; Muir, R. Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. *Nature* **1962**, *194*, 178–180.
6. van de Waterbeemd, H.; Smith, D. A.; Barry, C. J. Lipophilicity in PK design: methyl, ethyl, futile. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 273–286.
7. Grime, K. H.; Bartonm P.; McGinnity, D. F. Application of In Silico, In Vitro and Preclinical Pharmacokinetic Data for the Effective and Efficient Prediction of Human Pharmacokinetics. *Mol.Pharmaceutics* **2013**, *10*, 1191–1206.
8. Obach, R. S.; Lombardo, F.; Waters, N. J. Trend Analysis of a Database of Intravenous Pharmacokinetic Parameters in Humans for 670 Drug Compounds. *Drug Metabol. Disposition* **2008**, *36*, 1385–1405.

P-43: A new platform to solve the Computational Chemistry's BigData problem

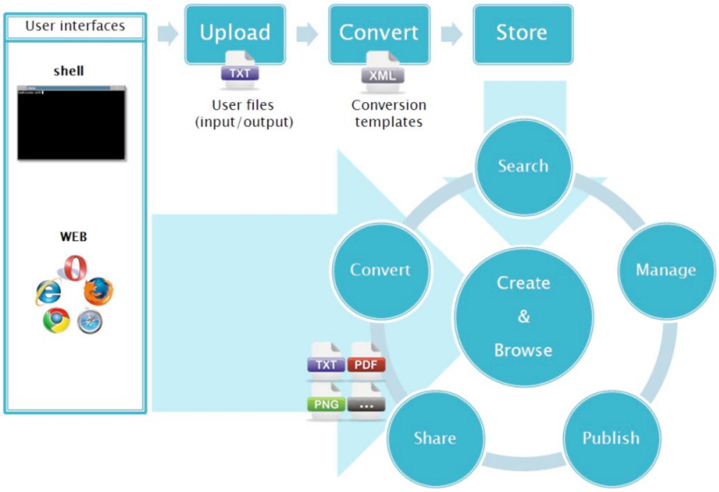
Carles Bo[1,2], Moises Álvarez[1,2], Feliu Maseras[1], Josep M. Poblet[2], Nuria López[1], Coen de Graaf[2]

[1]Institute of Chemical Research of Catalonia, ICIQ, Tarragona, Spain, [2]Department of Physical and Inorganic Chemistry, Universitat Rovira i Virgili, Tarragona, Spain

The massive use of simulation techniques in chemical research generates huge amounts of information, which starts to become recognized as the BigData problem [1]. The main obstacle for managing big information volumes is its storage in such a way that facilitates data mining as a strategy to optimize the processes that enable scientists to face the challenges of the new sustainable society based on the knowledge and the rational use of existent resources.

The present project aims at creating a platform of services in the cloud to manage computational chemistry. As other related projects [2], the concepts underlying our platform rely on well defined standards [3] and it implements treatment, hierarchical storage and data recovery tools to facilitate data mining of the Theoretical and Computational Chemistry's BigData. Its main goal is the creation of new methodological strategies that promote an optimal reuse of results and accumulated knowledge and enhances daily researchers' productivity.

This proposal automatizes relevant data extracting processes and transforms numerical data into labelled data in a database. This platform provides tools for the researcher in order to validate, enrich, publish and share information, and tools in the cloud to access and visualize data. Other tools permit creation of reaction energy profile plots by combining data of a set of molecular entities, or automatic creation of Suporting Information files, for instance. The final goal is to build a new reference tool in computational chemisty research, bibliography management and services to third parties. Potential users include computational chemistry research groups worldwide, university libraries and related services, and high performance supercomputer centers.



1. Lynch, C. Big data: How do your data grow? *Nature* **2008**, *455*, 28.
2. Chen, M.; Stott, A. C.; Li, S.; Dixon, D. A. Construction of a robust, large-scale, collaborative database for raw data in computational chemistry: The Collaborative Chemistry Database Tool (CCDBT). *J. Mol. Graph. Model.* **2012**, *34*, 67–75.
3. Adams, S.; de Castro, P.; Echenique, P.; Estrada, J.; Hanwell, M. D.; Murray-Rust, P.; Sherwood, P.; Thomas, J.; Townsend, J. The Quixote project: Collaborative and Open Quantum Chemistry data management in the Internet age. *J. Cheminf.* **2011**, *3*, 38.

P-45: Liberating Laboratory Data

Simon J. Coles, Colin Bird, Jeremy G. Frey, Richard Whitby

University of Southampton, Southampton, U.K.

Electronic Laboratory Notebooks (ELNs) are used routinely to capture and preserve procedures, materials, samples, observations, data, and analysis reports [1]. ELNs also retain the journal characteristics of traditional paper notebooks, thus enabling the capture of thoughts and deliberations. ELNs contain a wealth of data, information, and knowledge that is organised internally but is not necessarily readily accessible for external reuse.

Dial-a-Molecule [2] is a UK Grand Challenge Network that promotes research aimed at bringing about a step change in our ability to deliver molecules quickly and efficiently: *How can we make molecules in days not years?* To realise this vision, it will be essential to exploit the vast body of currently inaccessible chemical data and information held in ELNs, not only to make that data and information available but also to develop protocols for discovery, access and ultimately automatic processing.

We report research that based on a three-layer model that enables progressive discovery, with the knowledge layer as the entry point to the data and information embodied within ELNs. Core metadata identifying the material being made available and how it might be obtained is published [3] as an *elnItemManifest*, which is an XML file. Table 1 lists the top-level elements of this schema. If the material is potentially useful, another team can use the contact information in the elnItemManifest to obtain fuller contextual information held within the ELN. To reuse the material identified by exploring the information layer, the other team would request the detail metadata that describes the data itself and the process descriptions. The Dial-a-Molecule approach regards this detail as the third layer of a three-tier model that corresponds to the data-information-knowledge-wisdom (DIKW) hierarchy, a concept commonly first attributed to Ackoff [4]. If the Dial-a-Molecule vision is realised, chemists will have acquired the wisdom to *make molecules in days not years*.

Table 1: Top-level elements of the schema.

Element	**Datatype** (unqualified types defined within the schema)	**Brief description**
title	xs:string	Short, human-readable, text string to assist when viewing this record in a list: helps the reader to determine whether record is of interest
keywords	keywordSet	Text strings that might assist in searching or categorising
identifiers	identifierSet	Unique handles that identify this record
contact	contactOption	Specifies who or what to contact for more information. Could be: e-mail address, system URI, or brief instruction. Expect to get some sensible reply when contacting this person or system with the localIdentifier specified
licensingBasis	xs:string	[Optional] Indication of basis for licensing, e.g. Creative Commons
relatedItems	relatedItemSet	[Optional] List of items that might be related
contributors	contributorSet	List of contributory people, organisations, etc.
content	contentInformation	Items comprising the record that this describes
source	xs:string	String describing the system that generated this data. Generally identifies vendor, software package and version (resembles a browser user agent)
Dates	dateSet	One or more datestamps

1. Bird, C. L.; Willoughby, C.; Frey, J. G. Laboratory notebooks in the digital era: the role of ELNs in record keeping for chemistry and other sciences. *Chem. Soc. Rev.* **2013**, *42*, 8157–8175.
2. Dial-a-Molecule Grand Challenge. [Online] http://www.dial-a-molecule.org/wp/ (accessed Feb 13, 2014).
3. Coles, S. J.; Frey, J. G.; Bird, C. L.; Whitby, R. J.; Day, A. E. First steps towards semantic descriptions of electronic laboratory notebook record. *J. Cheminf.* **2013**, *5*, 52.
4. Ackoff, R. L. From data to wisdom. *J. Appl. Syst. Anal.* **1989**, *16*, 3–9.

P-47: Paper is not dead, and PDFs feel very well, thank you

Wolf D. Ihlenfeldt

Xemistry GmbH, Königstein; Germany

A gap continues to exist between chemical information designed for human consumption (text and structure/data renderings, nowadays mostly distributed as PDF files) and raw supporting data, which is normally attached separately in formats such as SD files, SMILES, InChI, or numeric tables in CSV or Excel format. These are separate realms – neither can data be reliably extracted from a PDF without resorting to exotic tools like Chemical OCR, nor can anybody obtain a general overview about the contents of a structure file without specialist viewer software usually not present on readers such as a Kindle or iPad which have become the standard means of perusing literature.

In response to this conundrum, we have developed a method to merge the technologies of a chemistry-aware SQL database, storing raw data, structures and reaction, and a PDF file with a corresponding visual data representation. We have thus created a unified medium for chemical data exchange by packing the the triple of database file, renderer control file, and PDF representation into a single PDF document. This file can both the viewed (or printed) with any standard PDF viewer, as well as queried, analysed, exported in standard formats or re-rendered with different layouts or information content as a whole or a user-selected subset by means of a standard Web application, user-designed toolkit scripts, or stream processors such as KNIME.

