

ICCS

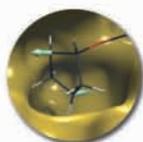
International Conference
on Chemical Structures

9th International Conference on Chemical Structures
June 5-9, 2011 ♦ Noordwijkerhout ♦ The Netherlands

Program & Abstracts

MOE is a fully integrated drug discovery software package. MOE runs on Windows, Linux, Unix, and Mac OS X. MOE contains a toolbox for adapting existing and creating new applications. With MOE, molecular modelers, medicinal chemists and occasional users can benefit from sharing the same software system.

MOE 2010.10 HIGHLIGHTS



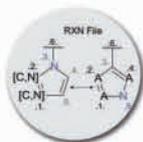
Streamlined Interactive Modeling Interface

- Toggle ligands, proteins and surfaces on/off.
- Analyze and optimize multiple ligand:receptor complexes.
- Create surfaces, calculate properties and display substitution points.



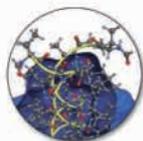
Integration of NAMD Engine in MOE

- Export parameters and scripts automatically.
- Import NAMD trajectories into MOE database.
- Run simulations on a cluster with restart capability.



Structure-Based Medicinal Chemistry Transformations

- Transform molecules in 3D using reaction style rules.
- Refine structures in an active site and apply 3D filters.
- Integrated with scaffold replacement, fragment linking, growing and BREED.



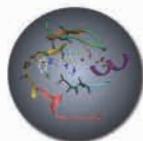
Enhanced Graphics

- Faster real-time GPU ray-tracing.
- 3D Stereo with anaglyph glasses.
- Clip molecular surfaces only.



Non-bonded Interaction Visualization

- Display H-bonds, CH...X, proton- π and VdW interactions.
- Show strengths or energies and set thresholds.
- Control visualization for ligand, receptor and solvent combinations.



Kinase Database and Explorer

- Search database of 3D aligned kinase structures.
- Add in-house structures with automated protocol.
- Browse kinases by core, pocket or canonical structural views.

Preface

Welcome to the Ninth International Conference on Chemical Structures.

With the Ninth International Conference on Chemical Structures we continue this well-established conference series that began in 1973 as a workshop on Computer Representation and Manipulation of Chemical Information sponsored by the NATO Advanced Study Institute and thereafter was held under its new name every third year starting in 1987. The 2011 conference continues the high standard of technical presentations and discussions that characterized all previous conferences. The response to the Call for Papers has produced an outstanding program of technical papers and posters and also attracted a sizable number of vendors and scientific institutions showing their newest software, content, and applications.

The conference was again chosen as the preferred venue to award the fourth CSA Trust Mike Lynch Award to Dr Englebert Zass of the Eidgenössische Technische Hochschule (ETH), Zürich. Dr Zass will open the conference by receiving the award and delivering the keynote address titled *The Intermediary Reloaded - On the Need for a "Go-Between" to Information Users and Producers* on Sunday evening. Prior recipients of the CSA Trust Mike Lynch Award include Professor Peter Willett of the University of Sheffield in 2002, Professor Johnny Gasteiger of the University of Erlangen-Nürnberg in 2005, and Professor Alexander Lawson of Elsevier Information Systems in 2008.

Once again, the scientific poster session has been divided into two sessions this year due to the large number of posters being presented. The posters have been divided into a red group and a blue group. All posters will be exhibited during the poster sessions; however, presenters from the red posters will be available during the Monday evening poster session and during the Tuesday evening poster session presenters from the blue posters will be available.

Following the conference you are encouraged to submit your talk/poster to the Journal of Chemical Information and Modeling (JCIM) for publication. Typically, a special issue of JCIM follows the conference containing accepted papers from the conference.

This year, the group excursion on Wednesday afternoon/evening returns to what has become viewed as the traditional ICCS sail boat ride on the IJsselmeer. The return of this excursion has proved popular and we are sure that you will enjoy the experience and welcome the valuable networking opportunity that it furnishes. The boats will sail the IJsselmeer and stop at the picturesque village of Volendam where you can choose between visiting a traditional eel-smoking facility, walking around and taking in the atmosphere, or simply relaxing at one of the cafes and bars that line the harbor front.

We hope that you enjoy the conference and if you ever need assistance during the week please contact one of the conference Organizing Committee or Scientific Advisory Board members or the assistants located at the conference desk.

Keith Taylor, Chair
Markus Wagener, Co-Chair

Contents

| | |
|---|-----|
| The Conference | 5 |
| Organizing Committee | 7 |
| Scientific Advisory Board | 7 |
| Sponsors | 9 |
| Exhibitors | 11 |
| Exhibitors..... | 13 |
| List of Exhibition Layout | 13 |
| Exhibitors | 13 |
| Exhibition Hours | 13 |
| Workshops | 15 |
| BioSolvIT | 15 |
| Tripos | 15 |
| Chemical Computing Group | 16 |
| Accelrys | 16 |
| Group Excursion: Sailing Cruise on the IJsselmeer | 17 |
| Itinerary | 17 |
| Volendam | 17 |
| Technical Program | 19 |
| Plenary Session | 21 |
| Poster Session – RED | 27 |
| Poster Session – BLUE | 31 |
| Plenary Session Abstracts | 35 |
| Poster Session Abstracts – RED | 65 |
| Poster Session Abstracts – BLUE | 99 |
| Participants | 131 |
| List of Participants | 133 |

The Conference

Organizing Committee and Scientific Advisory Board

Organizing Committee

| | |
|--|---|
| Dr Keith T Taylor, Chair | |
| Dr Markus Wagener, Co-Chair | |
| Dr Kimito Funatsu <i>Division of Chemical Information and Computer Science of the Chemical Society of Japan (CSJ)</i> |  |
| Dr Guenter Grethe <i>Division of Chemical Information of the American Chemical Society (CINF)</i> |  |
| Dr Phil McHale <i>Chemical Structure Association Trust (CSA Trust)</i> |  |
| Dr Frank Oellien <i>Chemistry-Information-Computer Division of the German Chemical Society (GDCh)</i> |  |
| Lutgarde Buydens <i>Royal Netherlands Chemical Society (KNCV)</i> |  |
| Dr Don Parkin <i>Chemical Information and Computer Applications Group of the Royal Society of Chemistry (RSC)</i> |  |
| Dr Peter Ertl <i>Swiss Chemical Society (SCS)</i> |  |

Scientific Advisory Board

- ◆ Dr Peter Ertl, Novartis Institutes for BioMedical Research
- ◆ Dr Kimito Funatsu, Department of Chemical System Engineering School of Engineering of The University of Tokyo
- ◆ Dr Val Gillet, University of Sheffield
- ◆ Dr Rajarshi Guha, NIH Center for Translational Therapeutics
- ◆ Dr Michael Lajiness, Eli Lilly and Company
- ◆ Dr Jordi Mestres, Universitat Pompeu Fabra
- ◆ Dr Matthias Rarey, Zentrum für Bioinformatik of the Universität Hamburg
- ◆ Dr Keith T Taylor, Accelrys Inc
- ◆ Dr Lothar Terfloth, Molecular Networks GmbH
- ◆ Dr Markus Wagener, MSD
- ◆ Dr W Patrick Walters, Vertex Pharmaceuticals Inc

Sponsors

Premier Sponsor



Platinum Sponsors



Gold Sponsors



Molecular Networks
Inspiring Chemical Discovery



Silver Sponsors



SCHRÖDINGER.



Publishing Sponsors

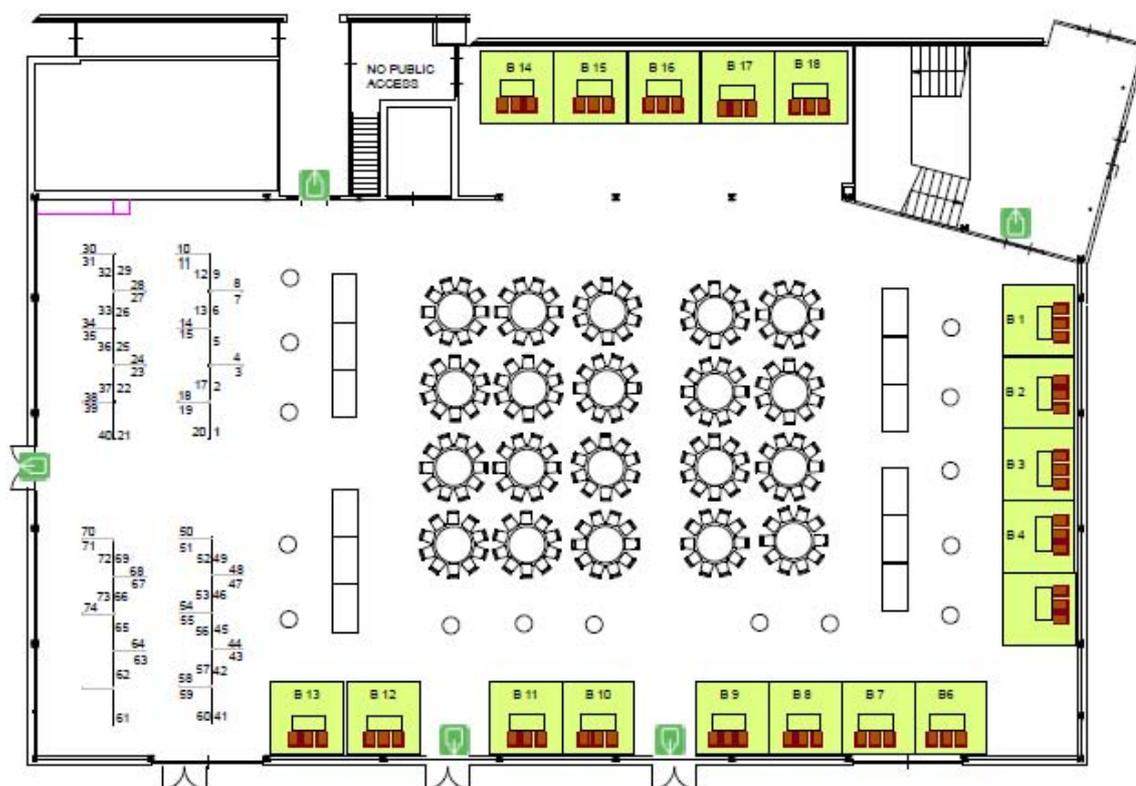


Exhibition

List of Exhibitors



Exhibition Layout (Atrium)



Exhibitors

| Exhibitor | Booth | Exhibitor | Booth |
|-------------------------------------|--------|---------------------------------|--------|
| Accelrys Inc | B - 7 | InteLigand GmbH | B - 10 |
| Advanced Chemistry Development, Inc | B - 5 | Keymodule Limited | B - 1 |
| BioSolveIT GmbH | B - 14 | Mesa Analytics & Computing Inc | B - 17 |
| CambridgeSoft Corporation | B - 4 | Molecular Networks GmbH | B - 2 |
| Chemical Abstracts Service | B - 3 | OpenEye Scientific Software Inc | B - 6 |
| Chemical Computing Group | B - 13 | Schrödinger Inc | B - 9 |
| Cresset BMD Limited | B - 12 | Simulations Plus Inc | B - 11 |
| Dotmatics Limited | B - 18 | Tripes | B - 8 |
| Evolvus | B - 16 | Xemistry GmbH | B - 15 |

Exhibition Hours

- Monday 14:30 – 19:30
- Tuesday 14:30 – 19:30

Workshops

Sunday June 5, 2011

BioSolveIT

Medicinal and computational chemists - team up!

The workshop will focus on solutions for scaffold hopping, fragment- and structure-based design problems. LeadIT combines the power of scaffold hopping (ReCore) with our well established docking engine (FlexX). Wrapped in an easy to use GUI, this team-building software brings Medicinal and Computational Chemists together in front of the screen. All we need to show you during the workshop is what you can do, the how-to is an effortless exercise with our GUI.

After the workshop, each participant will receive a FREE 3-month license to practice topics covered in the workshop.

Tripods

Informatics platforms for a more productive future

As the drug industry moves towards therapeutic focus within multi-disciplinary research teams, informatics platforms that permit access and analysis of data from various different areas or departments in a discovery unit are becoming vital in modern drug discovery.

One of the major challenges with this approach is to effectively leverage the vast amount of information being generated within research organizations. The challenge, more specifically, is to provide research scientists access to research data they need, regardless of where it resides; quickly and easily in order to make faster, better/more informed decisions. Thus, allowing large gains in efficiency and productivity at a time when our industry struggles against budget constraints and sparse drug discovery pipelines.

This FREE session will showcase the experiences Tripods has had implementing such informatics platform solutions at top pharma such as Pfizer to the smallest biotech companies around. Showing how company unique configurable platforms can be implemented spanning discovery research to provide research scientists with a single point of access to retrieve, analyze and share their scientific data. Eliminating time consuming, error prone, non-productive hours that scientists spend merging and manipulating data from multiple, disparate sources. Tripods will also feature their first steps undertaken in creating a fully functional translational informatics platform for a more productive future that potentially links discovery, pre-clinical and clinical data.

Thursday June 9, 2011

Chemical Computing Group

Computational approaches to fragment-based drug design and medchem exploration

CCG's MOE software contains a range of new applications for molecular scaffold replacement, growth of fragments within a binding site, and exploring medicinal chemistry possibilities for lead optimisation. All of these are carried out within the context of a 3-dimensional binding site, with assorted filters applied (molecular property and descriptor filters, QSAR models and molecular fingerprints, pharmacophore filters, and a synthesisability measure) to manage the output, before minimisation of the compounds produced within the context of the receptor, which may be kept rigid or allowed to relax, and ranking based on binding energy. Examples of the use of these applications will be given, illustrating the strategies which may be applied to derive potential ways forward in scaffold-hopping, fragment-based design, and lead optimisation projects.

Accelrys

Chemical representation of chemically modified biological sequences

Biologics are a hot area of research for many pharmaceutical companies; currently about 45% of new therapeutics entities are biologics. Companies need to characterize the materials and manage them much as they have managed conventional drugs, but this presents many challenges.

The normal way to represent biologics is as text strings with annotations. This presents difficulties in comparing the properties of different entities. In particular, structure activity comparisons are needed that take account of all the modifications that are present in the sequence.

In many cases a full sequence is unknown, and even if it is the history of the substance can have a critical impact on its activity. It is not uncommon for a committee to decide whether a substance is unique.

In this workshop, we will discuss the challenges in unambiguously characterizing chemically modified sequences, both natural Post Translational Modifications (PTM) and synthetic modifications. The capabilities of Accelrys' new, Self Contained Sequence Representation (SCSR) will be described and its ability to search for modifications and develop structure property tables will be demonstrated.

The workshop attendees will be encouraged to propose future developments of the representation.

Excursion: Sailing Cruise on the IJsselmeer (Lake IJssel)

Itinerary

| | |
|-------|---|
| 13:00 | Depart from the conference center, Noordwijkerhout |
| 14:00 | Arrive at Muiderslot castle, board the sailing boats Willem Barentsz or Abel Tasman |
| 17:00 | Arrive at Volendam , where there will be the possibility to <ul style="list-style-type: none"> • visit the eel smokehouse Smit-Bokkum • have a guided tour of the old fishing village • enjoy the bars and cafes at the quay |
| 18:30 | Return from the village and cast off for the way back, dinner will be served on board |
| 22:00 | Arrive at Blocq van Kuffeler, disembark and return to Noordwijkerhout |
| 23:00 | Arrive at the conference center, Noordwijkerhout |

Volendam

The destination of the sailing cruise on the IJsselmeer (Lake IJssel) is the picturesque fishing village Volendam, one of the best known ports on the former [Zuiderzee](#). Volendam is famous for its characteristic fisherman's cottages, the traditional women's costumes with a high, pointed bonnet, and also for its delicious smoked eel.

Attractions in Volendam include the old houses around the Doolhof (maze) which have been erected without any kind of recognizable organized street plan, the fish auction hall (Visafslag) which was in use until the end of last century, and the Stolphoevekerkje reformed church that dates from 1658, an almost square timber construction resting on a stone foundation. And last not least the lively atmosphere at the harbor with its many cafes and restaurants along the quay (De Dijk).

Please bring appropriate clothing as there is always the possibility for some rain or windy weather on the lake and it might get a little chilly after sundown. A good pair of walking shoes and sunscreen is also advised.

Technical Program

Plenary Session

Ninth International Conference on Chemical Structures

| Sunday June 5 | |
|---------------|--|
| 12:00 - 18:00 | Registration |
| 15:00 - 17:00 | Workshop: BioSolveIT, <i>Medicinal and computational chemists — team up!</i> Meeting Room Boston 13 |
| 15:00 - 17:00 | Workshop: Tripos, <i>Informatics platforms for a more productive future</i> Meeting Room Boston 11 |
| 17:00 - 18:00 | Free Time |
| 18:00 - 19:00 | Welcome and Keynote Address |
| 18:00 - 18:15 | ORG-1 : <i>Welcome and Introduction</i> Keith Taylor, ICCS Program Chair |
| 18:15 - 19:00 | Opening Session - Keynote Address, CSA Trust Mike Lynch Award K-1 : <i>The intermediary reloaded - On the need for a "Go-Between" to information users and producers</i> Engelbert Zass, ETH, Zurich |
| 19:00 - 20:00 | Welcoming Reception - Atrium |
| 20:00 - 22:00 | Reception Dinner - Atrium |

| Monday June 6 | |
|---------------|--|
| 8:30 - 14:30 | Structure-Activity and Structure-Property Prediction Michael Lajiness, Presiding |
| 8:30 - 9:00 | A-1 : <i>Beyond filters: assessing ADMET risk for multi-objective drug development</i> Robert Daniel Clark, Simulations Plus, Inc. |
| 9:00 - 9:30 | A-2 : <i>SeeSAR: see your structure-activity relationships</i> Jos Lommerse, MSD |
| 9:30 - 10:00 | A-3 : <i>Real world application of proteochemometric modeling for the design of enzyme inhibitors and ligands of G-protein coupled receptors</i> Gerard JP van Westen, Leiden University / LACDR |
| 10:00 - 10:30 | Break |
| 10:30 - 11:00 | A-4 : <i>QSAR Workbench: enabling best practices in large scale model building</i> Stephen D Pickett, GlaxoSmithKline |
| 11:00 - 11:30 | A-5 : <i>Using AutoQSAR to select the most predictive modeling methods</i> Sarah L Aaron, Accelrys/AstraZeneca |
| 11:30 - 12:00 | A-6 : <i>What should we make next?</i> Mark Mackey, Cresset BioMolecular Discovery Ltd |
| 12:00 - 13:00 | Lunch - Atrium |
| 13:00 - 13:30 | A-7 : <i>DrugscoreMaps – a graphical method for SAR analysis, active/inactive separation, binding mode analysis and docking pose evaluation using protein-ligand interaction fingerprints</i> Oliver Koch, Intervet Innovation GmbH |
| 13:30 - 14:00 | A-8 : <i>Mining of emerging structural patterns for identification of toxicophores</i> Richard J.Sherhod, University of Sheffield |
| 14:00 - 14:30 | A-9 : <i>Which substructures are interesting?</i> Jeroen Kazius, Curios-IT |
| 14:30 - 15:00 | Break |
| 14:30 - 19:30 | Exhibition and Posters – Atrium Val Gillet, Presiding |
| 15:00 - 17:00 | Poster Presentations RED (authors present) |
| 18:30 - 19:30 | Reception - Atrium |
| 19:30 - 21:30 | Dinner - Atrium |

Ninth International Conference on Chemical Structures

| Tuesday June 7 | |
|----------------|--|
| 8:30 - 12:00 | Structure-Based Drug Design and Virtual Screening Matthias Rarey, Presiding |
| 8:30 - 9:00 | B-1 : <i>Global free energy scoring functions based on distance-dependent atom-type pairs</i> Christian Kramer, Novartis Institutes for BioMedical Research |
| 9:00 - 9:30 | B-2 : <i>Development and validation of an in silico scaffold hopping protocol for identifying novel kinase inhibitors</i> Sarah R Langdon, The Institute of Cancer Research |
| 9:30 - 10:00 | B-3 : <i>Snooker: target-focused library selection using structure based pharmacophores</i> Marijn Sanders, Radboud University Nijmegen |
| 10:00 - 10:30 | Break |
| 10:30 - 11:00 | B-4 : <i>Cavity Knowledge Acceleration (CavKA) – metamorphosis in automatic pharmacophore elucidation</i> Florian Koelling, University of Technology Braunschweig |
| 11:00 - 11:30 | B-5 : <i>3D pharmacophore searching against ten trillion combinatorially accessible compounds</i> Qiang Zhang, Boehringer Ingelheim Pharmaceuticals, Inc. |
| 11:30 - 12:00 | B-6 : <i>Learning from the best: utilizing knowledge-based protein validation scores in receptor-ligand complex prediction.</i> Sander B Nabuurs, Radboud University Nijmegen |
| 12:00 | Group Photo |
| 12:00 - 13:00 | Lunch – Atrium |
| 13:00 - 14:30 | Integrated Chemical Information Markus Wagener, Presiding |
| 13:00 - 13:30 | C-1 : <i>Validation and characterization of chemical structures derived from names and images in scientific documents</i> John B Kinney, DuPont |
| 13:30 - 14:00 | C-2 : <i>Prediction of adverse drug reactions using systems biology</i> Cédric Merlot, drug design tech |
| 14:00 - 14:30 | C-3 : <i>Global mapping of ligand and target binding spaces</i> Felix A Krüger, European Bioinformatics Institute |
| 14:30 - 15:00 | Break |
| 14:30 - 19:30 | Exhibition and Posters – Atrium Lothar Terfloth, Presiding |
| 15:00 - 17:00 | Poster Presentations - BLUE (authors present) |
| 18:30 - 19:30 | Reception – Atrium |
| 19:30 - 21:30 | Dinner – Atrium |

| Wednesday June 8 | |
|------------------|---|
| 8:30 - 13:00 | Analysis Of Large Chemistry Spaces Pat Walters, Presiding |
| 8:30 - 9:00 | D-1 : <i>Pharmacophoric space: do targets segregate?</i> Andrea Zaliani, Evolvus |
| 9:00 - 9:30 | D-2 : <i>Extraction of useful bioisostere replacements from the PDB</i> Tina Ritschel, Radboud University Nijmegen |
| 9:30 - 10:00 | D-3 : <i>PubChem3D: diversity of shape space</i> Evan Bolton, U.S. National Center for Biotechnology Information (NCBI) |
| 10:00 - 10:30 | D-4 : <i>Comparison and visualization of large chemical spaces using unsupervised classification techniques</i> Alexander Böcker, Evotec AG |
| 10:30 - 11:00 | Break |
| 11:00 - 11:30 | D-5 : <i>WizePairZ: auto-curation of matched molecular pairs</i> David Wood, AstraZeneca |
| 11:30 - 12:00 | D-6 : <i>Mining for context-sensitive matched molecular pairs and bioisosteric replacements in large corporate chemical databases</i> George Papadatos, Eli Lilly |
| 12:00 - 12:30 | D-7 : <i>On the exclusion of unwanted chemical patterns from large fragment spaces</i> Hans-Christian Ehrlich, University of Hamburg |
| 12:30 - 13:00 | D-8 : <i>An analysis of fragment-spaces and their impact on fragment replacement</i> Geoff Skillman, OpenEye Scientific Software |
| 13:00 | Box Lunch |
| 13:00 - 23:00 | Excursion <i>Cruise the IJsselmeer on two traditional sailing boats visiting the picturesque fishing village Volendam where the old village can be explored. A banquet dinner will be served on the boats on the way back</i> |

Ninth International Conference on Chemical Structures

| Thursday June 9 | |
|-----------------|--|
| 7:30 - 8:30 | Hotel Check-out |
| 8:30 - 10:30 | Cheminformatics Keith Taylor, Presiding |
| 8:30 - 9:00 | E-1 : <i>ARChem route designer: the application of automated retrosynthetic rule generation to synthesis planning</i> Anthony P Cook, University of Leeds |
| 9:00 - 9:30 | E-2 : <i>De novo design of synthetically feasible compounds using reaction vectors and evolutionary multiobjective optimization</i> Benjamin Christopher Allen, University of Sheffield |
| 9:30 - 10:00 | E-3 : <i>Improving metabolite identification with chemoinformatics</i> Julio E Peironcelly, Leiden University |
| 10:00 - 10:30 | E-4 : <i>Efficient matching of multiple chemical subgraphs</i> Roger A Sayle, NextMove Software |
| 10:30 - 11:00 | Break and Hotel Check-out |
| 11:00 - 13:00 | Dealing with Biological Complexity Peter Ertl, Presiding |
| 11:00 - 11:30 | F-1 : <i>Targeting natural products for drug discovery by mining biomedical information resources</i> Eugene Muratov, University of North Carolina |
| 11:30 - 12:00 | F-2 : <i>Identifying and quantifying drug promiscuity by correlating ligand and target shape similarities</i> Violeta I Perez Nueno, INRIA Nancy – Grand Est (LORIA) |
| 12:00 - 12:30 | F-3 : <i>A knowledge-based approach to assessing the target promiscuity of chemical fragments</i> Xavi Jalencas, Institut Municipal d'Investigació Mèdica |
| 12:30 - 13:00 | F-4 : <i>Combining global and local measures for druggability predictions.</i> Andrea Volkamer, University of Hamburg |
| 13:00 - 13:15 | Closing Remarks, Markus Wagener |
| 13:15 - 14:00 | Lunch |
| 13:30 | Shuttle busses leave for Schiphol Airport |
| 14:30 | Shuttle busses leave for Schiphol Airport |
| 14:00 - 16:00 | Workshop: CCG, <i>Computational approaches to fragment-based drug design and medchem exploration</i> Meeting Room Boston 13 |
| 14:00 - 16:00 | Workshop: Accelrys, <i>Chemical representation of chemically modified biological sequences</i> Meeting Room Boston 11 |
| 16:15 | Shuttle busses leave for Schiphol Airport |

Poster Session

RED

Cheminformatics

- P - 2 : *PubChem3D: a significant new resource for scientists*
Evan Bolton, U.S. National Center for Biotechnology Information (NCBI)
- P - 4 : *A task-oriented comparison of multiple cheminformatics toolkits*
Andrew P Dalke, Andrew Dalke Scientific AB
- P - 6 : *Handling of tautomerism and stereochemistry in compound registration*
Alberto Gobbi, Genentech
- P - 8 : *Explora: a new language to define powerful structural queries*
Thierry G Hanser, Lhasa Limited
- P - 10 : *Revisiting the dataflow paradigm for chemical information processing*
Wolf D Ihlenfeldt, Xemistry GmbH
- P - 12 : *Managing chemical libraries using Screening Assistant 2.0*
Vincent le Guilloux, ICOA
- P - 14 : *High-throughput structure analysis and descriptor generation for crystalline porous materials*
Richard L Martin, Lawrence Berkeley National Laboratory
- P - 16 : *Electronic Laboratory Notebook – the academic point of view*
Felix Rudolphi, MPI fuer Kohlenforschung
- P - 18 : *CSRML – a new and open exchange format for chemical knowledge*
Christof H. Schwab, Molecular Networks GmbH
- P - 20 : *Comparison of protein structural motifs – challenges and algorithms. Improved approach and case studies*
Radka S Vařeková, Masaryk University
- P - 22 : *Barcode of small organic molecules and biological molecules*
Xianlong Wang, University of Electronic Science and Technology of China

Dealing with Biological Complexity

- P - 24 : *Multi-targeting in Alzheimer's disease: from in silico design of secretase inhibitors to in-vivo experiments*
Gloria Boursheh, Hebrew University
- P - 26 : *New structural alerts for phospholipidosis*
Lilia Fisk, Lhasa Limited
- P - 28 : *To hit or not to hit - that is the question! Structure-Based druggability predictions for Pseudomonas aeruginosa targets*
Aurijit Sarkar, University of Dundee

Integrated Chemical Information

- P - 30 : *Data integration and analysis – how a scientist would like it*
Paul Davis, Schrödinger
- P - 32 : *Harmonization of cheminformatics services after a recent acquisition: taking the best of two worlds*
Lars Ridder, Merck Research Laboratories

Analysis of Large Chemistry Spaces

- P - 34 : *Visual characterization and diversity quantification of chemical libraries: Creation and use of Delimited Reference Chemical Subspaces (DRCS)*
Lionel Colliandre, ICOA
- P - 36 : *Dataset overlap density analysis*
Andreas H. Goeller, Bayer Schering Pharma AG
- P - 38 : *Open source chemical structure generator*
Julio E Peironcely, Leiden University
- P - 40 : *NCI/CADD chemical identifier resolver: indexing and analysis of available chemistry space*
Markus Sitzmann, National Cancer Institute, National Institutes of Health
- P - 42 : *Handing of homology variation in structure representation, patent Markush search, enumeration and visualization*
Robert Wagner, Chemaxon Kft

Structure-Activity and Structure-Property Prediction

- P - 44 : *Boosting the predictive reliability of QSAR models*
Eugene Muratov, University of North Carolina at Chapel Hill
- P - 46 : *Fast methods of atomic charge calculation: the electronegativity equalization method for proteins*
Crina-Maria Ionescu, Masaryk University
- P - 48 : *Hole filling in medicinal chemistry libraries*
Vincent le Guilloux, ICOA
- P - 50 : *An open access hierarchy of molecular descriptors*
Jörg Marusczyk, Molecular Networks GmbH
- P - 52 : *Large-scale in silico model building*
Ulf Norinder, AstraZeneca R&D
- P - 54 : *3D-neighbourhood protein descriptors for proteochemometric modeling*
Remco F Swier, Leiden University, LACDR
- P - 56 : *Predictive data mining for the identification of CYP P450 isoform-specific sites of metabolism*
Alexios Koutsoukas, University of Cambridge
- P - 58 : *Structural and energetic aspects of oxazolidinone antibiotics binding to the ribosomal structure*
Jagmohan Saini, Heinrich-Heine-University Düsseldorf

Structure-Based Drug Design and Virtual Screening

- P - 60 : *An investigation of in silico methods to accurately simulate biologically-relevant conformations of drug-like macrocyclic molecules*
Nathan Brown, The Institute of Cancer Research
- P - 62 : *Exploring DNA topoisomerase I ligand space in search of novel anticancer agents*
Renate Griffith, University of New South Wales
- P - 64 : *Improved docking accuracy using 3D restraints derived from X-ray crystallography*
Richard J Hall, Astex Therapeutics

- P - 66 : *In silico identification of resistance-breaking inhibitors of influenza neuraminidase*
Johannes Kirchmair, University of Cambridge
- P - 68 : *Replacing structure-based pharmacophore filtering: high-throughput docking using constraints leads to TrxR inhibitors with high bioactivity on M. Tuberculosis*
Oliver Koch, Intervet Innovation GmbH
- P - 70 : *LoFT: design of combinatorial libraries for the exploration of virtual hits from fragment space searches*
Uta Lessel, Boehringer Ingelheim Pharma GmbH & Co. KG
- P - 72 : *PDB Ligands: high-level conformational energy calculations*
Marc Nicklaus, National Cancer Institute, NIH
- P - 74 : *In silico discovery of new types of allosteric inhibitors against bcr-abl pharmacophore models*
Nicole S Kast, Hebrew University
- P - 76 : *Comparison of various strategies in pharmacophore model generation – application to 5-HT_{1A} receptor ligands*
Dawid Warszycki, Institute of Pharmacology Polish Academy of Sciences
- P - 78 : *Fighting molecular obesity with sub-pharmacophore screening*
Modest von Korff, Actelion Ltd

Poster Session

BLUE

Cheminformatics

- P - 1 : *Continuous functions as a universal way of representing chemical structures in chemoinformatics*
Igor Baskin, Moscow State University
- P - 3 : *A treatment of stereochemistry in computer-aided synthesis planning*
Anthony Peter Cook, University of Leeds
- P - 5 : *Analysis of activity datasets using spectral clustering*
Sonny Gan, University of Sheffield
- P - 7 : *Fragment based de novo design and ADME/T analysis of dual binding site Acetylcholinesterase Inhibitors for Alzheimer's disease*
Shikhar Gupta, NIPER
- P - 9 : *Chemoinformatics approaches for identification of porous materials for industrial applications*
Maciej Haranczyk, Lawrence Berkeley National Laboratory
- P - 11 : *The file IO round robin game: on the development of a consistent chemical representation*
Adrian Kolodzik, University of Hamburg
- P - 13 : *Building an R&D chemical registration system*
Elyette Martin, Philip Morris Products S.A.
- P - 15 : *Optimizing metric properties of protein structure descriptors*
Peter Roegen, Technical University of Denmark
- P - 17 : *Visual chemical patterns: from automated depiction to interactive design*
Karen Schomburg, University of Hamburg
- P - 19 : Withdrawn
- P - 21 : *Condensing chemical reactions to pseudo-molecules: an efficient way of reaction mining*
Alexandre Varnek, University of Strasbourg
- P - 23 : *Similarity based virtual screening using frequency-based weighting-schemes: effect of the choice of similarity coefficient*
Hua Xiang, University of Sheffield
- P - 25 : *Reaction enumeration and machine learning enhancements for the open-source pipelining solution CDK-Taverna 2.0*
Achim Zielesny, University of Applied Sciences of Gelsenkirchen

Dealing with Biological Complexity

- P - 27 : *Models of substrates and inhibitors of P450 isoenzymes*
Maayan Elias, Hebrew University
- P - 29 : Withdrawn
- P - 31 : *Identification of alternative druggable targets involved in the prion disease*
Jorge M Valencia, University of Sheffield

Integrated Chemical Information

- P - 33 : *DEGAS – sharing and tracking target compound ideas with external collaborators*
Man-Ling Lee, Genentech, Inc.
- P - 35 : *Analysis of contents in proprietary and public bioactivity databases*
Pekka Tiikkainen, Merz Pharmaceuticals GmbH

Analysis of Large Chemistry Spaces

- P - 37 : *Mining chemical IP with OSRA*
Igor Filippov, SAIC Frederick
- P - 39 : *Efficient sparse and probabilistic binary classifier*
Robert Lowe, University of Cambridge
- P - 41 : *Improved chemical text mining of patents using automatic spelling correction and infinite dictionaries*
Roger A Sayle, NextMove Software
- P - 43 : *Automated extraction of chemical information from documents: recent advances in the CLiDE project*
Aniko T Valko, Keymodule Ltd.

Structure-Activity and Structure-Property Prediction

- P - 45 : *Quantifying model errors using similarity to training data*
Robert D Brown, Accelrys Inc
- P - 47 : *Proteochemometric modeling as a tool to predict clinical response to anti-retroviral therapy based on the dominant patient HIV genotype*
Alwin Hendriks, Leiden University / LACDR
- P - 49 : *Prediction of biological activities using semi-supervised and transductive machine learning methods*
Evgeny Kondratovich, University of Strasbourg
- P - 51 : *A global Class A GPCR proteochemometric model: a prospective validation*
E Bart Lenselink, Leiden University / LACDR
- P - 53 : *Ames toxicity: strategies to remove ames liability for anilines*
Jordi Munoz-Muriedas, Glaxo Smithkline
- P - 55 : *An approach toward the prediction of chemical degradation pathways*
Martin A Ott, Lhasa Limited
- P - 57 : *Turning 3D-QSAR weakness into strength with Open3DALIGN & Open3D*
Paolo Tosco, University of Turin
- P - 59 : *Effectiveness of fingerprint-based measures of multi-stage mass spectrometry similarity for virtual screening of chemical structures*
Miguel Rojas-Cherto, Leiden University

Structure-Based Drug Design and Virtual Screening

- P - 61 : *Homology modeling and binding site prediction of human IRAK-M*
Jiangfeng Du, Maastricht University
- P - 63 : *Key features for designing PPARgamma agonists: an analysis of ligand-receptor interaction by using a 3D-QSAR approach*
Laura Guasch, Universitat Rovira i Virgili

- P - 65 : *Understanding ligand binding affinity and specificity through analysis of hydration site thermodynamics*
David Rinaldo, Schrödinger
- P - 67 : *Insecticide and fungicide likeness: use of two-class Bayesian categorization models for the selection of molecules as screening inputs*
Carla J Klittich, Dow AgroSciences
- P - 69 : *The multi-conformations-receptor-based pharmacophore model generation schema and its potential applications in virtual screening*
Rafał Kurczab, Institute of Pharmacology Polish Academy of Sciences
- P - 71 : *ChemTattoo3D: an open source drug design analysis and visualization tool*
Norah MacCuish, Mesa Analytics & Computing, INC
- P - 73 : *Targeting of the tenase complex by rational design of factor viii-membrane interaction inhibitors*
Gerry A F Nicolaes, Maastricht University
- P - 75 : *A fragment—molecule alignment algorithm based on spherical Gaussians*
Nikolaus Stiefl, Novartis Institute for Biomedical Research
- P - 77 : *Virtual Ligand screen of human TRAF6: towards iPPI's as potential anti-inflammatory pharmaca*
Barbara Zarzycka, Cardiovascular Research Institute Maastricht
- P - 79 : *Ligand- and structure-based molecular invention using Benchware Muse*
Fabian Bös, Tripos International

Plenary Session Abstracts

K-1 : The intermediary reloaded - on the need for a "Go-Between" to information users and producers

E. Zass

Chemistry Biology Pharmacy Information Center, ETH Zürich, Switzerland

In the early times of chemical databases, intermediaries were indispensable to do the searches for chemists ("end-users"), due both to the complexity of the command-driven interfaces then used, and to the cost structure of these databases. With upcoming graphical user interfaces (GUIs), and flat-rate licensing models, such intermediaries seemed to be obsolete; at most, training and support for end-users was discussed as a minor remaining function for them.

This rather simplistic assessment, however, only considered the user interface, and ignored the complexity of the content, and the variations in the data structures and indexing policies of the underlying databases. This remained virtually unchanged, or even became more complex. While indeed most chemists nowadays do most of their searches themselves, they cannot do it all on their own. Major functions of intermediaries are thus not only training and support, but particularly analyses of search problems versus content and search facilities of databases, and communication of such analyses both to users for better searching, and to database producers to improve databases and interfaces.

A-1 : Beyond filters: assessing ADMET risk for multi-objective drug development

R. Clark, J. Zhang, R. Fraczkiwicz, M. Waldman, M. Bolger, W. Woltosz

Simulations Plus, Inc., Lancaster CA, USA

We have examined the distributions of predicted values for over 30 properties relevant to absorption, distribution, metabolism, excretion and toxicology (ADMET) across a large ($N = 2316$) and pharmaceutically pertinent subset of the World Drug Index (WDI), and have identified relevant 5% and 10% cutoff thresholds for each of those properties. These thresholds, combined with classification models for over 40 metabolic and toxicological liabilities, enable us to calculate an aggregate risk score (ADMET Risk™) for any molecule. This risk score is analogous to the "Rule of 5" formulated by Lipinski et al.² in that points are added for any prediction that violates a threshold value or is classified as problematic, but it incorporates many more potential liabilities. In addition, it is augmented by a list of mnemonic codes (ADMET Code™) that identifies each "rule" violated by a compound.

The subset of drugs we created was extracted using annotation filters similar to those used by Lipinski et al., but the WDI version we used was more current (2007). Besides identifying the thresholds themselves, we have also explored how the aggregate risk score is affected by applicability domain constraints and by the interdependencies^{2,3} found among estimated properties.

Taken together, the ADMET Risk value and its associated ADMET Code list provide a compact yet highly multidimensional view of the potential problems represented by any particular candidate structure. We believe this is a powerful tool to guide early steps in the multiobjective optimization process required to produce a successful drug or other biologically active molecule.

1. World Drug Index 2007/04; Thomson Reuters; http://thomsonreuters.com/products_services/science/science_products/a-z/world_drug_index/
2. Lipinski, C.A.; Lombardo, F.; Dominy, B.W.; Feeney, P.J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* 1997, 23, 3-25.
3. Oashi, T.; Ringer, A.L.; Raman, E.P.; MacKerrell, A.D, Jr. Automated selection of compounds with physicochemical properties to maximize bioavailability and druglikeness. *J. Chem. Inf. Model.* 2011, 51, 148-158.

A-2 : SeeSAR: see your structure-activity relationships

J. Lommerse¹, M. Wagener¹, S. Heisterkamp²

¹*Molecular Design & Informatics, Merck Research Laboratories, Oss, The Netherlands*

²*Biostatistics and Research Decision Sciences, Merck Research Laboratories, Oss, The Netherlands*

Pharmaceutical discovery programs, in which compounds are optimized towards pre-defined criteria, are often long-term. As such programs continue for several years, researchers working on the program may change or results which have been obtained in the past are no longer taken fully into account. However, for optimum project progress, an overview of the important aspects of all project compounds at any time is of utmost importance for all scientists involved.

For this, Free-Wilson analysis¹ can be a suitable solution, as it is an insightful method to quickly obtain an overview of a project which is focused on a particular chemical series. The analysis is based on the assumption that core and substituents of a compound series contribute in an additive manner to the total activity of the compounds. Using standard linear regression techniques as applied in the Free-Wilson analysis, the contribution of each substituent can be determined. Organizing a dataset by these contributions leads to a Structure-Activity Relationships (SAR) model that summarizes the effect of substituents at various positions of the core structure. The method links very well to concepts popular in medicinal chemistry which describe datasets in terms of better and worse performing substituents at certain positions of a scaffold.

We have implemented the Free-Wilson analysis methodology in an in-house intranet application, called SeeSAR. It automatically generates the SAR of congeneric series of compounds and suggests novel combinations of core and substituents with potentially superior activities or properties. The use of the system will be illustrated using several literature and in-house examples.