Smaller subsets of structures, which are for example subject of a group discussion in a MedChem project meeting, are still routinely put on paper for easy annotation. Paper hand-outs printed from smart PDF files are not necessarily dead media – they may contain Web links and specially prepared regions for Smartpen technologies. These enable interesting use cases, for example automatically capturing expert chemist opinions in discussions.

P-49: Peptide Line Notations for Biologics Registration and Patent filings

Roger Sayle

NextMove Software Limited, Cambridge, United Kingdom

The registration of peptide and peptide-like structures in chemical databases poses a number of technical challenges, as do other biological oligomers. Primary among these are the size and complexity of these compounds, often making it difficult or impossible for a biochemist to identify differences or similarities between compounds stored as all-atom representations. For biological (proteinogenic) sequences the solution is simple, to use strings of the one (or three) letter codes to represent the biopolymer. However, for post-translationally modified, D-, cyclic and non-standard peptides the way forward is less clear. One approach is to use an ever larger dictionary three letter codes to encode common non-standard amino acids. Unfortunately, this method quickly becomes unwieldy once the set of abbreviations exceeds those frequently encountered in the literature, requiring a chemist to consult a key or dictionary for entry or interpretation of structures. In this presentation, we propose the use of semi-systematic monomer names, based upon readily recognizable chemical line formulae, for the encoding and display of traditionally difficult to handle peptides. These rules lead to amino acid names such as N(Me)Ser(tBuOH) that are similar to those seen/used in vendor catalogs and scientific publications, though not formally ratified by IUPAC nor CAS.

P-51: An automated document classifier to retrieve ChEMBL-like papers

Gerard J. P. van Westen, George Papadatos, Simone Trubian, Rita Santos, Samuel Croset, John P. Overington

ChEMBL / Chemogenomics Group, European Molecular Biology Laboratory European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom

Introduction

The ChEMBL database stores 2D compound structures, bioactivity data and calculated molecular properties of drugs and drug-like molecules, primarily focussed on medicinal chemistry, chemical biology and drug discovery. Data contained in the resource is manually extracted from primary scientific literature and then curated to ensure correctness and consistency [1].

Scientific literature is written by and for humans, and hence carries all the inconsistencies of natural languages. Examples include differences in notations used for units, assay readouts, description of chemical structures as well as many synonyms for chemical and biological entities.

Due to these inherent inconsistencies, manual curation from text to data is necessary [2,3], which is expensive and time-consuming. Furthermore, it becomes increasingly difficult for curators to keep up with the scientific output produced. Therefore, biomedical researchers and text-mining efforts are in need of automated expert systems that can help with the initial steps of the curation process. This phase is known as triage, namely the selection of relevant scientific articles from large repositories, such as PubMed.

Results

We present an approach using the manually curated and comprehensive ChEMBL document corpus to train a Bag-of-Words (BoW) document classifier based on the abstract and titles of the articles. The strategy has already proven to be successful [4,5], and we wanted to extend and modify it in order to address a series of defined requirements, presented in this poster alongside the methodology behind the classifier.

We have employed two established classification methods, namely Naive Bayesian (NB) and Random Forests (RF). The resulting classification score (ChEMBL-likeness) helps prioritizing relevant documents for data extraction and curation during the triage process. The data pre-processing workflows and validated models are freely available online to the community as KNIME workflow and Pipeline Pilot protocols.

1. Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J. Davies, M.; Krueger, F. A.; Light, Y.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* **2014**, *42*, D1083–D1090.
2. Rebholz-Schuhmann, D.; Kirsch, H; Couto, F. Facts from text—is text mining ready to deliver? *PLoS Biology* **2005**, *3,*e65.
3. Burge, S.; Attwood, T. K.; Bateman, A.; Berardini, T. Z., Cherry, M; O'Donovan, C., Xenarios, I.; Gaudet, P. Biocurators and biocuration: surveying the 21st century challenges. Database: the journal of biological databases and curation. *Database* **2012**, bar059.
4. Davis, A. P.; Wiegers, T. C.; Johnson, R. J.; Lay, J. M.; Lennon-Hopkins, K.; Saraceni-Richards, C; Sciaky, D; Murphy, C. G.; Mattingly, C. J. Text mining effectively scores and ranks the literature for improving chemical-gene-disease curation at the Comparative Toxicogenomics Database. *PloS ONE* **2013**, *8*, e58201.

5. Vishnyakova, D.; Pasche E.; Ruch, P. Selection of relevant articles for curation for the comparative toxicogenomic database. In *Proceedings of the 2012 BioCreative Workshop*, Washington DC, USA. **2012**.

P-53: Open PHACTS: Solutions and the Foundation

Egon Willighagen[1,2]

[1]Department of Bioinformatics - BiGCaT, NUTRIM, Maastricht University, The Netherlands, [2]The Open PHACTS Consortium

Open PHACTS is a three year project of the Innovative Medicines Initiative, ending in September 2014. Its results reduce the barriers to drug discovery in industry, academia and for small businesses [1]. The Open PHACTS consortium has built a freely available platform, integrating data from a variety of information resources, and providing tools and services to query these integrated data to support life sciences research [2].

Currently, pharmaceutical companies expend significant and often duplicated efforts aligning and integrating internal information with public data sources. This process is difficult and inefficient and the vast majority of data sources cannot easily interoperate, often requiring additional steps to map identifiers or manually curate and correct the content. Open PHACTS's precompetitive infrastructure makes these data integration approaches available both to industry and to academia and smaller companies, who have historically not had access to large-scale integrated data resources.

Here we give an overview of the resulting platform, the semantic web solutions used in this, and describe the integration of the data into data analysis tools. The platform consists of components that communicate with each other using open standards and cover the full data lifecycle, from data loading to data sharing. Solutions include those for data provenance, normalization, standardization, and access. In particular, we have developed minimal reporting standards for provenance, technologies to express the level of equivalence of entities from different databases, a database identifier mapping infrastructure based on semantic web technologies, unit and end point normalization, as well as chemical structure normalization. For this we use open ontologies (BioAssay Ontology, QUDT, CHEMINF, etc), standards (RDF, SPARQL, REST, etc), and proposed solutions as outlined in published specifications.

On top of these approaches, user oriented solutions have been developed based on a number of research questions selected by the pharmaceutical industry [3]. Example questions include: "Give me all oxidoreductase inhibitors active <100 nm in human and mouse" and "Compounds that agonize targets in pathway X assayed in only functional assays with a potency <1 µm". The OPS platform provides a uniform route by which these questions can be addressed, exposed to the user by a novel pharmaceutical web service platform, called the Linked Data API (LDA).

With this LDA and a matching web-portal to access integrated data, the Open PHACTS platform supports an ecosystem of third-party applications addressing specialised needs such as polypharmacology, hit-selection, target validation and knowledge discovery. Additionally, more generic integrations have been developed too, like client libraries to the LDA in various programming languages, such as JavaScript, Scala, and Java, resulting in integration in generic data analysis platforms like PipeLine Pilot, KNIME, R, and Bioclipse.

Finally, results will be presented based on the sustainability work package, highlighting the future of Open PHACTS beyond the initial project. This is partly covered by the Open PHACTS Foundation which will ensure maintenance and continued development.

1. Williams, A. J.; Harland, L.; Groth, P.; Pettifer, S.; Chichester, C.; Willighagen, E. L.; Evelo, C. T.; Blomberg, N.; Ecker, G.; Goble, C.; Mons, B. Open PHACTS: semantic interoperability for drug discovery. *Drug Discov. Today* **2012**, *17*, 1188–1198.
2. Gray, A. J. G.; Groth, P.; Loizou, A.; Askjaer, S.; Brenninkmeijer, C.; Burger, K.; Chichester, C.; Evelo, C. T.; Goble, C.; Harland, L.; Pettifer, S.; Thompson, M.; Waagmeester, A.; Williams, A. J. Applying Linked Data Approaches to Pharmacology: Architectural Decisions and Implementation. *Semantic Web Journal.* **2014**, *5*, 101–113.
3. Azzaoui, K.; Jacoby, E.; Senger, S.; Cuadrado Rodríguez, E.; Loza, M.; Zdrazil, B.; Pinto, M.; Williams, A. J.; De la Torre, V.; Mestres, J.; Pastor, M.; Taboureau, O.; Rarey, M.; Chichester, C.; Pettifer, S.; Blomberg, N.; Harland, L.; Williams-Jones, B.; Ecker, G. K. Scientific competency questions as the basis for semantically enriched open pharmacological space development. *Drug Discov. Today* **2013**, *18*, 843–852.