Three special aspects of the SeeSAR implementation will further be discussed: First, to improve performance for small data sets as typically found in Lead Finding and early Lead Optimization programs Ridge Regression is used in place of the standard linear regression. This approach is validated using leave-one-out (LOO) predictions for 10 datasets extracted from the ChEMBL² database. The results indicate that Ridge Regression improves predictions for novel combinations of substituents (Fig. 1).

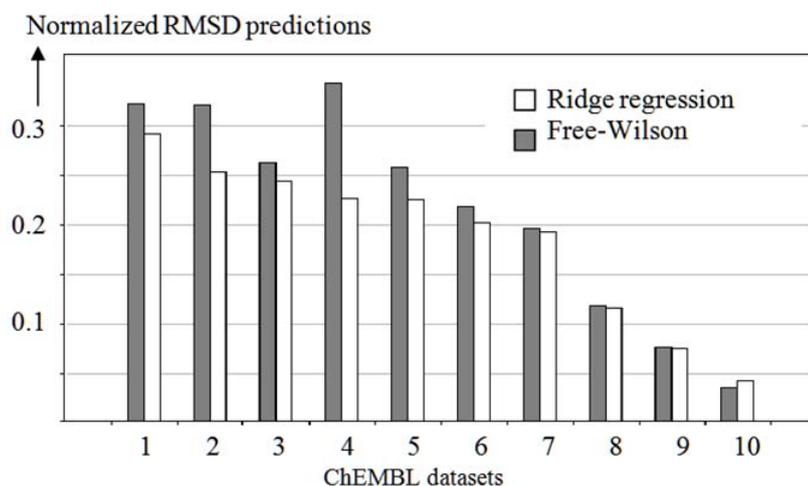


Figure 1: Comparison of Leave-One-Out predictions using Free-Wilson and Ridge Regression analysis for 10 different data sets extracted from ChEMBL.

The second aspect concerns the possibility to use Free-Wilson analysis for multiple related assay results simultaneously. Examples include the use of biochemical assay results to strengthen the Free-Wilson analysis model for (limited) cellular assay results. This approach has been evaluated for several projects targeting kinases, GPCR's and nuclear receptors.

Finally, a censoring algorithm has been implemented to also include in the analysis those compounds that are not active enough for an exact determination of assay results. Often, these compounds are neglected in QSAR analyses, although they carry valuable information and can contribute to the model.

- Free, S. M.; Wilson, J. W. A Mathematical Contribution to Structure-Activity Studies. *J. Med. Chem.* **1964**, 7, 395-399.
- Bender, A. Databases: Compound bioactivities go public. *Nature Chem. Biol.* **2010**, 6, 309-309.

A-3 : Real world application of proteochemometric modeling for the design of enzyme inhibitors and ligands of G-protein coupled receptors

G. van Westen¹, O. van den Hoven¹, J. Wegner², A. IJzerman¹, H. van Vlijmen^{1,2}, A. Bender^{1,3}

¹Division of Medicinal Chemistry, LACDR, Leiden, The Netherlands

²Tibotec BVBA, Beerse, Belgium

³Unilever Centre for Molecular Science Informatics, Cambridge, United Kingdom

The early phases of drug discovery often employ *in silico* models to rationalize structure activity relationships and to predict the activity of novel compounds. However, the predictive performance of these models is not always acceptable and the reliability of prospective predictions – both to novel compounds and to related protein targets, where possible – is in many cases limited.

Proteochemometric modeling¹ (PCM) attempts to remedy this situation by adding a target description, based on physicochemical properties of the binding site, to conventional, ligand-based bioactivity models. Our approach of PCM² is based on Scitegic circular fingerprints on the compound side and on a customized feature based protein fingerprint on the target side which is based on a selection of physicochemical descriptors obtained from the AAindex database.

In this work we then perform a large-scale validation of a PCM on 14 HIV reverse transcriptase (HIV RT) mutants, comprising 317 prospective data points (set A). In addition we present a smaller scale prospective validation on the four Adenosine receptor subtypes (set B).

Set A consisted of 451 non-nucleoside HIV RT inhibitors and 14 HIV RT mutants where a pEC₅₀ value was available for about 60 % of the ligand-mutant pairs. With our model we predicted pEC₅₀ values for the missing 40% of the pairs. Subsequently 317 data points were prospectively validated. We show that our model approaches the experimental limitations of the assay (Model RMSE was 0.62 log units, assay error was 0.50 log units) (Figure 1).

Set B consisted of 11,000 compounds and eight Adenosine receptors, four each from rat and human. Our PCM is able to make productive use of the added 4 rat receptors (RMSE 0.876 for human only and 0.817 for human and rat). The final virtual screen identified six novel high affinity compounds out of 55 compounds tested, that were tested to be active on one or more of the human Adenosine receptors.

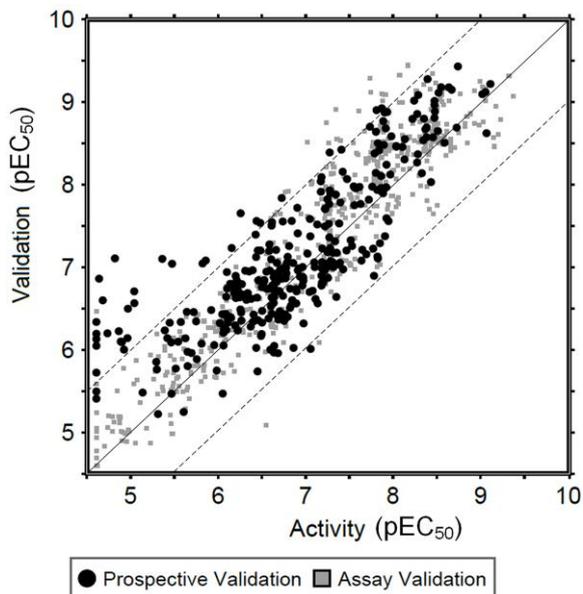


Figure 1: Prospective validation performance.

We conclude that PCM is able to reliably extrapolate the activity of compounds to new targets, on the datasets employed here and within the limitations of the training data provided. This ability makes it a useful tool to predict the activity of ligands on highly related proteins, such as in case of enzyme mutants or designing multi-target drugs.

- Freyhult, E.; Prusis, P.; Lapinsh, M.; Wikberg, J. E.; Moulton, V.; Gustafsson, M. G., Unbiased descriptor and parameter selection confirms the potential of proteochemometric modeling. *BMC Bioinformatics* **2005**, 6, 50-64.

- van Westen, G.J.P.; Wegner, J.K.; IJzerman, A.P.; van Vlijmen, H.W.T.; Bender, A., Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets. *Med. Chem. Commun.*, **2011**, 2, 16-30

A-4 : QSAR Workbench: enabling best practices in large scale model building

C. N. Luscombe¹, N. Malcolm², R. Cox², S. D. Pickett¹

¹ *GlaxoSmithKline, Stevenage, UK*

² *Accelrys, Cambridge, UK*

Development of local and global QSAR models remains a key requirement for many drug discovery programs; however increasing pressure on resources means that experts in specific statistical tools are not always available. At GSK we have had a long interest in the development and application of automated QSAR modeling methods to drug design^{1,2} Our experiences with previous in-house systems has guided us in the development of the QSAR Workbench, a collaborative project between GlaxoSmithKline and Accelrys that enables specialists to record and publish best practices through an easy to use graphical interface. The recorded workflows can be replayed by both QSAR expert and non-QSAR expert against new datasets. The workbench can be used to explore the massive potential model space defined through the combinatorial use of different statistical tools, descriptors-sets and training/test-set splits. The use of Pipeline Pilot to build the backend framework means that the system is easily extensible in terms of available statistical tools, descriptors and model analysis methods.

To explore the scientific validity of a semi-automated QSAR model building and testing approach, we have applied the QSAR Workbench to the five toxicological end-points chosen for the CAESAR project³: bioconcentration factor, skin sensitization, carcinogenicity, mutagenicity, developmental toxicity. These end-points were chosen because of their high relevance for REACH legislation. The automated model building steps available within the QSAR Workbench means that once the specialist in each method has recorded relevant workflows, then these best-practices can be re-applied by any user. In this way we have been able to build a very large pool of QSAR models for each end-point in a robust and consistent manner. Selection of the optimal model is a semi-automated process; statistical metrics, coupled with graphical analysis tools enable rapid identification of best models as well as potential outlier models that may be due to some inherent bias. The final models for the five end-points studied here are of equivalent quality to those produced by the CAESAR project. However the man-hours required to create the models with the QSAR Workbench is significantly lower.

- Stephen Pickett, EuroQSAR 2006
- Darren Green, <http://www.soci.org/News/Fine-Chemoinformatics-SAR> (2008)
- Benfenati, E. The CAESAR project for in silico models for the REACH legislation. *Chemistry Central Journal* 2010, 4(Suppl 1):I1.

A-5 : Using AutoQSAR to select the most predictive modeling methods

S. Rodgers¹, A. Davis², F. Brown³
¹AstraZeneca, Loughborough, UK
²AstraZeneca R&D, Loughborough, UK
³Accelrys, San Diego, USA

Predictive chemistry is a key component in the drug discovery process: having effective, up to date QSAR (quantitative structure-activity relationship) models helps chemists to identify those compounds with the desired pharmacological and physicochemical profiles. This reduces the number of compounds that need to be synthesized and hence makes the process more efficient.

There is an abundance of statistical methods and descriptors for generating predictive models. AstraZeneca's AutoQSAR system has automated the QSAR model generation process, allowing simple and fast exploration of model space. The system will build both global and local models using a range of statistical methods and descriptors. AutoQSAR's competitive workflow selects the most predictive model by assessing the ability of each to predict a temporal test set. The selected model is made available to chemists straight away, so that it can have an immediate impact.

We will describe the AutoQSAR system, focusing on the different statistical techniques that have been integrated. We will discuss the competitive workflow and how it identifies the most predictive methods (and demonstrate whether this is the case in simulation). We will present a full comparison of many statistical learning methods, and their suitability to different types of data.

A-6 : What should we make next?

M. Mackey¹, T. Cheeseright¹, J. Melville¹, R. Scoffin¹, C. Earnshaw²
¹Cresset BMD, Welwyn Garden City, UK
²Cambridge, UK

The process of lead optimization can be largely condensed into a single question: what compounds should we make next? The answer to this question is rarely obvious, and involves considering many aspects: synthesizability, patentability, activity, physicochemical properties, metabolism, pharmacokinetics and so forth. All of these are difficult to model, and true predictivity is rare, so that effective lead optimization still has considerable art involved.

The saving grace is that LO usually takes place in a series context, where rather than consider the whole 10^{60} space of chemical possibilities we are usually considering a relatively limited set of possible variations on our existing compounds. In this context, qualitative feedback from models can be more useful than attempts at quantitative prediction. Guidance on what parts of your molecule to change, and in which direction, can be invaluable in supporting decisions made by a compound design team.

In this talk we present a system for developing local 3D QSAR models, including information from the protein environment if available. The models are built based on molecular fields¹, using a novel sampling methodology. Although the models are often predictive in their domain of applicability, their primary use is to provide feedback to the modeler or chemist: why is this compound predicted to be active while that one is not? As well as providing this feedback in 3D, where conformational, steric and electrostatic effects can be examined directly, the results can be mapped to a 2D representation. This allows the chemist to design new compounds in familiar 2D space, while benefitting from a 3D model.

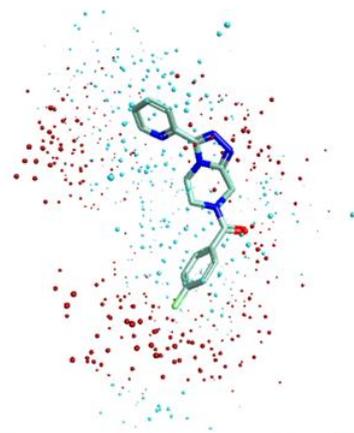


Figure 1 Electrostatic coefficients for an NK3 model

The results of this technique are discussed using not only existing literature data sets but also several series of NK3 inhibitors (Fig 1). The results on these demonstrate the benefits and pitfalls of building local vs global 3D QSAR models. The extension of this QSAR method to ADMET modeling is also explored.

1. Cheeseright, T.; Mackey, M.; Rose, S.; Vinter, A. Molecular Field Extrema as Descriptors of Biological Activity: Definition and Validation. *J. Chem. Inf. Model.* **2006**, 46, 665-676

A-7 : DrugscoreMaps – A graphical method for SAR analysis, active/inactive separation, binding mode analysis and docking pose evaluation using protein-ligand interaction fingerprints

O. Koch^{1,2}, G. Neudert³, G. Klebe³

¹ *Intervet Innovation GmbH, Schwabenheim, Germany and MOLISA GmbH, Magdeburg, Germany*

² *The Cambridge Crystallographic Data Centre, Cambridge, UK*

³ *Philipps-Universität Marburg, Marburg, Germany*

Protein-Ligand interaction fingerprints (PLIFs) are a smart way to determine similarities between protein-ligand binding modes. In combination with powerful cluster algorithms, PLIFs offer a wide spectrum of applications in drug design like SAR analysis, data mining in virtual screening, actives/inactives separation, binding mode analysis and docking pose evaluation. Here, we present DrugscoreMaps¹, a combination of Drugscore Fingerprints² and emergent self-organizing maps³. DrugscoreMaps benefits by straightforward visualization and an implicit weighting of the protein-ligand interactions.

DrugscoreFP² is the vectorial representation of atom-atom contact scores for protein-ligand interactions based on Drugscore^{CSD} pair potentials⁴. Thus, DrugscoreFP includes an implicit weighting of these interactions as in the Drugscore^{CSD} scoring function. Emergent self-organizing maps (ESOMs³) include thousands of neurons. Clusters are identified within a 3D landscape of the trained map using the U-matrix visualization. This leads to an intuitive way of visualizing the similarity in protein-ligand interactions as a 3D landscape of valleys and mountains.

In cases where activity-data are available, DrugscoreMaps can give a fast overview about the structure-activity landscape including protein-ligand interaction data. Figure 1 shows a successful application using a QSAR dataset for carbonic anhydrase II⁵. Clearly separated clusters can be identified that exhibit similar pK_i ranges. Using examples from the DUD dataset of useful decoys⁶, we will show that DrugscoreMaps can also be used to separate actives from inactives in a virtual screening campaign. Furthermore, this approach can also help to identify compounds with similar binding modes within large high-throughput screening results.

DrugscoreMaps is also useful for the evaluation of docking poses of single protein-ligand complexes. An extensive survey of the Astex Diverse Dataset⁷ exhibits that this approach also supports the search for the correct low energy binding mode in cases where no complex information from X-ray structures is available. High-ranked poses represent reasonable binding geometries if clearly separated clusters of high-ranked poses occur. If high-ranked poses are not observed within a cluster of high-ranked solutions, it is likely that these solutions should be neglected. Finally, the docking process needs further refinements when no cluster of high-ranked docking poses appears.

We will discuss the mentioned examples and applications in detail and will highlight the usefulness of clustering protein-ligand interaction fingerprints in the structure-based drug design process.

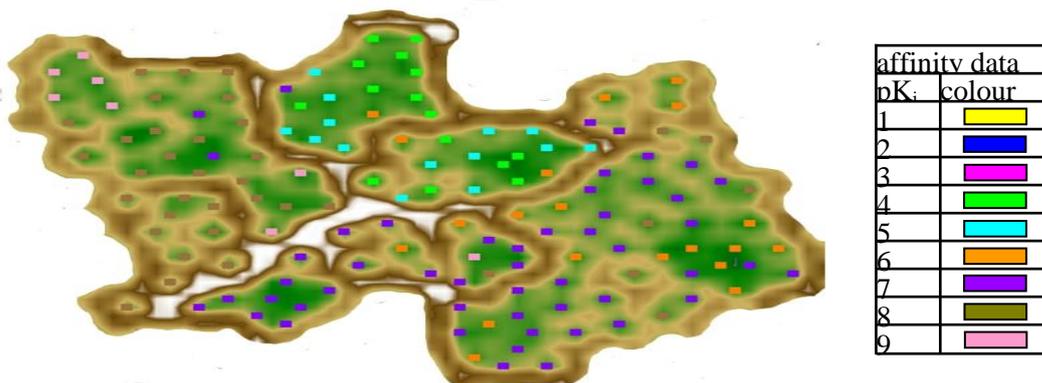


Figure 1: Drugscore^{Map} for carbonic anhydrase II QSAR dataset with cluster containing molecules with similar pK_i ranges [points: molecule, color: pK_i; cluster: within a “valley” surrounded by “mountains”]

1. Koch, O., Neudert, G., Klebe, G. *Chemistry Central Journal*. **2009**, 3(1): P61.
2. Pfeffer, P., Neudert, G., Klebe, G. *Chemistry Central Journal*. **2008**, 2(1): 16.
3. Ultsch, A. Proceedings of Workshop on Self-Organizing Maps; Kyushu, Japan **2003**: 225-230. [ESOM-Tools: <http://databionics-esom.sourceforge.net/>]
4. Velec, H.F.G., Gohlke, H., Klebe, G. *J. Med. Chem.* **2005**, 48(20): 6296-6303.
5. Hillebrecht, A., Supuran, C.T., Klebe, G. *Chem. Med. Chem.* **2006**, 1(8): 839-853.
6. Huang, N., Shoichet, B. and Irwin, J. J. *J. Med. Chem.* **2006**, 49(23), 6789-6801
7. Hartshorn, M.J., Verdonk, M.L., Chessari, G., Brewerton, S.C., Mooij, W.T., Mortenson, P.N., Murray, C.W. *J. Med. Chem.* **2007**, 50(4): 726-741.

A-8 : Mining of emerging structural patterns for identification of toxicophores

R. Sherhod¹, V. J. Gillet¹, P. Judson², J. Vessey²
¹ Information School, University of Sheffield, Sheffield, UK
² Lhasa Limited, Leeds, UK

The collection of structural features that result in toxicological activity is a time consuming task, even for toxicology and chemistry experts. In order to assist this process an automated methodology has been developed to aid toxicophore identification by mining descriptions of activating structural features directly from toxicity datasets. Such structural features may then be used to define new structural alerts, which are the basis of many knowledge-based toxicity prediction techniques.

The method is based on emerging pattern mining¹, a technique that is well known to computer science, but is relatively new to chemistry². For any data that can be expressed as a series of binary properties, emerging pattern mining can be used to extract patterns of those properties that occur more frequently in one dataset compared to another. For toxicological knowledge extraction, mining emerging patterns from structural fingerprints provides a means for generating patterns of structural features that distinguish toxicants from innocuous compounds. The current focus of the project is the mining of *jumping*-emerging patterns, which are emerging patterns that are exclusive to the toxicologically active class and therefore represent structural features that are only present in toxicants.

The Horizon-Miner algorithm³ and border-differential operation⁴ are applied to generate the minimal and maximal borders of a set of jumping-emerging patterns of structural descriptors. Using the minimal jumping-emerging patterns it is possible to cluster toxic compounds into groups defined by the presence of shared structural features that occur exclusively in the actives. From these clusters it is then possible to derive larger and more complete descriptions of distinguishing structural features that will be recognisable to toxicologists. A method has been developed to identify hierarchical relationships between clusters and

their associated jumping-emerging patterns, which has enabled families of structural feature descriptions to be arranged into trees. The root of each tree represents the most general and most commonly occurring structural feature description in the family. By inspecting clusters further down the tree, it is possible to identify variations in the significant structural features that further distinguish sets of toxic compounds. This may provide additional structural detail that is relevant to the toxicological endpoint, and useful for compiling a new alert.

The methodology has been tested on a number of datasets for various toxicity endpoints, including Ames mutagenicity, oestrogenicity and hERG channel inhibition. These tests have shown the method to be effective at clustering the datasets around minimal jumping-emerging structural patterns and finding larger descriptions of the significant structural features. The resulting descriptions of the significant structural features have been shown to be related to some of the known alerts for the tested endpoints.

1. Dong, G.; Li, J. In *Efficient mining of emerging patterns: discovering trends and differences*, The Fifth International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, 1999; Association for Computing Machinery Press: San Diego, CA, USA, 1999; pp 43-52.
2. Auer, J.; Bajorath, J. r., Emerging chemical patterns: a new methodology for molecular classification and compound Selection. *Journal of Chemical Information and Modeling* **2006**, 46, (6), 2502-2514.
3. Li, J.; Dong, G.; Ramamohanarao, K., Making use of the most expressive jumping emerging patterns for classification. *Knowledge and Information Systems* **2001**, 3, (2), 131-145.
4. Dong, G.; Li, J., Mining border descriptions of emerging patterns from dataset pairs. *Knowledge and Information Systems* **2005**, 8, (2), 178-202.

A-9 : Which substructures are interesting?

J. Kazius

Curios-IT, Leiden, the Netherlands

Substructure mining is a graph mining technique that performs substructures searches until all substructures are found that meet predefined criteria ¹. The resulting substructures can serve as intuitive precursors to chemical insights ², or even as predictors for mutagenicity and receptor binding ^{1,3}.

A substructure mining algorithm was built to search for the most discriminative (most interesting) substructures. The approach considers the chemistries of both actives and inactives during the data-driven extraction of structure-activity relationships (SARs). This presentation discusses the substructure mining results from affinity datasets of ~200 distinct biological targets.

Single, small substructures emerged to be unexpectedly strong descriptors of target affinity. Some activities are almost perfectly described with only one small fragment, which suggests either the importance of the fragment or a dataset bias. Moreover, inactivating fragments (low affinity) were often found to be as descriptive as activating fragments (high affinity).

The surprising strength of such simple models raises the question: do 'more sophisticated' substructures provide added predictive value? In this light, more complex types of substructures (namely SMARTS) were mined. For one, the consideration of substructures of any size and shape hardly increases the descriptive value of the resulting fragments. But including SMARTS-type chemical details in the search boosts the potential of the resulting substructures as affinity descriptors. In short, this presentation aims to link changes in affinity thresholds, fragment type and substructure size to SAR interpretation and predictive power.

1. 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.
2. Kazius, J.; McGuire, R.; Bursi, R. Derivation and Validation of Toxicophores for Mutagenicity Prediction. *J. Med. Chem.* **2005**, 48, 312-320.

- van der Horst, E.; Okuno, Y.; Bender, A.; IJzerman, A. P. Substructure Mining of GPCR Ligands Reveals Activity-Class Specific Functional Groups in an Unbiased Manner. *J. Chem. Inf. Model.* **2009**, *49*, 348-360.

B-1 : Global free energy scoring functions based on distance-dependent atom-type pairs

C. Kramer, P. Gedeck

Novartis Institutes for BioMedical Research, Computer-Aided Drug Design, Basel, Switzerland

The prediction of the absolute Free Energy of Binding is still an unsolved problem in computer-aided drug design. Despite the age of scoring function research, state-of-the-art scoring functions are mainly able to predict the correct binding geometry and in the best case give very coarse indications about the free energy of binding. Since a couple of years, large databases of protein-ligand crystals complemented with binding data are being assembled that can provide the basis for next generation scoring functions.

We have recently introduced a set of descriptors for protein-ligand interactions based on distance-dependent counts of atom type pairs.¹ Using these descriptors and classic QSAR techniques, we were able to generate free energy scoring functions that are able to predict free energies of binding with R^2 's of roughly 0.5 on very large out-of-bag validation sets of the PDBbind and the CSARdock data sets. We find that the major contributions come from descriptors representing very simply defined atom pairs weighted by their buriedness. In the CSARdock validation exercise, we were able to get a model that is based on three simple descriptors only and has an out-of-bag performance of $R^2 = 0.55$.² This is much higher and also much simpler than standard scoring functions.

The fact, that very simple scoring functions can perform much better than standard scoring functions, raises a couple of questions concerning future development of scoring functions that will be addressed in the talk. It will further be shown how these results translate into the area of structure-based ligand efficiency.

- Kramer, C.; Gedeck, P. Global Free Energy Scoring Functions based on distance-dependent Atom-type Pair Descriptors. *J. Chem. Inf. Model.* accepted
- Kramer, C.; Gedeck, P. A 3 descriptor model can predict 55% of the CSAR-NRC HiQ benchmark dataset. *J. Chem. Inf. Model.* submitted

B-2 : Development and validation of an in silico scaffold hopping protocol for identifying novel kinase inhibitors

S. R. Langdon^{1,2}, N. Brown², J. Blagg²

¹in silico Medicinal Chemistry and ²Medicinal Chemistry Team 1, Cancer Research UK Cancer Therapeutics Unit, The Institute of Cancer Research, 15 Cotswold Road, Sutton, Surrey, SM2 5NG, UK

Kinases have become one of the most pursued classes of drug target in recent years. As a result of this kinase inhibitor chemical space has been extensively mined leading to a crowding in intellectual property space. Scaffold Hopping is a technique used to identify compounds with similar activity to a known bioactive compound, but with a novel structure.¹ An *in silico* Scaffold Hopping Protocol has been developed to identify novel kinase inhibitors using a known kinase inhibitor as a probe.

An initial scaffold hopping protocol was developed and validated retrospectively. The protocol underwent several different implementations in an attempt to improve performance and its suitability for a prospective validation. The final implementation of the protocol represents each compound of a small molecule library with a scaffold from the Scaffold Tree.² This library is interrogated by a probe scaffold derived from a known well characterised active. Scaffolds, which mimic the probe, are selected using a variety of criteria. A diverse selection of compounds represented by the selected scaffolds is then selected for screening (Figure 1).

A retrospective validation of the protocol was carried out using a known kinase inhibitor from an inhouse project to select a set of compounds. An HTS assay has already been conducted against the kinase of interest, so we could obtain the hit rate for this selection and compare it to the hit rate of the original assay. The protocol achieved a hit rate for active scaffolds 2-fold higher than the original assay, indicating that it is useful for front loading a screening campaign.

For a prospective validation of the protocol a control experiment was designed. The control experiment uses commonly used virtual screening techniques to select compounds for screening (Figure 1). A retrospective validation of the scaffold hopping protocol versus the control shows that our scaffold hopping protocol selects more actives with unique scaffolds than the control.

Compounds have also been selected from our vendor collection using both the scaffold hopping protocol and control using the known kinase inhibitor as a probe. These compounds will be screened against the kinase of interest to assess how effective the scaffold hopping protocol is at finding actives with novel scaffolds in comparison to the control experiment. The results of this prospective validation will also be presented.

1. Langdon, S. R.; Ertl, P.; Brown, N., Bioisosteric Replacement and Scaffold Hopping in Lead Generation and Optimization, *Molecular Informatics* **2010**, *29*, 366–385.
2. Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzol, S.; Koch, M. A.; Waldmann, H., The Scaffold Tree – Visualization of the Scaffold Universe by Hierarchical Scaffold Classification, *Journal of Chemical Information and Modeling* **2007**, *47*, 47-58

B-3 : Snooker: target-focused library selection using structure based pharmacophores

M. Sanders¹, S. Verhoeven², C. de Graaf³, L. Roumen³, S. Nabuurs¹, B. Vroling⁴, J. de Vlieg^{1,2}, J. Klomp²

¹ Computational Drug Discovery Group, CMBI, Radboud University Nijmegen, Nijmegen, The Netherlands

² Merck Research Laboratories, MSD, Oss, The Netherlands

³ Division of Medicinal Chemistry, LACDR, VU University Amsterdam, Amsterdam, The Netherlands

⁴ Modeling and data mining Group, CMBI, Radboud University Nijmegen, Nijmegen, The Netherlands

High throughput screening (HTS) has been the leading technique in small molecule drug discovery for the past 2 decades resulting in lead compounds in ~50% of all screening campaigns. Lately, it has become recognized that target classes with mechanistic and structural information can be better addressed with more rational approaches. Structure based design can guide the hit to lead optimization process in a more informed manner than structure activity relationship exploration around HTS hits.

We present Snooker¹, a low resolution structure based approach to generate pharmacophore hypotheses for class A GPCRs. The method starts with the construction of a homology model of the transmembrane domains, next an ensemble of α -helix specific rotamers is added and residues are prioritized on ligand binding probabilities using multiple sequence alignment statistics². Pharmacophore features are generated by conversion of binding pocket properties to ligand space using protein-ligand interaction geometries and subsequent combination of these interaction properties from highly prioritized residues.

We show that Snooker pharmacophore hypotheses reproduce literature supported binding modes for ~75% of beta-2-adrenergic compounds fulfilling pharmacophore constraints and retrieve enriched sets of active compounds in retrospective screens. We furthermore show that it is possible to select a target focused compound set by using Snooker pharmacophore hypotheses.

For the community wide GPCR dock 2010 assessment Snooker³ was combined with Fleksy³ to predict the eticlopride binding mode in the human DRD3 receptor. Low resolution binding modes were generated using Snooker and optimized by Fleksy. All our 5 submitted eticlopride poses were amongst the top 10 predictions out of ~120 submitted predictions in this competition and our top pose ranked 2nd.

1. Sanders, M.; Verhoeven, S.; de Graaf, C.; Roumen, L.; de Vlieg, J.; Klomp, J. *Snooker: a structure based pharmacophore generation tool applied to class A GPCRs*. (in preparation)
2. Sanders, M.; Fleuren, W.; Verhoeven, S.; van den Beld, S.; Alkema, W.; Klomp, J.; de Vlieg, J. ss-TEA: Entropy based identification of receptor specific ligand binding residues from a multiple sequence alignment of class A GPCRs. (submitted)
3. Nabuurs, S.B., M. Wagener, and J. de Vlieg, A flexible approach to induced fit docking. *J Med Chem*, 2007. 50(26): p. 6507-18.

B-4 : Cavity knowledge acceleration (CavKA) – metamorphosis in automatic pharmacophore elucidation

F. Koelling, K. Baumann

University of Technology, Braunschweig, Germany

Three dimensional pharmacophore models can be considered as an ensemble of steric and electronic features in space, which are necessary to ensure intermolecular interaction with a specific target in order to trigger or to block biological activity¹. By identifying these features, a 3D pharmacophore model can be built in order to screen multi-conformational databases with the aim to detect compounds matching the pharmacophoric hypothesis and subsequently submit them to a biological testing. Even if a 3D crystal structure is at hand, the creation of a reliable pharmacophore model remains a challenging task.

CavKA (Cavity Knowledge Acceleration) is an in-house developed pharmacophore elucidation tool that employs the information of co-crystallised ligand-receptor complexes. Ligand features interacting with the binding site are detected and GRID² force field information is additionally taken into account as to weight and prioritize the identified features in question. The prioritized features are then transformed into a pharmacophore model.

CavKA is compared to LigandScout³ and an alternative approach studied by authors from Schrödinger⁴. LigandScout considers geometrical interaction rules. This is also one part of the rule set of CavKA. To augment the geometric analysis by an energetic term, Salam and co-workers from Schrödinger combine their automatic pharmacophore elucidation method Phase with the Glide XP scoring function. In contrast to this, CavKA employs GRID molecular interaction fields for the same reason.

The performance of CavKA and competitors is evaluated in a retrospective screening on the FieldScreen⁵ dataset outlining strengths, weaknesses and as well as similarities of each method for the scrutinized targets. It turns out that CavKA performs best in many cases. Furthermore, it is shown how alternative search templates affect the results and can be employed to improve the screening performance dramatically.

1. Wermuth C.G.; Ganellin C.R.; Lindberg P.; Mitscher, L.A. Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1998) *Pure Appl. Chem.* **1998**, 70, pp. 1129-1143.
2. GRID 22a, Molecular Discovery, www.moldiscovery.com.
3. Ligandscout 3.0, inte:ligand, www.inteligand.com.
4. Salam N. K; Nuti, R.; Sherman W. Novel method for generating structure-based pharmacophores using energetic analysis. *J. Chem. Inf. Model.* **2009**, 49, 2356-2368.
5. Cheeseright T.J.; Mackey M.D.; Melville J.M.; Vinter J.G. FieldScreen: virtual screening using molecular fields. Application to the DUD data set. *J. Chem. Inf. Model.* 2008, 48, 2108-2117.

B-5 : 3D pharmacophore searching against ten trillion combinatorially accessible compounds

Q. Zhang

Boehringer-Ingelheim Pharmaceuticals Inc, Ridgefield, CT, USA

PharmShapeCC has been developed to perform multi-conformational 3D pharmacophore and shape searches against a collection of virtual combinatorial libraries containing more than 10 trillion compounds. A PharmShapeCC query is created from a user-provided 3D pharmacophore model mapped onto one or more aligned template ligands. A composite inclusion shape component to the query is generated from the aligned ligands. Alternatively, inclusion shapes can be derived from protein structures. PharmShapeCC makes use of the combinatorial nature of the compound deck as well as the hypothesis that active library compounds are likely to bind in similar binding orientations to prune the search space. PharmShapeCC search results will be compared to those obtained from searching exhaustively enumerated compound spaces with the same query and algorithm. Examples of successful identification of novel chemical space will be presented.

B-6 : Learning from the best: utilizing knowledge-based protein validation scores in receptor-ligand complex predictionElmar Krieger^{1,2}, Bart Nijssse¹, Sander B. Nabuurs¹¹ *Computational Drug Discovery group, Centre for Molecular and Biomolecular Informatics, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands.*² *Yasara Biosciences, Vienna, Austria.*

An important but complex aspect of structure-based drug design is accurately modeling receptor flexibility. Over the past years several approaches to consider protein plasticity in receptor-ligand complex prediction have been developed^{1,2}. Unfortunately, a distinct shortcoming of these flexible docking approaches is that enlargement of the receptor's conformational space tends to yield more false-positive predictions¹.

This decrease in performance for a large part results from the utilized scoring functions^{1,2}. Most current scoring functions have been optimized towards scoring intermolecular receptor-ligand interactions, but usually do not take the receptor conformation into account. This does not pose a problem for rigid receptor docking, but can severely degrade performance of flexible docking approaches, which either select from or generate multiple receptor conformations.

Here, we investigate the feasibility of utilizing protein structure validation techniques^{3,4} to aid in the identification and ranking of realistic receptor-ligand binding geometries. We show that protein structure validation scores, derived from the highest quality X-ray structures and which evaluate for example the normality of induced side chain rotamers, non-bonded interactions or side chain planarity, can aid in the identification of incorrect or less probably receptor geometries. By implementing a select set of validation scores into the scoring function of our induced fit docking program Fleksy⁵, we show that more accurate solutions can be obtained. Additionally, we find that the number of false positive hits decreases when multiple receptor geometries are combined.

One of scoring functions developed as part of this work was applied to predict the binding mode of a small molecule ligand to the DRD3 receptor in the GPCR Dock 2010 competition⁶. All five submitted predictions for this target were in the top 10 and the best model was correctly ranked first.

1. Beier, C.; Zacharias, M., Tackling the challenges posed by target flexibility in drug design. *Expert Opinion Drug Discovery* 2010, 5, (4), 347-359.
2. Henzler, A. M.; Rarey, M., In Pursuit of Fully Flexible Protein-Ligand Docking: Modeling the Bilateral Mechanism of Binding. *Molecular Informatics* 2010, 29, 164-173.
3. Hooft, R. W.; Vriend, G.; Sander, C.; Abola, E. E., Errors in protein structures. *Nature* 1996, 381, (6580), 272.

4. Krieger, E.; Joo, K.; Lee, J.; Raman, S.; Thompson, J.; Tyka, M.; Baker, D.; Karplus, K., Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: Four approaches that performed well in CASP8. *Proteins* 2009, 77 Suppl 9, 114-22.
5. Nabuurs, S. B.; Wagener, M.; de Vlieg, J., A flexible approach to induced fit docking. *Journal of Medicinal Chemistry* 2007, 50, (26), 6507-18.
6. Kufareva, I.; Rueda, M.; Katritch, V.; GPCR Dock 2010 Participants; Stevens, R. C.; Abagyan, R., Status of GPCR modeling and docking as reflected by community wide GPCR Dock 2010 assessment, *submitted*.

C-1 : Validation and characterization of chemical structures derived from names and images in scientific documents

J. Kinney¹, M. Hermsmeier²

¹ *DuPont, Newark DE, USA*

² *Bristol Meyers Squibb, Lawrenceville NJ, USA*

High-quality software applications currently convert chemical names and images into chemical structures on a large scale to automatically curate large document sets such as the patent corpus. The challenge remains however, in the accuracy of the assigned structures. Errors can arise from inconsistencies in the quality of the original source documents. For example, OCR and typographical errors in the text or pixelated and fuzzy lines in the images all contribute to uncertainty in assigning structures. To address this challenge we are participating in a collaboration with IBM and several other companies to verify the structures of the millions of unique chemical entities extracted from these documents. This talk will discuss processes that we have developed in order to characterize and validate the structures that are identified by the image and text conversion algorithms.

C-2 : Prediction of adverse drug reactions using systems biology

C. Merlot, D. Zuaboni

drug design tech, Plan-les-Ouates, Switzerland

The aim of this work is to predict adverse drug reactions (ADRs) from the chemical structure through predicted drug-target interaction profiles and biological pathways.

| | | | | | | |
|--------------------|---|---------------------|---|------------------|---|-------------------------|
| Chemical Structure | ⇒ | Drug-Target Profile | ⇒ | Pathways Altered | ⇒ | Adverse Drug Reactions. |
|--------------------|---|---------------------|---|------------------|---|-------------------------|

The advantages of a multistep strategy over current toxicity prediction systems have recently been reviewed¹. When applied in the drug discovery process, the major advantages of this method are that:

- it proposes a mechanism of action,
- the mechanism of action can be translated into short term assays in order to validate or invalidate the prediction in a short timeframe,
- because it is based on more information than simply known drug-side effects associations, it is expected to have a greater sensitivity.

Prediction of ADRs from drug-target interaction profiles has been a focus of several groups^{2,3}. The use of pathways has been reported in the works of Scheiber⁴ and Wallach⁵. However there is currently no integrated system that takes all these methods to provide routinely ADR from the chemical structure on a large scale.

Our methodology starts with the construction of a large number of fragment-based models to predict the drug-target interaction profile: the Predicted Safety Pharmaceutical Profiling (PSPP). These models are based on the PubChem⁶ database, complemented with ad hoc datasets. The second step is to associate

drug-target interactions profiles to pharmacological effect through pathways. We are using KEGG pathways and the Sider⁷ database, a database listing side effects of known drugs, to train the system.

A real-life application of this new software in a drug discovery project will be presented.

1. Yamanishi, Y.; Kotera, M.; Kanehisa, M.; Goto, S. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics* 2010, 26 (12), 246-254.
2. Bender, A.; Scheiber, J.; Glick, M.; Davies, J. W.; Azzaoui, K.; Hamon, J.; Urban, L.; Whitebread, S.; Jenkins, J. L. Analysis of Pharmacology Data and the Prediction of Adverse Drug Reactions and Off-Target Effects from Chemical Structure. *ChemMedChem* 2007, 2, 861-873.
3. Fliri, A. F.; Loging, W. T.; Thadeio, P. F.; Volkmann, R. A. Analysis of drug-induced effect patterns to link structure and side effects of medicines. *Nature Chemical Biology* 2005, 1 (7), 389-397.
4. Scheiber, J.; Chen, B.; Milik, M.; Sukuru, S. C. K.; Bender, A.; Mikhailov, D.; Whitebread, S.; Hamon, J.; Azzaoui, K.; Urban, L.; Glick, M.; Davies, J. W.; Jenkins, J. L. Gaining Insight Off-Target Mediated Effects of Drugs Candidates with a Comprehensive Systems Chemical Biology Analysis. *Journal of Chemical Information and Modeling* 2009, 49, 308-317.
5. Wallach, I.; Jaitly, N.; Lilien, R. A Structure-Based Approach for Mapping Adverse Drug Reactions to the Perturbation of Underlying Biological Pathways. *PLoS ONE* 2010, 5 (8), 1-11.
6. Li, Q.; Cheng, T.; Wang, Y.; Bryant, S. H. PubChem as a public resource for drug discovery. *Drug Discovery today* 2010, 15 (23/24).
7. Kuhn, M.; Campillos, M.; Letunic, I.; Jensen, L. J.; Bork, P. A side effect resource to capture phenotypic effects of drugs. *Molecular Systems Biology* 2010, 6 (343), 1-6.

C-3 : Global mapping of ligand and target binding spaces

F. Krüger, J. Overington

European Bioinformatics Institute, Hinxton, United Kingdom

Increasing availability of pharmacological and high throughput screening data makes it possible to globally map the ligand and target binding spaces of bioactive compounds. Here we describe the challenges associated with large-scale integration of bioactivity data from the medicinal chemistry literature and report on our survey of observed deviations in binding affinity due to different assay conditions and reporting techniques. For about three quarters of all examined assays, observed deviations are less than tenfold and hence comparable.

To evaluate the robustness of compound performance in different species, binding affinity data measured against pairs of human targets and orthologous targets in other species was systematically compared. Our findings suggest that globally, the affinity of small molecule binding is conserved between human targets and their orthologs in rat and mouse.

Our study thus provides insight into the inter-assay precision of assays reported in the literature and the conservation of ligand binding between species.

D-1 : Pharmacophoric space: do targets segregate?