P-55: The role of negative evidence in fingerprint-based Naïve Bayes models

Nikolas H. Fechner, Sereina Riniker, Gregory A. Landrum

Novartis Institutes for BioMedical Research, Novartis Pharma AG, Novartis Campus, CH-4056 Basel, Switzerland

The representation of molecules by molecular fingerprints is a well-established approach to encode chemical information for data analysis. Although not generally necessary, such fingerprints are often encoded as fixed-sized bit vectors where each structural feature captured by the fingerprint leads to a bit being set to 1. This encoding allows the application of the plethora of machine-learning algorithms available for categorical (as opposed to numerical) vectorial data to molecules. An example is the Naïve Bayes approach to supervised learning approach, which is quite popular due to its simplicity and the remarkably good results it can provide (see for instance Bender et al. [1]).

When Naïve Bayes is applied in the field of cheminformatics, a Laplacian correction is typically applied and a new formulation was recently proposed that treats information about the presence of substructure patterns differently than their absence [2]. However, there is rarely discussion about which probability distribution is used and why this decision was made. The obvious choice for the probability distribution would be a Bernoulli coin-toss model, due to the binary nature of the fingerprint bits. Interestingly, the scenario of generating a probabilistic model using a representation that encodes a data item by the presence or absence of certain patterns resembles that in dictionarybased spam detection, which also often use Naïve Bayes models on bit vector-encoded data. In the spam-detection literature, the choice of the best probability model has been more systematically explored. In many cases multinomial probability distributions lead to better models than Bernoulli models, despite the boolean nature of the attributes [3]. In the spam detection context, this is attributed to the different treatment of *negative evidence* [4], i.e., a 0-bit being interpreted as the absence of a pattern, which is related to the effect of the Laplacian correction as suggested by Mussa et al [2].

The role of negative evidence in Naïve Bayes models is potentially of high relevance for data analysis based on molecular fingerprints. A prediction that is driven by the absence of structural patterns is at risk of being strongly influenced by the diversity of the molecules – generally higher for the inactive compounds in the training set – and not by the encoding of a true SAR by the model.

Here, we present an evaluation of the effect of the probability distribution on Naïve Bayes models trained with molecular fingerprints using a publicly available collection of benchmark datasets [5], and discuss evidence that Naïve Bayes models, especially if Bernoulli distributions are applied, might be strongly influenced by negative evidence.

1. Bender, A.; Mussa, H. Y.; Glen, R. C. Similarity Searching of Chemical Databases Using Atom Environment Descriptors (MOLPRINT 2D): Evaluation of Performance. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1708–1718.
2. Mussa, H.; Mitchel, J.; Glen, R. Full “Laplacianised” posterior naive Bayesian algorithm, *J. Cheminf.* **2013**, *5*, 37.
3. Metsis, V.; Androutsopoulos, I.; Paliouras, G. Spam Filtering with Naïve Bayes – Which Naïve Bayes? . In CEAS 2006 – 3rd Conference on Email and Anti-Spam, Mountain View, USA, 2006.
4. Schneider, M. On Word Frequency Information and Negative Evidence in Naïve Bayes Text Classification, Advances in Natural Language Processing, *LNCS* **2004**, *3230*, 474–485.
5. Riniker, S.; Fechner, N.; Landrum, G. Heterogeneous Classifier Fusion for Ligand-Based Virtual Screening: Or, How Decision Making by Committee Can Be a Good Thing. *J. Chem. Inf. Model.* **2013**, *53*, 2829–2836.

P-57: A Novel Mechanistic Approach to Free-Wilson SAR Analysis Enabling the Use of Potency Data in Computational Lead Optimization

Pranas Japertas[1,2], Andrius Sazonovas[1,2], Kiril Lanevskij[1,2], Remigijus Didžiapetris[1,2]

[1]VsI „Aukstieji algoritmai“, A.Mickeviciaus 29, LT-08117 Vilnius, Lithuania, [2]ACD/Labs, Inc., 8 King Street East, Toronto, Ontario, M5C 1B5, Canada

Lead optimization efforts are guided by a combination of factors, among which, the potency of the lead, as well as its ADME and safety-related properties play the major roles. As opposed to the latter global aspects influencing the viability of a drug candidate, each drug discovery project aims at optimizing activity against a specific target, represented by a specific endpoint. However, computational models for estimating affinities of the compounds to the multitude of possible targets are not readily available. As a result, the scope of conventional in silico lead optimization techniques is commonly limited to physicochemical and ADME/Tox profiling with potency rarely taken into account, and while there is objectively not much one can do about the first contributing factor, the development of new modeling techniques and refinement of the existing ones is still an area with ample room for improvements.

The proposed presentation would focus around new developments in the Free-Wilson type analysis that effectively enable the use of SAR approaches on the small sets of the user-defined potency data, rendering it available for the in silico compound profiling. The main idea behind the concept is that the structural constituent of the varying substituents in a series of known compounds with a common scaffold is replaced by their contributions to the major physicochemical properties of the molecule, such as size, lipophilicity, ionization, and hydrogen bonding potential. First of all this action provides a drastic reduction in the number of variables involved, making the class-specific SAR models from small data sets in the order of several tens of compounds feasible in general. In addition switching the perspective to the exploration of the physicochemical dependences is greatly beneficial towards the mechanistic interpretability of the resulting models, thus making them much more appealing to the medicinal chemists. Last but not least, a sensible amount of variables involved in modeling and their clear physicochemical meaning facilitates the use of the non-linear statistical analysis tools without sacrificing the ability to mechanistically interpret the results. This allows identification of potential non-linear trends, which despite being known to have a significant influence on various pharmacodynamic and pharmacokinetic processes, are at present rarely employed in the class-specific SAR analysis of the potency data, e.g., the bi-linear dependency of the compound’s activity on the lipophilicity of a substituent at a certain position.

Finally, we present a real-world scenario illustrating how the described mechanistic Free-Wilson SAR analysis could be applied to model target affinity for a small class-specific data set, and what insight could be gained from the results obtained. The effectiveness of the method is demonstrated using both reported EC_{50} values for the target receptor, and IC_{50} of hERG channel inhibition, as an example of a well defined endpoint associated with the potency of protein-ligand binding. The new method was able to successfully identify the optimal pattern of the substituent lipophilicities, and to account for the effects of ionization centers introduced in various regions of the molecule – all in agreement with the hERG channel and substrate binding modes reported in the literature.

Coupled with the full model development process automation this new method can be applied to the analysis of any in house small sets of potency data. The obtained class-specific models can then be utilized to gain better understanding of substituent effects, evaluate target activities of new compounds of the same class, and guide lead optimization efforts to the most favorable structural modifications that would yield the most promising candidates.

P-59: First-principles search for molecular structures with a genetic algorithm

Adriana Supady[1], Volker Blum[1,2], Carsten Baldauf[1], Matthias Scheffler[1]

[1]*Fritz-Haber-Institut der MPG, Berlin, Germany,* [2]*MEMS, Duke University, Durham, U.S.A.*

We present a genetic algorithm (GA) based search framework for structural searches of complex molecules based on empirical or first-principles (density-functional theory, DFT) energy functions. The aim is not just to find the single global minimum structure, but also to identify all conformers that appear in the low-energy conformational energy hierarchy and could be experimentally relevant. The use of DFT gives access to rather accurate energy functions and avoids the problem of parameterization for specific classes of chemical compounds.

In the GA, the geometry of a structure is encoded in a vector of torsional degrees of freedom (TDOF). The initial population of *N* individuals is randomly generated and evaluated by local relaxation. Two individuals are selected; the selection probability is a function of energy. Next, genetic operations are applied: (i) crossing over exchanges parts of the encoding vectors and (ii) mutations randomly assign new values to selected TDOFs. The resulting candidate structures are again evaluated by local relaxation and eventually replace individuals of the previous generation with a higher energy. The algorithm proceeds with a new selection round until a predefined number of iterations or a convergence criterion is met. Generated geometries are first checked for steric clashes and uniqueness (that is if they were computed already before) based on root mean square deviation (rmsd) of Cartesian coordinates. This greatly reduces the number of unproductive or redundant calculations, which is especially important when using a first principles energy function.

We demonstrate the principle for an azobenzene-based molecule (shown in Figure 1A), finding the conformational energy hierarchy for the *cis* and for the *trans* configurations. The accuracy of such GA prediction (by means of reproducing the conformational hierarchy from a systematic search) is critically linked to the search settings, for example, the number of repeats (illustrated in Figure 1B). While the global optimum is found very reliably with only a few repeats, the reproduction of the hierarchy is more demanding. Similar testing was performed for seven amino acid dipeptides (Ala, Gly, Val, Leu, Ile, Phe, Trp).