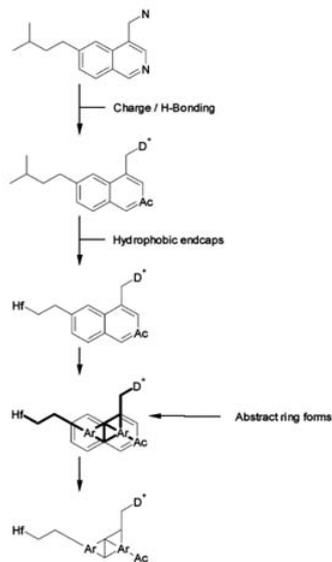
A. Zaliani¹, A. Ausekar²

¹ *Evolvus GmbH, Altenhoferallee 3, Frankfurt am Main 60438 Germany*

² *Evolvus Co., 88 Shukrawar Peth, Pune 411002 India*

Medicinal chemistry databases are more and more strategic to rationalize discovery approaches to new pharmacological targets. Bioinformatic and structural tools already exist to classify protein targets in many dimensions, from sequence to 3D-motifs. On the other side, chemical spaces where to search possible drugs

of the future has been, and actually still is, object of study, but has not yet been so deep described or characterized in terms of pharmacophore ¹.



Scheme 1

Fortunately, modern medicinal chemistry databases, when swiftly integrated with biological data and bioinformatics information, facilitate the analysis of this space. Here we focused on a special analysis aspect concerning the pharmacophoric space used by molecules (inhibitors or antagonists) active on same targets or class thereof.

The analysis has been conducted on ca. 23K compounds retrieved from Evolvus databases ². These compounds were active on their targets with at least published $pIC_{50} \geq 6$. In case of multiple activities on different targets, molecules were assigned to the target on which they showed maximal potency, provided they showed also at least 10x higher potency with respect to their second highest activity. If the second condition was not met, molecules were discarded. Actives have been tagged this way with their primary targets and their primary target class (e.g. GPCRA, NHR, Protease, Kinase, and Transporter).

Aim of the work was to show whether for each pharmaceutical relevant target, or class thereof, a consistent subset of pharmacophoric geometries or distances can be found, allowing simple recognition of target in literature molecule and/or automatic recognition and

prediction of possible target for untested catalogue molecules. To do so, we also defined the pharmacophoric space as the multidimensional space coded by the ErG's vector ³ of graph distances between couples of pharmacophoric groups found in the 2D molecular structures of each active (Scheme 1).

Briefly, ErG fingerprinting approach describes molecules as extended 2D-graphs where seven pharmacophoric groups can be used to extract a vector of graph distances between all possible combinations of pair of pharmacophoric groups. As a limit of 15 bond distances has been fixed for a usual drug-like compound and the possible combinations of pairs equals 21, we managed to code molecules with a compact vector of 315 graph distances (shortest path distance). Using this vector as fingerprint together with a Tanimoto metric ⁴ to measure similarity, we were able to obtain a rather precise picture of the pharmacophoric space.

Several statistical tools have been used to produce a convenient vision of the pharmacophoric space. First three component of PCA explained than 40% of total variance. Kohonen's maps helped us in better visualizing the fine differences among molecules classes. Finally, a simple Random Forest (RF) statistical model ⁵ developed within KNIME ⁶ allowed us to predict >90% of the removed actives in the correct Target Class (GPCRA, NHR, Kinase or Protease). Limitations of the description and of the entire approach are known, but we are confident that, with a more detailed exploitation of the method (e.g. weighting of particular subgraphs or knowledge-derived determinant distances), an increasing part of chemical space can be consistently be filtered according to the present segregation of pharmacophores.

1. Kirkpatrick, P.; C. Ellis (2004). "[Chemical space](#)". Nature 432 (432): 823–865.; Van Deursen R, Reymond J-L (2007). "Chemical Space Travel". ChemMedChem 2 (5): 636.;
2. Jacoby, B.; Bouhelal, R.; Gerspacher, M.; Seuwen, K.; ChemMedChem 2006, 1, 760 – 782
3. Stiefl, N.; Watson, I. A.; Baumann, K.; Zaliani, A. ErG: 2DPharmacophore Descriptions for Scaffold-Hopping. J. Chem. Inf. Model. 2006, 46, 208-220.
4. Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. J. Chem. Inf. Comput. Sci. 1998, 38, 983-996.
5. Anderson G.; Pfahringer B.(2006) Random relational rules. In Stephen Muggleton and Ramon Otero, editors, Extended Abstracts for 16th International Conference on Inductive Logic Programming;
6. Berthold, M. R.; Cebon, N.; Dill, F.; Gabriel, T.R.; Kötter, T.; Meinel, T.; Ohl, P.; Thiel, K.; Wiswedel, B.:(2009) [KNIME - The Konstanz Information Miner](#), SIGKDD Explorations, vol. 11, no. 1;

D-2 : Extraction of useful bioisostere replacements from the PDBT. Ritschel¹, D. J. Wood^{1,3}, J. de Vlieg^{1,2}, M. Wagener²¹ *Computational Drug Discovery Group, Centre for Molecular and Biomolecular Informatics, Radboud University Nijmegen Medical Centre, Nijmegen, NL*² *Molecular Design & Informatics, MSD, Oss, NL*³ *Current address: AstraZeneca, Mereside, Alderley Park, Macclesfield SK10 4TG, UK*

Bioisosteres are defined as structurally different molecules or substructures that can form similar intermolecular interactions with a macromolecular target structure. Ligand-based methods are extensively developed and applied to identify bioisosteres.^{1,2,3} In this work we present the new method KRIPO (Key Representation of Interactions in POckets), which uses the information of protein-ligand complexes to search for bioisosteres. Based on KRIPO a system has been put in place to suggest structural modifications to ligands solely based on the similarity of the binding pockets, not being biased by any ligand similarity.

The characteristics of binding pockets containing a ligand are encoded in 3D-pharmacophore fingerprints. The binding site fingerprints were optimized to improve their performance. A variety of attributes of the fingerprints were considered for the optimization, including the placement of pharmacophore features, whether or not the fingerprints are fuzzified, and the resolution and complexity of the pharmacophore fingerprints (2-, 3- and 4-point fingerprints). Finally, fuzzy 3-point pharmacophore fingerprints were chosen as an optimum representation of localized binding sites in a searchable fragment database.

The KRIPO fingerprint representation is key for fast and efficient processing of PDB scale crystal structure databases. Therefore, an on-the-fly searching of PDB scale crystal structure database for potential bioisosteric replacements is feasible. The architecture of the search system, details of the fingerprint definition and optimization as well as examples of bioisosteres suggested by KRIPO will be presented. The examples will discuss potential replacements in a structure-based context as it is shown for the adenine moiety bound to the hinge region in protein kinases in Figure 1.

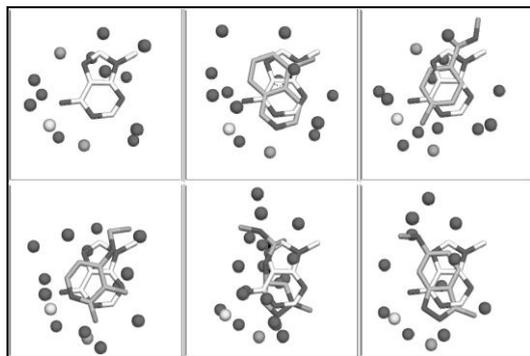


Figure 1. Bioisosteric replacements of the adenine moiety bound to the hinge region of protein kinases (PDB-code 1ATP). a) Query fragment (light gray) and the corresponding pharmacophore based on the properties of the protein binding site. b) – f) Hit fragments (dark gray) and the corresponding pharmacophore superposed to the query fragment (light gray). The feature of the pharmacophore are shown as spheres (dark gray: hydrophobe, light gray: H-bond acceptor, gray: Hbond donor).

1. Lima, L. M.; Barreiro, E. J. Bioisosterism: a useful strategy for molecular modification and drug design. *Curr Med Chem* 2005, 12, 23-49.
2. Wagener, M.; Lommerse, J. P. The quest for bioisosteric replacements. *J Chem Inf Model* 2006, 46, 677-685.
3. Olesen, P. H. The use of bioisosteric groups in lead optimization. *Curr Opin Drug Discov Devel* 2001, 4, 471-478.

D-3 : PubChem3D: diversity of shape space

E. Bolton

*US National Center for Biotechnology Information (NCBI)***Background**

The shape diversity of 16.4 million biologically relevant molecules from the PubChem Compound database and their 1.46 billion diverse conformers was explored as a function of molecular volume.

Results

The diversity of shape space was investigated by determining the shape similarity threshold to achieve a maximum on the count of reference shapes per unit of conformer volume. The rate of growth in shape space, as represented by a decreasing shape similarity threshold, was found to be remarkably smooth as a function of volume. There was no apparent correlation between count of conformers per unit volume and their diversity, meaning that a single reference shape can describe the shape space of many chemical structures. The ability of a volume to also describe the shape space of lesser volumes was also examined. It was shown that a given volume was able to describe 40-70% of the shape diversity of lesser volumes, for the majority of the volume range considered in this study.

Conclusion

The relative growth of shape diversity as a function of volume and shape similarity is surprisingly uniform. Given the distribution of chemicals in PubChem versus what is theoretically synthetically possible, the results from this analysis should be considered a conservative estimate to the true diversity of shape space.

D-4 : Comparison and visualisation of large chemical spaces using unsupervised classification techniquesA. Böcker¹, C. Kirchhoff¹, A. Müller¹, O. Barker², M. Whittaker², T. Hestekamp¹¹*Evotec AG, Hamburg, Germany*²*Evotec Ltd, Abingdon, United Kingdom*

HTS has brought a technical solution to screening large corporate compound collections with more than 100k molecules in a biological assay. One key challenge remains, the selection of an appropriate chemical collection for the identification of suitable hits for the target(s) under investigation. A question is how to compare and visualize large existing libraries with respect to physicochemical properties, chemical novelty and coverage of the available chemical space. This is mainly challenged by the fact that most computational methods are not applicable to large chemical libraries with millions of compounds. At Evotec we have systematically addressed this question and compared the Evotec Discovery Library (300K drug-like compounds) to the MDDR library (190K bioactive compounds)¹ and a commercially available set of 2.2 million compounds. We have applied principle component analysis (PCA), clustering² and emergent self-organizing maps (ESOM)³ in combination with MOE physicochemical descriptors⁴, EvoCATS pharmacophore descriptors⁵ and Unity 2D fingerprints.⁶ To the best of our knowledge this is the first time that ESOMs have been applied to the comparison of large libraries.

We could show with standard methods like PCA and clustering that the Evotec library provides a good coverage of the relevant bioactive reference space (MDDR) & the commercially available chemical space, respectively. While PCA provides an initial visualization of the data, clustering generates a more fine-grained (less visual) view on the data allowing for a quantitative assessment of the coverage of chemical series and the assessment of singletons.

An even better visualization was obtained using ESOMs. Maps were trained using a toroidal topology with 500x820 grid points on the entire compound set. This translated into a fine-grained visualization of the chemical space. It allowed for the identification of covered and exclusive chemical spaces. Moreover, by mapping historical HTS data on the maps bioactivity landscapes can be generated. We believe that ESOM is a unique visualization tool which is immediately applicable to large scale library analysis processes. At Evotec workflows have been implemented to automatically guide such library analyses which will provide added value for future drug discovery programs.

1. MDL Drug Data Report, Version December 2009; Symyx Technologies Inc.: Santa Clara, CA, 2009.
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, Vol. 46
4. Molecular Operating Environment (MOE), Version 2007.09; Chemical Computing Group Inc.: Montreal, Canada, 2007.
5. Schneider, G.; Neidhard, W.; Giller, T.; Schmid, G. "Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chemie Int. Ed.* 1999, 38, 2894-2896
6. SYBYLX version 1.2; Tripos International: St Louis, MO, 2009.

D-5 : WizePairZ: auto-curation of matched molecular pairs

D. Wood¹, C. Green¹, E. Griffen¹, M. Harrison¹, A. Rabow¹, S. St-Gallay¹, A. Ting¹, D. Warner²
¹ AstraZeneca, Alderley Park, Macclesfield, UK
² AstraZeneca, Montreal Research Centre, Montreal, Canada

Matched Molecular Pairs¹ (MMPs) are pairs of compounds that differ in their chemical structures in a single position, and analyses of MMP data may therefore provide information on the pharmacological or physical chemical effects of specific structural replacements. WizePairZ² is a system developed at AstraZeneca for identifying and extracting MMPs from sets of compounds. The approach can identify many thousands of potential transformations for a given pharmacological property; however, some of the transformations are more likely to be of use in drug discovery than others. The transformations of most value to drug design scientists are those with a high probability of having the predicted effect, and come from MMPs with many observations over a diverse range of chemistries. WizePairZ automatically validates and curates the WizePairZ transformations and provides users with only those that can be expected to make the desired property change to a new compound with a high degree of confidence. This is done with a Sample Size and Tanimoto Corrected Cumulative Distribution Function (CCDF), which represents the probability that a particular transformation will make a significant change to the property when applied to new compounds. The approach is demonstrated with a few examples, including a large hERG dataset of approximately 70,000 compounds.

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, 23, 6672-6682.
2. Warner, D. J.; Griffen, E. J.; St-Gallay, S. WizePairZ: A Novel Algorithm to Identify, Encode, and Exploit Matched Molecular Pairs with Unspecified Cores in Medicinal Chemistry. *J. Chem. Inf. Model.* 2010, 8, 1350-1357.

D-6 : Mining for context-sensitive matched molecular pairs and bioisosteric replacements in large corporate chemical databases

G. Papadatos, N. Fechner, D. Evans, M. Bodkin
 Eli Lilly UK, Erl Wood Manor, Windlesham, Surrey GU20 6PH, UK

Matched molecular pair analysis (MMPA) studies the effect of specific structural modifications on properties of interest.^{1,2} In contemporary drug discovery this analytical tool becomes very important, especially during the lead optimisation phase, when medicinal chemists synthesise and test iteratively hundreds of analogous compounds, which usually differ from each other by a small change. Furthermore, a

useful extension of MMPA includes the identification of (bio)isosteric replacements, i.e. structural modifications with minimal effect on the end point under examination.³

In this study, we have developed a novel methodology to assess the extent to which matched molecular pair analysis is applicable to the improvement of lead optimisation. Contrary to common practice, it is hypothesised here that the effect of a structural transformation is not global across chemical space; it rather depends on the invariable part of the molecule, where the modification took place, i.e. the *context*. Subsequently, transformations between “similar” contexts might have similar or otherwise predictable effect on a measured property.⁴

Starting with virtually all possible matched pairs in a data set identified by a fast and efficient algorithm,^{3,5} our hypothesis about the effect of the context of the transformation on the end point property has been systematically investigated: in particular, we have used several ways to represent/abstract the context of a matched molecular pair. Such approaches include the consideration of both the “local environment” around the attachment point, as well as whole molecule/context representations. In more detail, for the former approach we have used electro-topological state (E-state) values, Atom Environments and Reduced Graph nodes. For the whole molecule representation approach, RGs, Murcko frameworks and standard 2D fingerprints with Tanimoto similarity have been employed. Regarding the end point properties, we have investigated a number of important drug properties such as hERG inhibition and solubility, all derived from the Eli Lilly data archive.

The descriptors listed above enable the partition of the contexts for each transformation into “local” groups depending on the structure of the context. We have calculated the p-values for each local distribution in order to assess the statistical significance of them being different from the global one. After examination of the p-values, several local distributions are found to be significantly different to the corresponding global ones.

In conclusion, the further context-sensitive dissection of matched molecular pair and bioisosteric replacement data allowed for identification of trends that would otherwise be hidden in the global view of each transformation. Furthermore, we have surveyed several simple and interpretable descriptors in order to represent the contexts on a local or whole molecule level. Additional plans include the distillation of useful guidelines for several end-properties to lead optimisation and integration of this approach with workflow environments such as KNIME, which will undoubtedly make it more accessible to the medicinal chemists.

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* (23), 6672-6682.
2. Hajduk, P. J.; Sauer, D. R., Statistical analysis of the effects of common chemical substituents on ligand potency. *J. Med. Chem.* 2008, *51* (3), 553-564.
3. Wagener, M.; Lommerse, J. P. M., The quest for bioisosteric replacements. *J. Chem. Inf. Model.* 2006, *46* (2), 677-685.
4. Papadatos, G.; Alkarouri, M.; Gillet, V. J.; Willett, P.; Kadiramanathan, V.; Luscombe, C. N.; Bravi, G.; Richmond, N. J.; Pickett, S. D.; Hussain, J.; Pritchard, J. M.; Cooper, A. W. J.; Macdonald, S. J. F., Lead optimization using matched molecular pairs: Inclusion of contextual information for enhanced prediction of hERG inhibition, solubility, and lipophilicity. *J. Chem. Inf. Model.* 2010, *50* (10), 1872-1886.
5. Hussain, J.; Rea, C., Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. *J. Chem. Inf. Model.* 2010, *50* (3), 339-348.

D-7 : On the exclusion of unwanted chemical patterns from large fragment spaces

H.-C. Ehrlich¹, M. Rarey¹

¹ Center for Bioinformatics, University of Hamburg, Hamburg, Germany

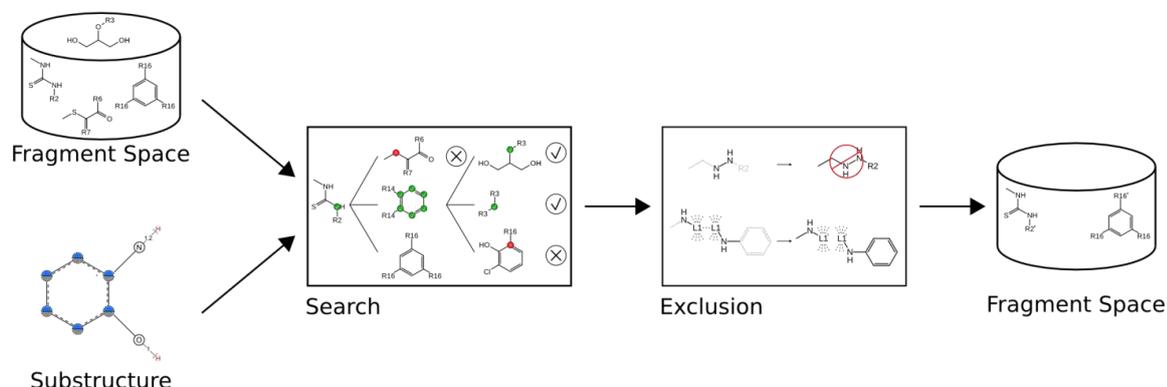


Figure 1: Workflow of fragment space modification excluding all products that include a predefined substructure.

Fragment spaces (FSs) are a compact and elegant form to represent a large number of chemical compounds. FSs are composed of molecular fragments and rules how to combine them to products. It has been recently shown that FSs can be used to model and search the synthetically accessible compound space in pharmaceutical research environments^{1, 2}. In contrast to molecular libraries, it is, however, much more difficult to exclude molecules having certain unwanted substructures. The removal of such products, especially when they consist of multiple fragments, would require an enumeration of all products which is inefficient or even impossible. Therefore, an efficient method must directly process fragments and joining rules. Even though methods directly processing FSs exist^{3, 4}, none of them can automatically redesign FSs such that all products having undesired substructures are removed.

We present an algorithm based on the VF2 subgraph isomorphism algorithm⁵, named Fragment Space Carver, excluding all molecules with undesired substructures from a FS. The algorithm uses a search procedure that detects substructures included in single fragments or in products that are formed by joining two or more fragments without explicitly enumerating products. The FS is modified by either removing the fragments or by changing the connection rules in a way that prohibits undesired substructure formations. The resulting FS can be further processed by any available FS method.

Fragment space carver was evaluated using the BRICS³ and KnowledgeSpace⁶ FSs. Both were analyzed for the presents of toxic functional groups and pan assay interference compounds (PAINS)⁷. The analysis shows that the toxic groups as well as all PAINS could be efficiently removed from both FSs.

1. U. Lessel, B. Wellenzohn, M. Lilienthal, H. Claussen. Searching Fragment Spaces with feature trees. *Journal of Chemical Information and Modeling*, **2009**, 49, 270-279
2. M. Boehm, T.-Y. Wu, H. Claussen, C. Lemmen. Similarity Searching and Scaffold Hopping in Synthetically Accessible Combinatorial Chemistry Spaces. *Journal of Medicinal Chemistry*, **2008**, 51, 2468-2480
3. J. Degen, C. Wegscheid-Gerlach, A. Zaliani, M. Rarey. On the art of compiling and using 'drug-like' chemical fragment spaces. *ChemMedChem*, **2008**, 3, 1503–1507.
4. M. Rarey and M. Stahl. Similarity searching in large combinatorial chemistry spaces. *Journal of Computer-Aided Molecular Design*, **2001**, 15, 497–520.

D-8 : An analysis of fragment-spaces and their impact on fragment replacementG. Skillman¹¹ *OpenEye Scientific Software, Inc., Santa Fe, USA*

Medicinal chemists often approach the multidimensional optimization of compound properties by holding a large portion of a molecule constant while exploring local SAR through fragment replacement. Though the concept of computational fragment replacement has been known since the work of Bartlett(1), recent years have seen a renewed interest in computational tools for the selection of suitable replacement fragments. Fragments replacement algorithms explore a large fragment-space to find a series of similar fragments, thus generating a series of molecules that sample a local space near the original molecule. While there are many algorithms for identifying suitable fragments, the quality of the generated molecules is ultimately dependent upon the fragment-space from which the replacement fragments are drawn. The quality is affected by both the fragments in the database as well as the conformers that represent each fragment. This study characterizes several fragment collections and explores their impact on successful fragment replacement. We report results comparing fragment databases generated from structural molecule datasets (CSD and PDB), large molecular databases (PubChem) and exhaustive graph enumeration. We will present a statistical analysis of the effect of database selection on fragment replacement in a large set of medicinal molecules.

1. Lauri, G.; Bartlett, P.A., CAVEAT: A Program to Facilitate the Design of Organic Molecules, *J. Comp.-Aided Mol. Design*, 1994, 8, 51-66.

E-1 : ARChem route designer: the application of automated retrosynthetic rule generation to synthesis planningA. Cook¹, A. P. Johnson¹, J. Law², M. Mirzazadeh², O. Ravitz², A. Simon²¹ *School of Chemistry, University of Leeds, Leeds, UK*² *SimBioSys Inc., Toronto, Canada*

The discipline of computer-aided chemical synthesis design is now 40 years old.¹ The original research programs of that period explored two approaches for knowledge acquisition of retrosynthetic reaction transformations.

In the original empirical rule based systems² transforms were collated by hand from the primary literature and distilled into coded reaction rules. This approach never acquired sufficient depth of knowledge due to the laborious and highly skilled nature of the task. The concurrent rapid discovery of new reactions, methods and selective reagents meant there was little hope of catching up and efforts to develop the knowledge bases waned.

The systems that chose to use formal methods to automatically generate reactions from first principles^{3,4} were limited by the need to apply rigorous constraints to remove unlikely reactions. These systems were capable of suggesting novel chemistry, but the lack of hard precedents, with accompanying scope and limitations, was not attractive for the end user.

In recognition of these fundamental problems, research into a third approach using automated production of transforms from curated reaction databases has begun to make inroads over the last few decades.⁵⁻⁸ Novel methodologies employed by the ARChem route designer program⁹ to generate reaction rules from available reaction databases are discussed. The organization of the automatically generated rules into hierarchies and their application to the synthesis design problem is demonstrated. The need for augmenting key reaction rules with manually curated data for improving the generated routes is discussed.

The requirement to produce enantiopure drugs¹⁰ poses significant challenges in planning a synthesis.¹¹ A new research project has been initiated at Leeds to study computer-aided enantioselective chemical synthesis planning. Novel algorithms for the recognition and treatment of stereochemistry in targets

molecules and transform rules are presented. Approaches for goal directed strategy selection for stereocontrolled synthesis is discussed.

1. Cook, A. P.; Johnson, A. P.; Law, J.; Mirzazadeh, M.; Ravitz, O.; Simon, A. Computer-aided synthesis design: 40 years on. In *Wiley Interdisciplinary Reviews: Computational Molecular Science*; in press, **2011**.
2. Corey, E.; Long, A.; Rubenstein, S. Computer-assisted analysis in organic synthesis. *Science*, **1985**, 228, 408-418.
3. Gasteiger, J.; Hutchings, M. G.; Christoph, B.; Gann, L.; Hiller, C.; Low, P.; Marsili, M.; Saller, H.; Yuki, K. A New Treatment of Chemical Reactivity: Development of EROS, an Expert System for Reaction Prediction and Synthesis Design. *Top. Curr. Chem.*, 1987, 137, 19-73.
4. Hendrickson, J. B.; Braun-Keller, E. Systematic synthesis design. 8. Generation of reaction sequences. *J. Comp. Chem.*, **1980**, 1, 323-333.
5. Röse, P.; Gasteiger, J. Automated derivation of reaction rules for the EROS 6.0 system for reaction prediction. *Anal. Chim. Acta*, **1990**, 235, 163-168.
6. Blurock, E. S. Computer-aided synthesis design at RISC-Linz: automatic extraction and use of reaction classes. *J. Chem. Inf. Comp. Sci.*, **1990**, 30, 505-510.
7. Satoh, K.; Funatsu, K. A Novel Approach to Retrosynthetic Analysis Using Knowledge Bases Derived from Reaction Databases. *J. Chem. Inf. Comp. Sci.*, **1999**, 39, 316-325.
8. Wang, K.; Wang, L.; Yuan, Q.; Luo, S.; Yao, J.; Yuan, S.; Zheng, C.; Brandt, J. Construction of a generic reaction knowledge base by reaction data mining. *J. Mol. Graphics Modell.*, **2001**, 19, 427-433.

E-2 : De novo design of synthetically feasible compounds using reaction vectors and evolutionary multiobjective optimization

B. Allen¹, V. J. Gillet¹, B. Chen¹, M. J. Bodkin², J. Cole³, J. Liebeschuetz³

¹ University of Sheffield, Sheffield, United Kingdom

² Eli Lilly, Erl Wood, United Kingdom

³ Cambridge Crystallographic Data Center, Cambridge, United Kingdom

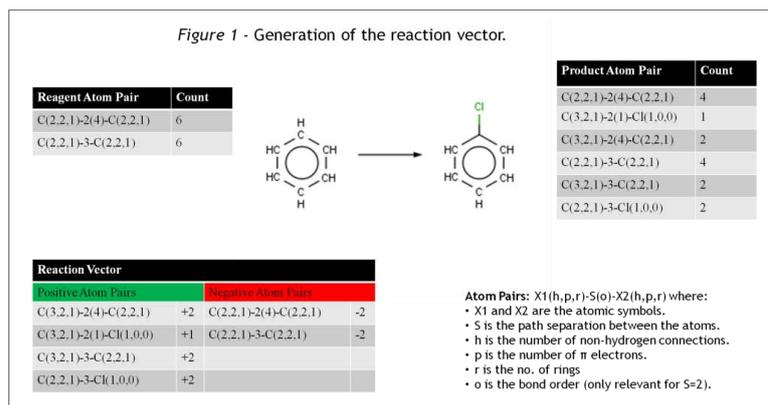
The goal of de novo design is to identify novel compounds with therapeutic potential from anywhere in chemical space, rather than being restricted to searching databases of previously synthesized compounds. The vast size of chemical space requires a highly efficient search strategy. Defining the therapeutic potential of a compound is an inherently multidimensional problem, as activity, bulk properties, ADME/Tox etc all need to be optimized simultaneously. Finally there is little point in identifying theoretically interesting compounds which cannot be synthesized. The de novo design tool described here uses reaction vectors to encode known reactions for use in generating novel compounds¹, and evolutionary multiobjective optimization to search the available chemical space.

The tool requires a knowledge-base of known reactions, which are preprocessed to encode the transformations in moving from reactants to products as reaction vectors, as shown in *Figure 1*. The reaction vectors can then be applied to suitable previously unseen reactants to generate novel compounds. Reaction vectors can be derived automatically from any reaction, with no limits on complexity, allowing the generation of a large and comprehensive knowledge-base. In particular, they can be generated from data from electronic lab notebooks. The advantage of this approach is that all molecules that are generated should be synthetically feasible, because they are constructed by the application of known reactions derived from the knowledge-base.

Although the use of the knowledge-base restricts the potential molecules to a synthetically feasible subset of chemical space, this subset is still vastly larger than can be searched exhaustively in realistic timescales. Our de novo design tool is implemented using evolutionary algorithm techniques which are based on Darwinian evolution and are designed to allow rapid exploration of large search spaces². This is achieved by representing a population of potential solutions as variables, or chromosomes, and then modifying the chromosomes to generate new solutions through reproductive processes. By ranking the solutions and preferentially selecting the higher ranked as the source of new solutions, the algorithm is able to converge on good regions of the solution space. For problems where multiple objectives need to be simultaneously

satisfied, Pareto ranking³ is used to combine the multiple objective scores into a single ranking. This approach has the additional benefit that it generates a set of optimal solutions covering the range of possible compromises between the various objectives.

The tool has been applied to several test sets including: replication of the development of sulmazole from its precursor molecule ARL57⁴; and the design of novel compounds that fit pharmacophore constraints consistent with thrombin inhibitors⁵, using a fragment-based drug design approach. The tool was able to automatically rediscover sulmazole, and additionally produced several other molecules with high similarity to ARL57 and improved drug like physiochemical properties. In the fragment-based test case it was able to generate sets of novel drug-like molecules, with short and plausible synthetic routes from the initial fragments. The thrombin inhibitor problem has been extremely well studied already, and the molecules generated by de novo design were shown to overlap in chemical space with the existing range of known thrombin inhibitors.



1. Patel, H.; Bodkin, M. J.; Chen, B.; Gillet, V. J., Knowledge-Based Approach to de Novo Design Using Reaction Vectors. *Journal of Chemical Information and Modeling* **2009**, *49* (5), 1163-1184.
2. Coello Coello, C. A., Evolutionary multi-objective optimization: a historical view of the field. *Computational Intelligence Magazine, IEEE* **2006**, *1* (1), 28-36.
3. Deb, K.; Pratap, A.; Agarwal, S.; Meyarivan, T., A fast and elitist multiobjective genetic algorithm: NSGA-II. *Evolutionary Computation, IEEE Transactions on* **2002**, *6* (2), 182-197.
4. Topliss, J. G., Some observations on classical QSAR. *Perspectives in Drug Discovery and Design* **1993**, *1* (2), 253-268.
5. Gurm, H. S.; Bhatt, D. L., Thrombin, an ideal target for pharmacological inhibition: A review of direct thrombin inhibitors. *American Heart Journal* **2005**, *149* (Supplement 1), S43-S53.

E-3 : Improving metabolite identification with chemoinformatics

J. Peironcely^{1,2,3}, M. Rojas-Cherto^{2,3}, P. Kasper^{2,3}, L. Coulier^{1,3}, R. Vreeken^{2,3}, T. Reijmers^{2,3}, A. Bender⁴, J. L. Faulon⁵, T. Hankemeier^{2,3}

¹TNO Quality of Life, Zeist, The Netherlands

²Leiden University, Leiden, The Netherlands

³Netherlands Metabolomics Centre, Leiden, The Netherlands

⁴University of Cambridge, Cambridge, United Kingdom

⁵University of Evry, Evry, France

To provide detailed information of biological phenotypes on a chemical basis metabolomics aims at profiling all sorts of metabolites, which form a chemically diverse group of substrates and products involved in enzymatic pathways. Current analytical platforms used in metabolomics produce a large amount of complex data, which require chemoinformatics tools to process and transform this data into meaningful information. For biological interpretation identification of metabolites (elucidating the chemical structure of the metabolites of interest) is essential. New analytical platforms and better software tools are required to advance in metabolite identification. Here we present a pipeline of software tools developed to facilitate identification of metabolites measured with Liquid Chromatography – Mass Spectrometry (LC-MS).

High-resolution multi stage MS spectra (MS^n data) were acquired for metabolite standards listed in the HMDB (Human Metabolome Database). Currently no tool exists that captures all relevant information present in MS^n data so a software tool was developed, integrating the Chemistry Development Kit (CDK) and XCMS, for preprocessing the spectral data. The Multi-stage Elemental Formula (MEF) tool automatically resolves the elemental composition of the parent compound, the fragment ions, and the neutral losses. This process of elemental formula assignment and fitting also removes artifacts of the spectra. The resulting enriched MS^n data of many metabolite standards are stored in XML format in a MS^n database, to allow structural elucidation of unknown metabolites by comparing the MS^n data of the unknowns with the MS^n data in the database. The database also enables the characterization of substructures from the unknown compound by querying and matching subsets of the MS^n data. A fingerprint based similarity search for MS^n data was developed to find out which trees in the database are most similar to an experimentally acquired MS^n data.

An open source chemical structure generator was implemented to generate candidate structures using the elemental formula and substructure information obtained with the previous tools. This structure generator combines concepts of graph theory and a chemistry library, the CDK, to exhaustively generate all non-isomorphic chemical structures for the input data. This input data is an elemental formula and optionally, one or multiple non-overlapping prescribed substructures. The output of the structure generator is a, usually large, list of structures which need to be further reduced. Therefore, models of Metabolite-Likeness were built to reject structures that do not resemble metabolites. Different molecular descriptors, fingerprints, and classifiers were evaluated, and the best combination employed to build a final model. Only candidate structures with a high Metabolite-likeness are kept in our Metabolite Identification pipeline.

In this work we demonstrate how this workflow of chemoinformatics tools improves the state of the art in metabolite identification using real life samples and how it helps to translate experimental data into chemical data.

E-4 : Efficient matching of multiple chemical subgraphs

R. Sayle

NextMove Software, Cambridge, UK

The use of SMARTS patterns¹ and MDL queries² for pattern matching has become ubiquitous in cheminformatics, and efficient implementations exist for identifying one or more instances of a user-defined substructure in a molecular graph.^{3,4} However, a very common usage of this functionality is in applications that test a number of patterns against each molecule. Examples include filtering of desirable/undesirable properties⁵, atom typing⁶, descriptor and physical property calculation⁷, pharmacophore perception, feature-based fingerprint generation and IUPAC name generation. In these use-cases, current practice is typically to match each of the (SMARTS) patterns independently and sequentially. This work describes significantly more efficient methods for matching multiple patterns simultaneously. Much like chemical database search systems use fingerprint prescreens to optimize searching a single pattern against multiple molecules; pre-processing analysis and pattern compilation can be used to optimize matching multiple patterns against a single target molecule. This approach, which often makes the match run-time independent of the number of patterns, enables software applications that require matching of thousands or tens of thousands of patterns.

Relative performance figures will be given for several real world applications, built on different cheminformatics toolkits, including generation of 166-bit MACCS keys and PubChem fingerprints⁸, and the automatic indexing of molecules extracted from patents using Derwent's CPI codes⁹.

1. Daylight Theory: SMARTS – A Language for Describing Molecular Patterns. <http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html> (accessed Jan 31, 2011)
2. Dalby, A.; Nourse, J.G.; Hounshell, W.D.; Gushurst, A.; Grier, D.L.; Leland, B.A.; Laufer, J. Description of Several Chemical-Structure File Formats Used by Computer Programs Developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, 32(3), 244-255.
3. Ullman, J.R. An Algorithm for Subgraph Isomorphism. *J. ACM.* **1976**, 23, 31-42.

- McGregor, J.J. Backtrack Search Algorithms and the Maximal Common Subgraph Problem. *Software – Pract. Exper.* **1982**, 12, 23-34.
- Charifson, P.S.; Walters, W.P. Filtering Databases and Chemical Libraries. *Molecular Diversity*. **2002**, 5, 185-197.
- Bush, B.L.; Sheridan, R.P. PATTY: A Programmable Atom Typer and Language for Automatic Classification of Atoms in Molecular Databases. *J. Chem. Inf. Comput. Sci.* **1993**, 33, 756-762.
- Lee, A.C.; Yu, J.; Crippen, G.M. pKa Prediction of Monoprotic Small Molecules the SMARTS Way. *J. Chem. Inf. Model.* **2008**, 28, 2042-2053.
- Henry, D.R.; Durant, J.L. Optimization of MDL Substructure Search Keys for the Prediction of Activity and Toxicity. *Chemometrics and Chemoinformatics*, ACS Symposium Series, Vol. 894, Chapter 10, 2005, 145-156.
- Derwent Information; Bryan, P. Derwent World Patents Index: CPI Chemical Indexing User Guide, 2000. http://science.thomsonreuters.com/m/pdfs/mgr/chemical_index_guidelines.pdf (accessed Jan 31, 2011)

F-1 : Targeting natural products for drug discovery by mining biomedical information resources

E. Muratov^{1,2}, N. Baker¹, N. Rice¹, D. Fourches¹, A. Tropsha¹

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.*

Parallel screening of Natural Products (NPs) is a typical approach for identifying drug candidates and their targets. However, biomolecular targets of NPs are often discovered serendipitously. We hypothesize that since many NPs may bind to similar biological targets in both plants and humans, we shall study similarities between plants and humans at the biochemical pathway level to discern novel drug-target-disease associations (Figure 1). We report on the use of Chemotext, a database of assertions extracted from biomedical literature that link chemicals, targets, and diseases¹ to rationalize the search for NP targets in the context of the Systems Chemical Biology paradigm². We have identified similar biochemical pathways that NPs are known to interact with in both plants and humans. Using Chemotext, we have collected and integrated cross-species NP-target associations. We present the case studies of *Diabetes mellitus* for predicting new compound-target interactions and Tacrolimus-Binding Proteins for detecting similar biochemical pathways in both plants and animals/humans.

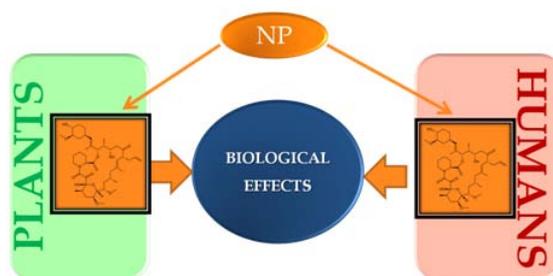


Figure 1. NPs can have similar biological effects at the biochemical pathway level for both plants and humans.

- Baker, N. C.; Hemminger, B. M. Mining connections between chemicals, proteins, and diseases extracted from Medline annotations. *J. Biomed. Informatics* **2010**, 43, 510-519.
- Oprea TI, May EE, Leitão A, Tropsha A. Computational systems chemical biology. *Methods Mol Biol.* **2011**; 672, 459-88.

F-2 : Identifying and quantifying drug promiscuity by correlating ligand and target shape similarities

V. I. Perez-Nueno, V. Venkatraman, L. Mavridis, D. W. Ritchie
 INRIA Nancy – Grand Est (LORIA), Vandoeuvre-lès-Nancy, France

Polypharmacology is becoming an increasingly important aspect in drug design. In the last 4 years, more than 30 drugs have been tested against more than 40 novel secondary targets based on promiscuity predictions. Pharmaceutical companies are discovering more and more cases in which multiple drugs bind to a given target (promiscuous targets) and in which a given drug binds to more than one target (promiscuous ligands). Both of these phenomena are clearly of great importance when considering drug side-effects. Current computational techniques for predicting drug pharmacological profiles relate targets to each other based on the similarity in the ligand chemical descriptor space ¹, protein target sequence space ², and more recently, in the pharmacophoric pocket descriptor space ³.

Here we present a shape-based approach which uses spherical harmonic (SH) representations ^{4,5} to compare molecular surfaces very efficiently. This approach compares targets by the SH similarity of their ligands and also of their binding pockets. This allows promiscuous ligands and targets to be predicted very rapidly.

We present details of our approach applied to the MDL Drug Data Report (MDDR) database containing 202,143 compounds distributed over 633 diverse pharmacological targets. To our knowledge, this is the largest all-against-all polypharmacological study to have been carried out using shape-based techniques. The correlation between binding pocket and ligand shapes is quantified as a novel expectation value (E-value) obtained from the similarity between drug compounds and the center molecule of a ligand set for a given target class. Our promiscuity predictions are compared with experimental activity values extracted from several public databases (*e.g.* BindingDB, PDBbind, ChEMBL, PDSP, and BRENDA).

1. Keiser, M. J. *et al.* Predicting new molecular targets for known drugs. *Nature* **2009**, 462, 175-181.
2. Weskamp, N.; Hüllermeier, E; Klebe, G. Merging chemical and biological space: Structural mapping of enzyme binding pocket space. *Proteins* **2009**, 76, 317-330.
3. Milletti, F.; Vulpetti, A. Binding pocket comparison using four-point pharmacophoric descriptors based on GRID. *J. Chem. Inf. Model.* **2010**, 50, 1418-143.
4. Lin, J.; Clark, T. An Analytical, Variable Resolution, Complete Description of Static Molecules and Their Intermolecular Binding Properties. *J. Chem. Inf. Model.* **2005**, 45, 1010-1016.
5. Ritchie, D. W.; Kemp, G. J. L. Protein Docking Using Spherical Polar Fourier Correlations. *Proteins: Struct. Func. Genet.* **2000**, 39, 178-194.

F-3 : A knowledge-based approach to assessing the target promiscuity of chemical fragments

X. Jalencas, J. Mestres
 Chemogenomics Laboratory, Research Unit on Biomedical Informatics (GRIB), IMIM – Hospital del Mar and Universitat Pompeu Fabra. Parc de Recerca Biomèdica, Barcelona, Catalonia, Spain

The number of structural entries in the Protein Data Bank containing a small molecule in the binding cavity of a protein has increased substantially in the recent years. Analysis of these data provides the structural basis for a better understanding of the levels of compatibility between chemical fragments and protein environments. Ultimately, the knowledge acquired in fragment-pocket interaction can be exploited in the hit and lead optimization process as well as in *de novo* drug design.

To this aim, the PDBbind database ¹, consisting on 4260 protein-ligand complexes, was analyzed. All simple chemical fragments containing rings were extracted from the ligands, and all protein environments binding those fragments were described in terms of their pharmacophorical properties in a similar way as described by Shulman-Peleg *et al.* ². For a further analysis of the data, a bipartite graph ³ was constructed, including all chemical fragments and protein environments. Topological analysis of the network shows not only that different fragments can bind to similar pockets, but also that different pockets can accommodate similar fragments.

This data can easily be exploited for drug design, either suggesting non-obvious replacements or in *de novo* drug design. This is achieved by populating known binding sites with chemical fragments that are co-crystallized in similar protein environments. Particularly interesting is the case where these similar protein environments belong to completely unrelated proteins, both in terms of sequence and fold. Being able to hop fragments between unrelated proteins provides a method to get an insight on which chemical fragments have a possibility to bind a particular binding site, even if there are no related solved structures available.

As a validation, this approach was applied to the reconstruction of known ligands for several protein structures, just by hopping chemical fragments coming from other unrelated proteins. Once this was shown to be successful, the method was further applied to the suggestion of new chemical fragment replacements. Some of these new replacements could be proven to be suitable, as they were also found in other known active compounds.

In summary, we introduce an approach to structure-based drug design based on that similar surface patches in proteins have a tendency to bind the same chemical fragments. This could be a suitable knowledge-based alternative to docking, as most of the needed descriptors can be pre-calculated and stored in databases.

1. Wang, R., Fang, X., Lu, Y. & Wang, S. The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem.* **2004**, 47, 2977-2980.
2. Shulman-Peleg, A., Nussinov, R. & Wolfson, H.J. Recognition of functional sites in protein structures. *J. Mol. Biol.* **2004**, 339, 607-33.
3. Vogt, I. & Mestres, J. Drug-Target Networks. *Mol. Inf.* **2010**, 29, 10-14.

F-4 : Combining global and local measures for druggability predictions

A. Volkamer ¹, T. Grombacher ², M. Rarey ¹

¹ *University of Hamburg, Center for Bioinformatics, Bundesstr. 43, 20146 Hamburg, Germany*

² *Bioinformatics, Merck Serono, Frankfurter Str. 250, 64293 Darmstadt, Germany*

Predicting druggability and prioritizing certain disease modifying targets for the drug development process is of high practical relevance in pharmaceutical research. Recently published methods affirm that hydrophobicity and size are important global pocket descriptors for automatic druggability prediction.^{3,4} While the classification works well for highly druggable and undruggable pockets, a classification especially in the intermediate area remains uncertain.

In our work, we developed a new procedure for automatic druggability prediction that takes into account global descriptors and adds local similarities between pockets. Pockets are predicted with DogSite¹ forming the basis of local shape and physiko-chemical descriptor calculation. Eventually, druggability scores are predicted by means of a support vector machine² (SVM). The SVM model is trained and tested on the (NR)DD³ dataset consisting of 1070 druggable, difficult and undruggable targets in total. Furthermore, a nearest neighbor method based on interaction histograms is used to take pocket similarities into account.

In over 90% of the cases, the SVM model correctly classifies a target as druggable or undruggable. Enrichment studies on the NRRD show that the method performs comparably to the recently published methods SiteMap⁴ and Fpocket³.

Our findings for global pocket descriptors coincide with already described properties. The hydrophobic character of a druggable pocket is indicated by a higher fraction of apolar pocket surface. A low surface to volume ratio, imitating the pocket enclosure, is another known indicator for druggability.

Additionally, we investigate into local properties of druggable pockets. Therefore, site interactions (SIACs⁵) present in the pockets are analyzed in terms of distance dependent histograms between SIAC pairs weighted by their accessibility. We found that druggable pockets tend to have less long range hydrophilic-hydrophilic SIAC pairs and more middle to long range lipophilic-lipophilic SIAC pairs compared to undruggable pockets.

The distances between SIAC profiles can be used for a nearest neighbor search on the NRDD pocket set. In 78% of the cases, the nearest neighbor and the structure itself conform in their druggability type. Such a neighborhood analysis can especially bring the annotation of intermediate targets forward.

Due to recent studies, some global druggability properties have been revealed. Analyzing the local pocket properties is another step towards a reliable descriptor-based druggability prediction. Nevertheless, the variety of pocket shapes and their flexibility upon ligand binding limit the automatic projection of druggable features onto descriptors.

1. A. Volkamer, A. Griewel, T. Grombacher, and M. Rarey. Analyzing the topology of active sites: On the prediction of pockets and subpockets. *Journal of Chemical Information and Modeling*, 50(11):2041-2052, 2010.
2. Chang CC, Lin CJ. LIBSVM -- A Library for Support Vector Machines. <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>
3. P. Schmidtke and X. Barril. Understanding and predicting druggability. a high-throughput method for detection of drug binding sites. *Journal of Medicinal Chemistry*, 53(15):5858-5867, 2010.
4. T. A. Halgren. Identifying and Characterizing Binding Sites and Assessing Druggability. *J. Chem. Inf. Model.*, 2009, 49(2):377-89.
5. I. Schellhammer and M. Rarey. Trixx: structure-based molecule indexing for large-scale virtual screening in sublinear time. *Journal of Computer-Aided Molecular Design*, 21(5):223-238, 2007

Poster Session Abstracts

RED

P-2 : PubChem3D: a significant new resource for scientists

E. Bolton

US National Center for Biotechnology Information (NCBI)

PubChem3D provides rapid access to 3-D similarity of chemical structures on a very large scale. A collection of tools, technologies, and interfaces, this relatively new global resource breaks down barriers to access and analyze biological and chemical structures in a single context. This poster will highlight key parts of this new system and its capabilities.

P-4 : A task-oriented comparison of multiple cheminformatics toolkits

A. Dalke

Andrew Dalke Scientific AB, Göteborg, Sweden

How do you begin to compare different cheminformatics toolkits? Some of the relevant factors include easy-of-use, performance, correctness, and support quality. I will present my analysis of a number of free and proprietary toolkits so that attendees will have an understanding of the background, strengths, and directions of the evaluated toolkits.

Most of the examples will be drawn from the Chemistry Toolkit Rosetta¹. It currently contains 18 examples of common tasks in cheminformatics with user-contributed solutions in OpenBabel², OpenEye's OEChem³, Xemistry's CACTVS⁴, the CDK⁵, RDKit⁶, GGA's Indigo⁷, Schrödinger's Canvas⁸, and the cinfony meta-toolkit⁹. It can be used as a set of recipes for common tasks, and in some cases it's the only non-vendor source of sample programs. It is designed, however, as a resource for comparing toolkits at the fundamental coding level. Someone who knows the CDK but needs to use OpenBabel can use the CTR "translations" as a starting-off point. A developer of a new cheminformatics toolkit can look to the CTR for use-cases and API insight. People who need to select a toolkit as the base for internal research projects can use the CTR solutions to judge how well each toolkit coding style fits the intended audience.

Developing solutions and reviewing project submissions has taught me a lot about toolkit differences. They can all have standard cheminformatics functionality but they have very different views in how to treat aromaticity and implicit hydrogens, manage SMARTS matches, do file I/O and format translation, and how to report errors. CACTVS and cinfony stand out because the former shows how modifying TCL's control structures can simplify toolkit use, while the latter shows that most other APIs can be simplified considerably. I will highlight the main differences.

The CTR explicitly limits itself to API comparisons. I will go beyond that and describe some of what I have learned from hands-on experience with most of the toolkits including performance numbers, implementation quality, and domain specialization.

1. Chemistry Toolkit Rosetta. <http://ctr.wikia.com/> (accessed January 2011).
2. OEChem. <http://www.eyesopen.com/oechem-tk> (accessed January 2011).
3. Ihlenfeldt, W. D.; Takahashi Y.; Abe, H.; Sasaki, S. Computation and Management of Chemical Properties in CACTVS: An Extensible Networked Approach toward Modularity and Compatibility. *J. Chem. Inf. Comp. Sci.* 1994, 34, 109-116.
4. Steinbeck, C.; Han, Y. Q.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E.L. The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics. *J. Chem. Inf. Comp. Sci.* 2003, 43, 493-500.
5. RDKit. <http://rdkit.org/> (accessed January 2011).
6. Indigo. <http://ggasoftware.com/opensource/indigo> (accessed January 2011).
7. Canvas. <http://www.schrodinger.com/products/14/23/> (accessed January 2011).
8. O'Boyle, N. M.; Hutchison, G.R. Cinfony - combining Open Source cheminformatics toolkits behind a common interface. *Chem. Cent. J.* 2008, 2, 24.

P-6 : Handling of tautomerism and stereochemistry in compound registration

A. Gobbi, M. Lee
Genentech Inc., South San Francisco, USA

Automated registration of compounds from external sources is a requirement dictated by the increasing number of compound acquisitions from external vendors, and by the increasing number of collaborations with external partners. A prerequisite for automating compound registration is a robust module for determining the structural novelty of input against previously registered compounds. Such a structure normalization module needs to be able to take uncertainty about stereochemistry into account, and to identify tautomeric forms of the same compound. It also needs to validate structures for potential mistakes.

Genentech has implemented a structure normalization module based on OpenEye Toolkits.¹ The module is incorporated in a graphical application for single compound registration as well as in scripts for bulk registration. It is also used for checking compounds submitted by our collaborators via partner-specific internet sites.

The widely used MOL2000 format is used to accommodate all collaborators. Additionally, a stereochemical assignment is required to denote how much information is known about the stereochemistry. A structural uniqueness check is performed using a canonical tautomer by generating its canonical SMILES.

The steps taken to validate the chemical structure and generate canonical smiles of the canonical tautomer will be discussed. The integration of the validation module in compound registration pathways will be described.

1. OpenEye Scientific Software, USA. <http://www.eyesopen.com>
2. Sayle, R. Canonicalization and Enumeration_of Tautomers **1999**
 (http://www.daylight.com/meetings/emug99/Delany/taut_html/index.htm).

P-8 : Explora: a new language to define powerful structural queries

T. Hanser, E. Rosser, S. Werner, P. Górný
Lhasa Limited, Leeds, UK

Structural query is a frequent and important activity in many Bio/Chemo-informatics applications. Database search, pattern recognition, fragment enumeration and clustering are only a few examples of where structural queries are key elements. At Lhasa Limited we use structural queries to define 2D toxicophores; they capture the expert knowledge in this field in the form of structural alerts. From an abstract view point, structural queries describe a domain in the chemical space that is of interest in a given context. In most contexts, the level of precision in defining the structural scope of a query is at the same time critical and challenging. Current approaches are based on generic atoms or bonds (e.g. A, Q, X, etc.) or more sophisticated descriptors like SMARTS strings. In some cases, these techniques are not flexible enough and they are often not intuitive. In order to create very accurate toxicophore alerts we have developed Explora, a powerful structural query language. This work is an extension of the development of the concept of L-Patterns¹ (logical Markush structures).

Explora is an expression language that is used to define constraints on individual atoms, bonds or the structure as a whole. The user defines a query structure that can be seen as a *hard coded* scaffold; this scaffold is then further refined with scripted constraints expressed in Explora. Each atom, bond or the structure as a whole can be associated with an Explora expression. Explora provides a collection of ready-to-use operators, functions and descriptors and can be easily extended with new functionalities written in Java. Explora has been designed to be very intuitive and flexible as shown in the following list of examples:

Atom expression: **charge** > 0 and **SP2**
 Atom expression: **type** == **O** and (not **ALCOHOL** or **alphaTo(AMINE)**)
 Bond expression: **type** == {**DOUBLE**, **AROMATIC**} and **fusion** == **NONE**

Structure expression: **molecularWeight** < 500 and **logP** < 5

Explora can be easily and seamlessly incorporated into any structural search algorithm.

The Explora expressions are compiled into Java objects and provide very fast evaluation times. We have written a user-friendly editor to facilitate the input of Explora expressions. Explora supports calls to predefined queries allowing a recursive definition of structural motifs in a more intuitive and more flexible way than SMARTS strings do.

Current development of Explora includes new physicochemical descriptors, topological accessibility, additional ring atom types (e.g. bridge head) and hybridization types. Explora's extensible library of functions combined with logical operators provides a new structural query paradigm with a rich, fast and intuitive expressiveness. Explora is not specific to toxicology and can thus be used in any Bio/Chemo-informatics context to drive efficient structural search.

In this communication we present the language and its development environment along with examples. We will also present the current effort to extend the language and a perspective in the general context of structural search.

1. Hanser Th.; Rosser E.; Werner S. L. -Patterns : A novel perspective on structure class definition and search in chemical structural spaces *5th Joint Sheffield Conference on Chemoinformatics*, Sheffield, UK 2010

P-10 : Revisiting the dataflow paradigm for chemical information processing

W.-D. Ihlenfeldt

Xemistry GmbH, Königstein, Germany

The use of dataflow processing software has become an indispensable component in the toolbox of the chemical information processing practitioner. Highly successful products such as Pipeline Pilot¹ or KNIME² are probably in daily use in most IT support sections in pharmacological research and related fields.

The quick assembly of re-useable components which are linked to processing pipelines has in many cases significant advantages over more traditional approaches such as the development of stand-alone custom software or the use of chemistry-aware scripting environments such as Cactvs/Tcl³, OpenEye/Python⁴ or Indigo/Python⁵, especially in addressing ad-hoc problems which do not reoccur on a regular basis.

Nevertheless, a problem often arises when there is a requirement for a processing step which is not covered by the collection of standard nodes. Standard systems such as those cited above do offer custom scripting nodes, but the scope of operations possible in these nodes, plus their oftentimes rather awkward interaction with the rest of the nodes and the overall system limits their usefulness in many usage scenarios.

Scripting toolkits which generally offer much more advanced capabilities for custom data object operations on the other hand possess a far steeper learning curve, and thus are often unsuited for the casual developer.

Interestingly, the very first example of a chemical information pipeline processing environment⁶ was based on a data object and object interaction model which is notably different from the current mainstream systems. It was implementing a desktop metaphor with identifiable and accessible individual objects of different classes instead of opaque, uniform data streams, and a scripting system which controlled the complete environment instead of just a local node. All nodes were actually composed of script snippets, instead of providing a special type of node that is customizable beyond adjustable parameter sets.

We have revisited this alternative approach toward dataflow system design and present a new prototype implementation of a dataflow processing system (project name Cactvs/Factory) for chemical information. It is built on advanced software design techniques such as multi-interpreter isolation, multi-threading between and within nodes, near complete platform independence, and enables universal script-controlled access

from the lowest chemistry sub-object attributes to the overall node assembly. This approach combines the benefit of simple node reusability with the power of classical scripting environments.

We will discuss numerous unique operations possible in this system.

<http://accelrys.com/products/pipeline-pilot/>

<http://knime.org/>

<http://www.xemistry.com>

<http://www.eyesopen.com>

<http://ggasoftware.com/opensource/indigo>

1. Ihlenfeldt, W. D., Takahashi, Y., Abe, H. Dataflow processing in a global networked context. A solution for the computational methods pool management problem, *Proc. 28th Annual Hawaii International Conference on System Sciences*, **1995**, 227-236

P-12 : Managing chemical libraries using Screening Assistant 2.0

V. Le Guilloux¹, S. Bourg², J. Dubois-Chevalier^{1,3}, L. Colliandre¹, A. Arrault⁴, P. Vayer⁴, L. Morin-Allory¹

¹ *Institut de Chimie Organique et Analytique (ICOA), Université d'Orléans, UMR CNRS 6005, B.P. 6759, rue de Chartres, 45067 Orléans Cedex 2, FRANCE*

² *Fédération de Recherche, "Physique et Chimie du Vivant" Université d'Orléans-CNRS; FR 2708, Avenue Charles Sadron, 45071 Orléans Cedex 2, FRANCE.*

³ *Laboratoire d'Informatique Fondamentale d'Orléans, Université d'Orléans, Rue de Chartres, 45067 Orléans Cedex 2, FRANCE*

⁴ *Technologie Servier, 27 rue Eugène Vignat, 45000 Orléans, FRANCE*

High throughput screening (HTS) is a routinely used method to discover new chemotypes having a desired biological profile¹. Up to millions of molecules can be screened in a single HTS campaign nowadays, and chemoinformatics and diversity analysis are appropriate tools to design screening libraries and manage the generated data. Despite recent valuable efforts to develop free and open-source chemoinformatic tools (e.g. Bioclipse, CDK-Taverna, Scaffold Hunter...), only few free softwares are dedicated to assist the screening process and facilitate the selection and management of chemical libraries and the associated biological data.

We present Screening Assistant 2.0 (SA2), a free and open-source JAVA software dedicated to the storage and analysis of small to very large chemical libraries intended to be used in screening campaigns. SA2 stores unique chemical structures in SDF format using a MySQL database, and associates various standard pre-computed descriptors as well as user-defined properties that can be imported in a flexible way. Various chemoinformatics functions are available in the software, including: management of chemical providers, sub-structure search, similarity search, filtering, interactive visualization of chemical spaces in 2D using principal component analysis or Self-organizing maps, dynamic creation and storage of new chemical spaces based on the Delimited Reference Chemical Subspaces methodology², diversity analysis and diverse subset extraction, etc. SA2 is based on the professional and modular NetBeansPlatform that eases the addition of new functionality to the software. We illustrate the use of SA2 to describe and compare various chemical libraries. The program and source code are made freely available under the GNU General Public License.

1. Mayr, L.M.; Bojanic, D.; Novel trends in high-throughput screening, *Current Opinion in Pharmacology* **2009**, 9, 580–588
2. Le Guilloux, V.; Colliandre, L.; Bourg S.; Guénegou, G.; Dubois-Chevalier, J.; Morin-Allory, L.; Visual characterization and diversity quantification of chemical libraries. 1) Creation of Delimited Reference Chemical Subspaces, Submitted to *J. Chem. Inf. Model*

P-14 : High-throughput structure analysis and descriptor generation for crystalline porous materials

R. L. Martin, T. Willems, C. H. Rycroft, Prabhat, M. Kazi, M. Haranczyk
Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

Porous materials such as zeolites and metal organic frameworks have been of growing importance as materials for energy-related applications such as CO₂ capture, hydrogen storage, and as catalysts for oil refinement. Very large databases of these structures are being developed, and so there is a requirement for tools to analyze and screen structure libraries – which can contain millions of entries – to discover materials with certain properties important to these applications. Critical to the success of this endeavor are tools and approaches which enable automatic, unsupervised analysis of these materials, as well as development of suitable descriptors which encode chemical information, and which can be used for screening.

We present two approaches to analyze structures and calculate relevant descriptors in a high-throughput manner. These approaches – Voronoi decomposition and the Fast Marching Method - rely on computational geometry and partial differential equation (PDE)-based methods respectively.

The Voronoi decomposition performed for a given arrangement of atoms in a periodic domain provides a graph representation of the void (accessible) space (see Figure 1). The resulting Voronoi network is analyzed to obtain the diameter of the largest included sphere and the largest free sphere, which are two geometrical parameters that are frequently used to describe pore geometry. Accessibility of nodes in the network is also determined for a given guest (probe) molecule and the resulting information is later used in Monte Carlo sampling of accessible surfaces and volumes (see Figure 2). Further analysis provides information on topology of void space, in particular dimensionality of channel systems present in a material. The obtained structural descriptors can be used in screening.

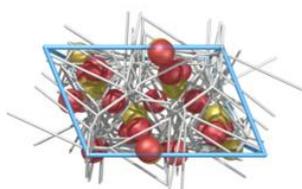


Figure 1: Voronoi network for YUG zeolite

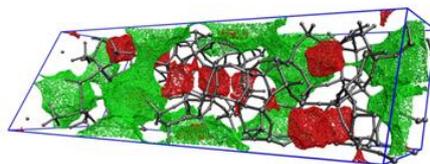


Figure 2: Monte Carlo sampling of accessible surface for DDR zeolite (probe accessible surface: green, lighter gray; not accessible: red, darker gray)

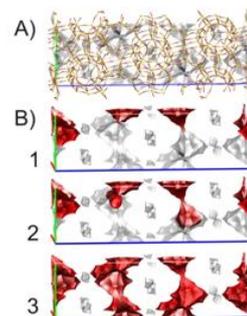


Figure 3: Propagating front explores the void space of DDR zeolite

Tools for automatic structure analysis are essential in order to enable high-throughput characterization of porous materials using molecular simulations. Such characterization is needed not only to perform brute force screening but also to generate reference datasets required to build and tune chemoinformatics screening approaches. A key aspect of this analysis is determination of the accessibility of pores to guest molecules. In our approach, the void space within the structure is explored by performing a PDE-based front propagation on a discrete grid overlaid upon the system (see Figure 3). The positions which a guest molecule can occupy are determined at each discrete position and orientation. The Fast Marching Method is then applied to segment this grid into distinct marches, which are post-processed to detect channels and inaccessible pockets. Descriptors resulting from this approach include channel and pocket locations and volumes, traversable dimensionality of the structure, and an approximation of the entropy of the guest molecule. The unique feature of our approach is that we can model the guest molecule as a real, flexible molecule rather than a spherical probe as assumed in other available approaches.

The complete spectroscopic data of an experiment is attached to the experiment. It can be transferred into the database directly from the measuring instrument if connected to the network, even automatically when following a naming scheme. In many cases (PDF, Office, images, proprietary formats), a preview image is generated instantly, allowing a quick inspection. The (centrally stored) data can be further interpreted by the scientist using (locally installed) software.

Researchers can search their own data, open data created by colleagues and data of former group members within the ELN, making the collected data much more valuable than handwritten journals. Search criteria include reaction substructures, search within the full texts, reaction parameters and the availability of spectra. The program can compare similar experiments and highlight their differences. Screening tables for publications representing a small subset of a large number of experiments can be designed within the application and exported to Excel.

In conclusion, *open inventory* facilitates the creation and analysis of experimental data with an emphasis on the requirements of the academic world. Users can extend the software to the requirements of their field of research and are not bound to a commercial software that stores their data in a black box. It is already in use at various universities and research institutes, but also companies and security authorities².

1. Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.*, **1989**, 29, 97-101.
2. (a) Rudolphi, F.; Gooßen, L. J. Ein elektronisches Laborjournal als Open- Source- Software. *Nachr. Chem.*, 2010, 58, 548-550; (b) Rudolphi, F.; Gooßen, L. J.; Stesycki, M. Leistungsstark und durchdacht. *Chemie Report* 2010, 5, 16-17; (c) Rudolphi, F.; Gooßen, L. J. Ein elektronisches Laborjournal Chem. Unserer Zeit, 2011, in print.

P-18 : CSRML – a new and open exchange format for chemical knowledge

C.H. Schwab¹, B. Bienfait¹, J. Gasteiger¹, T. Kleinöder¹, J. Maruszyk¹, S. Ringeissen², O. Sacher¹, A. Tarkhov¹, L. Terfloth¹, C. Yang³

¹ *Molecular Networks GmbH, Erlangen, Germany*

² *L'Oréal Recherche, Aulnay-sous-Bois, France*

³ *Altamira LLC, Columbus, OH, USA*

Chemical substructures (or subgraphs) store chemical knowledge and are widely used since a long time in cheminformatics applications such as substructure searching, fingerprinting or for structural alerts. However, the existing standards still have some insufficiencies. Although, most of these standards provide means for defining complex queries from a chemical structure point of view, the inclusion of physicochemical properties in the query or the annotation of any meta-information are some of the limitations. Furthermore, a seamless interconversion of the different standards is usually not possible and no mechanisms to validate a query definition and its semantics are specified.

In order to overcome these limitations of existing standards, the well-structured, XML-based standard specification, the Chemical Subgraph Representation Markup Language (CSRML), is proposed and presented. CSRML supports a flexible mechanism to annotate meta-data and properties at each level of a substructure definition and provides means for user-defined extensions. The open specification of CSRML also includes mandatory test cases and can be used as an exchange format. Furthermore, a graphical editor, Chemical Subgraph Editor (CSE), has been developed and is freely available to conveniently define, edit and annotate CSRML-based queries and documents.

The presentation provides insights into the specification of the open CSRML format, briefly introduces CSE and shows an application of CSRML in the area of chemical safety assessment, in particular a workflow algorithm developed as an alternative method to animal testing to predict from the structure of a molecule its potential to trigger skin irritancy.

P-20 : Comparison of protein structural motifs – challenges and algorithms. Improved approach and case studiesR. Svobodová Vařeková^{1,2}, D. Sehnal¹, L. Pravda¹, J. Oppelt¹, C. Ionescu¹, S. Geidl¹, J. Koča¹¹National Centre for Biomolecular Research, Masaryk University, Brno, Czech Republic²ANF DATA, a Siemens company, Prague, Czech Republic

Nowadays, a large amount of information about the 3D structure of proteins is available, and more and more structures are being solved every year due to the advances and increased availability in experimental techniques. This richness of data has given rise to novel kinds of analyses and interpretation. For example, we can study protein structural motifs like binding sites, secondary structure elements, cavities and tunnels. The data processed this way can help identify the main characteristics of important protein motifs, which can consequently be used as patterns in drug design¹, to understand the relationship between structure and function, to classify proteins and research their evolutionary conservation, etc.

In order to perform these analyses, we need to compare large sets of protein structural motifs, i.e. superpose them, quantify the differences among them, and eventually calculate their model. This process includes extracting the motifs from the complete structures, finding the most relevant pairing between the atoms therein and finally optimally fitting the motifs².

The methods of searching for motifs closely depend on the nature of the motif in question – from a simple extraction of secondary structure elements, to a sophisticated search for tunnels. Finding the proper atom pairing is equivalent to a graph isomorphism problem, and therefore has NP complexity². Heuristic approaches are used to solve the problem – i.e., dividing atoms into classes, employing descriptors, subgraph matching etc.. State of art algorithms for fitting molecules or motifs use quaternion algebra³, which has linear time complexity relative to the number of atoms. The next level of the challenge is to process multiple motifs, in which case we may resort to different pivoting, scoring or common core approaches.

We present an overview of methods and software tools for superposition (denoted also superimposition) of structural motifs and we discuss their strong and weak points. We also present one of the latest developments in the field, the web server SiteBinder. SiteBinder provides an effective, intuitive and user-friendly solution for the superposition of multiple protein structural motifs. This software performs the atom pairing via heuristics tailored to the protein environment, applies quaternion algebra approaches for fitting the motifs, and includes recent algorithms for processing multiple motifs⁴, also adapted to specific features in protein motifs. Additionally, we include a discussion on the performance of some of the currently available software tools, supported by three case studies based on the comparison of large sets of biochemically important motifs. The first case study focuses on the analysis of all zinc finger structures contained in the Protein Data Bank (more than 500 motifs). The second case study compares the structures of 66 PA-IIL sugar binding sites containing six different sugars, in the search of specific structural trends. In the last case study we attempted to superimpose 10 BH3 domains from several pro-survival and pro-apoptotic proteins in order to determine if the current classification of these apoptotic factors is related with BH3 domain structure differences.

1. Baran, I.; Svobodová Vařeková, R.; Parthasarathi, L.; Suchomel, S.; Casey, F.; Shields, D.C. Identification of Potential Small Molecule Peptidomimetics Similar to Motifs in Proteins. *J. Chem. Inf. Model.* **2007**, *47*, 464-474.
2. Eidhammer, I.; Jonassen, I.; Taylor, W. R. Protein Bioinformatics: An algorithmic approach to sequence and structure analysis. J. Wiley Sons Ltd.: Chichester, 2004.
3. Coutsias, E. A.; Seok, C.; Dill, K. A. Using quaternions to calculate RMSD, *J. Comp Chem.* **2004**, *25*, 1849-1857.
4. Wang, X.; Snoeyink, J.: Defining and Computing Optimum RMSD for Gapped and Weighted Multiple-Structure Alignment. *IEEE/ACM Trans. Comput. Biology Bioinform.* **2008**, *5*, 525-533.

P-22 : Barcode of small organic molecules and biological molecules

X. Wang, M. Jiang, B. He

Key Laboratory for NeuroInformation of Ministry of Education, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China

Network analysis of molecular structure of both small molecules and large biological molecules has been well studied and it constitutes the theoretical basis of the molecular representation formats of SMILES and IhCHI. Graph theory based network analysis is also the foundation of many topological molecular descriptors. In this work, we apply the recent developments of the persistent homology theory to examine the persistent homological properties of molecular networks.

Barcode here refers to a representation of persistent ranks of homology groups (i.e. Betti numbers) during a filtration process of a network and the filtration could be thought as a growing process of a network. Visual inspection of barcodes distinguishes long-lived topological features during the filtration process from short-lived ones (considered as topological ‘noise’). Different filtration schemes were examined based on the topological path, paired distance and interactions for the generation of barcodes of both small organic molecules and large biological molecules, such as proteins and DNA. An illustrative example is given in Figure 1 to show the barcode of caffeine based on the Rip’s filtration. The possible applications of the persistent barcodes are investigated in the study of the structural diversity, molecular descriptors, molecular dynamics simulation analysis and other areas.

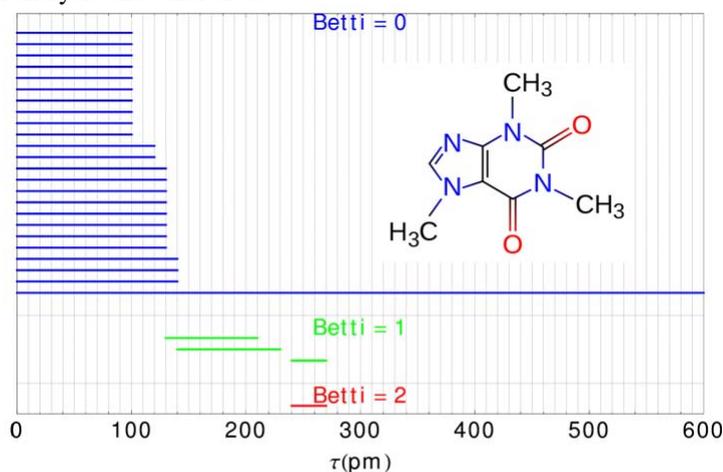


Figure 1. Barcode of the molecule of caffeine based on the Rip’s filtration process. The bars with a Betti number of 0 encode the bond lengths, and the first two bars with a Betti number of 1 show the birth and death of two ring structures, while the other bars show the features not available from the covalent-bonded structure.

P-24 : Multitargeting in Alzheimer's disease: from in silico design of secretase inhibitors to in-vivo experimentsG. Boursheh, D. Marcus, A. Rayan, A. Goldblum
Institute for Drug Research, Jerusalem, Israel

Alzheimer’s disease (AD) is one of the most common neurodegenerative diseases today. AD is characterized by two main proteins that accumulate in the Central Nervous System; hyper-phosphorylated tau proteins that form tangles inside the axons and amyloid beta that accumulates outside of the nerve cells. ¹ Amyloid-beta ($\text{A}\beta$) peptide is a major cause of the pathogenic cascade leading to the development of AD, both as soluble oligomers or as aggregates. The formation of these $\text{A}\beta$ aggregates occurs as a result of the degradation of Amyloid Precursor Protein (APP) by a number of sequential proteolysis steps, starting with beta amyloid cleaving enzyme 1 (BACE-1) that cleaves the APP and activates the toxic pathway, continuing with the gamma secretase complex protein (GS) that releases the $\text{A}\beta$ fragment, which is a

partially soluble end-product. This end product both in its soluble form and as it starts to accumulate, triggers the toxic cascade leading eventually to neuronal loss.²

Simultaneous blocking of both enzymes that initiate and continue the process of formation of A β could be a path to prevent or stop the neurodegeneration of this particular source, as well as supply molecular tools for studying more in depth the exact pathways that lead to such neurodegeneration. Moreover, this approach has been recently shown to be effective in Tg mice.³

Our lab developed a generic algorithm to find the best solutions among an enormously huge number of possibilities. Iterative Stochastic Elimination (ISE)⁴ has been employed to find optimal states of protein structures and interactions and to discover novel molecules based on their properties. Based on known BACE-1 and GS inhibitors, we constructed an ISE model which has the ability to "fish" molecules with inhibitory characteristics from databases containing millions of molecules. Molecules found by that method have diverse structures, are different than those that we "learned" from and may eventually be developed into drugs.

Molecules that are discovered by the model will subsequently undergo -

- Docking: using *in silico* methods we evaluate the affinity of the potential inhibitors to their targets, if the target structures are available.
- Assigning the best molecules and purchasing them
- *In vitro* experiments of BACE-1 and GS inhibitory activity, and lowering of Amyloid Beta level activity.
- *In vivo* behavioral experiments in Tg-mice of AD models

Following that process, and having recent experience from discovering novel acetylcholinesterase inhibitors, we expect to discover lead compounds for developing bi-functional molecules that would have simultaneous modulating properties towards both secretases.

1. Hardy, J.; Allsop, D. Amyloid deposition as the central event in the aetiology of Alzheimer's disease. *Trends Pharmacol. Sci.* **1991**, 10, 383–88.
2. Nistor, M.; et al. Alpha- and beta-secretase activity as a function of age and beta-amyloid in Down syndrome and normal brain. *Neurobiol Aging* **2007**, 28, 1493–1506.
3. Chow, V.W.; et al. Modeling an anti-amyloid combination therapy for Alzheimer's disease. *Science Translational Medicine* **2010**, 2, 1-23.
4. Glick, M.; Rayan, A. and Goldblum, A. A stochastic algorithm for global optimization and for best populations: A test case of side chains in proteins. Proceedings of the National Academy of Sciences of the United States of America **2002**, 99, 703-708

P-26 : New structural alerts for phospholipidosis

L. Fisk, R. Williams

Lhasa Limited, 22-23 Blenheim Terrace, Woodhouse Lane, Leeds LS2 9HD, United Kingdom

The loss of drug candidates during the later stages of pre-clinical development is a matter of concern to the pharmaceutical industry. One finding that threatens drug development is phospholipidosis, which is observed for many cationic amphiphilic drugs during repeat-dose toxicity studies. Phospholipidosis (PLOSIS) is an intracellular accumulation of phospholipids, often in lysosomes, manifested as lamellar bodies detected by electron microscopy. PLOSIS may be an adaptive response and has not been unequivocally associated with toxic effects, however a PLOSIS positive result could prevent further development of a drug candidate or prompt additional testing to prove that the effects observed in animals are not relevant to humans. To overcome this problem, a lot of effort was put in developing methods to predict potential ability of chemicals to induce PLOSIS at an early stage of drug development.

A model built on the physicochemical properties pKa and ClogP was developed for prediction of PLOSIS [Ploemen et al]. Improvement in its predictive performance was achieved in two ways: a) manual adjusting of Ploemen model rules to increase sensitivity [Pelletier et al] ; b) refinement by consideration of pharmacokinetics, measured by the volume of distribution (V_d) [Hanumegowda et al]. Additionally a

structural alert for alkylamines based loosely on Ploemen's model was implemented in a knowledge base system, where predictive rules are derived and implemented by human experts. Other techniques such as machine learning methods for building a PLOsis prediction model have also been described. A Bayesian model for PLOsis that uses a number of molecular descriptors evaluated and selected by experts was able to increase correct predictivity for a data set of 125 compounds in comparison to Ploemen's model [Pelletier et al]. A support vector machine model utilising circular fingerprints descriptors, successfully predicted the activity of a larger data set of 185 chemicals [Lowe et al]. Despite the availability of several PLOsis models, the activity of some chemical classes are consistently poorly predicted.

Further developments of a knowledge base system that already incorporates the structural alert discussed above have been investigated using the previously mentioned phospholipidosis data set consisting of 185 compounds [Lowe et al]. The existing structural alert displays a sensitivity of 25% for this data set – this could be improved by the “relaxation” of structural requirements, however this would lead to decrease in specificity. Visual analysis identified three potential classes (piperazine-containing compounds, aminoglycosides and erythromycin-like macrolide antibiotics), which are known to cause PLOsis in vivo but were shown to be problematic to predict by previous models [Ploemen et al, Pelletier et al]. Projected improvement of sensitivity from 25 to 45% and concordance from 59 to 69% is expected after investigation of SAR of these classes followed by implementation of new structural alerts for PLOsis.

1. Ploemen, J. P.; Kelder, J.; Hafmans, T.; van de Sandt, H.; van Burgsteden, J. A.; Salemink, P. J.; van Esch, E. Use of physicochemical calculation of pKa and CLogP to predict phospholipidosis-inducing potential: a case study with structurally related piperazines. *Exp Toxicol Pathol.* **2004**, *55*, 347-355.
2. Pelletier, D. J.; Gehlhaar, D.; Tilloy-Ellul, A.; Johnson, T. O.; Greene, N. Evaluation of a published in silico model and construction of a novel Bayesian model for predicting phospholipidosis inducing potential. *J Chem Inf Model.* **2007**, *47*, 1196-1205.
3. Hanumegowda, U. M.; Wenke, G.; Regueiro-Ren, A.; Yordanova, R.; Corradi, J. P.; Adams S. P. Phospholipidosis as a function of basicity, lipophilicity, and volume of distribution of compounds. *Chem Res Toxicol.* **2010**, *23*, 749-755.
4. Lowe, R.; Glen, R. C.; Mitchell, J.B. Predicting phospholipidosis using machine learning. *Mol Pharm.* **2010**, *7*, 1708–1714.

P-26 : To Hit or Not To Hit - That Is The Question! Structure-Based Druggability Predictions for *Pseudomonas aeruginosa* Targets

A. Sarkar, A. Krasowski, R. Brenk *

Division of Biological Chemistry & Drug Discovery, College of Life Sciences, University of Dundee, Dundee, UK

Structural genomics initiatives have been invaluable in producing information about disease-related proteins. Drug discovery initiatives benefit immensely from this information. However, with limited financial and human resources available, it becomes necessary to be selective about which targets to pursue. Druggability, which is commonly defined as the ability of a protein binding site to bind drug-like, orally bioavailable molecules with high affinity [1-3], is an important concept in target selection. We recently developed a method called Drug-Pred to predict druggability solely based on the 3D structure of the target. Drug-Pred was derived by correlating descriptors describing the size and polarity of a binding site to the druggability of a protein binding site using Partial Least Squares-Discriminant Analysis (PLS-DA). The resulting model was capable of discerning druggable binding sites from non-druggable ones with > 90% accuracy. Here, we apply this method for genome analysis of *Pseudomonas aeruginosa* targets.

P. aeruginosa is a major pathogenic agent in opportunistic and nosocomial infections. Over 500 crystal structures of potential targets from *P. aeruginosa* (aeropath.lifesci.dundee.ac.uk) have been determined in the RCSB Protein Data Bank (PDB). Several challenges had to be overcome in order to apply Drug-Pred in an automated way to such a large set of target structures, including identification of appropriate binding sites on proteins and classifying the protein as druggable/non-druggable by consolidating predictions for multiple structures of the same protein. Methods were developed to address all these problems and resulted in druggability predictions for all protein structures currently available for the *P. aeruginosa* proteome. The

current status of this project, a comparison of the results with a ligand-based druggability prediction method, along with future plans shall be discussed herein.

1. Cheng, A.C.; Coleman, R.G.; Smyth, K.T.; Cao, Q.; Soulard, P.; Caffrey, D.R.; Salzborg, A.C.; Huang, E.S. Structure-based maximum affinity model predicts small-molecule druggability. *Nature Biotech.* **2007**, *25*, 71-75.
2. Hajduk, P.J.; Huth, J.R.; Fesik, S.W. Druggability indices for protein targets derived from NMR-based screening data. *J. Med. Chem.* **2005**, *48*, 2518-2525.
3. Sheridan, R.P.; Maiorov, V.N.; Holloway, M.K.; Cornell, W.D.; Gao, Y.-D. Drug-like density: a method of quantifying the “bindability” of a protein target based on a very large set of pockets and drug-like ligands from the protein data bank. *J. Chem. Inf. Model.* **2010**, *50*, 2029-2040.

P-30 : Data integration and analysis – how a scientist would like it

P. Davis

Schrödinger, Inc., 120 West 45th Street, 17th Floor, New York, New York 10036, US

Drug discovery scientists need an integrated view of chemical, biological and preclinical data (corporate and public) in order to facilitate rapid and accurate analysis, interpretation and communication of project results to the widest possible audience. As each project iteration feeds into the next new compounds, test results, data extrapolations, team decisions, workflow adjustments and scheduling of resources must be taken into account so that the “next steps” in the project are the “right” ones.

In this talk we will look at a selected case studies highlighting how enterprise discovery informatics software has been successfully used in this context. We will examine the set of characteristics software applications should exhibit in order to have the best chance of meeting the diverse needs of the project team scientists they aim to support.

P-32 : Harmonization of cheminformatics services after a recent acquisition: taking the best of two worlds

L. Ridder¹, J. Wertenbroek¹, R. van Schaik¹, J. Voigt², B. Sherborne², J. de Vlieg^{1,3}, M. Wagener¹

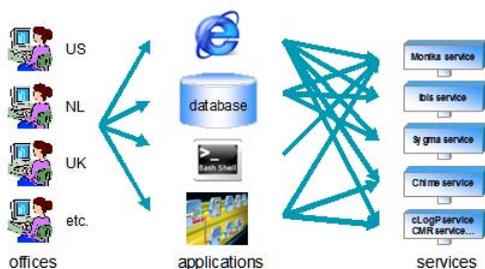
¹ *Molecular Design and Informatics, Merck Research Laboratories, Oss, The Netherlands*

² *Chemistry Modeling and Informatics, Merck Research Laboratories, USA*

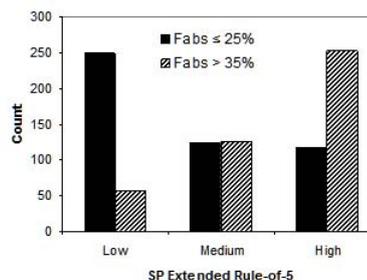
³ *Computational Drug Discovery, Center for Molecular & Biomolecular Informatics, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands*

After Organon was acquired by Schering-Plough, the use of *in silico* calculators in the combined pharmaceutical R&D organization had to be harmonized. The main objective was to achieve consistency in calculated properties used in project decision making. It involved calculated properties in data visualization tools, documentation and databases, and their usage across different disciplines and different geographic locations. A further objective was to optimize the usage of *in silico* calculations by maximizing quality, relevance, usability and accessibility of the models.