Post-processing of the data, for example the evaluation of the geometrical similarity of the structures and taking into account their energetic relation, allows for visualization of the topology of the potential energy landscape in form of a graph. Such information can be further utilized to identify which pairs of states are likely to be connected by a low-energy barrier. We will use the described strategy to predict functional molecules (by adding a library of side groups) that are tuned for a specific use, e.g., as switchable catalysts of a target chemical reaction.

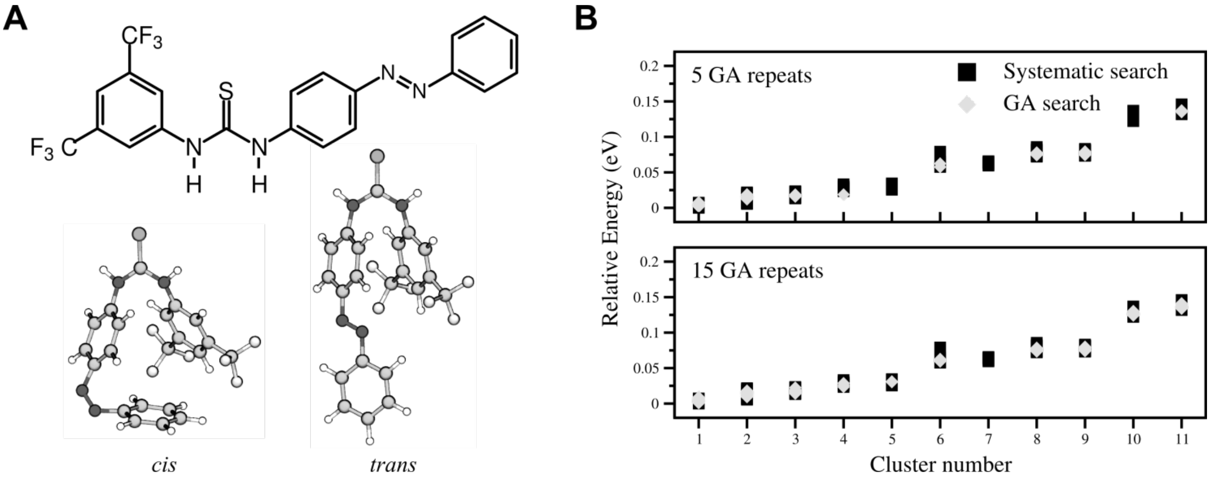


Figure 1: A) Chemical structure of an azobenzene based molecule and the two global minima with *cis* and *trans* azo-group. **B)** Comparison of a systematic search (black squares) and the GA search (gray diamonds) for the *trans* conformation. The systematic search yielded 11 clusters (representing 3,894 structures) within a relative energy of 0.2 eV. The 125 structures from 5 GA repeats (upper panel B) and the 375 structures from 15 GA repeats (lower panel B), respectively, were sorted into the same clusters as the systematic search results.

P-61: Exploratory Data Analysis & Visualization Applied to Structure-Activity Relationships

Ismail Ijjaali, Mary Donlan, Pierre Morieux

PerkinElmer, 16 Avenue du Québec, Bât Lys, 91140 Villebon sur Yvette, France

In medicinal chemistry, the analysis of structure-activity relationships (SAR) is of fundamental importance in understanding the structural determinants of biological activity, and it underpins lead generation for drug development. Small changes in molecular structure and properties can have diverse effects on biological efficacy and potency, bioavailability and metabolic stability. In addition, due to the accumulation of large amounts of screening data, visualization & data mining are key aspects of both the analysis and understanding of SAR data.

Using a case study dataset, we will illustrate how interactive visualizations and data analysis can help medicinal chemists to explore their SAR data and extract hidden structure-activity patterns. Among multiple applications of exploratory visualizations in SAR analysis we will be illustrating in the presented case study:

- Chemical diversity analysis by chemical property distribution & clustering
- Correlating chemical features to biological activity,
- Identification of key compounds with privileged chemical scaffolds,
- Biological activity profiling,
- Compound selectivity assessment.

P-63: VAMMPIRE-LORD: an open access web server for targeted lead optimization based on Matched Molecular Pairs

Julia Weber, Janosch Achenbach, Daniel Moser, Rene Blöcher, Sandra K. Wittmann, Ewgenij Proschak

Institute of Pharmaceutical Chemistry, Goethe University Frankfurt Max-von-Laue-Str. 9, D-60438 Frankfurt am Main, Germany

A Matched Molecular Pair (MMP) is defined as a pair of molecules that differ only "by a particular, well defined, structural transformation" [1]. The valuable information of an MMP along with the subsequent effect on pharmaceutical properties can be used to improve a lead compound regarding solubility, plasma protein binding and oral exposure [1–3]. Considering the chemical environment of the structural transformation, MMPs might also be used to improve the affinity of a lead compound to a specific target. The recently described VAMMPIRE database [4] was designed to link MMP transformations with the relevant receptor environment and the corresponding effect on ligand affinity. Based on VAMMPIRE database and using a pharmacophore based shape descriptor [5–7] to represent the substitution environment, we developed the prediction tool LORD (**L**ead **O**ptimization by **R**ational **D**esign). LORD operates on the principle that molecular transformations cause similar effects in similar substitution environments and is therefore able to extrapolate the knowledge of a given substitution effect to any similar system. LORD was implemented as an easy-to-use web server that guides the user step-by-step through the optimization process of a defined lead compound.

1. Leach, A. G.; Jones, H. D.; Cosgrove, D. A.; Kenny, P. W.; Ruston, L.; MacFaul, P.; Wood, J. M.; Colclough, N.; Law, B. Matched Molecular Pairs as a Guide in the Optimization of Pharmaceutical Properties; a Study of Aqueous Solubility, Plasma Protein Binding and Oral Exposure. *J. Med. Chem.* **2006**, *49*, 6672–6682.
2. Dossetter, A. G.; Griffen, E. J.; Leach, A. G. Matched Molecular Pair Analysis in Drug Discovery. *Drug Discov. Today* **2013**, *18*, 724–731.
3. Griffen, E.; Leach, A.; Robb, G.; Warner, D. Matched Molecular Pairs as a Medicinal Chemistry Tool. *J. Med. Chem.* **2011**, *54*, 7739–7750.
4. Weber, J.; Achenbach, J.; Moser, D.; Proschak, E. VAMMPIRE: A Matched Molecular Pairs Database for Structure-Based Drug Design and Optimization. *J. Med. Chem.* **2013**, *56*, 5203–5207.
5. Ballester, P. J.; Richards, W. G. Ultrafast Shape Recognition to Search Compound Databases for Similar Molecular Shapes. *J. Comput. Chem.* **2007**, 28, 1711–1723.
6. Ballester, P. J. Ultrafast Shape Recognition: Method and Applications. *Future Med. Chem.* **2011**, *3*, 65–78.
7. Schreyer, A. M.; Blundell, T. USRCAT: Real-Time Ultrafast Shape Recognition with Pharmacophoric Constraints. *J. Cheminf.* **2012**, *4*, 27.

P-65: A-WOL Ligand Based Screening combined with HTS

N. Berry[1], Jaclyn Bibby[1], P. O'Neill[1], S. Ward[2], M. Taylor[2]

[1]Robert Robinson Laboratories, Department of Chemistry, University of Liverpool, Crown Street, Liverpool, L69 7ZD, U.K., [2]Liverpool School of Tropical Medicine, Pembroke Place, Liverpool, L3 5QA, U.K.

Filariasis inflicts serious health problems throughout tropical communities, with infection occurring when filarial parasites are transmitted to humans through mosquitoes. The major disease causing species including those responsible for lymphatic filariasis (Elephantiasis), *Wuchereria bancrofti* and *Bru-*

gia malayi, and onchocerciasis (River blindness), *Onchocerca volvulus*, together infect more than 150 million people, ranking filariasis as one of the leading causes of global morbidity [1]. River blindness is the second leading cause of blindness caused by infection with currently 500,000 people blinded. 120 million people are currently infected by Lymphatic filariasis, with about 40 million people disfigured it ranks as the second leading cause of global disability. Filarial species that infect people co-exist in mutualistic symbiosis with Wolbachia bacteria, which are essential for growth. Antibiotic anti-Wolbachia (A-WOL) therapy delivers safe macrofilaricidal activity with superior therapeutic outcomes compared to all standard anti-filarial treatments.

Our present work to identify compounds with A-WOL activity involves an iterative combination of HTS and ligand based virtual screening, with the results of screening campaigns feeding back into models to select subsequent rounds of compounds (Figure 1). An initial diversity screen of 10,000 compounds taken from the Biofocus library yielded 50 actives (a hit rate of 0.5%). This formed the basis for a similarity search of the 500,000 compound MMV library to select 17,000 compounds (Request 1). The similarity search prioritised hit-like compounds led to an increase in the initial detection of actives (365 hits) with 53 hits (0.3%) 17 of which were very high potency hits (0.1%).

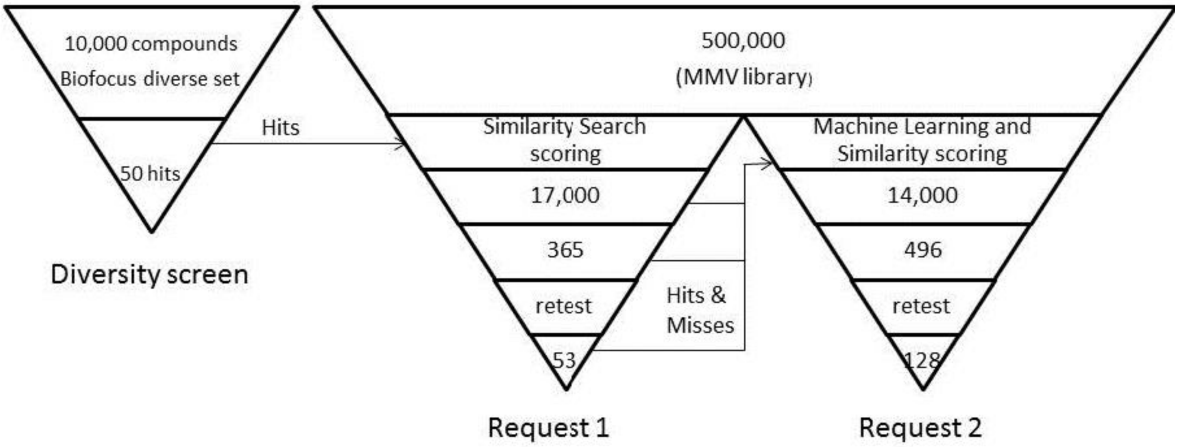


Figure 1: Overview of virtual screening and HTS.

As the compounds of the MMV library are arranged on plates, the aim of the MMV screen was to identify the 'best' plates, in terms of containing a high number of virtual screening hits. Methods to prioritise individual compounds would be insufficient as all compounds on a plate need to be taken into account. A range of data fusion methods (including sum, rank, reciprocal rank, parallel, Z2) were used to assign a score or rank to each plate rather than to an individual compound.

The results of Request 1 (both hits and misses) were fed back into computational modeling, improving their reliability and scope. For this second screen of compounds from the MMV library (Request 2), a range of machine learning tools were employed (including Random Forest and SVM), in addition to the similarity searching. Random forests and SVMs were trained on the data using 15 different physicochemical descriptors (e.g. LogP, number of rotable bonds, H bond donors etc.) and also fingerprint descriptors (ECFP4, FCFP4 and MDL MACCS). Robust models (as assessed by high AUC scores) were used to select the next set of compounds to be screened. A-WOL activity screening of this set demonstrated a further improvement in hit detection, the number of high potency hits nearly doubling to 0.18%, and a threefold increase in moderate hits (0.9%). In addition to the increased hit rate between each successive round of screening, some methods proved to be affective at achieving scaffold hopping with the identification of novel chemotypes, demonstrating the effectiveness of machine learning tools. We will present an analysis of the methods.

Our studies demonstrate the effectiveness of our ligand based methods in combination with HTS for the identification of novel antibacterial compounds. The methods provided an iterative increase in hit rate and have identified six new tractable chemotypes, four of which yielding compounds with nM activities, that are promising candidates for future work.

1. Taylor, M. J.; Hoerauf, A.; Bockarie, M. Lymphatic filariasis and onchocerciasis. *Lancet* **2010**, *376*, 1175–1185.

P-67: MOARF: A novel workflow for the multiobjective optimisation of drug-like molecules

<u>Nicholas Firth</u>, Nathan Brown, Julian Blagg

Cancer Research UK Cancer Therapeutics Unit, Division of Cancer Therapeutics, The Institute of Cancer Research, London, SM2 5NG, U.K.

Designing a molecule *de novo* is an efficient alternative to high throughput screening and yields novel chemical entities and yet computer-based *de novo* design has only recently become useful in drug discovery [1]. Similarly, multiobjective optimisation has seen a recent renewal of interest [2]. The evaluation and optimisation of multiple pharmaceutically relevant objectives simultaneously has become particularly important in the progression of *de novo* design [3,4]. In this work we describe a novel workflow: MultiObjective Automated Replacement of Fragments (MOARF) for multiobjective optimisation by *de novo* design. This workflow uses fragment replacement to propose novel chemical entities, which are then scored using multicriteria decision analysis as part of an evolutionary algorithm.

The novelty of MOARF stems from the procedures used to overcome the key issues encountered in fragment replacement and d*e novo* design. We will present two of the workflow components: Fragment database design and fragment alignment. These two components enable exploration of synthetically feasible and structurally novel landscapes. The program Synthetic Disconnection Rules (SynDiR) is used to fragment molecular libraries into relevant building blocks and the Rapid Alignment of Topological Scaffolds (RATS) is applied to align R-groups in fragment replacement.

An example application is presented in which we virtually mimic an in-house drug discovery project, optimizing the small molecule roscovitine in order to retain activity against CDK2 whilst improving the pharmacokinetic properties [5]. This optimisation project was scored using a bespoke scoring function, which consists of four distinct scoring methods, including a statistical classifier, for which the training set was designed and tested to enable accurate prediction. These scoring methods are then fused using z-Score ranking [6] to rank the compounds in a population at each generation. The workflow is able to produce compounds with high similarity to those synthesized in the lead optimisation program. Furthermore, a set of compounds with properties predicted to outperform those synthesized in the program has been suggested by the workflow.

1. Schneider, G.; Baringhaus, K. H. *De Novo* Design: From Models to Molecules, In *De novo Molecular Design*; Schneider, G. Ed; *Wiley-VCH,* **2013**.
2. Nicolaou, C. A.; Brown, N. Multi-objective optimisation methods in drug design. *Drug Discov. Today: Technol.*, **2013**, *10*, 427–435.
3. Gillet, V. J.; Bodkin, M. J.; Hristozov, D. Multiobjective *De Novo* Design of Synthetically Accessible Compounds, In *De novo Molecular Design*, Schneider, G., Ed.; *Wiley-VCH,* **2013**.
4. Besnard, J. et al. Automated design of ligands to polypharmacological profiles. *Nature* **2012,** *492*, 215–220.

5. Wilson, S. C. et al. Design, synthesis and biological evaluation of 6-pyridylmethylaminopurines as CDK inhibitors. *Bioorg. Med. Chem.* **2011**, *19*, 6949–6965.
6. Sastry, G.; Madhavi, V. S.; Sandeep, I.; Sherman, W. Boosting virtual screening enrichments with data fusion: Coalescing hits from 2D fingerprints, shape, and docking. *J. Chem. Inf. Model.* **2013,** *53*, 1531–1542.

P-69: Interactions of Strongly and Weakly Bound Water Molecules to Protein Binding Sites and Propensity for Replacements

Stefan Güssregen[1], Hans Matter[1], Gerhard Hessler[1], Evanthia Lionta[1], Stefan Kast[2]

[1]Sanofi, Frankfurt am Main, Germany, [2]TU Dortmund, Dortmund, Germany

Experimental evidence suggests that water molecules play a crucial role to mediate interactions between ligands and protein binding sites. Displacement of specific water molecules by ligand moieties is possible in some cases and could in addition favourably contribute to the free energy of binding. Therefore an understanding of the thermodynamic nature of these water molecules might guide further structure-based drug design. Here it is of critical importance to determine the propensity of such a structurally conserved water molecule within a macromolecular binding site to be replaced by a part of a novel ligand.

The nature of these water interactions in several protein binding sites was studied by using 3D-RISM calculations towards an understanding of their thermodynamic features and the possibility for replacement by ligand parts.

The 3D reference interaction site model (RISM) [1,2] constitutes an implicit solvent model that is based on classical density functional theory. It solves the converged H and O atom densities on a 3D grid (g-function) and directly provides equilibrium thermodynamic quantities. The results are equivalent to those of infinite Monte Carlo or Molecular Dynamics simulations at a fraction of the computational cost.

In our investigations, we are focussing on two types of water molecules in X-ray structures of relevant protein binding sites, namely streptavidin, COX-2, factor Xa and factor VIIa: Those which can be replaced by a part of the ligand and those, which are almost integral part of the receptor and cannot be replaced, but should potentially be targeted in structure-based design. Our approach allows the semi-quantitative assessment of whether some given structural water molecule indeed could potentially be targeted for replacement in structure-based design.