A five stage harmonization strategy was implemented including identification of calculators, collection of evaluation datasets, evaluation and selection of calculators per property and a consistent update protocol. Based on this harmonization strategy a set of preferred calculators was defined and globally adopted. A web-service oriented infrastructure was implemented which allows flexible access to the selected set of preferred calculators from different applications and geographic sites. Finally, a new web application was developed to provide direct and easy access to the *in silico* calculators for users from all disciplines and sites.



Web-service oriented infrastructure



Results of the combined classification scheme applied to the rat PK dataset.

To illustrate the harmonization process, the evaluation and comparison of the well known Rule-of-5 used in Schering-Plough and a set of less stringent rules implemented in Organon is presented in more detail. It was based on rat pharmacokinetic data of 917 compounds originating from more than 50 discovery projects in the combined company. The results identified complementary strengths of both sets of rules, leading to the implementation of a combined classification scheme with enhanced applicability especially in projects in which finding active compounds with good physico-chemical properties is a challenge.

P-34 : Visual characterization and diversity quantification of chemical libraries: Creation and use of Delimited Reference Chemical Subspaces (DRCS)

L. Colliandre¹, S. Bourg², J. Dubois-Chevalier^{1,3}, Luc Morin-Allory¹, V. Le Guilloux¹

¹ Institut de Chimie Organique et Analytique (ICOA), Université d'Orléans, UMR CNRS 6005, B.P. 6759, rue de Chartres, 45067 Orléans Cedex 2, FRANCE

² Fédération de Recherche, "Physique et Chimie du Vivant" Université d'Orléans-CNRS; FR 2708, Avenue Charles Sadron, 45071 Orléans Cedex 2, FRANCE.

³ Laboratoire d'Informatique Fondamentale d'Orléans, Université d'Orléans, Rue de Chartres, 45067 Orléans Cedex 2, FRANCE

High Throughput Screening (HTS) is routinely used to explore internally or commercially available chemical libraries in order to discover new chemotypes having a desired biological profile. Despite the increasing capabilities of HTS to test compounds, only a few million of compounds can be tested making diversity analysis a method of choice to design chemical libraries and prioritize their screening.

Various methodologies can be used to explore, visualize and compare chemical libraries.^{1,2} In this poster, we present a new method based on the creation of Delimited Reference Chemical Subspaces (DRCS).³ A set of 16 million screening compounds from 73 chemical providers has been gathered, resulting in a database of 6.63 million standardized and unique molecules. These molecules have been used to create three representative spaces based on three different sets of chemical descriptors. A robust PCA model for each space has been determined, whereby molecules are projected in a reduced 2D viewable space. The specificity of our approach is that each reduced space has then been delimited by a representative contour encompassing a very large proportion of molecules, reflecting its overall shape, and creating the DRCS. This allows a rapid and easy visual comparison of chemical libraries.

Moreover, the DRCS methodology has been applied to compare the relative molecular diversity of chemical libraries using a DRCS based diversity index which is cell-based and independent of the size of the library.⁴ The delimitation of the chemical subspaces makes it possible to use numerous mathematical methods to compute this diversity. Various chemical libraries has been mapped, characterized and compared in terms of diversity.

This methodology is bundled in Screening Assistant 2.0,⁵ a free and open-source JAVA software.

1. Dubois, J.; Bourg, S.; Vrain, C.; Morin-Allory, L., Collections of Compounds - How to Deal with them? *Current Computer - Aided Drug Design* **2008**, 4, (3), 156-168.
2. Verheij, H. J., Leadlikeness and structural diversity of synthetic screening libraries. *Mol Divers* **2006**, 10, (3), 377-88.
3. Le Guilloux, V.; Colliandre, L.; Bourg S.; Guéneqou, G.; Dubois-Chevalier, J.; Morin-Allory, L.; Visual characterization and diversity quantification of chemical libraries. 1) Creation of Delimited Reference Chemical Subspaces. Submitted to *J. Chem. Inf. Model*.
4. Colliandre, L.; Le Guilloux, V.; Bourg S.; Morin-Allory, L.; Visual characterization and diversity quantification of chemical libraries. 2) Creation of DRCS based diversity indices; In preparation.
5. Screening Assistant 2.0: <http://www.univ-orleans.fr/icoa/screeningassistant>

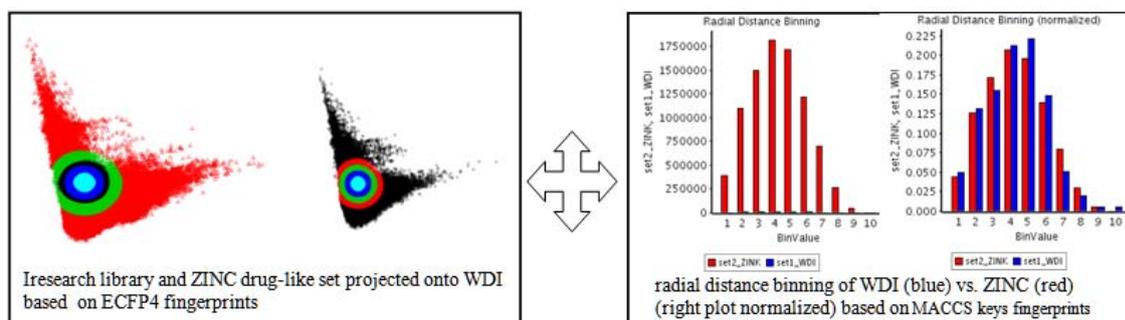
P-36 : Dataset overlap density analysis

A. H. Göller

Bayer Schering Pharma AG, Computational Chemistry, Wuppertal, Germany

The need to compare datasets arises from various scenarios, like mergers, library extension programs, gap analysis, combinatorial library design, or estimation of QSAR model applicability domains. Whereas it is relatively easy to find identical compounds in two datasets, the quantification of the overlap is not straightforward. The various approaches described include pairwise nearest neighbor comparisons,¹ clustering and mixed cluster statistics,² or binning of e.g. rule-of-five property space distributions. The BCUT methodology³ creates a binned N-dimensional space and allows to assess the amount of mixed cells. ChemGPS creates a PCA reference projection based on drug-like and satellite molecules in property space to classify new compounds.⁴

But how does one quantify the overlap of two datasets in a single interpretable number? PCA map projections of dataset A on to dataset B in principle inherently contain this information on overlaps and gaps, but the visual inspection is hampered by the crowdedness of maps of large sets.



A single dataset overlap value is introduced here. A PCA map of the World Drug Index as drug-like reference space is created based on any descriptor set, as exemplified by MACCS, ECFP4, estate or Lipinsky-like descriptors. Any second dataset is projected onto this map. The distance to the center and the sector of the circular 2D projection for each point is calculated. The populations of each such PCA area for each dataset are counted and compared as bar charts of absolute and compound number-normalized populations for each distance bin for the overall radial or sector populations.

By summing up the overlap ratios of all distance bins of the two datasets one finally comes up with single number of the dataset overlap density in a particular descriptor space. By doing so in a set of descriptor spaces one finally creates a signature vector for dataset overlap. The approach can be adjusted for N-dimensional mappings or cube-binning schemes, and as fine-grained as needed for a particular application. It allows to quantify local gaps or overlaps. Proprietary datasets can be compared just by the first N principal components without even seeing the descriptors behind.

1. Turner, D. B.; Tyrrell, S. M. & Willett, P. Rapid Quantification of Molecular Diversity for Selective Database Acquisition. *J. Chem. Inf. Comput. Sci.*, 1997, 37, 18-22.

- Engels, M.F.M.; Gibbs, A.C.; Jaeger, E.P.; Verbinnen, D.; Lobanov, V. S. & Agrafiotis, D. K., A Cluster-Based Strategy for Assessing the Overlap between Large Chemical Libraries and Its Application to a Recent Acquisition . *J. Chem. Inf. Comput. Sci.*, 2006, 46, 2651-2660.
- Pearlman, R. & Smith, K. M. Novel software tools for chemical diversity. *Perspectives in Drug Discovery and Design*, 1998, 9/10/11, 339-353.
- Oprea, T.I. & Gottfries, J. Chemography: The Art of Navigating in Chemical Space *J. Comb. Chem.*, 2001, 3, 157-166.

P-38 : Open source chemical structure generator

J. Peironcely^{1,2,3}, M. Rojas-Cherto^{2,3}, L. Coulier^{1,3}, T. Reijmers^{2,3}, J.L. Faulon⁴, T. Hankemeier^{2,3}

¹TNO Quality of Life, Zeist, The Netherlands

²Leiden University, Leiden, The Netherlands

³Netherlands Metabolomics Centre, Leiden, The Netherlands

⁴University of Evry, Evry, France

Computer Assisted Structure Elucidation is one of the classical fields working at the intersection of chemoinformatics, chemistry, and mathematics. It has been used for decades to discover the chemical structure of unknown compounds. A standard structure elucidation system receives experimental chemistry data of an unknown molecule as input, and outputs a list of hypothetical chemical structures. In this work we introduce, to the best of our knowledge, the first general purpose open source structure generator, which for a given elemental formula produces all non-isomorphich chemical structures that match the formula.

Almost all structure generators have close connections with graph theory to produce their desired output. Interestingly, compounds can be regarded as molecular graphs where atoms and bonds are translated into vertices and edges, to which theorems and algorithms proposed by graph theory can be applied. This ensures that the output is correct, exhaustive, and free of isomorphs. Such methods can be the homomorphism principle (1) used by MOLGEN, or the “canonical augmentation path” proposed by McKay (2) and applied to the generation of some families of graphs and also to generate the chemical universe up to 11 atoms (3). We effectively combine the algorithm described by the canonical augmentation path approach with an open source chemoinformatics library, the CDK (4), to exhaustively generate all non-isomorphich chemical structures for a given elemental formula.

Furthermore, this generator can accept one or multiple non overlapping prescribed substructures. The resulting tool generates all possible non-duplicate chemical structures for a given elemental formula, with the option to generate only those that contain one or multiple non-overlapping fragments. In order to benchmark the tool, we generated chemical structures for the elemental formulas of different molecules, using our tool and the commercially available generator MOLGEN. A discussion comparing the results of both tools is presented.

- Brinkmann, G. *Discrete Mathematical Chemistry*. **2000**, 51, 25 - 38.
- McKay, B. *Journal of Algorithms*. **1998**, 26, 306-324.
- Fink, T.; Reymond, J.-L. *Journal of chemical information and modeling*. **2007**, 47, 342-53.
- Steinbeck, C.; Hoppe, C.; Kuhn, S.; Floris, M.; Guha, R.; Willighagen, E. L. *Current Pharmaceutical Design*. **2006**, 12, 2111-2120.

P-40 : NCI/CADD chemical identifier resolver: indexing and analysis of available chemistry space

M. Sitzmann¹, W-D. Ihlenfeldt², M. C. Nicklaus¹

¹ *Chemical Biology Laboratory, Center for Cancer Research, National Cancer Institute/Frederick, NIH, DHHS, 376 Boyles Street, Frederick, MD 21702, USA*

² *Xemistry GmbH, Auf den Stieden 8, D-35094 Lahntal, Germany*

We present recent developments of our Chemical Identifier Resolver (CIR) and other services available at our web server (<http://cactus.nci.nih.gov>) of the NCI/CADD Group. We will give an overview on how these resources can be integrated and linked in into one's own web applications. CIR also includes open access to the underlying Chemical Structure DataBase (CSDB) which currently indexes approx. 120 million chemical structure records representing about approx. 75 million unique chemical structures from more than 150 chemical structure databases. The de-duplication of chemical structures has been performed on basis of our NCI/CADD Structure Identifiers (FICTS, FICuS and uuuuu),¹ as well as InChI/InChIKey.²

For the set of unique structure records in CSDB we will also report on an analysis of scaffolds, fragments and tautomers. Specifically, tautomerism was found to be possible for more than two-thirds of the unique structures in CSDB.³ A total of 680 million tautomers were calculated from the unique structure record set. Tautomerism overlap within the same individual database (i.e. at least one other entry was present that was really only a different tautomeric representation of the same compound) was found at an average rate of 0.3% of the original structure records, with values as high as nearly 2% for some of the databases in CSDB. Tautomeric overlap across all constituent databases in CSDB was found for nearly 10% of the records in the collection.

1. Sitzmann, M.; Filippov, I. V.; Nicklaus, M. C. Internet Resources Integrating Many Small-Molecule Databases. *SAR & QSAR in Env. Res.* 2008, 19, 1-9.
2. The IUPAC International Chemical Identifier (Version 1.03) - Download. <http://www.iupac.org/inchi/release103.html> (accessed Oct 27, 2010).
3. Sitzmann, M.; Ihlenfeldt, W.; Nicklaus, M. C. Tautomerism in Large Databases. *J. Comput. Aided Mol. Des.* 2010, 24, 521-551.

P-42 : Handling of homology variation in structure representation, patent Markush search, enumeration and visualization

R. Wagner, Sz. Csepregi, N. Máté, A. Baharev, T. Csizmazia, F. Csizmadia
ChemAxon Ltd., Budapest, Hungary

Markush structures are indispensable in combinatorial library design and chemical patent applications for the description of compound classes. The presentation will discuss how an existing molecule drawing tool (Marvin) and chemical database engine (JChem Base/Cartridge) were extended to handle generic features, including: R-group definitions, atom and bond lists, link nodes and repeating units, position and homology variation with and without additional properties¹. The listed Markush features make it possible to support Thomson Reuters patent Markush structures from the Merged Markush Service

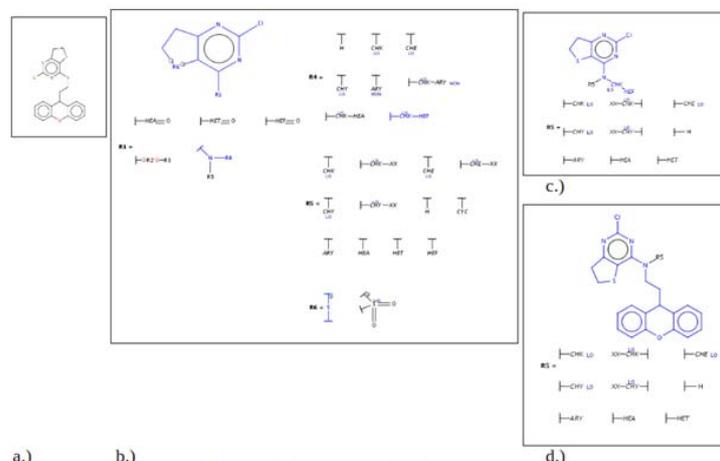


Figure 1: Different Markush search result visualization options. a.) shows the query, b.) colored (blue) Markush diagram, c.) colored Markush diagram with Markush reduction to the hit and d.) similar to c.) but homology groups are also expanded according to the query. CHK, XX, ARY ... etc are Thomson's abbreviations for homology groups.

(MMS) database ² – including editing, storage, search and enumeration and various visualization techniques.

In this presentation, a special focus will be given to homology variation, which can represent a large - possibly infinite - number of specific (sub)structures. Examples of homology groups include alkyl, aryl, aliphatic heterocycle, any ring or ring segment, etc. The database search engine handles homology groups in two ways depending on their nature:

Groups with limited number of represented structures (e.g. halogen) are translated to R-groups. These groups are customizable.

Groups with theoretically unlimited number of structures (e.g. fused heterocycle) are handled by the searching algorithm based on the structural features that define such groups. Any specific query fragment fulfilling the required criteria can match the given group provided that the structural context is appropriate.

Additional homology properties refine the matching and enumeration behavior by restricting the represented structures. These include various chemical properties like atom counts (C1-10), ring size, ring fuseness, saturation, etc. This technology also allows the use of homology groups as part of queries against specific structure databases.

Various visualization techniques illustrate the relationship between Markush structures and the query structures: a.) Coloring and alignment of the matched atoms and bonds. b.) Markush reduction to hit: It displays a simpler Markush formula by reducing Markush features that are fixed by one particular match of the query. c.) During hit reduction homology groups can optionally be chosen to be expanded according to the query. Figure 1. shows an example of searching a query on a Thomson MMS Markush structure with different hit visualization options

Further help of visualization is provided by different enumeration methods which allow the analysis of Markush structures and their enumerated libraries. These methods include full, partial and random enumerations as well as calculation of the library size. Homology groups are enumerated using a sample set of substructures which both fulfill the properties of the group and are chemically reasonable. These sets were obtained by data mining a 100K drug-like database and taking the most frequently occurring fragments.

Examples and validation results will be shown. Enumerates of a given structure need to be found by the search process in the same Markush structure. Therefore searching the Markush with its enumerate is a suitable technique for measuring the correctness of searching. Results showing the amount of correctly found enumerates of several thousand Thomson Reuters MMS Markush structures will be shown.

1. Sz. Csepregi, Representation of Markush structures: From molecules toward patents, ACS meeting, Boston, 2010 <https://www.chemaxon.com/library/scientific-presentations/markush-search/representation-of-markush-structures/>
2. Markush DARC User Manual, Derwent World Patents Index [Online] 2008, 1.9-1.11 http://science.thomsonreuters.com/scientific/m/pdfs/mgr/Markush_Darc_User_Manual.pdf

P-44 : Boosting the predictive reliability of QSAR models.

D. Fourches¹, E. Muratov^{1,2}, A. Tropsha¹

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.*

Maximizing the predictive power of QSAR models has always been a critical goal for cheminformaticians. We posit that careful and thoughtful processing of experimental data should increase the QSAR model accuracy. To this end, we describe a workflow integrating good practices (Figure 1) for both curation and

representation of chemical data prior to computational modeling¹. We also discuss the influence of (i) experimental data variability, (ii) the presence of structural outliers and activity cliffs, and (iii) in-depth neighborhood analysis to efficiently detect prediction outliers. We demonstrate with examples that QSAR models rigorously developed using the above practices can be used to correct erroneous biological data associated with certain compounds, which further increases model accuracy. In summary, we show that thorough data curation and processing are mandatory steps in QSAR model development helping to achieve models with improved predictive power.

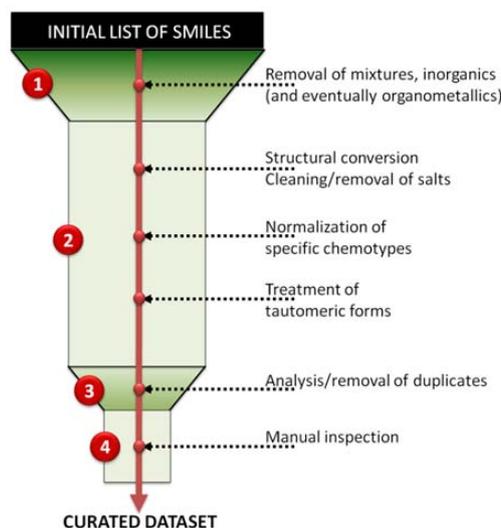


Figure 1. General workflow for chemical dataset curation.

1. Fourches, D.; Muratov, E.; Tropsha, A. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J. Chem. Inf. Model.* 2010, 50, 1189-1204.

P-46 : Fast methods of atomic charge calculation: the electronegativity equalization method for proteins

C. M. Ionescu¹, R. Svobodová Vařeková^{1,2}, D. Sehnal¹, O. Skřehota^{1,2}, J. Koča¹

¹National Centre for Biomolecular Research, Masaryk University, Brno, Czech Republic

²ANF DATA, a Siemens company, Prague, Czech Republic

The concept of atomic charge is a useful way to quantify electronic density, and has been successfully used in explaining and predicting various trends in the chemical behavior of molecules. Atomic charges are essential parameters in various kinds of molecular simulations (molecular dynamics, docking, conformational searches, binding site prediction, etc.), as they are used in computing the electrostatic term of the interaction, and they are involved in QSAR analyses as simple or composite indices¹. Therefore it is desirable to have good knowledge of atomic charges. However, since they are not physical observables, atomic charges must be computed using various methodologies and formalisms. The quantum chemical approach can be very precise, but extremely time demanding, thus its applicability to biomolecules is highly restricted. Fortunately, empirical methods that combine accuracy and speed have been developed. One of them is the Electronegativity Equalization Method (EEM)², which allows for the fast calculation of partial atomic charges with remarkable precision, provided that the proper parameters have been previously determined.

EEM has progressed a lot since its original development by various modifications to the formalism, an increase in the number of supported atom types and charge schemes, implementation within modeling software or on parallel platforms³⁻⁶, etc. However, because of the computational complexity, these studies

worked with very small systems. We are here interested in how to apply this fast method of atomic charge calculation to the biomolecular level.

We have parameterized EEM on large fragments of real proteins. The elements covered by our parameters are those commonly found in proteins (C, H, N, O, S) plus Ca^{++} as a ligand. All the training and test molecules are fragments of real proteins, whose structures are available in the Protein Data Bank and can be directly used in molecular dynamics and docking studies. We predict that our parameters are able to handle full-sized proteins to a reasonable approximation. The charges generated are conformationally dependent, thus extremely suitable for studying phenomena based solely on conformational changes, something other kinds of empirical charge sets are not very good at. We present the complete process of generating our EEM parameters for proteins.

1. Jelfs, S.; Ertl, P.; Selzer, P. Estimation of $\text{p}K_a$ for Druglike Compounds Using Semiempirical and Information-Based Descriptors. *J. Chem. Inf. Model.* **2007**, *47*, 450-459.
2. 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.
3. Rappe, A. K.; Goddard III, W. A. Charge Equilibration for Molecular Dynamics Simulations. *J. Phys. Chem.* **1991**, *95*, 3358-3363.
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 Charge Schemes. *J. Phys. Chem. A.* **2002**, *106*, 7895-7901.
5. 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.
6. Svobodová Vařeková, R.; Jiroušková, Z.; Vaněk, J.; Suchomel, S.; Koča, J. Electronegativity Equalization Method: Parameterization and Validation for Large Sets of Organic, Organohalogen and Organometal Molecules. *Int. J. Mol. Sci.* **2007**, *8*, 572-582.

P-48 : Hole filling in medicinal chemistry libraries

V. le Guilloux¹, A. Arrault², L. Morin-Allory¹, P. Vayer²

¹ Institut de Chimie Organique et Analytique (ICOA), Université d'Orléans, UMR CNRS 6005, B.P. 6759, rue de Chartres, 45067 Orléans Cedex 2, FRANCE

² Technologie Servier, 27 rue Eugène Vignat, 45000 Orléans, FRANCE

Modern drug discovery involves the simultaneous optimization of both biological activities and several physicochemical parameters of candidate molecules. With the emergence of polypharmacology, ADME-tox profiling and high throughput assays, the number of biological data associated with each compound is constantly increasing. Technical limitations, experimental error or solubility issues often lead to datasets containing lots of missing values, making SAR / SPR analysis even more difficult to conduct. In such cases, in-silico and statistical methods, in particular QSAR modeling, are interesting alternatives allowing predicting biological properties based on quantitative description of molecules. Conversely, so-called "hole filling" methods have been widely applied in the Bioinformatics field since 10 years now to replace missing values in micro-array datasets. A large number of methods have been developed to this end, and recent benchmark revealed the best performing ones¹. Hole filling methods have some common paradigms with QSAR methods, in the sense that they exploit local or global correlations to replace missing values by extrapolated ones. Interestingly, they have never been applied to medicinal chemistry datasets.

In this poster, we test the performances of hole filling methods used to predict ADME properties on in-house datasets. Furthermore, we compare the performances of predictions obtained using in-silico and / or in-vitro data as molecular descriptors.

1. Celton, M.; Malpertuy, A.; Lelandais, G.; de Brevern, A.G.; Comparative analysis of missing value imputation methods to improve clustering and interpretation of microarray experiments, *BMC Genomics*, **2010**, Jan 7, 11:15.

P-50 : An open access hierarchy of molecular descriptors

J. Maruszczyk¹, B. Bienfait¹, J. Gasteiger¹, T. Kleinöder¹, O. Sacher¹, C.H. Schwab¹, A. Tarkhov¹,
L. Terfloth¹, C. Yang²

¹ *Molecular Networks GmbH, Erlangen, Germany*

² *Altamira LLC, Columbus, OH, USA*

Molecular descriptors play a crucial role in many areas of Chemoinformatics, such as predictive model building, similarity perception or profiling of chemical compounds. Within the past years Molecular Networks developed a hierarchy of molecular descriptors on different levels of sophistication to represent chemical structures on a sound physicochemical and geometric basis and combined them in the program package ADRIANA.Code¹. The descriptors of ADRIANA.Code have been successfully applied to solve challenges in the area of drug design, ADME and toxicity prediction but also for modeling chemical reactivity or simulating IR and NMR spectra for structure elucidation purposes.

To support the use of computational tools in risk assessment, Molecular Networks will make a standard set of the most versatile and widely applicable descriptors of ADRIANA.Code publicly available to the scientific community for non-commercial use under an open access license. The standard set was selected based on the collaboration with US FDA CFSAN and will also be used within publicly funded projects, e.g., the COSMOS consortium within the EU FP7/Colipa cluster.

This presentation will provide information about the open access descriptors of ADRIANA.Code and demonstrate a recent application of the descriptors for classification models on *Salmonella typhimurium* reverse mutation and rodent carcinogenicity.

1. Gasteiger J. Of Molecules and Humans. *J. Med. Chem.* **2006**, *49*, 6429-6434.

P-52 : Large-scale in silico model building

U Norinder et.al.

AstraZeneca Research and Development | Innovative Medicines, Södertälje, Sweden

Drug discovery is a time-consuming and costly process.¹ Consequently, predictive *in silico* methods are increasingly used in order to not only lower the costs of drug development but to also obtain more accurate estimations of various biopharmaceutical properties, e.g. solubility, hERG liability and membrane permeability, for (virtual) compounds under investigation.² AstraZeneca has developed platforms³ and machine learning procedures⁴ to try to address these issues and needs. The presentation will describe some of the methods and procedures and discuss lessons learned in relation to modeling large datasets in more automated industrialized settings. Furthermore, the presentation will show how various types of interpretations with respect to predictions can aid chemists in the design process of novel compounds.

1. van de Waterbeemd, H.; Gifford E. ADMET in silico modelling: towards prediction paradise?, *Nat. Rev. Drug. Discov.* **2003**, *2*, 192–204.
2. Wagener, M.; van Geerestein, V. Potential Drugs and Nondrugs: Prediction and Identification of Important Structural Features, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 280–292.
3. The AutoQSAR and AZOrange platforms.
4. <http://www.his.se/infusis> (accessed Jan 25, 2011).

P-54 : 3D-neighbourhood protein descriptors for proteochemometric modelingR. Swier¹, G. van Westen¹, J. Wegner², A. IJzerman¹, H. van Vlijmen^{1,2}, A. Bender^{1,3}¹ Division of Medicinal Chemistry, LACDR, Leiden, The Netherlands.² Tibotec BVBA, Beerse, Belgium³ Unilever Centre for Molecular Science Informatics, Cambridge, United Kingdom

Similar to QSAR, Proteochemometric Modeling (PCM) can be used to predict compound-target affinities. However PCM can calculate the activity toward several proteins by including target descriptors in addition to ligand descriptors.¹ The advantage is that one global model can be trained on a larger dataset and often a more robust model is created compared to several single-target models.²

PCMs use protein descriptors that can describe a binding pocket by the physicochemical properties of its amino acids. Inclusion of neighbouring residues is expected to improve the predictability of the trained model by adding information about sequence order and physicochemical neighbourhood. This principle is demonstrated in Table 1. The two sequences contain identical amino acids (described by descriptor x), but in a different order. By combining the neighbouring residues of the amino acid in focus, the new descriptor captures more information about the order than a descriptor based on individual receptor residues. Benchmark experiments based on the Adenosine A₁, A_{2A}, A_{2B} and A₃ receptors showed that the neighbour included descriptor performed better compared to the same descriptor without neighbours, improving the R₀² slightly from 0.61 to 0.63 and the RMSE from 0.69 to 0.68.

Further improvement is expected by making use of 3D crystal structures, when available. We applied this extension of the concept by selecting the neighbouring residues in a sphere of x Å around each amino acid that was part of the binding pocket. These neighbours were then described by their average physicochemical properties, hence creating a 'physicochemical neighborhood' of a particular residue.

We applied this concept to a HIV-1 Reverse Transcriptase Non-Nucleoside inhibitor (NNRTI) dataset consisting of 451 analogues and 14 mutants (including wild type) to create robust models able to extrapolate in target space. With the advent of crystal structures of GPCRs, it is expected that this extension of protein binding pocket descriptors can also be applied to this broad drug target class *via* homology modeling.

1. Lapinsh, M.; Prusis, P.; Gutcaits, A.; Lundstedt, T.; Wikberg, J.E.S. Development of proteochemometrics: a novel technology for the analysis of drug-receptor interactions. *Biochim. Biophys. Acta.* **2001**, 1525, 180-190.
2. Westen, G. J. P. van; Wegner, J. K.; IJzerman, A. P.; Vlijmen, H. W. T. van; Bender, A. Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets. In *Med. Chem. Commun.* **2011**, 2, 16-30.

P-56 : Predictive data mining for the identification of CYP P450 isoform-specific sites of metabolism

J. Tyzack, J. Kirchmair, A. Koutsoukas, D. Murrell, M. Williamson, S. E. Adams, A. Bender, P. Bond, R. C. Glen

Unilever Centre for Molecular Sciences Informatics, Department of Chemistry, University of Cambridge, Cambridge, United Kingdom

Despite considerable progress in automation technology, experimental elucidation of the metabolism of xenobiotics is highly demanding in terms of laboratory equipment and expertise.¹ Even more problematic is the time lag arising from the experimental approaches, impeding fast and efficient optimization of

Table 1. Concept of neighbour included protein descriptor.

| Descriptor x_i | Sum of x_{i-1} , x_i and x_{i+1} . |
|------------------|--|
| Sequence 1 | |
| 5 | 17 |
| 4 | 13 |
| 4 | 15 |
| 7 | 19 |
| 8 | 20 |
| Sequence 2 | |
| 5 | 13 |
| 4 | 16 |
| 7 | 19 |
| 8 | 19 |
| 4 | 17 |

metabolic properties. Therefore, *in silico* prediction of xenobiotic metabolism has gained substantial interest in recent years.² Approximately 75% of the total metabolism is facilitated by CYP P450s, with CYPs 3A4, 1A2 and 2D6 being the most important isoforms.³ While the majority of today's *in-silico* tools focus on the global prediction of sites of metabolism, several techniques have been reported to model substrate specificity, such as QSAR and machine learning methods,⁴⁻⁸ pharmacophores^{9, 10} and structure-based approaches.¹¹

MetaPrint2D^{12,13} is designed to mine huge collections of biotransformation data. The database used for model training is the Accelrys Metabolite Database, which includes pairs of substrates and products of experimentally observed metabolic reactions. Based on this data, the probability of an atom with defined atom environment being involved in a metabolic reaction is calculated. The proximate atom environment is thereby encoded using circular fingerprints. Any small molecule can then be submitted to MetaPrint2D in order to predict the liability of atom positions to metabolic reactions.

The prediction of isoform-specific transformations is a non-trivial problem, given the much more limited data to train the model (about 13% of the biotransformation data is available for the specific isoform as annotated) and the unfavorable signal-to-noise ratio. In the current study, we investigate the feasibility of this data mining approach to generate isoform-specific models for the most prominent isoforms of the CYP P450 protein family. While our data indicate the accuracy of predictions of atom positions involved in biotransformations to be largely sustainable, the specificity of the models is an issue. In order to improve specificity of the individual models we employ different data preprocessing approaches to increase the quality of the training data and investigate the applicability of property-activity relationships (PARs) as post-processing filters and for consensus voting.

1. Reichman, M.; Gill, H. In *Automated drug screening for ADMET properties*, 2009; John Wiley & Sons, Inc.: 2009; pp 129-166.
2. Refsgaard, H. H. F.; Jensen, B. F.; Christensen, I. T.; Hagen, N.; Brockhoff, P. B. *In silico* prediction of cytochrome P450 inhibitors. *Drug Dev. Res.* **2006**, *67*, 417-429.
3. Lewis, D. F. V.; Ito, Y. Human CYPs involved in drug metabolism: structures, substrates and binding affinities. *Expert Opin. Drug Metab. & Toxicol.* **2010**, *6*, 661-674.
4. Yap, C. W.; Chen, Y. Z. Prediction of cytochrome p450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines. *J. Chem. Inf. Model.* **2005**, *45*, 982-992.
5. Terfloth, L.; Bienfait, B.; Gasteiger, J. Ligand-based models for the isoform specificity of cytochrome P450 3A4, 2D6, and 2C9 substrates. *J. Chem. Inf. Model.* **2007**, *47*, 1688-1701.
6. Sheridan, R. P.; Korzekwa, K. R.; Torres, R. A.; Walker, M. J. Empirical regioselectivity models for human cytochromes P450 3A4, 2D6, and 2C9. *J. Med. Chem.* **2007**, *50*, 3173-3184.
7. Yamashita, F.; Hara, H.; Ito, T.; Hashida, M. Novel hierarchical classification and visualization method for multiobjective optimization of drug properties: Application to structure-activity relationship analysis of cytochrome P450 metabolism. *J. Chem. Inf. Model.* **2008**, *48*, 364-369.
8. Michielan, L.; Terfloth, L.; Gasteiger, J.; Moro, S. Comparison of multilabel and single-label classification applied to the prediction of the isoform specificity of cytochrome P450 substrates. *J. Chem. Inf. Model.* **2009**, *49*, 2588-2605.
9. Schuster, D.; Laggner, C.; Steindl, T. M.; Langer, T. Development and validation of an *in silico* P450 profiler based on pharmacophore models. *Curr. Drug Discov. Technol.* **2006**, *3*, 1-48.
10. de Groot, M. J.; Ekins, S. Pharmacophore modeling of cytochromes P450. *Adv. Drug Delivery Rev.* **2002**, *54*, 367-383.
11. Cruciani, G.; Carosati, E.; De Boeck, B.; Ethirajulu, K.; Mackie, C.; Howe, T.; Vianello, R. MetaSite: Understanding metabolism in human cytochromes from the perspective of the chemist. *J. Med. Chem.* **2005**, *48*, 6970-6979.
12. Carlsson, L.; Spjuth, O.; Adams, S.; Glen, R. C.; Boyer, S. Use of historic metabolic biotransformation data as a means of anticipating metabolic sites using MetaPrint2D and Bioclipse. *BMC Bioinf.* **2010**, *11*.
13. Boyer, S.; Arnby, C. A.; Carlsson, L.; Smith, J.; Stein, J.V.; Glen, R.C. Reaction Site Mapping of Xenobiotic Biotransformations. *J. Chem. Inf. Model.*, 2007, *47* (2), pp 583-590.

P-58 : Structural and energetic aspects of oxazolidinone antibiotics binding to the ribosomal structure

Jagmohan Saini, Simone Fulle, Holger Gohlke

Institute of Pharmaceutical and Medicinal Chemistry, Heinrich-Heine-University, Duesseldorf

The availability of high resolution crystal structures of antibiotics bound to ribosomal structures of different species provides crucial insight regarding binding sites, binding modes, and mechanisms of action of these antibiotics. Such information provides opportunities for rational structure-based drug design approaches to improve or obtain novel antibiotics. However, structural determination by X-ray crystallography only provides *static* views of the binding processes but does not reveal the *dynamics* involved with antibiotics binding, or structural and energetic determinants of binding. Theoretical and computational approaches such as molecular dynamics (MD) simulations in combination with free energy calculations are suitable to fill this gap.

In the present study, we aim at investigating the determinants of binding of antibiotics of the oxazolidinone class; this class is one of the only three new classes of synthetic antibiotics that have entered the market during the last 30 years¹. In particular, we investigate linezolid, its derivative radezolid, and the structurally related oral anticoagulant drug rivaroxaban in complex with the *H. marismortui* (archaea) structure² to obtain insights into the determinants of binding of oxazolidinones. The molecular mechanics Generalized-Born surface area (MM-GBSA) method is used to determine the effective free energy of binding.³

Our initial results show that radezolid forms more stable hydrogen bonds with binding site residues than linezolid, which may be the reason for its 100-fold stronger binding affinity. However, no hydrogen bond at all is observed in the case of rivaroxaban. Furthermore, the root mean square atomic fluctuations demonstrate that rivaroxaban moves inside the binding site whereas the core of linezolid and radezolid are largely immobile. The movement of rivaroxaban is in agreement with missing hydrogen bond and stacking interactions.

The results for the effective binding energies combined with a free energy decomposition on a per-residue level show critical residues contributing favourably to aromatic stacking interactions. Notably, the calculations identify radezolid as the most potent ligand. However, further analysis is required to investigate whether these findings can be linked to the effect of mutants that lead to the development of resistance against these antibiotics.

- 1) R. C. Moellering, *Ann. Intern. Med.* (2003) 138, 135-142.
- 2) J. A. Ippolito *et. al.*, *J. Med. Chem.* (2008) 51, 3353-3356.
- 3) H. Gohlke, D. A. Case, *J. Comput. Chem.* (2004) 25, 238-250.

P-60 : An investigation of in silico methods to accurately simulate biologically-relevant conformations of drug-like macrocyclic moleculesN. Brown¹, J. Mallinson², I. Collins²¹ *In Silico Medicinal Chemistry, Cancer Research UK Cancer Therapeutics Unit, The Institute of Cancer Research, 15 Cotswold Road, Sutton, SM2 5NG, UK*² *Medicinal Chemistry, Cancer Research UK Cancer Therapeutics Unit, The Institute of Cancer Research, 15 Cotswold Road, Sutton, SM2 5NG, UK*

Macrocycles have been identified as an underexploited area of chemistry space in drug discovery. Macrocyclisation of bioactive molecules has been shown to improve potency and alter physicochemical properties that are particularly important during the discovery and development stages of drug design.¹ However, current molecular modelling approaches and protocols have largely been designed to consider more classical drug-like molecules and are not directly applicable to modelling drug-like macrocycles.²

Here, we investigate the application of existing *in silico* methods to modelling conformations of known drug-like macrocycles, aiming to find successful methods that reproduce experimentally observed small

molecule crystal or biological conformations. It is intended that this *in silico* protocol will provide guidance on the modelling of novel macrocycles *a priori*.

Small molecule crystal structures and protein-ligand complexes of macrocycles identified from the literature will be modelled using a variety of conformation exploration methods, including: CORINA, OMEGA, MacroModel, LowModeMD and MMFF94x. Docking approaches (*e.g.* GOLD and Glide) are also used to understand whether these methods can also explore the conformational space of macrocycles with explicit reference to the protein environment. Care is taken to ensure that the original geometric information is not involved in the calculation of the modelled conformations.

The modelled macrocycle conformations will be compared to the experimentally-observed reference conformations, to identify their similarity in shape and calculated energy. This will provide guidelines on those methods that tend towards recreating the reference conformation and will therefore aid in the design of future drug-like macrocyclic molecules.

1. Driggers, E. M.; Hale, S. P.; Lee, J.; Terrett, N. K., The exploration of macrocycles for drug discovery – an underexploited structural class. *Nat. Rev. Drug Discov.* **2008**, *7*, 608-624.
2. Bonnet, P.; Agrafiotis, D. K.; Zhu, F.; Martin, E. J. Conformational analysis of macrocycles: finding what common search methods miss. *J. Chem. Inf. Model.* **2009**, *49*, 2242-2259.

P-62 : Exploring DNA topoisomerase I ligand space in search of novel anticancer agents

R. Griffith¹, M. Drwal¹, K. Agama², L. Wakelin¹, Y. Pommier²

¹ *Department of Pharmacology, School of Medical Sciences, University of New South Wales, Sydney, Australia*

² *Laboratory of Molecular Pharmacology, Center for Cancer Research, National Cancer Institute, Bethesda, USA*

DNA topoisomerases are over-expressed in tumour cells and are thus an important target in anticancer therapy. Topoisomerase I (topo I) relaxes DNA torsional strain generated during DNA processing by introducing transient single-strand breaks and allowing the broken strand to rotate around the intermediate topo I – DNA covalent complex. This complex can be trapped by a group of anticancer agents interacting with the DNA bases and the enzyme at the cleavage site, preventing further topoisomerase activity.

The focus of this study was the identification of novel topo I inhibitors as potential anticancer agents using a combination of structure- and ligand-based molecular modelling methods. In particular, pharmacophore models have been developed based on the molecular characteristics of derivatives of the alkaloid camptothecin (CPT), which represent potent antitumour agents and the main group of topo I inhibitors. The models generated were used for *in silico* screening of the National Cancer Institute (NCI, USA) compound database, leading to the identification of a set of structurally diverse molecules. The strategy was validated by the observation that amongst these molecules were several known topo I inhibitors and agents cytotoxic against human tumour cell lines. An expert selection step was used to assess the top 20 structures of two hitlists. This step used the following criteria: (1) compound is not a camptothecin derivative; (2) compound has not been tested for topoisomerase inhibition; and (3) compound is dissimilar to other compounds already selected for further investigation. The potential of the selected hits to inhibit topo I activity was further evaluated by docking into the binding site of a topo I – DNA complex, resulting in a selection of 10 compounds for biological testing. Figure 1. below gives an overview of the procedure used to select compounds for biological testing.

Limited by the compound availability, seven compounds have been tested *in vitro* for their topo I inhibitory activity, 5 of these displayed mild to moderate topo I inhibition. A further compound, found by similarity search to the active compounds, also showed mild activity.

In conclusion, the combination of pharmacophores, docking methods and expert assessment can be successfully applied in database screening to retrieve known topoisomerase I inhibitors, compounds with

anticancer activity as well as structurally new compounds with topo I inhibitory activity and anticancer activity combined.

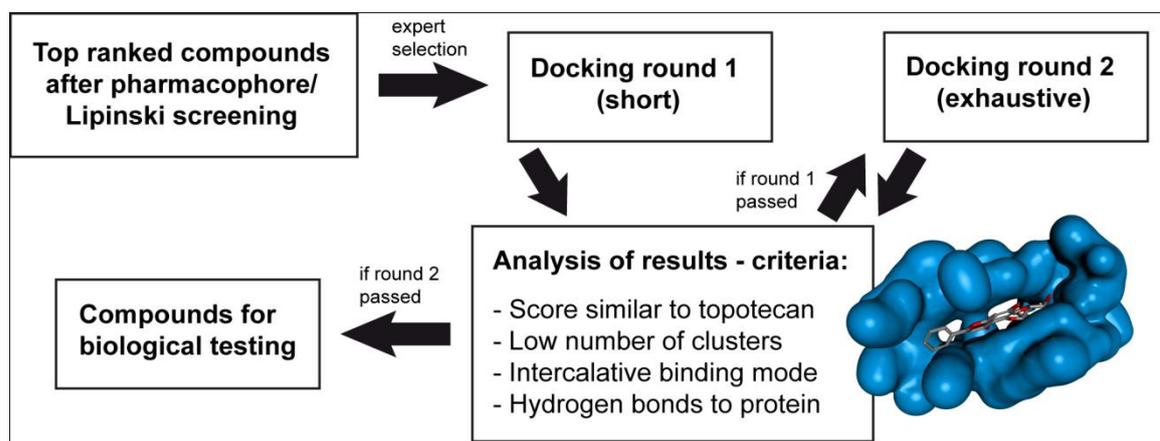


Figure 1. Overview of the procedure used to select compounds for biological testing. On the lower right hand corner, the result of one of the docking runs is illustrated. A small molecule in stick representation is shown docked into the binding pocket of a topoisomerase I crystal structure in complex with topotecan (PDB ID 1k4t [1]). The surface of the binding pocket is shown and this is comprised of both DNA bases and amino acid sidechains. 22 compounds were subjected to short docking runs (round 1), and 16 passed and were docked again in a more exhaustive fashion (round 2). 10 compounds were selected for biological testing.

1. Staker, B. L.; Hjerrild, K.; Feese, M. D.; Behnke, C. A.; Burgin, A. B. Jr.; *et al.* (2002) The mechanism of topoisomerase I poisoning by a camptothecin analog. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, 99, 15387-15392.

P-64 : Improved docking accuracy using 3D restraints derived from X-ray crystallography

R. Hall

Astex Therapeutics, Cambridge, UK

This presentation describes a novel algorithm that can be used to improve ligand docking performance during a structure based optimization campaign. At Astex all optimisation projects start with a fragment screen in which multiple fragments are co-complexed with a protein target. The solved crystallographic geometry of fragments enables the identification of the key binding regions of an active site. All available ligands are then used to automatically build pharmacophore models that reward docking into these key regions. As the project progresses, these models are continually updated with new structures, reflecting the direction of the medicinal chemistry effort. The technique has been applied retrospectively to a number of Astex projects and we shall present results that demonstrate a substantial improvement in cross docking accuracy.

P-66 : In silico identification of resistance-breaking inhibitors of influenza neuraminidaseJ. Kirchmair¹, J. M. Rollinger², U. Grienke², N. Seidel³, A. Krumbholz³, M. Schmidtke³, K. R. Liedl⁴¹ *Unilever Centre for Molecular Informatics, University of Cambridge, Cambridge, United Kingdom*² *Institute of Pharmacy/Pharmacognosy, University of Innsbruck, Austria*³ *Institute of Virology and Antiviral Therapy, Jena University Hospital, Jena, Germany*⁴ *Institute of Theoretical Chemistry, University of Innsbruck, Austria*

Influenza A viruses have been reported to become increasingly resistant to ion channel blockers and approved neuraminidase inhibitors, asking for the development of novel resistance-breaking compounds.^{1,2} Recently, we have identified katsumadain A, a plant constituent from *Alpinia katsumadai*, to exhibit significant inhibitory activity on influenza neuraminidase. The potential binding mode of this compound was investigated using a combined approach of classic molecular dynamics simulations of the protein coupled with a downstream protein-ligand docking protocol on representative conformations of the protein.³

In the current study⁴ we employed the shape-focused virtual screening engine ROCS to screen for molecules with potential activity on influenza neuraminidase, using katsumadain A as the template structure. From the National Cancer Institute (NCI) database we identified and purchased five interesting compounds, three of them being flavonoids – a class of plant metabolites that was repeatedly identified to exhibit interesting antiviral and also NA-inhibiting activity. The NA-inhibiting potential of these compounds was tested with influenza virus A/PR/8/34, 3 clinical H1N1v isolates, and an oseltamivir-resistant H1N1 isolate from the season 2008/09 using a chemiluminescence-based enzyme inhibition assay. All five compounds strongly inhibited the NA of the oseltamivir-susceptible H1N1 and H1N1v strains. Artocarpin, a twofold isoprenylated flavon present in different species of the genus *Artocarpus*, exhibited highest activity and its 50% inhibitory concentration was found to be 10-times lower compared to katsumadain A. The compound shows also high activity against the oseltamivir-resistant H1N1 isolate.

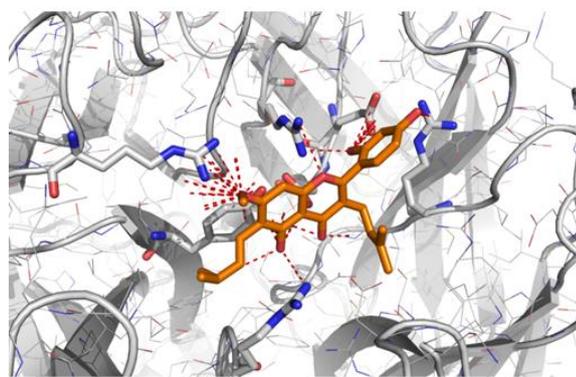


Figure 1: Potential binding mode of artocarpin to influenza neuraminidase.

Independent models of the potential binding mode derived from a ligand-based (pharmacophores) and a structure-based (protein-ligand docking on representative frames of the protein derived from molecular dynamics simulations, Figure 1) perspective show remarkable consistency. Based on these new insights on the SAR of inhibitors of influenza neuraminidase, we are developing novel inhibitors of influenza NA with improved resistance profiles.

1. Moscona, A. Global transmission of oseltamivir-resistant influenza. *N. Engl. J. Med.* **2009**, 360, 953-956.
2. Weinstock, D. M.; Zuccotti, G. The evolution of influenza resistance and treatment. *JAMA, J. Am. Med. Assoc.* **2009**, 301, 1066-1069.
3. Grienke, U.; Schmidtke, M.; Kirchmair, J.; Pfarr, K.; Wutzler, P.; Durrwald, R.; Wolber, G.; Liedl, K. R.; Stuppner, H.; Rollinger, J. M. Antiviral potential and molecular insight into neuraminidase inhibiting diarylheptanoids from *Alpinia katsumadai*. *J. Med. Chem.* **2010**, 53, 778-786.
4. Kirchmair, J.; Rollinger, J. M.; Liedl, K. R.; Seidel, N.; Krumbholz, A.; Schmidtke, M. Identification, biological evaluation and investigation of the binding mode of resistance-breaking inhibitors of influenza neuraminidase. *Future Med. Chem.* **2011**, in press.

P-68 : Replacing structure-based pharmacophore filtering: high-throughput docking using constraints leads to TrxR inhibitors with high bioactivity on *M. tuberculosis*

O. Koch^{1,2}, N. Doil³, K. Heller², F. Stuhlmann², S. Schmitt², P. C. Khandavalli², D. Schinzer², L. Flohé², F.-C. Bange³, T. Jaeger², P. M. Selzer¹

¹ *Intervet Innovation GmbH, Schwabenheim, Germany*

² *MOLISA GmbH, Magdeburg, Germany*

³ *Hannover Medical School, Hannover, Germany*

Using high-throughput docking of millions of compounds in a structure-based virtual screening campaign is still not feasible. The main limitation is the computational power to generate accurate results in a reasonable time. Structure-based pharmacophore filtering is a standard procedure to massively reduce the number of compounds¹ and, thus, the computational time for docking.

High-throughput docking using constraints combines the strength of both worlds and can be applied as pre-filtering step before accurate docking. The constraints give the opportunity to dock with very fast settings. Using GOLD², it is possible to filter for specific protein-ligand hydrogen bonding or hydrophobic interactions, protein-ligand atom distances or specific positions of substructures in the binding pocket. As advantage over pharmacophore filtering, hydrogen bond strength cut-offs can be set and all binding site properties are included. A discussion about the right conformer generation, like in the case of pharmacophore approaches, is also no longer required.

As a test case, high-throughput docking using constraints was applied for targeting the thioredoxin reductase (TrxR) / Thioredoxin (Trx) System in *Mycobacterium tuberculosis*. In *M. tuberculosis*, the Trx system contributes to peroxide reduction and is pivotal to ribonucleotide reduction and, thus, guarantees the survival within macrophages and proliferation³. *M. tuberculosis* lacks the common glutathione system and the *Mt*TrxR is substantially different in sequence, mechanism and structure from human TrxRs. This makes the TrxR a promising target of drugs for the treatment of tuberculosis.

Two important hydrogen bonding interactions were identified at the protein-protein interface of the TrxR-Trx complex. A high-throughput docking using constraints was applied to filter the Intervet's supplyable compound library (~6.5 million compounds) for hits that interact with these two hydrogen bonding acceptors. The resulting compounds (~150,000) were redocked without any constraints and most accurate docking settings. Rescoring of the created docking poses using the same constraints yielded ~11,000 compounds for final compound selection based on an automated ranking using a normalization and consensus scoring strategy.

Out of first 170 compounds tested, 17 inhibited *Mt*TrxR with an IC₅₀ value in the low μ M range. Four different scaffolds with a low molecular weight were selected as promising candidates for further development. Lead optimization has so far yielded compounds that inhibit *Mt*TrxR within the low nM range. The *in vitro* inhibitory effect of these leads on *M. tuberculosis* was tested using the MGIT 960 system (Becton Dickinson). The best MIC was observed down to 0.7 μ g ml⁻¹, which is close to the MIC of Rifampicin (0.25 μ g ml⁻¹)⁴, one of the most effective antituberculous drug.

The constraint-docking approach will be discussed in detail together with the underlying target mechanism and the development and results of the promising leads.

1. Klebe, G., Virtual ligand screening: strategies, perspectives and limitations. *Drug Discovery Today*. **2006**, 11, 580 – 594
2. Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W., Taylor, R. D. Improved Protein-Ligand Docking Using GOLD. *Proteins*. **2003**, 52, 609-623
3. Jaeger, T., Flohé, L., The thiol-based redox networks of pathogens: unexploited targets in the search for new drugs. *Biofactors*. **2006**, 27, 109-120
4. Rastogi, N., Goh, K. S., Berchel, M., Bryskier, A., *Journal of antimicrobial chemotherapy*. **2000**, 46, 565-570

P-70 : LoFT: design of combinatorial libraries for the exploration of virtual hits from fragment space searchesU. Lessel¹, R. Fischer², M. Rarey²¹ *Boehringer Ingelheim Pharma GmbH & Co. KG, Department of Lead Identification and Optimization Support, 88397 Biberach an der Riss, Germany*² *Center for Bioinformatics Hamburg, University of Hamburg, Bundesstrasse 43, D-20146 Hamburg, Germany*

At Boehringer Ingelheim fragment spaces based on reactions from combinatorial chemistry are successfully screened with Feature Trees for the detection of new lead classes. Once interesting scaffolds have been detected, corresponding combinatorial libraries have to be designed. For this purpose the software LoFT (Library Optimizer using Feature Trees) was developed. LoFT enables the design of focused libraries optimized according to Feature Trees similarity to one or more queries and/or Feature Trees dissimilarity to one or more anti-queries. In addition, a desired range of physico-chemical properties for library compounds can be taken into account.

Routine application of LoFT showed that it is essential for the final reagent selection to assess and influence the Feature Tree matches. For this purpose we make use of a recently introduced extension of the Feature Trees which allows to assess and to visualize the Feature Trees similarity via FlexS. Additionally, a new option was added to LoFT for the restriction of Feature Tree matches to predefined positions.

Finally, the Feature Trees algorithm within LoFT was extended enabling the differentiation between regioisomers at aromatic rings which is particularly important for focused library design.

In the presentation we share a possible workflow for post-processing results from fragment space searches and application examples are shown to illustrate the usefulness of the LoFT approach.

P-72 : PDB Ligands: high-level conformational energy calculations

M. C. Nicklaus, M. Sitzmann, I. V. Filippov

Chemical Biology Laboratory, Center for Cancer Research, National Cancer Institute/Frederick, NIH, DHHS, 376 Boyles Street, Frederick, MD 21702, USA (IVF: [Contr./SAIC])

We report on the cheminformatics procedures applied to more than 350,000 ligand instances in the RCSB Ligand Expo collection to generate a “high-quality” subset of small-molecule ligand crystal structures, and the subsequent high-level energy calculations to determine, as reliably as possible, conformational energy changes of small-molecule ligands when binding to proteins. We obtained a subset of about 1,300 ligand instances fulfilling the criteria of being crystallographically as high-quality as possible, as unambiguous in their connectivity as possible, and as being true, non-covalent ligands in the medicinal chemistry sense. A de-duplicated subset of 640 ligand instances was submitted to a protocol of step-wise optimization of internal degrees of freedom in Gaussian 03 jobs at the DFT B3LYP/6-311++G(3df,2p)//B3LYP/6-31G(d) level of theory. Both vacuum and the IEFPCM solvent model were explored. We present and discuss the conformational energies found in the bound vs. the unbound state. These energies quite evenly populated a range of 0 - ~30 kcal/mol independently of crystallographic resolution and other crystallographic quality parameters, with some outliers at yet higher values.

P-74 In Silico discovery of new types of allosteric inhibitors against Bcr-AblN. S. Kast¹, J. Mahajna², Y. Najajreh³, M. Ruthardt⁴ and A. Goldblum¹¹ *Molecular Modeling and Drug Design Laboratory, Institute for Drug Research, The Hebrew University of Jerusalem,*² *Cancer Drug Discovery, Migal-Galilee Technology Center, Kiryat Shmona, Israel*³ *Anticancer Drugs Research Lab, AL-Quds University, Jerusalem P.O.Box 20002, Palestinian Authority*⁴ *Medizinische Klinik II/Abtl. Hematology, Johann Wolfgang Goethe University, Frankfurt, Germany*

C-abl is a kinase that has important roles in the normal process of the cell cycle. It has a typical kinase structure¹ and a unique ability to auto-regulate itself by inserting a myristolated group at its N-terminus to a specific hydrophobic cleft near its C-terminus (the *myristoyl pocket*) thus resulting in structural changes that block the kinase activity².

The loss of self-regulation is a result of the genetic translocation known as the "Philadelphia chromosome" which is the direct cause of Chronic Myelogenous Leukemia (CML), a type of blood cancer, due to a fusion protein, Bcr-Abl that possesses C-abl kinase activity but lacks the myristolated N-terminus.

Clinically used inhibitors of Bcr-Abl that are aimed at the ATP - binding site. Unfortunately, in a later phase of the treatment, most CML patients develop resistance to the available drugs, as a result of point mutations in and around that site.

Two allosteric inhibitors of Bcr-Abl were developed recently - GNF2 and GNF5³, and were later expanded into a series of chemically related compounds⁴.

Recent studies have shown to produce a major reduction in cell proliferation rate (including cell lines that express the resistant form of Bcr-Abl) due to a combination of treatment with both an ATP binding site inhibitor and an allosteric inhibitor⁵.

In order to develop allosteric inhibitors to Bcr-Abl that significantly differ structurally from the GNF variants, we incorporated several approaches to design pharmacophore models based on information from different ligand- interactions with the protein⁶. We use a novel approach, in which the pharmacophore models are based on a search by our Iterative Stochastic Elimination algorithm⁷. Applying this kind of models to large databases leads to the discovery of novel candidate allosteric inhibitors of Bcr-Abl.

1. Bikker, J. A.; Brooijmans, N.; Wissner, A.; Mansour, T. S.; Kinase Domain Mutations in Cancer: Implications for Small Molecule Drug Design Strategies. *J. Med. Chem.* 2009, 52 (6), 1493-1509.
2. Nagar, B.; Hantschel, O.; Young, MA.; Structural Basis for the Autoinhibition of c-Abl Tyrosine Kinase. *Cell.* 2003, 112, 737-40.
3. Adrián, FJ.; Ding, Q.; Sim, T. Allosteric inhibitors of Bcr-Abl-dependent cell proliferation *Nat Chem Biol.* 2006, 2, 63-4.
4. Deng, X.; Okram, B.; Ding, Q.; Zhang, J. Expanding the diversity of allosteric bcr-abl inhibitors *J. Med. Chem.* 2010, 53, 6934-46.
5. Hassan, AQ.; Sharma, SV.; Warmuth, M. Allosteric inhibition of BCR-ABL. *Cell Cycle.* 2010, 9, 3710-4.
6. Zou, J.; Xie, HZ.; Yang, SY. Towards more accurate pharmacophore modeling: Multicomplex-based comprehensive pharmacophore map and most-frequent-feature pharmacophore model of CDK2. *J Mol Graph Model.* 2008, 27, 430-8.
8. Glick, M.; Rayan, A.; Goldblum, A., A stochastic algorithm for global optimization and for best populations: A test case of side chains in proteins. *PNAS* 2002, 99, 703-708.

P-76 : Comparison of various strategies in pharmacophore models generation – application to 5-HT_{1A} receptor ligands

D. Warszycki¹, K. Kristiansen², R. Kurczab¹, I. Sylte², A. J. Bojarski¹

¹ Institute of Pharmacology, Polish Academy of Sciences, 12 Smetna Street, 31-343 Kraków, Poland

² Medical Pharmacology and Toxicology, Department of Medical Biology, Faculty of Health Sciences, University of Tromsø, N-9037 Tromsø, Norway

In ChEMBL database¹ there are over 6 thousands 5-HT_{1A} receptor ligands (both, active compounds and decoys) extracted from about 520 papers. Among them 3616 are relatively strong binders, with Ki (or equivalent) below 100 nM. Those ligands were clustered by three different approaches: using 3D pharmacophore or MOLPRINT 2D fingerprints (as implemented in Canvas software²) or based on a classical method grouping compounds by a common core (aminotetralines, arylpiperazines, ergolines, etc.). For representative compounds selected from each cluster separate pharmacophore hypotheses were

developed and tested using Phase software³. Next, they were reduced by grouping similar models and the less effective ones were discarded. The remaining hypotheses and their linear combinations were tested on external test set consisting of 200 active compounds (unused in pharmacophore development), 200 decoys (extracted from ChEMBL database) and 200 assumed inactives (already used drugs with confirmed inactivity against 5-HT_{1A} receptor). The described study discusses efficiency of single hypotheses in comparison to their linear combinations and relation between quality of pharmacophore models and methodology of generating clusters for their development. Created pharmacophore models will be used in further studies in multistep virtual screening for searching new compounds acting on 5-HT_{1A} receptor.

Acknowledgments

This study were partly supported by grant PNRF-103-AI-1/07 from Norway through Norwegian Financial Mechanism

1. <https://www.ebi.ac.uk/chembl/>
2. Canvas, version 1.3, Schrödinger, LLC, New York, NY, 2010.
3. Phase, version 2.1, Schrödinger, LLC, New York, NY, 2010

P-78 : Fighting molecular obesity with sub-pharmacophore screening

M. von Korff¹, J. Freyss¹, T. Sander¹, C. Boss², C-L. Ciana²

¹ *Department of Research Informatics*

² *Drug Discovery Chemistry*

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

Pharmacophore based virtual screening is a commonly used methodology to investigate scaffold hopping. Often molecules developed in lead optimization programs are too large to be good seeds. These molecules are complex, exhibiting multiple pharmacophore points. But the probability of detecting biosimilars in commercially available compound collections decreases as the number of pharmacophore points grows. In this contribution, we demonstrate how automatically generated sub-pharmacophore models can be used to increase the number of virtual screening hits worth testing in biological assays. A 3D pharmacophore descriptor developed in-house was used to generate the sub-pharmacophore models². The DUD dataset was used to compare the pharmacophores with docking and chemical fingerprints. The DUD is the first published dataset providing active molecules, decoys and references for crystal structures of ligand-target complexes¹. It contains 2,950 active compounds against a total of 40 target proteins. Furthermore, the dataset contains 36 structurally dissimilar decoy compounds with similar physicochemical properties for every ligand.

The ligands were extracted from the target proteins to extend the applicability of the dataset to include ligand based virtual screening. Of the 40 target proteins, 37 contained ligands that were used as query molecules for virtual screening evaluation. With this dataset, a comparison between the pharmacophores, three different chemical fingerprints and docking was done. In terms of enrichment rates, the chemical fingerprint descriptors performed better than the pharmacophores and the docking tool. After removing molecules chemically similar to the query molecules, the pharmacophores outperformed the chemical descriptors.

Encouraged by these results, the sub-pharmacophores were applied to some in-house drug discovery projects. In one project, only one highly active but large (800 Daltons) molecule was available as seed. Neither the full pharmacophore model nor the chemical fingerprints were able to detect any similars in a database of seven million commercially available molecules. However, a sub-pharmacophore search resulted in the detection of hundreds of interesting molecules. These molecules were purchased and biologically tested. Biological assay results for this and other virtual screening experiments will be reported here. Resulting was new series of molecules with much higher ligand efficiency than the seed molecule.

1. Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* 2006, 9, 6789-6801.

2. 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.

**Poster Session
Abstracts
BLUE**

P-1 : Continuous functions as a universal way of representing chemical structures in chemoinformaticsI. Baskin^{1,2}, N. Zhokhova¹¹ *Department of Chemistry, Moscow State University, Moscow, Russia*² *Laboratoire d'Informatique, UMR 7177 CNRS - University of Strasbourg, Strasbourg, France*

One of the main distinctions of chemical data is that they are functional by their nature. Molecules can be more naturally described in terms of electronic density and molecular fields derived from it than by means of commonly used molecular descriptors and graph kernels. So, the goal of this presentation is to present continuous functions as the third major data type in chemoinformatics (in addition to or instead of descriptor vectors and graph kernels). In this case chemical information is processed by means of functional data analysis (FDA)¹ instead of commonly used multivariate data analysis techniques. FDA deals with infinite number of descriptors organized into continuous functional form. It operates with: functions (in functional Hilbert space) instead of data vectors (in Euclidean space) in multivariate data analysis, linear operators instead of matrices, integration instead of summation, integral equations instead of matrix equations, etc.

In this presentation we show the way how continuous functions describing molecules can be encapsulated into kernels and through this universal mechanism be used to: build 3D-QSAR (see preliminary communication²), 4D-QSAR and 2D-QSPR regression models; build 2-class and multi-class classification SAR models, perform ligand-based virtual screening based on similarity of molecular fields (shapes) using 1-class classification mechanism; perform 3D alignment of molecules; analyze and visualize fields of model parameters; clusterize diverse chemical databases and map chemical space; elaborate scoring functions for supramolecular interactions and analyze molecular recognition, etc. In all cases the advantages of using FDA with continuous molecular functions (fields), in comparison with the use of multivariate statistical data analysis in conjunction with traditional vectors of molecular descriptors, are pointed out.

1. Ramsay, J. O.; Silverman, B. W. *Functional Data Analysis*. Springer: New York, 2005.
2. Zhokhova, N. I.; Baskin, I. I.; Bakhrinov, D. K.; Palyulin, V. A.; Zefirov, N. S. *Method of Continuous Molecular Fields in the Search for Quantitative Structure-Activity Relationships*. *Doklady Chemistry* 2009, 429, 273-276.

P-3 : A treatment of stereochemistry in computer-aided synthesis planningA. Cook¹, A. P. Johnson¹, H. Muller², A. Simon³¹ *School of Chemistry, University of Leeds, Leeds, UK*² *School of Computing, University of Leeds, Leeds, UK*³ *SimBioSys Inc., Toronto, Canada*

A new research project has been initiated at the University of Leeds to study aspects of computer-aided enantioselective chemical synthesis planning as part of the ARChem Route Designer software.¹

The application of computers in aiding modern enantioselective synthesis planning requires the handling of a broad category of stereogenic centres types in target molecules, starting materials, reagents and catalysts at all stages of the development of the synthesis plan. The recognition of, and perceived relationships between stereogenic centres aids the selection of appropriate stereochemical goal directed strategies,² the selection of simplifying stereoselective and enantioselective reaction transforms and the recognition and retrieval of chiral and *meso* compound starting materials.

A number of existing methods for the recognition, representation and manipulation of stereogenic centres have been described,^{3,4} but are limited to handling asymmetric carbon and olefins. A novel generalised pattern based approach that is capable of extending the recognition scope to allenes, atropisomers, asymmetric hetero atoms (S, P, N) and metal coordination complexes in both Natta and perspective style molecule diagrams is described. An efficient screening algorithm is applied to significantly reduce the number of patterns tried at each atom and bond.

A novel fast and generalised algorithm for comparing stereogenic centres has been developed. The embedment of this algorithm within a subgraph isomorphism algorithm enables efficient retron recognition via a stereochemical substructure search of the ARChem reaction rule database.¹

As part of the stereo recognition pattern screening process, a novel fast ring perception algorithm has been developed employing a cycle vector space method. This significantly improves on the running times of the Gibbs algorithm,⁵ and improves the Syslo algorithm⁶ by extending the scope to both planar and non-planar graphs. An application of this algorithm in a fast approximate Huckel aromatic ring perception algorithm is demonstrated.

1. Law, J.; Zsoldos, Z.; Simon, A.; Reid, D.; Liu, Y.; Khew, S. Y.; Johnson, A. P.; Major, S.; Wade, R. A.; Ando, H. Y. Route Designer: A Retrosynthetic Analysis Tool Utilizing Automated Retrosynthetic Rule Generation. *J. Chem. Inf. Model.*, **2009**, *49*, 593-602.
2. Corey, E. J. The logic of chemical synthesis. John Wiley: New York, **1989**.
3. Wipke, W. T.; Dyott, T. M. Simulation and evaluation of chemical synthesis. Computer representation and manipulation of stereochemistry. *J. Am. Chem. Soc.*, **1974**, *96*, 4825-4834.
4. Corey, E. J.; Howe, W. J.; Pensak, D. A. Computer-assisted synthetic analysis. Methods for machine generation of synthetic intermediates involving multistep look-ahead., *J. Am. Chem. Soc.*, **1974**, *96*, 7724-7737.
5. Gibbs, N. E. A Cycle Generation Algorithm for Finite Undirected Linear Graphs. *J. ACM*, **1969**, *16*, 564-568.
6. Syslo, M. M. An Efficient Cycle Vector Space Algorithm for Listing All Cycles of a Planar Graph., *SIAM J. Comput.*, **1981**, *10*, 797.

P-5 : Analysis of activity datasets using spectral clustering

S. Gan¹, V. J Gillet¹, E. Gardiner¹, D. Cosgrove²
¹Information School, University of Sheffield, Sheffield, UK
²AstraZeneca, Macclesfield, UK

The application of traditional clustering algorithms chemical datasets is well established and has been reviewed extensively¹⁻². Despite their widespread application, new advances in clustering methodologies have been limited.

In recent years, clustering methods which utilize the eigenvectors of an input matrix to partition a dataset have gained considerable attention in the field of computer vision. These spectral clustering algorithms have shown good results in a variety of tasks³⁻⁴, often outperforming traditional clustering algorithms⁵. To our knowledge, the application of spectral clustering to chemical datasets is confined to the work of Brewer⁶, who implemented a spectral clustering algorithm to quantify both the contribution that molecular features make to a cluster and the degree of intermolecular similarity within a cluster.

A spectral clustering algorithm which produces non-overlapping clusters has been developed for use with chemical data. This non-overlapping spectral clustering algorithm (NOSC) builds on the previous work of Brewer⁶, using a simple eigen-decomposition algorithm to find the eigenvalues and eigenvectors of a similarity matrix. The NOSC algorithm determines which eigenvectors define “meaningful” clusters based upon the 95% positive eigenvector rule⁴, subsequently classifying molecules into clusters according to their largest eigenvector component.

The outcome of the eigen-decomposition algorithm depends upon the calculated similarity scores between pairs of molecules which are represented as elements within the similarity matrix. Thus, the composition of the similarity matrix plays a hugely important role in spectral clustering. Therefore the parameters which affect the performance of the NOSC algorithm were optimized for five two-dimensional molecular fingerprint types (ECFP_4, Daylight, Unity, BCI and MDL Public Keys).

The ability of the NOSC algorithm to cluster several activity datasets described by the five fingerprints has then been evaluated using various activity prediction measures, including the Quality Clustering Index⁷, and

the performance compared to the leading traditional clustering algorithms, i.e. the Ward's and K-means algorithms, with promising results.

1. Downs, G. M.; Barnard, J. M., Clustering methods and their uses in computational chemistry. *Reviews in Computational Chemistry, Vol 18* **2002**, 1-40.
2. Willett, P., *Similarity and clustering in chemical information systems*. Research Studies Press Letchworth: 1987.
3. Weiss, Y., Segmentation Using Eigenvectors: A Unifying View. In *Proceedings of the International Conference on Computer Vision-Volume 2 - Volume 2*, IEEE Computer Society: 1999; p 975.
4. Sarkar, S.; Boyer, K., Quantitative measures of change based on feature organization: Eigenvalues and eigenvectors. *Computer Vision and Image Understanding* **1998**, 71 (1), 110-136.
5. von Luxburg, U., A tutorial on spectral clustering. *Statistics and Computing* **2007**, 17 (4), 395-416.
6. Brewer, M. L., Development of a spectral clustering method for the analysis of molecular data sets. *Journal of Chemical Information and Modelling* **2007**, 47 (5), 1727-1733.
7. Varin, T.; Bureau, R.; Mueller, C.; Willett, P., Clustering files of chemical structures using the Szekely-Rizzo generalization of Ward's method. *Journal of Molecular Graphics and Modelling* **2009**, 28 (2), 187-195.

P-7 : Fragment based de novo design and ADME/T analysis of dual binding site Acetylcholinesterase Inhibitors for Alzheimer's disease

S. Gupta, C. G. Mohan

Department of Pharmacoinformatics, National Institute of Pharmaceutical Education & Research (NIPER), S.A.S. Nagar, Mohali, Phase X, 160062 Punjab, India.

Alzheimer's disease (AD) is the most common form of dementia. Interestingly, research on acetylcholinesterase (AChE) enzyme has increased due to findings supporting this enzyme involvement in the β -amyloid peptide fibril formation during AD pathogenesis¹. Compound binding towards the active gorge site has two binding sub-sites, a catalytic sub-site (CS) consisting of the catalytic triad (S200, H440 and E327) together with W84 and F330, and peripheral anionic sub-site (PAS) including Y70, Y121 and W279 of AChE enzyme. Inhibitors showing both the anti-aggregating $A\beta$ and anti-cholinesterase effect bind to these sub-sites, the so called dual binding site, which is separated by 14 Å apart. Docking pose of one of the dual binding site inhibitor within the active site of AChE enzyme recently identified by our group is presented in Fig. 1(a)². Here, we propose a comprehensive workflow of structure and fragment based *de novo* design of dual binding site AChE inhibitors by taking the advantage of large number of available crystal structures of AChE enzyme with structurally diverse inhibitors. LigBuilder, which has been developed for structure-based de novo drug design was used for designing dual inhibitors for the AD. Based on the structural constraints of the target protein, LigBuilder builds up ligands iteratively by using a library of organic fragments. The program provides growing and linking strategies to build up ligands on the provided seed structures and the whole construction process is controlled by a genetic algorithm.

We have used 31 fragments as seed structure from the dissection of 20 co-crystallized inhibitors of the AChE enzyme from protein data bank (PDB), by retaining their bioactive conformation for the 3D interactions with the query binding pocket. Compared to other de novo methods in fragment based drug design, our strategy is the first protocol to take advantage of both the interaction information from the PDB and chemical diversity, for the generation of new active compounds. We have also calculated the physicochemical properties of 31 fragment seeds. The results indicated that on average fragment seeds possessed properties consistent with the "Astex Rule of 3" (Number of hydrogen-bond donors ≤ 3 , Number of hydrogen-bond acceptors ≤ 3 , and $\log P \leq 3$). In addition, it was noted that molecular weight (MW) was ≤ 250 , number of rotatable bonds was on average ≤ 3 and polar surface area was ($PSA \leq 60 \text{ \AA}^2$) respectively.

The first step to design new ligands was to analyze the binding pocket of the active site using the POCKET module. The co-crystallized structure of Bis-(7)Tacrine (PDB id: 2CKM) was used to define the pocket and a grid of 10 Å was defined. Ligand compounds for the target protein were constructed by GROW module, from a "seed" structure and they are developed and evolved with a genetic algorithm procedure implemented in the software. The compounds generated by GROW module were collected in a LigBuilder

and PROCESS module were used for extracting the desired compounds, and converting them to viewable Mol2 files. Based on 23 seed structures, LigBuilder generated 3097 compounds which were further used for virtual screening (VS). The generated 3097 compounds were filtered by our developed AChE inhibitors multi parameter optimization (AChEIs MPO) algorithm, which was developed by thorough analysis of properties for 42 highly active AChE inhibitors (leads) and 4 AChE inhibitor drug candidates (Donepezil, Galanthamine, Rivastigmine and Tacrine) for AD.

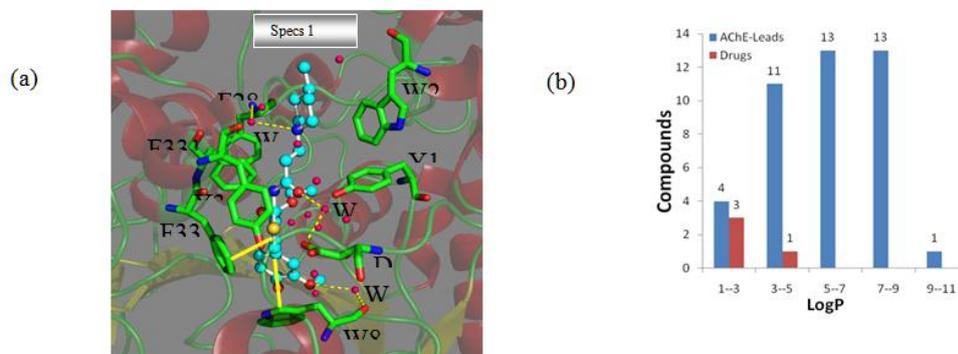


Fig. 1: (a) Docking pose of the VS lead compound **Specs1** in the dual binding site (CS and PAS) of AChE enzyme. (b) LogP distribution for AChE inhibitor drugs and leads.

We described a detailed analysis of seven fundamental physicochemical properties associated with these two sets of compounds: (a) lipophilicity, partition coefficient (logP); (b) distribution coefficient at pH=7.4 (logD); (c) molecular weight (MW); (d) polar surface area (PSA); (e) Blood-Brain-Barrier (BBB) penetration (CNS); (f) number of hydrogen bond donors (HBD); (g) number of hydrogen bond acceptors (HBA)³. The logP distributions for the AChE inhibitor leads and the drugs in the data set analyzed in this study are reported in Fig. 1(b). Hence, the sequential AChEIs MPO filtering reduced the hits to 245 compounds. These 245 compounds were further docked into the active site of AChE by FlexX (BioSolveIT) and toxicity was predicted through DEREK software. Finally, we got 20 leads showing good docking score in AChE and no major toxicity as predicted by DEREK. These 20 leads were further evaluated for its synthetic accessibilities by the SYLVIA program. Schematic representation of the present workflow is shown in Fig.2.

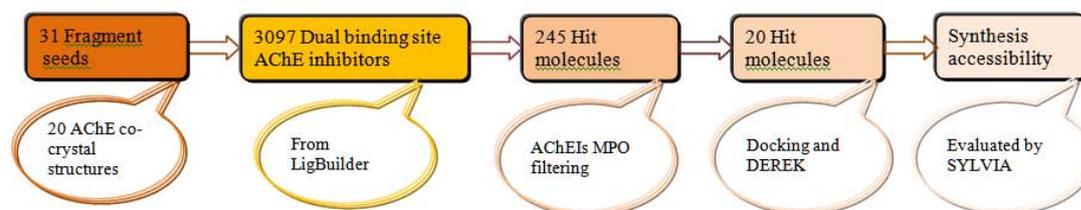


Fig. 2: Schematic representation of workflow performed in the present study.

1. Torrero, D.M. Acetylcholinesterase Inhibitors as Disease-Modifying Therapies for Alzheimer's Disease. *Curr. Med. Chem.* **2008**, *15*, 2433-2455.
2. Gupta, S.; Fallarero, A.; Järvinen, P.; Karlsson, D.; Johnson, M.S.; Vuorela, P.M.; Mohan, C. G. Discovery of Dual Binding Site Acetylcholinesterase Inhibitors Identified by Pharmacophore modelling and Sequential Virtual Screening Techniques. *Bioorg. Med. Chem. Lett.*, **2011**(In Press).
3. Travis, T. W.; Ramalakshmi, Y.C.; Xinjun, H.; Matthew, D.T.; Patrick, R.V.; Anabella, V.; Yvonne W. Defining Desirable Central Nervous System Drug Space through the Alignment of Molecular Properties, in Vitro ADME, and Safety Attributes. *ACS Chem. Neurosci.* **2010**, *1*, 420-434.

P-9 : Chemoinformatics approaches for identification of porous materials for industrial applicationsM. Haranczyk¹, J. Swisher², T. Willems¹, K. Jariwala,¹ R. L. Martin¹, B. Smit²¹ Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA² Department of Chemical Engineering, University of California, Berkeley, CA 94720, USA

Porous materials, such as zeolites and metal organic frameworks, have many applications in the chemical industry. Among others, they can be used for gas separations. The number of possible zeolite structures has been estimated to be larger than 2.5 million. Databases of hypothetical zeolite structures are being developed and they could in principle be screened to discover the best structures for a given industrial application. Our work focuses on the task of CO₂ capture. The current state-of-the-art molecular simulations allow for accurate *in silico* prediction of zeolite properties, but the computational cost of such calculations prohibits their application in the characterization of the entire database of hypothetical structures, which would be required to perform brute-force screening for novel structures with useful properties.

Our work focuses on the development of chemoinformatics approaches that allow an efficient screening of such databases. Our techniques rely on the similarity principle exploited in a combination of similarity searching and diversity selection approaches. We have developed novel structure descriptors and have defined appropriate similarity measures to describe complex, periodic chemical structures such as zeolites. We use concepts from modern mathematics, mainly topology and computational geometry to characterize the pore structure of a porous material. Our similarity measure takes into account both topology of a material as well as geometrical parameters to efficiently quantify similarities between different pore structures. The resulting screening approach requires expensive characterization only for carefully selected and statistically relevant subset of a database.

P-11 : The file IO round robin game: on the development of a consistent chemical representation

A. Kolodzik, S. Urbaczek, R. Fischer, T. Lippert, M. Rarey

Center for Bioinformatics, University of Hamburg, Hamburg, Germany

Every application program in cheminformatics depends on external data which is most commonly provided by chemical file formats. These data is converted into an internal molecular representation and then used in subsequent calculations. Since molecules are represented in many different ways in the common file formats, a consistent chemical model is needed to ensure the appropriate interpretation of the provided data. Most widely used software tools allow to import various file formats but unfortunately, this often results in inconsistencies. As a simple test, one can read-write molecules in different formats and analyze the resulting structures.

We present a new model for the representation of molecules and its application in a converter tool called NAOMI. The main focus is to provide a reliable and accurate internal representation of molecules that allows a consistent calculation of molecular descriptors regardless of the provided input file format.