This approach was extended to describe ligands in their crystallographic conformation and to compute molecular interaction fields capturing information about Gibbs free energy of solvation (ΔG) factorization into respective enthalpic and entropic terms. PLS-based regression models from those solvation fields capturing thermodynamical properties of ligands fields are shown to provide a significant advantage to understand SAR features. This will be demonstrated for a congeneric series of serine protease inhibitions, for which several X-ray structures in addition provide insight into replacement structural water.

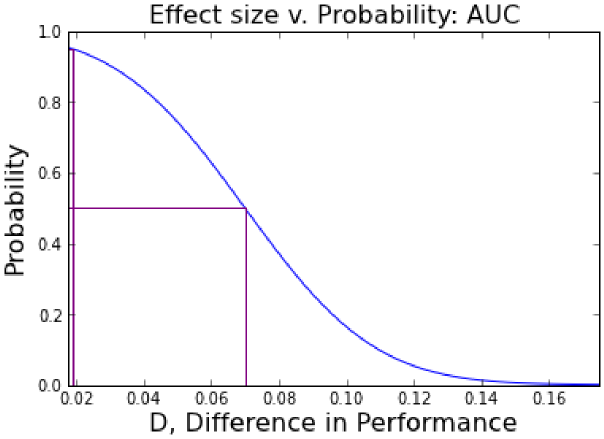
1. Beglov, D.; Roux, B. An Integral Equation To Describe the Solvation of Polar Molecules in Liquid Water. *J. Phys. Chem. B* **1997**, *101*, 7821.
2. Kovalenko, A.; Hirata, F. Three-dimensional density profiles of water in contact with a solute of arbitrary shape: a RISM approach. *Chem. Phys. Lett.* **1998**, *290*, 237.

P-71: Maximising recovery in virtual screening using information entropy

Paul Hawkins, Matthew Geballe, Gregory Warren

OpenEye Scientific, Santa Fe, USA

An enormous variety of methods for virtual screening have been developed over recent years [1]. Over time different studies have arrived at different conclusions as to the relative performance of different tools or different methodologies in virtual screening (2D v. 3D, ligand v. structure-based) [2,3]. A large number of methods have also been developed to fuse results from different virtual screening methods to maximise performance [4]. In this paper we present novel methods to assess the complementarity of structure-based VS and 3D ligand-based VS. We approach the problem from two different angles: comparing absolute performance of individual methods and their fused results using appropriate statistical methods and comparing the hits from different methods to quantitate their overlap. In attacking the latter problem we use both simple set-based approaches and a more sophisticated approach using information entropy [5].



$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i)$$

1. Virtual Screening: Principles, Challenges and Practical Guidelines; Sotriffer, C., Ed.; Wiley-VCH Verlag GmbH: Weinheim, 2011.
2. Hu, G. et al. Performance Evaluation of 2D Fingerprint and 3D Shape Similarity Methods in Virtual Screening. *J. Chem. Inf. Model.* **2012**, *52*, 1103–1113.
3. Venkatraman, V. et al., Comprehensive Comparison of Ligand-Based Virtual Screening Tools Against the DUD Data set Reveals Limitations of Current 3D Methods. *J. Chem. Inf. Model.* **2010**, *50*, 2079–2093.
4. Tan, L. et al. Integrating Structure- and Ligand-Based Virtual Screening: Comparison of Individual, Parallel, and Fused Molecular Docking and Similarity Search Calculations on Multiple Targets. *Chem. Med. Chem.* **2008**, *3*, 1566–1571.
5. Shannon, C. E. A Mathematical Theory of Communication. *Bell System Tech. J.* **1948**, *27*, 379–423.

P-73: Knowledge-Based Potentials in Protein-Protein Docking

Dennis M. Krüger[1,2], **Holger Gohlke**[1]

[1] Institute for Pharmaceutical and Medicinal Chemistry, Department of Mathematics and Natural Sciences, Heinrich-Heine-University, Universitätsstr. 1, 40225 Düsseldorf, Germany; [2] present address: Chemical Genomics Centre of the Max Planck Society, Otto-Hahn-Str. 15, 44227 Dortmund, Germany

Solving the problem of structure prediction of protein-protein complexes would allow the generation of structures for ~65% of the interactions in the human interactome. A key ingredient of computational approaches for the structure prediction of protein-protein complexes are so-called scoring and objective functions, which quantify the structural correctness of the predictions.

Here, we evaluate the DrugScore^{PPI} statistical potentials previously developed for *in silico* alanine scanning and hot spot prediction on given structures of protein-protein complexes [2] as such a scoring and objective function. We do so in connection with the computationally efficient protein-protein docking algorithm FRODOCK [3]. Our results show that the DrugScore^{PPI} potentials balance well different types of interactions important for protein-protein recognition, which is remarkable given that these potentials have not been specifically designed for the prediction of protein-protein complex structures [1].

The results are discussed in view of the influence of crystal packing and the type of protein-protein complex docked. Finally, a simple criterion is provided with which to estimate *a priori* if unbound docking with DrugScore^{PPI}/FRODOCK will be successful [1].

The ability to accurately and efficiently predict protein-protein complex structures is expected to impact the development of drugs and diagnostics that interfere with such complexes, and to foster our basic understanding of evolutionary processes and signaling in biological systems.

1. Krüger, D. M.; Ignacio-Garzon, J.; Chacon, P.; Gohlke, H. DrugScore^{PPI} knowledge-based potentials used as scoring and objective function in protein-protein docking. *PLoS ONE* **2014**, *9*, e89466.
2. Krüger, D. M.; Gohlke, H. DrugScore^{PPI} webserver: fast and accurate *in silico* alanine scanning for scoring protein-protein interactions. *Nucleic Acids Res.* **2010**, *38*, W480–486.
3. Garzon, J. I.; Lopez-Blanco, J. R.; Pons, C.; Kovacs, J.; Abagyan, R. FRODOCK: a new approach for fast rotational protein-protein docking. *Bioinformatics* **2009**, *25*, 2544–2551.

P-75: Identification of novel tubulin inhibitors by parallel virtual screening protocol of reaction-based combinatorial library of combretastatin CA-4 derivatives

Rafał Kurczab[1], Zbigniew Dutkiewicz[2], Tomasz Stefański[2], Renata Mikstacka[2], Stanisław Sobiak[2]

[1]Department of Medicinal Chemistry, Institute of Pharmacology PAS, Kraków, Poland, [2]Department of Chemical Technology of Drugs, Poznań, Poland

Microtubules are cytoskeletal filaments consisting of αβ-tubulin heterodimers and are involved in a wide range of cellular functions. In the mitotic phase, microtubules are in dynamic equilibrium with tubulin dimers by assembling the tubulin into microtubules or disassembling microtubules to tubulin [1]. Disruption of the dynamic equilibrium can induce arresting cell cycle and lead to apoptosis. Hence, the compounds that could inhibit tubulin polymerization or interrupt microtubule depolymerization would be useful in the treatment of cancer [2]. In recent decades, mostly natural products, targeting tubulin have been discovered and developed; some of them are already in clinical use, such as epothilone, paclitaxel, and combrestatin A-4 (CA-4) [3].

We present here an application of parallel virtual screening protocol to identification of novel CA-4 analogs. Based on the elaborated synthesis protocol for the CA-4 analogs, the virtual combinatorial library (VCL) was created using CombiGlide. The library consisted of 1339 structures was obtained by the direct linking of four different cores (*cis*-stilbenes, α-phenylcinnamic acids, N-methylimidazoles and oxazoles as *cis*-restricted analogs) with available reagents. Next, the potency of each VCL compound was evaluated by means of developed screening protocol, which combined different evaluation models, such as 3D pharmacophore, QSAR models and post-docking scoring method based on Structural Interaction Fingerprints (SIFt) for two tubulin crystals (PDB ID: 1SA0 and 1SA1). Final ranking list was obtained by combining all rankings using consensus scoring method [4]. Based on the final scores the 16 diverse structures, with different level of predicted activity, were selected and synthesized.

The ability of compounds to inhibit tubulin polymerization was evaluated experimentally and showed good correlation with our predictions.

Acknowledgements: The study was supported by a grant of the Polish National Science Center No. 2011/03/B/NZ7/00509.

1. Valiron, O.; Caudron, N.; Job, D. Microtubule dynamics. *Cell. Mol. Life Sci*. **2001**, *58*, 2069–2084.
2. Sengupta, S.; Thomas, S. A. Drug target interaction of tubulin binding drugs in cancer therapy. *Expert Rev. Anticancer Ther*. **2006**, *6*, 1433–1447.
3. Carlson, R. O. New tubulin targeting agents currently in clinical development. *Expert Opin. Invest. Drugs* **2008**, *17*, 707–722.
4. Clark, R. D.; Strizhev, A.; Leonard, J. M.; Blake, J. F.; Matthew, J. B. Consensus scoring for ligand/protein interactions. *J Mol. Graph. Model.* **2002**, *20*, 281–295.