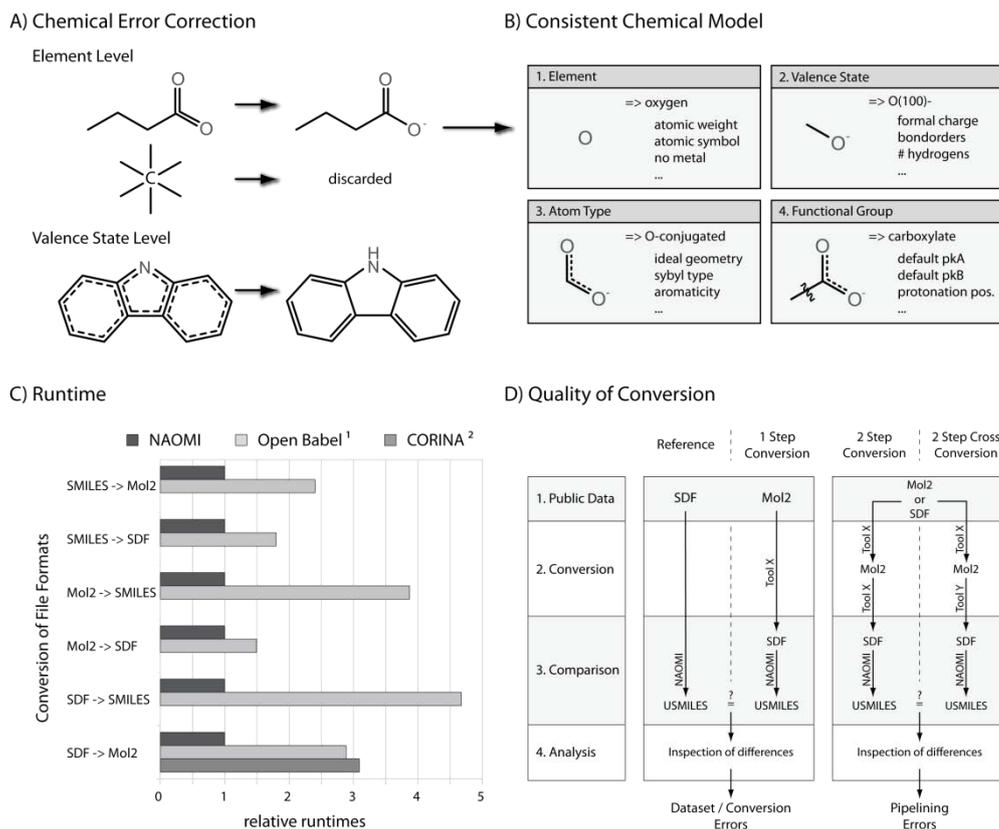


Figure 1: A) Correction of common chemical errors B) Consistent chemical model that is independent of input file formats C) Runtime of NAOMI compared to OpenBabel and CORINA for conversion of the ZINC-everything data set to SMILES³, Mol2⁴ and SDF⁵ D) Validation strategy for conversion of file formats

When interconverting SMILES, Mol2 files and SDF files, NAOMI outperforms commonly used tools with respect to quality of generated files and runtime (see Figure 1). Molecular structures in converted files are guaranteed to have a valid valence bond structure and contain valid formal charges.

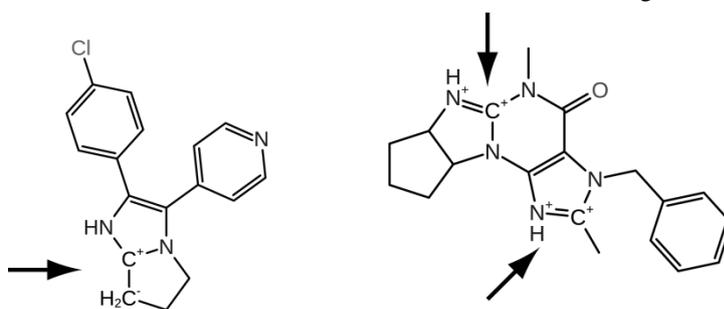


Figure 2: Examples for incorrectly converted molecules

Table 1: Differently converted molecules for conversion of the DUD⁶ decoys from MOL2 to SDF compared to NAOMI.

| Name of tool | DUD ligands | DUD decoys |
|--------------|-------------|------------|
| Open Babel | 898 (23%) | 8580 (7%) |
| CORINA | 449 (11%) | 5473 (4%) |
| MOE | 103 (3%) | 1859 (2%) |

NAOMI's consistent chemical model and well-defined output allows to easily check large databases for chemically correct molecular representations. Surprisingly, frequently used public data sets still contain

errors. If these data sets are converted from Mol2 to SDF format, output significantly differs for CORINA, MOE and Open Babel if compared to NAOMI (see Table 1). Manual inspection of output files lead to the identification of several classes of errors (see Figure 2):

- incorrectly converted aromatic systems containing charged nitrogen atoms
- incorrectly charged carbon atoms of amidinium and guanidinium groups
- incorrectly charged carbon atoms of guanidinium- and amidinium-like groups in five-membered aromatic rings

Using NAOMI for preprocessing of input files lowers error rates of third party tools. Therefore, NAOMI can be seen as a data cleaner and should be used in pipelines of consecutively applied tools in cheminformatics.

1. Guha, R.; Howard, M.T.; Hutchison, G.R.; Murray-Rust, P.; Rzepa, H.; Steinbeck, C.; Wegner, J.; Willighagen, E.L. The Blue Obelisk - Interoperability in Chemical Informatics. *J. Chem. Inf. Model.* **2006**, *46*, 991-998
2. Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of Automatic Three-Dimensional Model Builders Using 639 X-ray Structures. *J. Chem. Inf. Comput. Sci.*, **1994**, *34*, 1000-1008.
3. Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.*, **1988**, *28*, 31-36.
4. Tripos, TRIPOS Mol2 File Format. <http://tripos.com/data/support/mol2.pdf> (accessed 29.01.2011)
5. Symyx, Symyx CTfile Formats. <http://www.symyx.com/downloads/public/ctfile/ctfile.jsp> (accessed 20011-01-27)
6. Huang, N.; Shoichet, B. K.; Irwin, J. J.; Benchmarking sets for molecular docking. *J Med Chem*, **2006**, *49*(23), 6789-6801.

P-13 : Building an R&D chemical registration system

E. Martin¹, Monge², J. A. Duret¹, P. Pospisil¹

¹Philip Morris International R&D, Philip Morris Products S.A., Neuchâtel, Switzerland

²blue-infinity, Geneva, Switzerland

There are thousands of different chemical constituents in both the tobacco plant and in smoke from conventional lit-end cigarettes. In order to store and manage this information, we have used state-of-the-art chemoinformatic technologies to build a centralized master chemical database with unique structure identifiers. We call this the Unique Compound Database (UCD).

Structural uniqueness, normalization, and molecular structures are drawn using Symyx cartridge technologies. In a two-stage quality control process, the automated control is reviewed and cross-checked manually. In brief, structural molecules are entered as neutral entities which can be associated with the salt. The salts are listed in a dictionary and bound to the molecule with the appropriate stoichiometric coefficient in an entity called "substance." The substances are associated with batches. The UCD enables the management of user errors in the structure entry by reassigning or archiving the batches. The system also allows updating of the records to include newly discovered properties of individual structures.

As our research spans a wide variety of scientific fields, the database was designed to enable registration of mixtures of compounds, enantiomers, tautomers, and compounds with unknown stereochemistries; it also includes links to other chemical and biological databases. The UCD currently contains more than 8000 unique individual structures that can be searched in a matter of seconds, thus, simplifying the identification of chemical components in tobacco smoke. The future inclusion of analytical spectra in the database is under development.

P-15 : Optimizing metric properties of protein structure descriptors

P. Røgen

Technical University of Denmark, Department of Mathematics, Kgs. Lyngby, Denmark

Background and aim: The number of known protein and RNA 3d-structures grows very fast. As the number of structure pairs grows faster than computer power in time - it gets harder and harder to use pairwise similarity measures to study the known structures. An alternative is to use structural descriptors whose main calculation time is linear in the number of structures. The aim of this work is to optimize the Euclidean metric used in the descriptor space to locally be close to RMSD without destroying the descriptors superior ability to give large distances between folds.

Method: Relatively low dimensional geometric descriptor vectors are sufficient to recognize protein (1, 2) and RNA (3) folds. In these works a descriptor vector v_i is (pre)-calculated for each chain molecule M_i and the standard Euclidean distance $\|v_i - v_j\|$ is used as a similarity measure (in fact a pseudo metric) between the original chain molecules M_i and M_j . The descriptor space is Euclidean making all-against-all comparisons, automatic classification (1) and clustering of both native and model protein structures (4) very fast and efficient. The goal of this work is to maintain the benefits of a Euclidean descriptor vector space and to choose the optimal Euclidean metric $\|v_i - v_j\|_Q = \sqrt{(v_i - v_j)^t Q (v_i - v_j)}$ in it. Here Q is a positive semi definite matrix.

For small structural deformations of one chain, most similarity measures are highly correlated with RMSD. Therefore *Problem 1* is: If protein structure M_i is a smaller deformation of structure M_j we want to have an isometric representation locally, i.e., $\|v_i - v_j\|_Q \cong RMSD(M_i, M_j)$.

Problem 2: As Problem 1 and if proteins M_i and M_j belong to different folds we want fold separation $\|v_i - v_j\|_Q > RMSD(M_i, M_j)$.

Problem 3: As Problem 2 but also optimizing the automatic classification procedure presented in (1).

Results: We present an optimal way to formulate *Problems 1-3* as linear semi definite optimization problems. On protein data our results are: We can improve the local metric properties of the descriptor space while maintaining the ability to automatically classify close to the 96% of all chains reported in (1). Furthermore, the dimension of the resulting descriptor space drops such that less information is needed to be stored and the method may be seen as a way to combine and weight information of different type.

1. Røgen, P.; Fain, B. Automatic classification of protein structure by using Gauss integrals, *P. Natio. Acad. Sci.*, **2003**, 100(1), 119-124,
2. Røgen, P.; Karlsson, P. W. Parabolic section and distance excess of a space curve applied to protein structure classification. *Geon. Dedicata* **2008**, 134, 91-107.
3. Kirillova, S.; Tosatto, S. C.; Carugo, O. FRASS: the web-server for RNA structural comparison. *BMC Bioinformatics* **2010**, 11, 327.
4. Harder T.; Borg, M.; Boomsma, W.; Røgen, P.; Hamelryck, T. Clustering very large amounts of protein structures using Gauss Integrals, *In preparation*

P-17 : Visual chemical patterns: from automated depiction to interactive design

K. Schomburg, M. Rarey

Center for Bioinformatics, University of Hamburg, Germany

Chemical patterns are employed in a growing number of various pharmaceutical, chemical and computational applications. Among others, a typical use case of chemical patterns is the filtering of a database of molecules. The patterns used for filtering represent either unwanted reactive functional groups that are excluded from the database or are used to extract compounds that are similar to a structure with

some desired activity. Although more and more chemoinformatic applications depend on chemical pattern expressions, the possibilities to represent these patterns are limited. There are linear chemical pattern languages which are designed for efficient computational processing like MQL¹, SLN² and SMARTS³ but are hardly suited for being created and interpreted by humans. For being able to express generic chemical patterns, these languages extend the concept of molecule representations with chemical properties as well as logic expressions like AND, OR and NOT. Since these features are mapped to alphanumeric characters, the patterns result in very complicated strings that have to be deciphered to understand the meaning of a pattern. So far, no representation exists, that allows a straightforward interpretation and design of chemical patterns.

Structure diagrams, 2D depictions of molecules, allow scientists to easily analyze the structure and function of compounds. Therefore, a comparable representation of chemical patterns assists the understanding, editing, creating and validating of pattern expressions. Due to the complexity of chemical pattern languages, a significant extension of the visualization concept for molecules is needed to depict patterns. With SMARTSviewer 4, a concept is introduced that depicts SMARTS expressions visually. Based on the structure diagram concept, a fully automated procedure visualizes a chemical pattern like a molecular structure. However, new graphic elements are introduced to depict property definitions and logical operators used in patterns. Furthermore, the visualization of the pattern is enhanced by a textual legend of the pattern making it easily interpretable. Examples of the SMARTSviewer visualization are shown in Figure 1 where a filter for an unwanted reactive group is depicted and in Figure 2 where a pattern defining a rotatable bond is shown together with the textual legend.

Visualizing patterns is only the first step in supporting scientists in formulating and working with patterns. Similar to molecular editors, chemical pattern editors can assist scientists and erase the need to employ difficult linear pattern languages. Here we introduce the first step towards a visual SMARTS editor. It supports the design of SMARTS pattern significantly by interactively selecting and editing parts of the pattern. The final aim of a SMARTS visualization with an editor is to hide the linear pattern languages from the user completely.

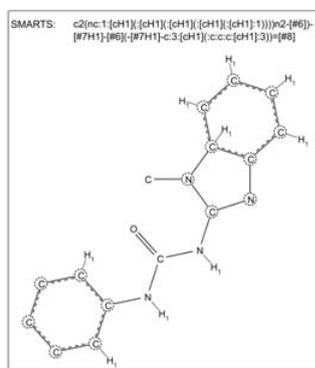


Figure 1: A pattern used as a filter for virtual screening [5] depicted by the SMARTSviewer in the element-symbols modus.

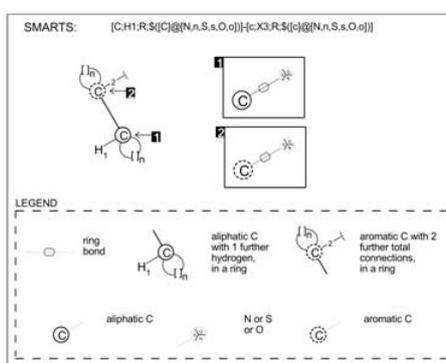


Figure 2: A SMARTS pattern representing a rotatable bond, shown together with the textual legend of each distinct atom and bond.

1. Proschak, E., Wegner, J. K., Schüller, A., Schneider, G., Fechner, U. Molecular Query Language (MQL) – A Context-Free Grammar for Substructure Matching. *J. Chem. Inf. Model.* 2007, 47, 295-301.
2. Homer, R.W., Swanson, R., Jilek, R. J., Hurst, T., Clark, R. D. SYBYL Line Notation (SLN): A Single Notation To Represent Chemical Structures, Queries, Reactions, and Virtual Libraries. *J. Chem. Inf. Model.* 2008, 48, 2294-2307.
3. James, C. A.; Weininger, D. Daylight Theory Manual. Daylight Chemical Information Systems, Inc of Aliso Viejo, CA [online] 2008, <http://www.daylight.com/dayhtml/doc/theory/index.pdf> (accessed Jan 26, 2011)
4. Schomburg, K., Ehrlich, H.-C., Stierand, K., Rarey, M. From Structure Diagrams to Visual Chemical Patterns. *J. Chem. Inf. Model.* 2010, 50, 1529-1535
5. Baell, J. B., Holloway, G. A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J. Med. Chem.* 2010, 53, 2719-2740

P-19 : withdrawn

P-21 : Condensing chemical reactions to pseudo-molecules: an efficient way of reactions mining

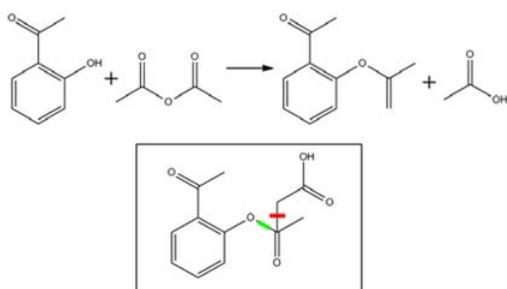
A. Varnek

Laboratoire d'Infochimie, UMR 7177 CNRS - Université de Strasbourg

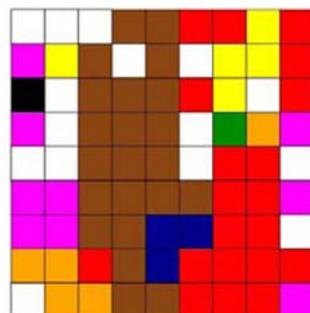
Chemical reactions are difficult objects in chemoinformatics because they involve several species of two different types: reactants and products. The “Condensed Graph of Reaction” (CGR) approach opens new perspectives in the mining of reaction databases since it allows one to transform several 2D molecular graphs describing a chemical reaction into one single graph (see Figure). Thus, a chemical reactions database can be transformed into a set of “pseudo-compounds” which can be easily encoded by descriptors or by fingerprints. Thus, most of chemoinformatics methods developed for individual molecules could be applied to chemical reactions.

Here, we discuss an application of the CGR approach to the (i) reactions classification in large reaction databases, (ii) reaction similarity search and (iii) QSPR modelling of reaction parameters. Particular attention will be paid to classification of the metabolic reactions (predictions of EC numbers), and predictions of metabolites for oxidation reactions catalysed by CYP1A2 and CYP3A4 enzymes.

In these studies ISIDA fragment descriptors extracted from CGR have been used. The advantages of CGR over conventional approaches will be illustrated on several “difficult” case studies.



Phenol acetylation reaction and related Condensed Graph of Reaction. Dynamical bonds marked with green and red colors correspond, respectively, to formation and breaking a single bond.



The Kohonen map generated from CGRs separates different types of reactions (Diels-Alder, Sonogashira, metathesis, ...)

1. I. Baskin and A. Varnek Fragment Descriptors in SAR/QSAR/QSPR Studies, Molecular Similarity Analysis and in Virtual Screening. In “Chemoinformatics Approaches to Virtual Screening”, A. Varnek and A. Tropsha, Eds., RSC Publishing, 2008
2. A. Varnek, D. Fourches, D. Horvath, O. Klimchuk, C. Gaudin, P. Vayer, V. Solov'ev, F. Hoonakker, I. V. Tetko, G. Marcou, ISIDA - Platform for virtual screening based on fragment and pharmacophoric descriptors, Current Computer-Aided Drug Design, 2008, 4 (3), 191-198.

P-23 : Similarity based virtual screening using frequency-based weighting-schemes: effect of the choice of similarity coefficient

H. Xiang, J. Holliday, P. Willett
Information School, University of Sheffield, United Kingdom

Any similarity measure that is to be used for similarity-based virtual screening (SBVS) has three principal components: the structure representation, the weighting scheme, and the similarity coefficient¹. The many previous studies of SBVS that have been carried out have demonstrated that effective screening can be achieved using binary fingerprints and the Tanimoto coefficient.

In recent work, we sought to determine whether increases in performance can be achieved by weighting the bits in a fingerprint that represent the presence or absence of 2D substructural fragments, rather than using just the presence or absence that is encoded in a binary fingerprint. Two types of weighting-scheme were investigated, these being the weighting of fragments on the basis of their frequencies of occurrence within an individual molecule² and on the basis of their inverse frequencies of occurrence within the entire database that was to be searched³. Analysis of our extensive results demonstrated that the precise level of performance that could be expected was strongly dependent on the relative magnitudes of the three components of the Tanimoto coefficient: the bits that describe the reference structure's fragments, the bits that describe a database structure's fragments, and the bits that describe the fragments common to both molecules. This suggests that a similarity coefficient that uses these components in a different way might respond differently to the two types of weighting-scheme.

The work reported here seeks to ascertain whether this is, indeed, the case and whether it is possible to identify a similarity coefficient that might (either in certain circumstances or more generally) be superior to the Tanimoto coefficient when either or both of our frequency-based approaches to fragment weighting are used. Our initial experiments have involved using the cosine coefficient to study occurrence-based weighting² on sets of bioactive molecules selected from the MDDR and WOMBAT databases. These experiments demonstrate clearly that the cosine coefficient retrieves greater numbers of active molecules than does the Tanimoto coefficient, when averaged over multiple searches and multiple types of bioactivity. Moreover, the cosine coefficient is noticeably less affected by changes in the nature of the weighting scheme that is used, whereas the Tanimoto coefficient can give very low levels of performance with some types of weighting scheme. An analysis of the characteristics of the two coefficients provides a rationalization for the cosine coefficient's superior retrieval behavior.

Subsequent work will extend the present study to consider other types of similarity coefficient when used for occurrence-based weighting, and, in the longer term, the use of inverse frequency weighting schemes.

1. Willett, P. Similarity Methods in Chemoinformatics. *Ann. Review Inf. Sci. Technol.* **2009**, 43, 3-71
2. Arif, S. M.; Holliday, J. D.; Willett, P. Analysis and Use of Fragment Occurrence Data in Similarity-Based Virtual Screening". *J. Comput.-Aided Mol. Design* **2009**, 23, 655-668.
3. Arif, S. M.; Holliday, J. D.; Willett, P. Inverse Frequency Weighting of Fragments for Similarity-Based Virtual Screening. *J. Chem. Inf. Model.* **2010**, 50, 1340-1349.

P-25 : Reaction enumeration and machine learning enhancements for the open-source pipelining solution CDK-Taverna 2.0

A. Truszkowski^{1,3}, S. Neumann², A. Zielesny³, E. Willighagen⁴, C. Steinbeck¹
¹ *Chemoinformatics and Metabolism, European Bioinformatics Institute (EBI), Cambridge, UK*

² *GNWI – Gesellschaft für naturwissenschaftliche Informatik mbH, Oer-Erkenschwick, Germany*

³ *University of Applied Sciences of Gelsenkirchen, Institute for Bioinformatics and Chemoinformatics, Recklinghausen, Germany*

⁴ *Department of Pharmaceutical Biosciences, Uppsala University, Uppsala, Sweden*

Pipelining or workflow tools allow for the LegoTM-like, graphical assembly of I/O modules and algorithms into a complex workflow which can be easily deployed, modified and tested without the hassle of

implementing it into a monolithic application. The CDK-Taverna project aims at building an open-source pipelining solution through combination of different open-source projects such as Taverna ¹, the Chemistry Development Kit (CDK) ^{2,3} or Bioclipse ⁴. A first integrated version 1.0 of CDK-Taverna was recently released to the public ⁵.

CDK-Taverna 2.0 provides a refactored setup and worker architecture on the basis of the latest versions of Taverna (2.2) and CDK (1.3.5) which themselves introduce major improvements to the whole platform. Specific new functions and options extend the reaction enumeration abilities based on a reaction template and corresponding reactant libraries (e.g. multi-match detection, arbitrary numbers of reactants, products and generic groups as well as variable R-groups, ring sizes or atom definitions). The machine learning capabilities are significantly enhanced by new workers that provide access to the open-source WEKA library ⁶. The new version also supports 64-bit computing that allows a fast processing of large data volumes in memory.

1. Oinn, T, Addis, M and Ferris, J, Marvin, D, Senger, M, Greenwood, M, Carver, T, Glover, K, Pocock, MR, Wipat and A, P Li, Taverna: a tool for the composition and enactment of bioinformatics workflows, *Bioinformatics* 2004, 20(17), 3045-3054.
2. Steinbeck, C, Han, YQ, Kuhn, S, Horlacher, O, Luttmann, E and E Willighagen, The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics, *Journal of Chemical Information and Computer Sciences* 2003, 43(2), 493-500.
3. Steinbeck, C, Hoppe, C, Kuhn, S, Guha, R and EL Willighagen, Recent Developments of The Chemistry Development Kit (CDK) - An Open-Source Java Library for Chemo- and Bioinformatics, *Current Pharmaceutical Design* 2006, 12(17), 2111-2120.
4. Spjuth, O, Helmus, T, Willighagen, EL, Kuhn, S, Eklund, M, Steinbeck, C and JE Wikberg, Bioclipse: An open rich client workbench for chemo- and bioinformatics, *BMC Bioinformatics* 2007, 8, 59.
5. Kuhn, T, Willighagen, EL, Zielesny, A and C Steinbeck, CDK-Taverna: an open workflow environment for cheminformatics, *BMC Bioinformatics* 2010, 11, 159.
6. Hall, M, Frank, E, Holmes, G, Pfahringer, B, Reutemann, P and IH Witten, The WEKA Data Mining Software: An Update, *SIGKDD Explorations* 2009, 11(1), 10-18.

P-27 : Models of substrates and inhibitors of P450 isoenzymes

M. Elias, D. Marcus, A. Goldblum

Molecular Modeling and Drug Design Laboratory, Institute for Drug Research, School of Pharmacy, Faculty of Medicine, The Hebrew University of Jerusalem, Israel

There is increasing interest in optimizing molecules at early stages of drug development, for efficacy as well as for pharmacokinetic and toxicological properties. Toxicology issues account for 20% of late development stage dropouts, therefore implementing early screening may help to enrich desired properties ¹. The Cytochrome P450 enzyme family is a major target for studies due to its wide participation in most metabolic transformations of marketed drugs ². The potential interactions of newly developed molecules with this enzyme family, either as substrates, inhibitors or inducers, can dramatically affect their chances for further clinical development.

We used our Iterative Stochastic Elimination (ISE) ³ algorithm to build a ranking classification model for any molecule's potential to become a ligand of this enzyme. The learning set for this model's construction is based on a hand-curated diverse (Tanimoto < 0.8) database of P450 ligands, for increased robustness ⁴. Molecular descriptors were computed for 3D ligand structures from constrained-docking. Docking was performed with published 3D crystal structures of the CYP isoforms.

The combined information from conformations and from 2D molecular descriptors was used to build a model by choosing a small optimal subset of properties and their ranges from a very large initial pool of possible descriptors ⁵. The combined model demonstrates an advantage for including 3D descriptors that increases accuracy by 5-15% over the 2D accuracies of 62%-72%.

We show that ISE is very powerful for finding the optimal models among a huge number of options that are not amenable to any exhaustive optimization. ISE produces the ability to quantify the potential of any

molecule to be a ligand (substrate, inhibitor or activator) of P450 isoforms by using the success of alternative filters of the model, each of them being a result of classification. The combined model integrates structural information and ligand based information that can improve the prediction ability for newly investigated compounds and can serve as a powerful screening tool for new potential candidates for further development in drug discovery.

1. Muster, W. G.; Breidenbach, A.; Fischer, H.; Kirchner, S.; Muller, L.; Pahler, A., Computational toxicology in drug development. *Drug discovery today* **2008**, 13 (7-8), 303-310.
2. Danielson, P. B., The cytochrome P450 superfamily: Biochemistry, evolution and drug metabolism in humans. *Current drug metabolism* **2002**, 3 (6), 561-597.
3. Glick, M.; Rayan, A.; Goldblum, A., A stochastic algorithm for global optimization and for best populations: A test case of side chains in proteins. *P Natl Acad Sci Usa* **2002**, 99 (2), 703-708.
4. Elias, M.; Marcus, D.; Goldblum, A., Predicting metabolic transformation by Cytochrome P450 main isoforms. *The ACS 238th National Meeting & Exposition* **2009**. SCI-MIX (28).
5. Rayan, A.; Marcus, D.; Goldblum, A., Predicting Oral Druglikeness by Iterative Stochastic Elimination. *J Chem Inf Model* **2010**, 50 (3), 437-445.

P-29 : withdrawn

P-31 : Identification of alternative druggable targets involved in the prion disease

J. M. Valencia¹, B. Chen², P. C. Wright³, C. Evans³, J. Louth², V. J. Gillet¹

¹Information School, ²Department of Chemistry, ³Biological and Environmental Systems Group, The University of Sheffield, Sheffield, UK

Human prion diseases are progressive and fatal neurological disorders occurring in a variety of animals and humans. They can be transmitted by an infection-like mechanism causing Transmissible Spongiform Encephalopathies (TSEs) such as Scrapie and variant Creutzfeldt-Jakob disease (vCJD)¹, or they can occur spontaneously, such as sporadic Creutzfeldt–Jakob disease (sCJD)², possibly caused by somatic mutations.

The responsible etiological agent is the proteinaceous infectious particle (PrP^{Sc}), which is able to modify the non-infectious prion isoform (PrP^C), adopting the aberrant insoluble infectious conformation. The infective PrP^{Sc} form is resistant to denaturation and digestion with protease and it is aggregated and accumulated as extracellular deposits leading to neural disorder and thereafter the death of the animals and humans affected. Thus far, no effective therapeutic compounds exist, although there are some compounds capable of slowing the progression of the disease³.

It is known that during the conversion of the cellular isoform PrP^C to the PrP^{Sc} disease form, there are several biological pathways involved⁴. Thus it is possible to implement a systems biology approach to identify protein–protein interaction networks, biological pathways and other informational molecules⁵ altered during this process in order to propose alternative druggable targets to disrupt these interactions.

We report a three stage approach to the target deconvolution of anti-prion compounds discovered in-house. The first stage was to carry out a meta-analysis of the literature to obtain all the possible genes and proteins related to prion transformations and integrate them in the MetaCore program⁶ in order to identify possible targets. The second stage involved an isobaric tag for relative and absolute quantitation (iTRAQ) proteomics analysis of the in-house anti-prion compounds in a scrapie mouse brain cell model⁷ and integration of the data in order to corroborate the targets proposed in the literature analysis. In parallel with the experimental work, we have also explored chemoinformatics approaches to target deconvolution, including inverse docking, as further validation of the identified targets.

1. Brown, K.; Mastrianni, J. A., The Prion Diseases. *J Geriatr Psych Neur* **2010**, *23* (4), 277-298.
2. Weissmann, C., The state of the prion. *Nat Rev Microbiol* **2004**, *2* (11), 861-71.
3. Rigter, A.; Langeveld, J. P.; van Zijderveld, F. G.; Bossers, A., Prion protein self-interactions: a gateway to novel therapeutic strategies? *Vaccine* **2010**, *28* (49), 7810-23.
4. Hwang, D.; Lee, I. Y.; Yoo, H.; Gehlenborg, N.; Cho, J. H.; Petritis, B.; Baxter, D.; Pitstick, R.; Young, R.; Spicer, D.; Price, N. D.; Hohmann, J. G.; Dearmond, S. J.; Carlson, G. A.; Hood, L. E., A systems approach to prion disease. *Mol Syst Biol* **2009**, *5*, 252.
5. McDermott, J., *Computational systems biology*. Humana ; Springer [distributor]: Totowa, N.J. London, 2009; p xviii, 587 p.
6. Schuierer, S.; Tranchevent, L. C.; Dengler, U.; Moreau, Y., Large-scale benchmark of Endeavour using MetaCore maps. *Bioinformatics* **2010**, *26* (15), 1922-1923.
7. Thompson, M. J.; Louth, J. C.; Greenwood, G. K.; Sorrell, F. J.; Knight, S. G.; Adams, N. B. P.; Chen, B. N., Improved 2,4-Diarylthiazole-Based Antiprion Agents: Switching the Sense of the Amide Group at C5 Leads to an Increase in Potency. *Chemmedchem* **2010**, *5* (9), 1476-1488.

P-33 : DEGAS – sharing and tracking target compound ideas with external collaborators

M. Lee, I. Aliagas, J. Dotson, A. Gobbi, T. Heffron
Genentech Inc., South San Francisco, USA

Today's drug discovery teams are required to propose clinical candidates with good physicochemical properties to reduce the risk of failure in clinical studies. On the other hand, the pressure to deliver clinical candidates in shorter time has kept increasing. Therefore, sharing target compound ideas and tracking their synthesis progress has become important, especially if drug discovery teams are on different continents.

The goal of the presentation is to show how Genentech has implemented DEGAS to ensure that only the best target compounds are made and that they are made without duplicate effort. This application is accessible by Genentech users and users outside Genentech, enabling collaborative prioritization of target compounds and synthesis progress tracking. The ease of use has kept the effort for training and supporting external users to a minimum.

DEGAS is an integration of a Pipeline Pilot web-port protocol¹ and a specific configuration of AEREA, a reporting application presented at the ICCS 2008². The Pipeline Pilot protocol allows the users to register new target compounds ideas. In addition, it takes care of computing properties such as cLogD, cLogP, cpKa³, and predicting microsomal and hepatic stability⁴. The data are stored in a relational database. The data model is designed to link the DEGAS data to the in-house database containing the registered compounds via structure identity.

The AEREA configuration allows the users to review the structures, computed physicochemical properties and predictions. The link to the database with registered compound enables the inclusion of identifiers for already registered compounds and is thus useful for the novelty check of target ideas. AEREA also provides the functionality to enter annotations such as rationales on target compounds, and the status of a synthesis. A Pipeline Pilot protocol for docking of selected target compounds can also be invoked and the results are available in AEREA to all team members.

This presentation will give an overview of the application and on how access by our external partners is enabled. It will show examples of how DEGAS is used by project teams and discuss the impact of DEGAS on Genentech's small molecule drug discovery that is heavily relying on contract research organizations.

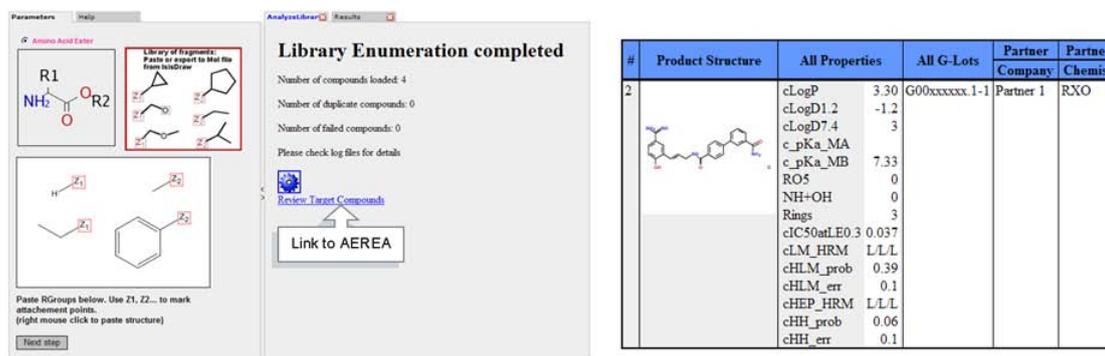


Fig. 1 The screenshot on the left shows the UI for registering target compound ideas. The screenshot on the right is an example report with a subset of computed properties and stability predictions, the “All G-Lots” column containing any existing Genentech lot identifiers and columns showing who is working on the given target compound.

1. Accelrys Inc., USA. <http://accelrys.com/>
2. Lee, M.; Aliagas, I.; Gobbi, A. Scientific Database Application Without Borders: Empowering The Scientists. In *Program & Abstracts of 8th International Conference on Chemical Structures*; Noordwijkerhout, The Netherlands, 2008, 70-71.
3. Molecular Discovery Ltd., United Kingdom. <http://www.moldiscovery.com/>
4. Aliagas, I. Delivering Integrated Results From Liver Microsome Stability Models and Property Calculations Using Pipeline Pilot. 2009 Accelrys User Group Meeting, San Diego, USA.

P-35 : Analysis of contents in proprietary and public bioactivity databases

P. Tiikkainen, L. Franke

Merz Pharmaceuticals GmbH, Frankfurt am Main, Germany

Activity data for small molecules is invaluable in chemoinformatics. It is required when predicting target relations for novel molecules – be it repositioning of existing drug molecules or developing novel chemical entities.

Various bioactivity databases exist containing detailed information of target proteins and quantitative binding data for small molecules. Larger of these contain hundreds of thousands to millions of data points normally collected from scientific papers and patents. Until recently, these databases were commercial requiring substantial investment. In the past few years the situation has changed with public domain databases such as Pubchem¹ and ChEMBL².

In the current work, we have combined several bioactivity databases by both commercial and public vendors into one bioactivity metabase. The molecular presentation, target information and activity data of the vendor databases were standardized. The main motivation of the work was to combine all distinct data sources available to us into a single relational database which allows fast and simple data retrieval by in-house scientists. Secondly, we wanted to know the amount of overlap between databases by commercial and public vendors in order to decide if a continued investment in the former is justifiable. Thirdly, we quantified the degree of inconsistency between data sources by comparing data points derived from the same scientific article cited by more than one vendor.

We found that each data source contains unique data which is due to different scientific articles cited by the vendors. When comparing data derived from the same article we found that inconsistencies between the vendors are common. Mostly these were due to a different stereoisomer reported for a small molecule, discrepancies on the reported molecular target and miscellaneous causes such as decimal errors in the activity value.

In conclusion, using databases by different vendors is still justifiable since the data overlap is not complete. Though, it should be noted that this can be partially explained by the inconsistencies and errors we have found in the source data.

1. The Pubchem Project. <http://pubchem.ncbi.nlm.nih.gov>
2. ChEMBL. <https://www.ebi.ac.uk/chembl/index.php>

P-37 : Mining chemical IP with OSRA

I. V. Filippov¹, M. C. Nicklaus²

¹ *Chemical Biology Laboratory, SAIC-Frederick, Inc., NCI-Frederick, Frederick, Maryland 21702, USA*

² *Chemical Biology Laboratory, NCI, NIH, DHHS, NCI-Frederick, Frederick, Maryland 21702, USA*

Chemical information by its very nature is often hard to convey by text alone. Until the widespread use of personal computers, the most popular way of depicting a molecule in a journal article or a patent document was by means of a structure diagram. Such images, while easy to understand for a human, present tremendous obstacles to a computer-led data mining process. The use of machine learning techniques is becoming more important as computing power becomes more available and the number of potential sources of chemical information soars. The patents alone now comprise millions of documents, and to manually process them using human curators is prohibitively expensive. Optical Structure Recognition Application is a software tool developed by the Computer Aided Drug Design group at the Chemical Biology Laboratory, National Cancer Institute - Frederick. It allows for extraction and recognition of molecular structure images from patents and other documents, and conversion into several widely-used cheminformatics formats, such as SMILES, SDF and others. We present the application of OSRA to extract chemical information from the United States Patent Office, Japanese Patent Office, and WIPO documents. New functionality leading to improved recall rates based on a systems combination approach is reported.

P-39 : Efficient sparse and probabilistic binary classifier

H. Mussa, R. Lowe, R. Glen

Unilever Centre for Molecular Sciences Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, CB2 1EW, U.K.

Developing accurate and computationally efficient binary classifiers is an open challenge in Cheminformatics (and in Machine Learning in general).

Typically the available data set is noisy and finite; and sometimes the two classes may overlap intrinsically. In these scenarios, employing probabilistic binary classifiers (PBC) is essential. However, PBC suffer from well documented severe computational bottle-necks.¹ Platt arguably partially addressed these computational bottle-necks² by coercing PBC from Support Vector Machines, which are by design deterministic.

Tipping argued that classification estimates yielded by Platt's approach can be unreliable.³ He [Tipping] then introduced a sparse probabilistic binary classifier³ based on Bayesian techniques: A sparse classifier in the sense that only few of the available training data points are required to generate the classifier. In other words, his approach, which he called the Relevance Vector Machine (RVM), can become computationally efficient when it comes to classifying a new molecule. However, in practice, yielding the estimated RVM binary classifiers requires some conceptual approximations and the estimation process may also entail intensive computations.

In this talk we briefly describe a binary classifier which is probabilistic and sparse, but unlike the RVM approach, its estimation does NOT require intensive computation.

We also present some results of realistic chemical classification problems solved with the algorithm we proposed. A comparison between the performance of our binary classifier and other binary classifiers based on the RVM method is presented.

1. Friedman, J. H., *Journal of Classification*, 2006; p 175-197
2. Platt J. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In *Advances in Large Margin Classifiers*; Smola, A.; Bartlett, P.; Scholkopf, B.; and Schuurmans, D., Ed.; Cambridge MA, 2000
3. Tipping, M.E. The Relevance Vector Machine. In *Advances in Neural Information Processing Systems*; Solla, S.A.; Leen, T.K.; and Muller, K.R., Ed.; MIT Press, 2000; Vol. 12; p 652-658

P-41 : Improved chemical test mining of patents using automatic spelling correction and infinite dictionaries

R. Sayle¹, P. Xie², P. Petrov², J. Winter³, S. Muresan²

¹*NextMove Software, Cambridge, UK*

²*AstraZeneca, Mölndal, Sweden*

³*AstraZeneca, Alderley Park, UK*

The text mining of patents and patent applications for chemical structures of interest to medicinal chemists poses a number of unique challenges not encountered in other fields of text analytics. Traditional text mining relies on the co-occurrence of common terms between documents to provide similarity measures that can be used to cluster and rank related documents. The more words shared between two documents, the more similar they are, and the greater the probability that they discuss the same topic. By contrast, in pharmaceutical “composition of matter” patents the novel and unique chemical entities are far more significant than those that can be found elsewhere. Although the text of a pharmaceutical patent may explicitly name thousands of individual compounds, and via generic Markush structures claim an infinite number, the role of these patents is to protect the intellectual property of only one or perhaps two drug candidates.

In this poster, we present an analysis of the “quality not quantity” of structures extracted by automatic Chemical Named Entity Recognition (CNER) methods, using a number of name-to-structure programs¹, both on a small hand-curated benchmark set of 50 US top-selling drug patents and a large-scale analysis of a comprehensive database of 12 million patents^{2,3}. Our results show the limited value of traditional lexicon/dictionary based approaches in extracting “key” compounds, and that the major impediment is not the performance of the name-to-structure software used, but the high rate of OCR errors, typos and lexicographic problems found in patent-office data feeds. To address this problem, novel algorithms for automatic spelling correction of chemical names have been developed, that take advantage of the grammar used in IUPAC-like nomenclature. This forms a preprocessing pass, independent of the name-to-structure software used, and is shown to greatly improve results in our study.

1. Sayle, R. Foreign Language Translation of Chemical Nomenclature by Computer. *J. Chem. Inf. Model.* **2009**, 49, 519-530.
2. Rhodes, J.; Boyer, S.; Kreulen, J.; Chen, Y.; Ordonez, P. Mining Patents using Molecular Similarity Search. *Pac Symp Biocomput.* **2007**, 12, 304-315.
3. Suriyawongkul, I.; Southan, C.; Muresan, S. The Cinderella of Biological Data Integration: Addressing the Challenges of Entity and Relationship Mining from Patent Sources. In *Data Integration in the Life Sciences. Lect. Notes Bioinf.* Vol. 6254, Springer 2010.

P-43 : Automated extraction of chemical information from documents: recent advances in the CLiDE projectA. Valko¹, P. Johnson²¹*Keymodule Ltd., Leeds, UK*²*University of Leeds, Leeds, UK*

CLiDE, an acronym for Chemical Literature Data Extraction ¹, is a software application that interprets molecular images found in a variety of sources and extracts meaningful chemical information such as structure connection tables that can be stored in standard electronic formats. The tool is commonly used to extract chemical structures from scientific literature, patents, legacy corporate documents and miscellaneous image files.

During the last few years, CLiDE has seen extensive development aimed at making it an easy-to-use aid for experimental chemists. The system can be used either interactively or in batch mode for unsupervised extraction. The interactive versions are now equipped with a new document browser style user interface, akin to the Adobe Acrobat Reader. The structures extracted by CLiDE can be seamlessly transferred to drawing software such as ChemDraw, ISIS/Draw or Accelrys Draw, edited where necessary and saved in a variety of chemical formats.

The presentation will focus on a number of difficult problems faced in this ongoing work and the techniques used to solve some of them. These include (a) automated detection of regions of a document which contain chemical structures; (b) automated identification and storage of possible errors in interpretation in order to flag the need for manual editing (especially important for batch processing); (c) structural motifs which frequently cause errors; (d) capture of data associated with structures.

CLiDE has been tested on a large number of images and documents originating from various sources. The results of these tests will be summarized to show the level of accuracy of recognition that is achievable with CLiDE.

1. Valko, A. T; Johnson, A. P. CLiDE Pro: The Latest Generation of CLiDE, a Tool for Optical Chemical Structure Recognition. *J. Chem. Inf. Model.*, 2009, 49(4), 780-787.

P-45 : Quantifying model errors using similarity to training dataR. Brown¹, D. Honeycutt¹, S. Aaron²¹*Accelrys Inc, San Diego, CA USA*²*Accelrys Ltd, Cambridge, UK*

When making a prediction with a statistical model, it is not sufficient to know that the model is "good", in the sense that it is able to make accurate predictions on test data. It is also important to ask: How good is the model for a specific sample whose properties we wish to predict? Stated another way: Is the sample within or outside the model's domain of applicability or what is the degree to which a test compound is within the model's domain of applicability? A variety of studies have been done on determining appropriate measures to address this question ¹⁻⁴.

In this talk we focus on a derivative question: Can we determine an applicability domain measure suitable for deriving quantitative error bars -- that is, error bars which accurately reflect the expected error when making predictions for specified values of the domain measure? Such a measure could then be used to provide an indication of the confidence in a given prediction (i.e. the likely error in a prediction based on to what degree the test compound is part of the model's domain of applicability).

Consistent with recent work by others ⁵⁻⁶, the measures we have seen that best correlate with model prediction error are distances to individual samples in the training data or distances to the centroid of the training data. In this talk we describe our attempts to construct a recipe for deriving error bars for regression models and predicted ROC AUC scores for classification models as functions of these distances.

1. Eriksson, L. Jaworska, J., Worth, A.P., Cronin, M.T.D, McDowell, R.M., Gramatica, P. Methods for Reliability and Uncertainty Assessment and for Applicability Evaluations of Classification- and Regression-Based QSARs. *Environmental Health Perspectives*, **2003**, 111, 1361
2. Tropsha, A., Gramatica, P, Gombar, V.K. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models, *QSAR Comb. Sci.*, **2003**, 22, 69
3. Jaworska, J., Nikolova-Jeliazkova, N., Aldenberg, T. QSAR applicability domain estimation by projection of the training set descriptor space: a review. *Altern. Lab Anim.* **2005**, 33, 445-59.
4. Stanforth, R.W., Kolossov, E. Mirkin, B. A Measure of Domain of Applicability for QSAR Modelling Based on Intelligent K-Means Clustering. *QSAR & Combinatorial Science*, **2007**, 26, 837.
5. Sheridan, R.P., Feuston, B.P., Maiorov, V.N., Kearsley, S.K. Similarity to Molecules in the Training Set Is a Good Discriminator for Prediction Accuracy in QSAR. *J. Chem. Inf. Comput. Sci.*, **2004**, 44, 1912.
6. Horvath, D, Gilles, M., Alexandre, V. Predicting the Predictability: A Unified Approach to the Applicability Domain Problem of QSAR Models. *J. Chem. Inf. Model.*, **2009** 49, 1762-1776.

P-47 : Proteochemometric modeling as a tool to predict clinical response to anti-retroviral therapy based on the dominant patient HIV genotype

A. Hendriks¹, G. van Westen¹, J. Wegner², H. van Vlijmen^{1,2}, A. IJzerman¹, A. Bender^{1,3}

¹ Division of Medicinal Chemistry, LACDR, Leiden, The Netherlands

² Tibotec BVBA, Mechelen, Belgium

³ Unilever Centre for Molecular Science Informatics, Cambridge, United Kingdom

Although several clinical drugs inhibiting the Human Immunodeficiency Virus (HIV) are on the market, HIV remains a challenging opponent due to fast arising resistance. Resistance inferring mutations occur during the transcription of the viral genome into the host's genome and the question which drug to use against which viral strain (and thus which patient) is only partially answered today. Personalized medicine by screening of the viral genetic code is available and largely based on in vitro susceptibility assays, like PhenoSenseTM and Antivirogram[®]. Some predictive computer models, like VirtualPhenotypeTM-LM¹ and virco[®]TYPE HIV-1², have also been developed. However, these models still have a considerable shortcoming: As they are trained solely on protein sequence information they are unable to extrapolate the activity of known drugs to unknown mutants. This is precisely where Proteochemometric Modeling (PCM) can contribute to improve the prediction of a clinical response to a given drug treatment. PCM is a robust multi target modeling technique taking both ligand and protein information into account.³

In this work we trained a PCM on the VIRCO⁴ database, consisting of about 60,000 unique HIV mutant genomes comprising mutation patterns of both protease (P) and reverse transcriptase (RT) sequences. The VIRCO database also contains activity data of all clinical drugs currently on the market. In the current work we modeled bioactivity data from 4 drug classes, namely NRTIs, NtRTIs, NNRTIs and PIs, comprising a total of 21 drugs. Drugs were described using standard chemical descriptors while proteins were described by the physicochemical properties of the amino acids making up the drug binding site. The final model was able to give a personalized prediction of the activity of clinical inhibitors on individual viral mutants, and hence provides a step towards personalized treatment of viral infections.

1. Vermeiren, H.; Van Craenenbroeck, E.; Alen, P.; Bacheler, L.; Picchio, G.; Lecocq, P. Prediction of HIV-1 drug susceptibility phenotype from the viral genotype using linear regression modeling. *J. Virological Methods* **2007**, 145, 47-55.
2. Van Houtte, M.; Picchio, G.; Van Der Borgh, K.; Pattery, K.; Lecocq, P.; Bacheler, LT. A Comparison of HIV-1 Drug Susceptibility as Provided by Conventional Phenotyping and by a Phenotype Prediction Tool Based on Viral Genotype. *J. Med. Virol.* **2009**, 81, 1702– 1709.
3. van Westen, G.J.P.; Wegner, J.K.; IJzerman, A.P.; van Vlijmen, H.W.T.; Bender, A., Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets. *Med. Chem. Commun.*, **2011**, 2, 16-30.
4. <http://www.vircolab.com/faq/virco-type-hiv-1-faq#one>

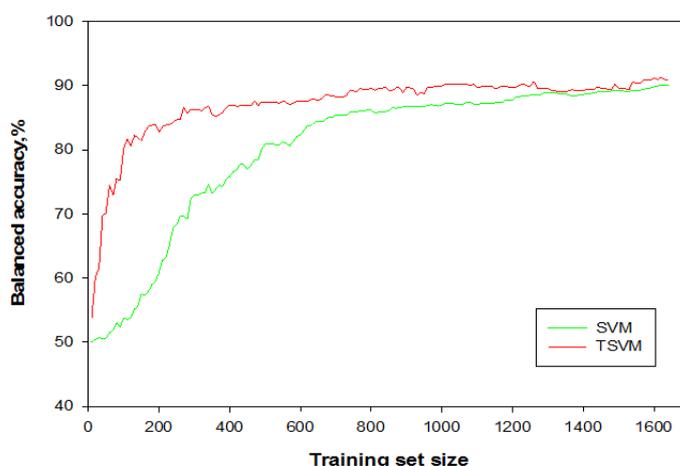
P-49 : Prediction of biological activities using semi-supervised and transductive machine learning methods.

E. Kondratovich^{1,2}, I. Baskin^{1,2}, A. Varnek¹

¹ *Laboratoire d'Informatique, UMR 7177 CNRS - University of Strasbourg, France*

² *Department of Chemistry, Moscow State University, Russia*

Semi-supervised and transductive approaches use unlabelled data (*i.e.* compounds for which properties values are not available) for leveraging predictions. In this work, this ability is demonstrated to predict activities of organic molecules from the DUD database toward some biological targets. Two methods have been used: Transductive Support Vector Machine (TSVM)¹ and Spectral Graph Transducer (SGT)². TSVM builds a separation hyperplane traversing low-density regions of chemical space. SGT performs predictions using spectral decomposition of the Laplacian of the neighborhood graph, built on the objects both from training and test sets. The dependence of the "transductive effect" (a gain of the predictive performance compared to conventional techniques) upon, the size of the training and test sets, their diversity and the "clustering assumption" is reported.



Predictive performance of the SVM and TSVM classification models for acetylcholine esterase inhibitors as a function of the training set size. One may see that TSVM models significantly outperform those of SVM.

P-51 : A global class A GPCR proteochemometric model: a prospective validation

E Lenselink¹, G van Westen¹, J. Wegner², A. IJzerman¹, H. van Vlijmen^{1,2}, A. Bender^{1,3}

¹ *Division of Medicinal Chemistry, LACDR, Leiden, The Netherlands.*

² *Tibotec BVBA, Beerse, Belgium*

³ *Unilever Centre for Molecular Science Informatics, Cambridge, United Kingdom*

Proteochemometrics (PCM) is a robust multi target modeling technique taking both ligand and protein information into account¹. In addition, proteochemometrics models – since they also include target relationships – are conceptually able to also make use of bioactivity data measured against related targets.

In this work we constructed a predictive GPCR PCM model based on all receptors of the class A subtype that we have available. The final model we obtained

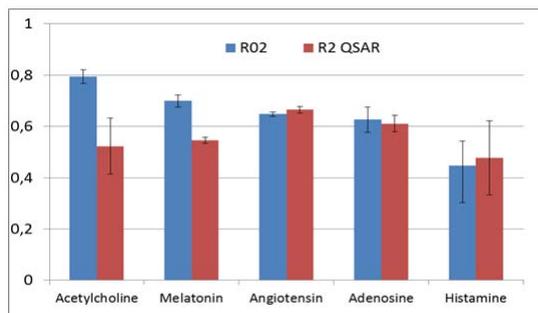


Figure 1: showing performance from different local PCM models versus QSAR models.

comprises 195 different class A GPCRs. This global model was validated and compared with different local models which were constructed on GPCR receptor subfamilies. Bioactivity data from different species were included in the model generation step because it has been shown that this improves model quality. Structures and K_i values were retrieved from both ChEMBL₀₈ and BindingDB^{2,3} resulting in 77004 compounds with 159032 affinities towards the 195 receptors. The GPCR binding pocket in the global model was defined as in Gloriam *et al.*⁵. Both Random forest trees as well as Support Vector Machines (continuous and categorical) models were constructed. For the local models, crystal structure information was taken into account if available. Receptor and chemical similarity was combined to determine the individual GPCR receptor subfamilies.

Depending on the receptor family either local or the global model performed best. In almost all circumstances PCM outperformed QSAR. The best results for regression were from a local subfamily model validated on the muscarinic acetylcholine receptor 5 (externally cross validated R^2_0 was 0.81 RMSE was 0.377) (Figure 1).

Finally the models were used in a prospective screening of the Chemdiv database. Based on model predictions compounds were selected and tested in bio-assay based experimental validation. Compounds were selected for the adenosine and chemokine receptors.

1. van Westen, G.J.P.; Wegner, J.K.; IJzerman, A.P.; van Vlijmen, H.W.T.; Bender, A., Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets. *Med. Chem. Commun.*, **2011**, 2, 16-30.
2. Chen, X.; Lin, Y.; Liu, M.; and Gilson, M.K. The Binding Database: Data Management and Interface Design *Bioinformatics* 18, **2002**, 130-139
3. Liu, T.; Lin, Y.; Wen, X.; Jorriens, R.N.; and Gilson, M.K. *BindingDB*: a web-accessible database of experimentally determined protein-ligand binding affinities *Nucleic Acids Research*, **2007**, 35, D198-D201
4. Gloriam, D.E.; Foord S.M.; Blaney F.E.; and Garland S.L. Definition of the G protein-coupled receptor transmembrane bundle binding pocket and calculation of receptor similarities for drug design. *J Med Chem*, **2009**, 52(14): p. 4429-42.

P-53 : Ames toxicity: strategies to remove Ames liability for anilines

J. Hussain, J. Munoz-Muriedas
Glaxo Smithkline, Stevenage, United Kingdom

Anilines and their substituted derivatives are commonly used for the synthesis of chemicals in the pharmaceutical industry. Examples of marketed drugs containing an aniline moiety include the leprosy treatment dapson and the antibacterial sulphapyridine. However, anilines are also potential genotoxic agents. Their toxicity can be attributed to the oxidizability of their amino group by Phase I metabolic enzymes leading to reactive aromatic hydroxylamines.

Regulatory agencies require all drug candidates to be tested for the potential of causing genetic toxicity. One of the obligatory in vitro tests is the Ames test, assessing the potential of a compound to cause genetic mutations in bacteria. While some anilines appear to be toxic to bacteria, others appear to be safe. The underlying aim of this study is to identify structural changes that can be made to anilines to make them non-toxic.

The study was carried out by extracting all the anilines from an Ames dataset and identifying all the matched molecular pairs (and associated structural changes) within the extracted anilines. The Ames results associated with the compound pairs were then used to identify any structural changes that produce a negative Ames result (i.e. structural changes that result in a toxic aniline becoming non-toxic). A number of structural changes were identified and these will be presented.

In addition, a subset of the interesting structural changes identified were further analysed using QM calculations and the results will be used to help explain why the structural changes result in removing the Ames liability (eg reactivity or steric effects). The findings will be very useful for medicinal chemists to

help synthesise non-toxic anilines and help suggest structural changes to transform a toxic aniline into a non-toxic one.

P-55 : An approach toward the prediction of chemical degradation pathways

M. A. Ott, R. L. Toy, W. G. Button
Lhasa Limited, Leeds, UK

In forced degradation studies, drugs (formulated or pure) are exposed to harsh conditions to study their degradation behaviour. Many degradation products may be formed, some of them unexpectedly. A major challenge is posed by the elucidation of degradation pathways and structural identification of the products, making a predictive expert system a very useful tool.

Major developments have been made in the development of the Meteor system for metabolism prediction.^{1,2} Given the problem stated above, Meteor's functionality provided an obvious and excellent starting point for the development of a new degradation prediction system. Meteor already contained a chemical engine allowing the description and application of degradation transformations, a reasoning engine allowing the assessment of transformation likelihoods,³ and a graphical interface allowing the entry of query structures and the display of prediction results. A number of major changes and extensions were carried out to arrive at a degradation prediction system: functionality for bimolecular reactions (reactions between two query molecules), handling of reaction conditions including their manipulation by the reasoning engine, and the capability to perform a number of related predictions (*e.g.*, with different reaction conditions) automatically. Other capabilities are the same as Meteor's: assessment of likelihoods (called absolute reasoning) and competing reactions (called relative reasoning), filtering of results (by exact mass and/or molecular formula) and display of results (as detailed trees and tables).

The new system, called Zeneth, comprises (1) a program with chemical structure awareness, a reasoning engine and a graphical user interface, and (2) a knowledge base of chemical degradation transformations. It predicts degradation under the influence of reaction conditions and optionally in the presence of other compounds such as excipients. Two of the main advantages of a system such as Zeneth are total recall and the absence of bias. A further major benefit is the steady accumulation of knowledge about degradation chemistry in an accessible form, which can also be a major asset in training. Zeneth is under continuous development, both in program functionality and in knowledge base content. Recent developments are presented, including automatic processing under various conditions, filtering of duplicate structures, calculation of formula differences, and determination of pathway likelihoods. A number of illustrative applications are also given.

1. Marchant, C. A.; Briggs, K. A.; Long, A. In *Silico Tools for Sharing Data and Knowledge on Toxicity and Metabolism: Derek for Windows, Meteor, and Vitic*. *Toxicol. Mech. Methods* **2008**, 18, 177-187.
2. Balmat, A.-L.; Judson, P.; Long, A.; Testa, B. Predicting Drug Metabolism - An Evaluation of the Expert System METEOR. *Chem. Biodivers.* **2005**, 2, 872-885.
3. Button, W. G.; Judson, P. N.; Long, A.; Vessey, J. D. Using Absolute and Relative Reasoning in the Prediction of the Potential Metabolism of Xenobiotics. *J. Chem. Inf. Comput. Sci.* **2003**, 43, 1371-1377.

P-57 : Turning 3D-QSAR weakness into strength with Open3DALIGN & Open3DQSARP. Tosco¹, T. Balle²¹ *Department of Drug Science, University of Turin, Torino, Italy*² *Department of Medicinal Chemistry, University of Copenhagen, Copenhagen, Denmark*

The traditional 3D-QSAR workflow starts with an alignment of ligands in their putative bioactive conformation. The initial step is then followed by calculation of molecular interaction fields (MIFs), extraction of relevant information from the latter by statistical analysis and correlation with activity. The resulting model may have the power to predict the activity of new molecules before they are synthesised and tested.¹

The major weakness of 3D-QSAR methodologies is their dependency on the underlying alignment. Even when the bioactive conformation of a template molecule is known, usually from an experimentally determined structure of a ligand-target complex, the alignment procedure itself is a difficult and time-consuming operation, especially in the presence of flexible or structurally heterogeneous ligands. When the structure or even the identity of the target is not known, it becomes very difficult to hypothesize a univocal and reliable alignment. Unfortunately, the lack of knowledge of the target's structure is also the situation where a ligand-based approach would be most desirable, since it becomes basically the only option for computer-aided drug design.

Herein we present Open3DALIGN,² an open-source tool capable of running conformational searches (by a TINKER³-based QMD engine) and generating unsupervised ligand alignments. The initial alignment bottleneck in 3D-QSAR may be overcome by ranking a large number of possible alignments on the basis of their consistency and the predictive performance of the corresponding 3D-QSAR models, built and evaluated with Open3DQSAR.⁴ This procedure allows to formulate unbiased hypotheses on the bioactive conformation of a series of ligands in the absence of prior knowledge of the target's structure or ligand's SAR. Therefore, the 3D-QSAR alignment dependency is turned from a weakness into strength: binding mode hypotheses may be challenged according to their ability to explain experimental activities from a purely ligand-based perspective. Promising results were obtained applying this methodology on eight benchmark literature datasets.⁵

Most recent developments of Open3DALIGN are discussed, namely the implementation of a novel, all-atom alignment algorithm in addition to the pharmacophore-based one relying on Pharao.⁶

1. Cross, S.; Cruciani, G. Molecular fields in drug discovery: getting old or reaching maturity? *Drug Disc. Today* **2010**, *15*, 23-32.
2. Tosco, P.; Balle, T. Open3DALIGN: an open-source software aimed at unsupervised molecular alignment; <http://open3dalign.org> (accessed Jan 31, 2011).
3. TINKER - Software Tools for Molecular Design, version 5.1; <http://dasher.wustl.edu/tinker/> (accessed Jan 31, 2011).
4. Tosco, P.; Balle, T. Open3DQSAR: an open-source software aimed at high-throughput chemometric analysis of molecular interaction fields; <http://open3dqsar.org> (accessed Jan 31, 2011).
5. Tosco, P.; Balle, T. Poster communication at the 6th German Conference on Chemoinformatics, Goslar, Germany, 7-9 November 2010; http://va.gdch.de/wwwdata/abstracts/5412/5412_0066.pdf (accessed Jan 31, 2011).
6. Pharao version 3.0.3; <http://www.silicos.be/pharao.html> (accessed Jan 31, 2011).

P-59 : Effectiveness of fingerprint-based measures of multi-stage mass spectrometry similarity for virtual screening of chemical structures

M. Rojas-Cherto^{1,2}, J. Peironcely^{1,2,3}, P. Kasper^{1,2}, R. Vreeken^{1,2}, T. Hankemeier^{1,2}, T. Reijmers^{1,2}
¹*Netherlands Metabolomics Centre, Leiden, the Netherlands*
²*Leiden University, Leiden, the Netherlands*
³*TNO Quality of Life, Zeist, the Netherlands*

Multi-stage mass spectrometry (MSⁿ) is a technique that since a few years is started to be used more often for the identification of unknown compounds. The technology used in MSⁿ permits, by consecutive isolation and fragmentation of ions under low-energy collision-induced dissociation (CID), the creation of a set of hierarchical linked mass spectral data. This MSⁿ data provides a more enriched description of the chemical structure of the analyzed compound than the data obtained from single-stage MS or tandem MS. Furthermore, it is known that similar compounds will generate similar MSⁿ data. After organizing the MSⁿ data in a database, methods for *virtual screening* need to be developed in order to extract of a small subset of the database useful or meaningful information, e.g. molecules with high structural similarity.

We present here a new fingerprint-based algorithm for *virtual screening* of MSⁿ databases where there is a need to identify similar compounds using the concept that similar MSⁿ data originate from compounds having similar chemical structure. We also report an investigation of the effectiveness of the algorithm when used for *virtual screening* and compare the results obtained with those from fingerprint-based measures used to define similarities between chemical structures.

P-61 : Homology modeling and binding site prediction of human IRAK-M

J. Du¹, M. Kulharia¹, C. Van't Veer^{2,3}, G. A.F. Nicolaes¹
¹*Department of Biochemistry, Cardiovascular Research Institute Maastricht, Maastricht University, the Netherlands,*
²*Center for Experimental and Molecular Medicine,* ³*Center for Infection and Immunity Amsterdam Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands*

Interleukin-1 receptor-associated kinase M (IRAK-M) is a novel and tissue specific member of the IRAK family with a molecular mass of 68 kDa. IRAK-M consists of two important domains: a so-called death domain and a kinase domain. The death domain has no confirmed interaction partners, even though it contains a predicted protein-protein interaction motif. The kinase domain is reportedly inactive in IRAK-M, as it lacks any measurable kinase activity. Homology models of the death domain and kinase domain, with a high packing quality, were built by use of the protein modeling packages YASARA, WHATIF, ICM-Pro and LOOPY.

Data mining, structure analysis, molecular dynamics simulations and protein-protein interaction predictions (ODA (Molsoft), Cons-PPISP and PPI-Pred) were used to identify regions in IRAK-M that are potentially involved in protein binding interactions. Based on these predictions site-directed mutagenesis studies were performed that verified the correctness of our predictions.

Based on our model of the kinase domain of IRAK-M, we could further confirm that IRAK-M likely contains an inactive serine kinase domain and we have provided a structural explanation for this inactivity, despite the presence of an ATP binding pocket.

P-63 : Key features for designing PPAR γ agonists: an analysis of ligand-receptor interaction by using a 3D-QSAR approach

L. Guasch, Sala, M. Mulero, G. Pujadas, S. Garcia-Vallvé

Grup de Recerca en Nutrigenòmica, Departament de Bioquímica i Biotecnologia, Universitat Rovira i Virgili, Tarragona, Catalonia, Spain

PPAR γ is a ligand-activated transcription factor, member of the nuclear receptor superfamily that plays an important role in adipogenesis and glucose homeostasis. Some PPAR γ agonists, such as the thiazolidinediones (TZD), have a powerful insulin-sensitizing action and are used as anti-diabetic drugs. Unfortunately, as anti-diabetic drugs, TZD present various side effects, including weight gain, increased adipogenesis, renal fluid retention, bone fracture and increased incidence of cardiovascular events.¹ Other compounds with poor agonist activities for PPAR γ , called PPAR γ modulators or PPAR γ partial agonists, retain very good anti-diabetic effects without side effects. It has been described that the different degree of stabilization of the PPAR γ helix 12 (H12) induced by ligands determines its behavior as full or partial agonists.² Recently Choi et al.³ have revealed a new mechanism of action for the anti-diabetic effect of some PPAR γ agonists. This mechanism is completely independent of the classical PPAR γ transcriptional activity and consists on the inhibition of the phosphorylation of PPAR γ on serine 273, thereby preventing the dysregulated expression of some genes, including adipin and adiponectin.³ This alternative mechanism could clarify a long standing paradox: why PPAR γ activation by wide range of ligands does not always correlate with the ligands' in vivo efficacy. With this new knowledge, many research groups have to change the past drug discovery efforts which focused exclusively on potency and agonist activity.

In the present study, we use several structure activity relationship (SAR) studies of synthetic PPAR γ partial agonists to explore the features of PPAR γ agonists. Our approach consists of building two 3D-QSAR models: one for the transcriptional activity of PPAR γ and another for the binding affinity between PPAR γ and its ligands. These models let us to attribute which regions are favorable and unfavorable for binding or for the transcriptional activity. Finally, the combination of both models let us to describe which features must have an effective PPAR γ partial agonist.

1. Feldman, P. L.; Lambert, M. H.; Henke, B. R. PPAR modulators and PPAR pan agonists for metabolic diseases: the next generation of drugs targeting peroxisome proliferator-activated receptors? *Curr Top Med Chem* **2008**, *8*, 728-749.
2. Pochetti, G.; Godio, C.; Mitro, N.; Caruso, D.; Galmozzi, A.; Scurati, S.; Loiodice, F.; Fracchiolla, G.; Tortorella, P.; Laghezza, A.; Lavecchia, A.; Novellino, E.; Mazza, F.; Crestani, M. Insights into the mechanism of partial agonism: crystal structures of the peroxisome proliferator-activated receptor gamma ligand-binding domain in the complex with two enantiomeric ligands. *J. Biol. Chem* **2007**, *282*, 17314-17324.
3. Choi, J. H.; Banks, A. S.; Estall, J. L.; Kajimura, S.; Boström, P.; Laznik, D.; Ruas, J. L.; Chalmers, M. J.; Kamenecka, T. M.; Blüher, M.; Griffin, P. R.; Spiegelman, B. M. Anti-diabetic drugs inhibit obesity-linked phosphorylation of PPAR γ by Cdk5. *Nature* **2010**, *466*, 451-456.

P-65 : Understanding ligand binding affinity and specificity through analysis of hydration site thermodynamicsD. Rinaldo¹ D. Robinson¹, C. Higgs¹, T. Beuming¹, W. Sherman¹, R. Abel¹, R. Farid¹, R. Friesner²¹ *Schrödinger, Inc., 120 West 45th Street, 17th Floor, New York, New York 10036, US*² *Dept. Chemistry, Columbia University, New York City, U.S.A*

Understanding the underlying physics of the binding of small-molecule ligands to protein active sites is a key objective of computational chemistry and biology. It is widely believed that placement of water molecules from the active site by the ligand is a principal (if not the dominant) source of binding free energy. Although continuum theories of hydration are routinely used to describe the contributions of the solvent to the binding affinity of the complex, it is still an unsettled question as to whether or not these continuum solvation theories describe the underlying molecular physics with sufficient accuracy to reliably

rank the binding affinities of a set of ligands for a given protein. Here we develop a novel, computationally efficient method, called Watermap, to assess the contribution of the solvent to the binding free energy of a small molecule and its associated receptor that captures the effects of the ligand displacing the solvent from the protein active site with atomic detail. We will present successful applications of this method to a number of different protein ligand systems such as GPCR, FXa, PDZ domain etc in understanding the ligand binding SAR. Additionally we will also show that detailed analysis of thermodynamics and locations of binding site waters in highly conserved kinase families can yield insight into previously inexplicable selectivity and structure-activity relationship.

P-67 : Insecticide and fungicide likeness: use of two-class Bayesian categorization models for the selection of molecules as screening inputs.

D. A. Demeter, C. Klittich

Discovery Research, Dow AgroSciences, Indianapolis, IN 46268, USA

Pharmaceutical and agricultural chemical companies are constantly in search of new chemicals worthy of testing in biological screening assays. Currently, there are over 11 million unique 3rd party vendor compounds¹ to choose from. Budget constraints and the logistics of handling millions of compounds means that for many companies, only a tiny fraction of the available compounds can be purchased annually, and methods to enrich the potential of compounds to generate hits are critical. At Dow AgroSciences, we have developed two-class Bayesian categorization models based on our own 2nd tier fungicide and insecticide assay results. Over 70,000 compounds (tested identically) were used to derive the individual statistical models. These have been applied as filters along with ag-like criteria in the purchase of compounds from commercial vendors. Hit rates for in vivo activity has been improved by 2-4X for insecticides and 6-15X for fungicides over previous methods.

1. Irwin and Shoichet, J. Chem. Inf. Model. 2005;45(1):177-82
All purchasable (2010-12-05). <http://zinc.docking.org/subset1/>

P-69 : The multi-conformations-receptor-based pharmacophore model generation schema and its potential applications in virtual screening

R. Kurczab, A. J. Bojarski

Department of Medicinal Chemistry, Institute of Pharmacology PAS, Krakow, Poland

The pharmacophore modelling technique is well-know and one of the major tools used in drug discovery for a long time. Up to date, various different approaches to pharmacophore model generation and its application in virtual screening, *de novo* design, lead optimization and multitarget drug design have been reported¹. Depending on existing information about given biological target, the pharmacophore modelling approaches could be divided into two main types: ligand-based and structure-based pharmacophore modelling. In recent years, it also appeared a hybrid approaches^{1,2} mixing those information in according to get pharmacophore model recognizing broader range of ligands and try to describing the dynamic of ligand-receptor complex.

We present here a methodology of generation of non static pharmacophore model based on docking of a collection of diverse know 5-HT7 antagonists to the set of different conformations of 5-HT7 receptor. For the different receptor geometries with the best enrichment factors, we obtained the static pharmacophore models using to build general dynamic one. The obtained averaged pharmacophore model was then validated by an external set build with known actives/decoys ligands³ in order to assess it performance in virtual screening.

Acknowledgements

The study was partly supported by a grant PNRF-103-AI-1/07 from Norway through the Norwegian Financial Mechanism.

1. Sheng-Yong, Y. Pharmacophore modelling and applications in drug discovery: challenging and recent advances, *Drug Disc. Today*, 2010, 15, 444-450.
2. Deng, J.; Leo, K. W.; Sanchez, T.; Cui, M.; Neamati, N.; Briggs, J. M. Dynamic receptor-based pharmacophore model development and its application in designing novel HIV-1 integrase inhibitors. *J. Med. Chem.* 2005, 48, 1496-1505.
3. <https://www.ebi.ac.uk/chembl> (accessed Jan 23, 2010).

P-71 : ChemTattoo3D: an open source drug design and analysis tool

N. MacCuish, J. MacCuish, M. Chapman
Mesa Analytics & Computing, Inc., Santa Fe, NM, USA

ChemTattoo3D is an analysis tool, which analyzes structures with similar shape overlays to find chemical features relevant to drug design. An interactive tool allows the user to visualize compounds which share both shape and pharmacophore features, defined by SMARTS patterns, within a certain distance in 3d space from a defined or perceived target, that the algorithm determines based on a frequency of occurrence modal. This program has been released under the GNU General Public License. This presentation will describe the development of ChemTattoo3D from its 2D predecessors, Stigmata and ChemaTattoo2D. It will also demonstrate applications in the area of lead hopping, shape cluster analysis, and hypothesis generation for structure based drug design.

1. Shemetulskis, N.E.; Weininger, D.; Blankley, C.J.; Yang, J.J.; Humblet, C. Stigmata: An Algorithm to Determine Structural Commonalities in Diverse Datasets. *J. Chem. Inf. Comput. Sci.* **1996**,36(4), 862-871.
2. MacCuish, J.; MacCuish, N.E.; Chapman, M.; Hawrylycz, M. quasi-Monte Carlo methods for the generation of molecular shape fingerprints, in preparation

P-73 : Targeting of the tenase complex by rational design of factor viii-membrane interaction inhibitors

M. Kulharia¹, O. Sperandio², J. Voorberg², S. Wielders¹, B. O. Villoutreix², G. A.F. Nicolaes¹
¹ *Cardiovascular Research Institute Maastricht, Maastricht, The Netherlands*

² *University of Paris V, Paris, France*

³ *Department of Blood coagulation, CLB, Plesmanlaan 125, 1066 CX Amsterdam, The Netherlands*

Coagulation Factor VIII (FVIII) is an essential cofactor molecule of the tenase complex. FVIII binds to the serine protease factor IXa and its substrate factor X, on a membrane surface in the presence of calcium. Formation of the tenase complex can be prevented by preventing the VIIIa-membrane association. The membrane interaction interface of the C2 domain in FVIII consists of solvent exposed, hydrophobic residues resident on two β -turns and an ω -loop. At the base of these hydrophobic protrusions a small pocket is encircled by a group of positively charged residues. For identification of possible inhibitors for the membrane binding activity of FVIII via its C2 domain we adopted a bipronged strategy. One approach involved the use of Omega software to create a conformer database for commercially available, drug-like molecules followed by their rigid docking using FRED. The top scoring molecules were clustered on the basis of various scores to short-list a diversified representative set of interacting molecules which were screened by flexible ligand docking using Surflex to further narrow down the list. In the second approach, structure based pharmacophores were created from a set of known inhibitors. Using these pharmacophores the database of small drug-like molecules was filtered and docking analysis was carried out for the selected molecules. In total 1000 potential lead compounds were purchased from a commercial vendor were tested in a direct interaction assay using the Biacore T100 biosensor instrument, wherein molecules were screened

for their ability to disrupt the FVIII-membrane interaction. The highest scoring compounds were next tested *in vitro* in the more physiological plasma environment to ascertain their use as potential drugs candidates. A quantitative structure-activity relationship (QSAR) model was developed from the activity data which was used in *de novo* molecule design for identification of high affinity FVIII inhibitors. The success rate of our approach is at least 16-fold higher than high throughput approaches that have targeted similar proteins.

P-75 : A fragment - molecule alignment algorithm based on spherical gaussians

N. Stiefl¹, G. Landrum¹, V. Konyukhovskiy², V. Schwartz²
¹Novartis Institute for Biomedical Research, Basel, Switzerland
²GGA Software Services, Cambridge, US

Typical shape-based molecular alignment algorithms perform poorly when applied to molecules of different sizes: the usual result is a fragment aligned somewhere near the middle of the larger molecule.

Here we present a new algorithm for molecular alignment based on spherical gaussians¹ and its extension to fragment-based alignment. The algorithm for fragment-based alignment was inspired by the subshape methodology for aligning shapes encoded on numeric grids²: terminal point sets are generated and then triangle pairs are used to match subshapes onto target molecules. Next, these alignments are used as starting points to optimize the alignment based on gaussian overlap. Using this procedure, good alignments can be achieved in relatively short calculation times.

We will present the results of a number of validation studies which demonstrate the computational speed of the method and the quality of the resulting alignments.

1. Grant, J. A.; Pickup, B. T. Gaussian Description of Molecular Shape *J. Phys. Chem.* **1995**, *99*, 3503-3510.
2. Putta, S.; Eksterowicz, J.; Lemmen, C.; Stanton R. A Novel Subshape Molecular Descriptor, *J. Chem. Inf. Comp. Sci.* **2003**, *43*, 1623-1635.

P-77 : Virtual Ligand Screen of human TRAF6: towards iPPI's as potential anti-inflammatory pharmaca

B. Zarzycka¹, M. Kulharia¹, E. Lutgens², T. Seijkens², S. B. Nabuurs³, G. Vriend³, G. A. F. Nicolaes¹
¹Department of Biochemistry, ²Department of Pathology, Cardiovascular Research Institute Maastricht, Maastricht University, The Netherlands
³Radboud University Medical Centre Nijmegen, Nijmegen, The Netherlands

Tumour Necrosis Factor (TNF) receptor associated factor 6 (Traf6) is involved in dendritic cell differentiation and maturation, cytokine production, osteostasis, cell fate determination and adaptive immunity via T cells¹. TRAF6 act as adaptor protein for the trans-membrane CD40 protein in cell signalling². Recent studies on TRAF6-CD40 interaction suggest that TRAF6 is a potential therapeutic target for the treatment of atherosclerosis³. In our study, in order to identify possible inhibitors of protein-protein interaction (iPPI), for the inhibition of TRAF6-CD40 interactions, structure-based virtual ligand screening was carried out. First a druggable pocket was identified using a consensus pocket prediction method. Nearly a million drug-like molecules from the ChemBridge compound collection were subjected to ADME/Tox filtering with FAF-Drugs2⁴ to remove non-druglike molecules. Next, 579200 molecules were subjected to multiple conformer generation with OMEGA (Openeye) with a maximum of 50 conformers for each molecule. All conformers were screened by the rigid-body docking package FRED (Openeye) into the druggable pocket. The top 60,000 molecules were ranked on the basis of structural, energetic and interaction mode parameters in order to obtain a diversified representative set of interacting molecules. Subsequently, compounds were screened by flexible ligand docking using Surflex⁵ which resulted in a final set of 800 potential binding compounds. The 800 compounds were purchased and subjected to *in vitro*

analysis. At a final concentration of 10 μM in a cell based assay, 48 molecules completely inhibited cell signalling, presumably by inhibition of the Traf6- CD40 interaction. Based on these hits, 2D and 3D similarity searches were performed and resulting similar compounds were screened by flexible docking. To this end, after manual investigation of top scoring molecules a new set of 150 potential lead compounds was chosen for further *in vitro* and *in vivo* investigation.

1. Wu, H.; Arron, J. R. TRAF6, a molecular bridge spanning adaptive immunity, innate immunity and osteoimmunology. *BioEssays*. **2003**, 25, 1096-1105.
2. Pullen, S. S.; Dang, T. T. A.; Crute, J. J.; Kehry, M. R.; CD40 signaling through tumor necrosis factor receptor-associated factors (TRAFs). Binding site specificity and activation of downstream pathways by distinct TRAFs. *J. Biol. Chem.* **1999**, 274, 14246-14254.
3. Lutgens, E.; Lievens D.; Beckers, L.; Wijnands, E.; Soehnlein, O.; Zerneck, 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. Deficient CD40-TRAF6 signaling in leukocytes prevents atherosclerosis by skewing the immune response toward an antiinflammatory profile. *J. Exp. Med.* **2010**, 207, 391-404.
4. Lagorce, D.; Sperandio, O.; Galons, H.; Miteva, M. A.; Villoutreix, B. O. FAF-Drugs2: Free ADME/tox filtering tool to assist drug discovery and chemical biology projects. *BMC Bioinformatics*. **2008**, 9, 396.
5. Jain, A. N. Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.* **2002**, 46, 499-511.

P-79 : Ligand- and structure-based molecular invention using Benchware Muse

F. Bös¹, B. Masek², J. Damewood³

¹ *Tripos International, Martin-Kollar-Straße 17, München, Germany*

² *Tripos International, 1699 South Hanley Road, St. Louis, United States*

³ *AstraZeneca Pharmaceuticals, 1800 Concord Pike, Wilmington, United States*

Successful drug discovery often requires optimization against a set of biological and physical properties^{1,2}. We describe our work on multi-parameter approaches to ligand- and structure-based molecular invention and studies that demonstrate its ability to successfully generate lead hops or scaffold hops between known classes of ligands as well as its application in the field of computational fragment growing.

The software Benchware Muse is based on an evolutionary algorithm that operates on an initial population of structures to invent new structures with improved scores. Muse is unique in that it has the ability to work with any user-defined scoring function that provides results in the form of a numerical evaluation of designed structures^{3,4}.

For ligand-based molecular invention, we describe a multi-criteria scoring function incorporating molecular shape similarity, molecular fingerprint similarity, and a number of popular “Lipinski-like” molecular properties. The structure-based scoring function uses Surflex-Dock to pose and score invented structures in context of the protein and is combined with a number of molecular properties to focus the design on medically relevant chemistries. Coupled with placed substructure constraints, this scoring function enables growing and optimizing attachments to fragments inside the target’s active site.

Several retrospective studies on various targets using the ligand-based scoring function demonstrate the ability of the above mentioned approach to generate novel ideas that are not only appealing to design scientists but are also validated by comparison to compounds known to demonstrate activity at the desired biological target. Specific examples where this approach has generated either significant modifications of existing molecular frameworks or structurally new molecular templates relative to design starting points (i.e., lead hopping) will be provided. The structure-based scoring function has been applied to grow fragments generated from co-crystallized ligands into full-size molecules for various targets. Examples will be shown where invented molecules are shape similar to known ligands and feature similar hydrogen bond patterns although this was specified as one of the design criteria.

1. Feher M.; Gao Y.; Baber J. C.; Shirley W.A.; Saunders J. The use of ligand-based design for scaffold hopping and side-chain optimization: Two case studies *Bioorg. Med. Chem.* 2008, 16, 422-427.
2. Nugiel D.A.; Krumrine J. R.; Hill D. C.; Damewood J. R.; Bernstein P. R.; Sobotka-Briner C. D.; Liu J. W.; Zacco A.; Pierson M. E. Denovo Design of a Picomolar Nonbasic 5-HT1b Receptor Antagonist *J. Med. Chem.* **2010**, 53, 1876–1880.
3. Damewood J. R., Lerman C. L.; Masek B. B. NovoFLAP: A Ligand-Based De Novo Design Approach for the Generation of Medicinally Relevant Ideas *J. Chem. Inf. Model.* **2010**, 50, 1296–1303.
4. Liu Q.; Masek B. B.; Smith K.; Smith J. A Tagged Fragment Method for Evolutionary Structure-Based de novo Lead Generation and Optimization *J. Med. Chem.* **2007**, 50, 5392-5402.



Sponsoring Societies

- Division of Chemical Information (CINF),
American Chemical Society (ACS)
- Royal Netherlands Chemical Society (KNCV)
- Chemistry-Information-Computer Division,
German Chemical Society (GDCh)
- The Chemical Structure Association Trust
(CSA Trust)
- Chemical Information and Computer Applications Group,
Royal Society of Chemistry (RSC)
- Division of Chemical Information and
Computer Science of the Chemical Society
of Japan (CSJ)
- Swiss Chemical Society (SCS)