P-77: PTP1B – Combining Inhibitory Activity with Selectivity

Alexandra Naß, Gerhard Wolber

Pharmaceutical and Medicinal Chemistry, Institute of Pharmacy, Freie Universität Berlin, Berlin, Germany

Protein Tyrosine Phosphatase 1B (PTP1B) has now been known for about 20 years as a promising target. However, the search for selective and potent inhibitors is still ongoing [1,2]: While PTP1B is important for regulating insulin signaling in healthy humans, its activity increases with type 2 diabetes and obesity as well as mammary and ovarian carcinoma [3,4]. Especially selectivity over the closely related TCPTP remains a challenge; the impact of TCPTP inhibition in vivo is still not thoroughly discovered, but judged on the fate of TCPTP knockout mice which die soon after birth selectivity over TCPTP seems advisable to avoid severe side effects [5]. We present a work-flow that aims at increasing selectivity while maintaining high affinity levels with high ligand efficiency (LE).

By dividing the binding site into an activity and a selectivity relevant part, available protein-ligand crystal structures and known inhibitors are exploited to build 3D pharmacophore models for each site separately, which were then used to virtually screen for fragments with high LE. The molecular building blocks discovered that way are used to create a combinatorial database of potentially highly active and selective inhibitors of PTP1B. Ensemble docking experiments are used to investigate the way of binding for the resulting assembled molecules. Additionally, molecular dynamics simulations for both PTP1B and TCPTP are performed in order to identify subtle differences in binding site rigidity that can be used to increase inhibitor selectivity.

1. Groves, M. R.; Yao, Z.-J.; Roller, P. P.; Burke, T. R., Jr. Structural Basis for Inhibition of the Protein Tyrosine Phosphatase 1B by Phosphotyrosine Peptide Mimetics. *Biochemistry* **1998**, *37*, 17773–17783.
2. Bialy, L.; Waldmann, H. Inhibitors of Protein Tyrosine Phosphatases: Next-Generation Drugs? *Angew. Chem., Int. Ed.* **2005**, *44*, 3814–3839.
3. Johnson, T. O.; Ermolieff, J.; Jirousek, M. R. Protein Tyrosine Phosphatase 1B Inhibitors for Diabetes. *Nat. Rev. Drug Discov*. **2002**, *1*, 696–709.

4. Östmann, A.; Hellberg, C.; Böhmer, F. D. Protein-tyrosine Phosphatases and Cancer. *Nat. Rev. Cancer.* **2006**, *6*, 307–320.
5. Koren. S.; Fantus, I. G. Inhibition of the Protein Tyrosine Phosphatase PTP1B: Potential therapy for Obesity, Insulin Resistance and Type-2 Diabetes Mellitus. *Best Pract. Res. Clin. Endocrinol. Metab.* **2007**, *21*, 612–640.

P-79: Rational Design Supported by Ligand-Based NMR Data

Ionut Onila[1], Kai Fredriksson[1], Luca Codutti[2], Adam Mazur[3], Oliver Korb[1], Teresa Carlomagno[2], Christian Griesinger[3], Thomas E. Exner[4]

[1] University of Konstanz, Konstanz, Germany, [2] European Molecular Biology Laboratory, Heidelberg, Germany, [3] Max Planck Institute for Biophysical Chemistry, Göttingen, Germany, [4] Eberhard Karls University of Tübingen, Tübingen, Germany

In structure-based drug design, the experimental elucidation of protein-ligand complexes plays a central role in the design of high-affinity drug candidates from weakly bound lead compounds. In most cases this is done by X-ray crystallography, but when this is not possible, NMR methods can alternatively be applied. One method, the INPHARMA [1,2] (Internuclear NOEs for Pharmacophore Mapping) approach, determines the relative orientation of two competitive ligands in the receptor binding pocket. It is based on the observation of interligand transferred NOEs mediated by spin diffusion through protons of the protein and is, therefore, sensitive to the specific interactions of each of the two ligands with the protein [3].

In a recent publication [4] it was demonstrated that INPHARMA can successfully predict the poses of 5 weakly bound cAMP-dependent Protein Kinase (PKA) ligands, using a set of 10 pairwise experiments. This was done in a two-step procedure in which first trial poses were generated with a docking program and the correct poses were then identified by a rescoring with the experimental data. Here, we present that it is even more beneficial and efficient to combine these steps. The docking program PLANTS [5,6] (Protein-Ligand ANT System) developed in our group was extended to directly include the experimental information. The standard scoring function ChemPLP [6] is augmented with an INPHARMA score which describes the agreement between the experimental spectrum and a back-calculated spectrum using the full relaxation matrix approach. We will also demonstrate that, besides this improved function, the careful preparation of the input structures and the docking setup including protonation states and flexible side chains is essential.

1. Sanchez-Pedregal, V. M.; Reese, M.; Meiler, J.; Blommers, M. J. J.; Griesinger, C.; Carlomagno, T. The INPHARMA Method: Protein-Mediated Interligand NOEs for Pharmacophore Mapping. *Angew. Chem., Int. Ed.* **2005**, *44*, 4172–4175.
2. Orts, J.; Tuma, J.; Reese, M.; Grimm, S. K.; Monecke, P.; Bartoschek, S.; Schier, A.; Wendt, K. U.; Griesinger, C.; Carlomagno, T. Crystallography-Independent Determination of Ligand Binding Modes. *Angew. Chem., Int. Ed.* **2008**, *47*, 7736–7740.
3. Reese, M.; Sanchez-Pedregal, V. M.; Kubicek, K.; Meiler, J.; Blommers, M. J. J.; Griesinger, C.; Carlomagno, T. Structural Basis of the Activity of the Microtubule-Stabilizing Agent Epothilone A Studied by NMR Spectroscopy in Solution. *Angew. Chem.* **2007**, *119*, 1896.
4. Skjaerven, L.; Codutti, L.; Angelini, A.; Grimaldi, M.; Latek, D.; Monecke, P.; Dreyer, M. K.; Carlomagno, T.; Accounting for Conformational Variability in Protein-Ligand Docking with NMR-Guided Rescoring. *J. Am. Chem. Soc.* **2013**, *135*, 5819–5827.
5. Korb, O.; Stützle, T.; Exner, T. E. An ant colony optimization approach to flexible protein- ligand docking. *Swarm Intell.* **2007**, *1*, 115–134.

6. Korb, O.; Stützle, T.; Exner, T. E. Empirical Scoring Functions for Advanced Protein-Ligand Docking with PLANTS. *J. Chem. Inf. Model.* **2009**, *49*, 84–96.

P-81: Pharmacophore based Virtual-Screening for the Identification of Covalent Coxsackievirus 3C Protease Inhibitors

Robert Schulz, **Gerhard Wolber**

Pharmaceutical Chemistry, Institute of Pharmacy, Freie Universität Berlin, Berlin, Germany

Viral proteases are promising targets for the treatment of viral infections due to the important role they play in the replication circle. However, targeting proteases is challenging as reversible inhibitors often show limited affinities. One possibility to address this challenge, is the design of covalently binding substructures, so-called warheads, which result in high affinity and provide an experimental tool for lead optimization through dynamic ligation screening [1][1]. This way, the rational development of reversible non-peptidic ligands becomes possible.

We have developed a new 3D pharmacophore feature, called "Residue Bonding Point", which resembles the covalent interaction between a warhead and reactive amino acid site chains. It was implemented into the pharmacophore modeling software LigandScout [2]. As a use case, the coxsackievirus B3 3C protease was chosen. This cysteine protease is the main polyprotein processing protease in the large family of picornaviridae, to which also well-known human pathogenic viruses such as hepatitis a virus, rhinoviruses or poliovirus belong. The active site cysteine147 has been shown to be prone to covalent modification by warhead containing ligands [3].

A structure-based pharmacophore model has been developed and validated for the 3C protease, including the "Residue Bonding Point" feature on the cysteine147. The subsequent virtual screening of commercially available compound libraries followed by covalent docking in Gold led to a compound selection of which a subset was purchased and tested in a protease inhibition assay. In this work we have shown the potential of 3D pharmacophores to identify covalent inhibitors and so expanded the scope of the design of irreversible inhibitors.

1. Schmidt, M. F.; Isidro-Llobet, A.; Lisurek, M.; El-Dahshan, A.; Tan, J.; Hilgenfeld, R.; Rademann, J. Sensitized detection of inhibitory fragments and iterative development of non-peptidic protease inhibitors by dynamic ligation screening. *Angew. Chem.* **2008**, *47*, 3275–3278.
2. Wolber, G.; Langer, T., LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J. Chem. Inf. Model.* **2005**, *45*, 160–169.
3. Dragovich, P. S.; Webber, S. E.; Babine, R. E.; Fuhrman, S. A.; Patick, A. K.; Matthews, D. A.; Reich, S. H.; Prins, T. J.; Marakovits, J. T.; Littlefield, E. S.; Zhou, R.; Tikhe, J.; Ford, C. E.; Wallace, M. B.; Bleckman, T. M.; Meador, J. W.; Ferre, R. A.; Brown, E. L.; Binford, S. L.; DeLisle, D. M.; Worland, S. T., Structure-based design of irreversible human rhinovirus 3C protease inhibitors. *Abstr. Pap. Am. Chem. S.* **1998**, *215*, U863–U863.

P-83: Use of Back-Calculated Protein Chemical Shift Perturbations in Fragment Docking

Tim ten Brink, Clementine Aguirre, Isabelle Krimm

Institute des Sciences Analytiques – CNRS UMR5280 , Universite Claude Bernard - Lyon 1, Villeurbanne, F-69622, France

The resolution the the binding mode is an essential step in fragment based drug design. X-ray crystallography is usually the method of choice but can fail due to badly diffracting crystals or affinity problems. Computational docking of fragment like molecules is on the other hand often handicapped by the fact that different fragment orientations are often ranked equally by the scoring functions [1]. Filtering the docking results by agreement with experimental data can be a way to reveal the fragments true binding modes. Chemical Shift Perturbations (CSP) occurring for the protein protons upon ligand binding can be obtained from HSQC experiments performed during NMR screening of fragment libraries. These CSP are long known to contain enough structural information to orientate ligands into protein binding sites by comparing back-calculated CSP for docked ligand orientations with experimental CSP [2,3]. However questions like the choice of nuclei (amide proton versus aliphatic protons) or post docking filtering versus back-calculation included into the docking process, along with influences in the CSP not directly caused by the binding of the ligand, have until now prevented the method of being routinely used.

Here we try to answer some of these question by directly comparing the performance of a post docking filter, that uses back-calculated proton CSP to rescore ensembles of docked fragment poses to an approach were the CSP calculation was directly included as a new term into the scoring function of the PLANTS [4,5] program. Additionally we evaluate the performance of amide proton CSP and aliphatic proton CSP for a protein target (human peroxiredoxin 5) with a set of analogous fragments. Here we lay special empathises on the question wheter the additional experimental effort to obtain the $^{1}H^{13}C$ data is justified by better performance of the CSP filtering compared to the more readily available $^{1}H^{15}N$ data.

1. Verdonk, M. L.; Giangreco, I.; Hall, R. J.; Korb, O.; Mortenson, P. N.; Murray, C. W. Docking Performance of Fragments and Druglike Compounds *J. Med Chem*. **2011**, *54*, 5422–5431.
2. González-Ruiz, D.; Gohlke, H. Steering Protein–Ligand Docking with Quantitative NMR Chemical Shift Perturbations. *J. Chem. Inf. Mod.* **2009**, *49*, 2260–2271.
3. McCoy, M. A.; Wyss, D. F. Alignment of weakly interacting molecules to protein surfaces using simulations of chemical shift perturbations. *J. Biomol. NMR* **2000**, *18*, 189–198.
4. Korb, O.; Stützle, T.; Exner, T. E. PLANTS: Application of Ant Colony Optimization to Structure-Based Drug Design Ant Colony Optimization and Swarm Intelligence, Dorigo, M.; Gambardella, L. M.; Birattari, M.; Martinoli, A.; Poli, R.; Stützle, T., Eds.; 5th International Workshop, ANTS 2006, 2006, 247–258.
5. Korb, O.; Stützle, T.; Exner T. E. Empirical scoring functions for advanced protein-ligand docking with PLANTS. *J. Chem. Inf. Mod.* **2009**, *49*, 84-89.

P-85: Small molecule inhibitors of CD40-TRAF6 interaction reduce atherosclerosis by targeting its inflammatory nature.

Barbara Zarzycka[1,*], Tom Seijkens[2,*], Oliver Soehnlein[3,4], M. Hoeksema[2], Linda Beckers[2], E. Smeets[2], S. Meiler[2], M. Gijbels[2,5], R. Schrijver[1], Louis Boon[6], T. Hackeng[1], N. Gerdes[2,4], Menno de Winther[2], Christian Weber[4,7], G Vriend[8], S. B. Nabuurs[8,9], Esther Lutgens[2,4#], G. A. F. Nicolaes[1#]

[1]Department of Biochemistry, Cardiovascular Research Institute Maastricht, Maastricht University, Maastricht, Netherlands, [2]Department of Medical Biochemistry, Academic Medical Center (AMC), University of Amsterdam, Amsterdam, Netherlands, [3]Department of Pathology, Academic Medical Center (AMC), University of Amsterdam, Amsterdam, Netherlands, [4]Institute for Cardiovascular Prevention (IPEK), Ludwig Maximilians University, Munich, Germany, [5]Department of Pathology, Cardiovascular Research Institute Maastricht (CARIM), Maastricht University, Maastricht, Netherlands, [6]Bioceros BV, Utrecht, Netherlands, [7]DZHK (German Centre for Cardiovascular Research), partner site Munich Heart Alliance, Munich, Germany, [8]Center for Molecular and Biomolecular Informatics, Radboud University Medical Center, Nijmegen, Netherlands, [9]Lead Pharma Medicine, Nijmegen, Netherlands

[#]These authors contributed equally to this work*

Atherosclerosis is a chronic inflammatory disease of the arterial wall that may lead to stroke or myocardial infarction [1]. The interaction between CD40 protein and its ligand CD40L plays pro-inflammatory role in atherosclerosis. In order to elicit intracellular signalling, CD40 needs to recruit adaptor proteins: the tumour necrosis factor receptor-associated factors (TRAFs) [2]. The CD40-TRAF6 but not CD40-TRAF2/3/5, interactions are critically involved in atherosclerosis progression; therefore CD40-TRAF6 interaction is more downstream target which can inhibit inflammatory nature of atherosclerosis [3].

To discover small drug-like inhibitors of CD40-TRAF6 interaction we used hierarchical protocol of structure-based virtual ligand screening (SBVLS). The top scored compounds were tested in an *in vitro* cell-based NFϰB (RAW-NFkB-luc) screening assay. The top 7 compounds dose-dependently reduced NFϰB activation and CD40-induced IL1β and IL6 expression in primary macrophages. Surface plasmon resonance (SPR) experiments confirmed direct binding of the compounds to the TRAF6 C-domain.To determine the in vivo effect of the compounds we treated atherosclerotic ApoE-/- mice with two selected compounds for 6 wks starting at the age of 12 wks. Compound treatment reduced total plaque area per aortic arch by 47.1% and 66.8%, and changed the plaque phenotype by lowering number of leukocytes per plaque by 43.1% and 52.6%, respectively. Intravital microscopy of the carotid artery in ApoE-/- mice revealed that monocyte and neutrophil adhesion were reduced by 40.1% and 51.2% respectively. Accordingly, chemokine (CCL2, CCL5) expression of compound treated macrophages was decreased.

The compounds inhibited CD40-TRAF6 interaction by direct binding to TRAF6 C-domain and reduced progression of atherosclerosis by inhibition of chemokine-mediated leukocyte influx into the arterial wall. Our results indicate possibilities of long-term therapeutic inhibition of CD40-TRAF6 interactions in atherosclerosis and possibly other inflammatory diseases.

1. Hansson, G. K.; Hermansson, A. *Nat. Immunol.* **2011**, *12*, 204–212.
2. Engel, D.; Seijkens, T.; Poggi, M.; Sanati, M.; Thevissen, L.; Beckers, L.; Wijnands, E.; Lievens, D.; Lutgens, E. *Semin. Immunol.* **2009**, *21*, 308–312.
3. Lutgens, E.; Lievens, D.; Beckers, L.; Wijnands, E.; Soehnlein, O.; Zernecke, A.; Seijkens, T.; Engel, D.; Cleutjens, J.; Keller, A. M.; Naik, S. H.; Boon, L.; Oufella, H. A.; Mallat, Z.; Ahonen, C. L.; Noelle, R. J.; de Winther, M. P.; Daemen, M. J.; Biessen, E. A.; Weber, C. J. *Exp. Med.* **2010**, *207*, 391–404.

Sponsoring Societies

- Chemical Information and Computer Applications Group, Royal Society of Chemistry (RSC)
- Chemical Structure Association Trust (CSA Trust)
- Chemistry-Information-Computer Division (CIC), German Chemical Society (GDCh)
- Division of Chemical Information (CINF), American Chemical Society (ACS)
- Division of Chemical Information and Computer Science, Chemical Society of Japan (CSJ)
- Royal Netherlands Chemical Society (KNCV)
- Swiss Chemical Society (SCS)