ICCS

International Conference on Chemical Structures

8th International Conference on Chemical Structures

June 1-5, 2008 ◆ Noordwijkerhout ◆ The Netherlands

Program & Abstracts

www.int-conf-chem-structures.org

Preface

Welcome to the Eighth International Conference on Chemical Structures!

With the Eighth International Conference on Chemical Structures we continue this well-established conference series that begun 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 2008 conference will assuredly continue 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 chosen as the preferred venue to award the third CSA Trust Mike Lynch Award to Professor Alexander Lawson at Elsevier Information Systems. Professor Lawson will kick-off the conference by receiving the award and delivering the keynote address titled *Challenges/opportunities for chemical structure databases in the 21st century* on Sunday evening. Prior recipients of the CSA Trust Mike Lynch Award include Professor Johnny Gasteiger at the University of Erlangen-Nürnberg in 2005 and Professor Peter Willett at the University of Sheffield in 2002.

The scientific poster session has been divided into two sessions this year due to the large number of posters being presented. All posters will be exhibited during the poster sessions; however, presenters from the odd-numbered posters will be available during the Monday evening poster session and during the Tuesday evening poster session presenters from the even-numbered posters will be available.

Following the conference you are encourage 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 there will be a change from prior years for the group excursion on Wednesday afternoon/evening in that we will be exploring nearby Amsterdam. Buses will transport the conference attendees to the Museumplein in the center of Amsterdam. Free tickets will be provided which allow you to visit the Van Gogh Museum. Afterwards you are free to explore Amsterdam until 18:00, at which time we will all hop aboard a boat at the Blue Boat Company and view Amsterdam by traveling through its many canals. Following the tour the boats will drop us off at the restaurant d'Vijff Vlieghen where we will have dinner and drinks. At the conclusion of dinner, the buses will pick us up at the restaurant and transport us back to the conference center. We welcome your feedback on the new excursion as we deviate from the traditional ICCS sail boat ride on The IJsselmeer.

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.

Bob Snyder, Chair
Markus Wagener, Vice Chair

Contents

Organizing Committee and Scientific Advisory Board

Organizing Committee

Dr. Robert W. Snyder, Chair	
Dr. Markus Wagener, Vice Chair	
Dr. Guenter Grethe, Division of Chemical Information of the American Chemical Society (CINF)	
Dr. Christoph Steinbeck, Chemical Structure Association Trust (CSA Trust)	CSA TRUST
Dr. Kimito Funatsu, Division of Chemical Information and Computer Science of the Chemical Society of Japan (CSJ)	CSJ
Dr. Frank Oellien, Chemistry-Information-Computer Division of the Society of German Chemists (GDCh)	GDCh
Dr. Lutgarde Buydens, Royal Netherlands Chemical Society (KNCV)	KNCV
Dr. Don Parkin, Chemical Information Group of the Royal Society of Chemistry (RSC)	RS•C ROYAL SOCIETY OF CHEMISTRY
Dr. Peter Ertl, Swiss Chemical Society (SCS)	

Scientific Advisory Board

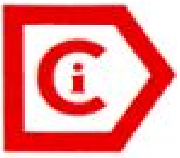
- Dr. Dimitris Agrafiotis, Johnson & Johnson Pharmaceutical Research & Development
- Dr. Kimito Funatsu, Department of Chemical System Engineering of the University of Tokyo
- Dr. Val Gillet, University of Sheffield
- Dr. Michael Lajiness, Eli Lilly and Company
- Dr. Matthias Rarey, Zentrum für Bioinformatik of the Universität Hamburg
- Dr. Robert W. Snyder, FEI Company
- Dr. Lothar Terfloth, Molecular Networks GmbH
- Dr. Markus Wagener, N.V. Organon, a part of Schering-Plough Corporation

Sponsors

List of Sponsors

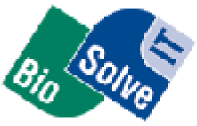
Premier Sponsors

Platinum Sponsors



Gold Sponsors

Silver Sponsors



Exhibition

List of Exhibitors

CHEMICAL COMPUTING GROUP
Scalable Software. Scalable Science.

OpenEye Scientific Software

CambridgeSoft
www.cambridgesoft.com

Cresset
BioMolecular Discovery Limited

simulations plus, inc.

COSMOlogic GmbH & Co. KG
Your Competent Partner for Computational Chemistry and Fluid Thermodynamics

www.ccdc.cam.ac.uk
40
1965 - 2005

cas
A Division of the American Chemical Society

digital chemistry

Molecular Networks
Inspiring Chemical Discovery

ChemAxon

eidrux
PHARMINFORMATICS

Bio Solve IT

keymodule

idbs

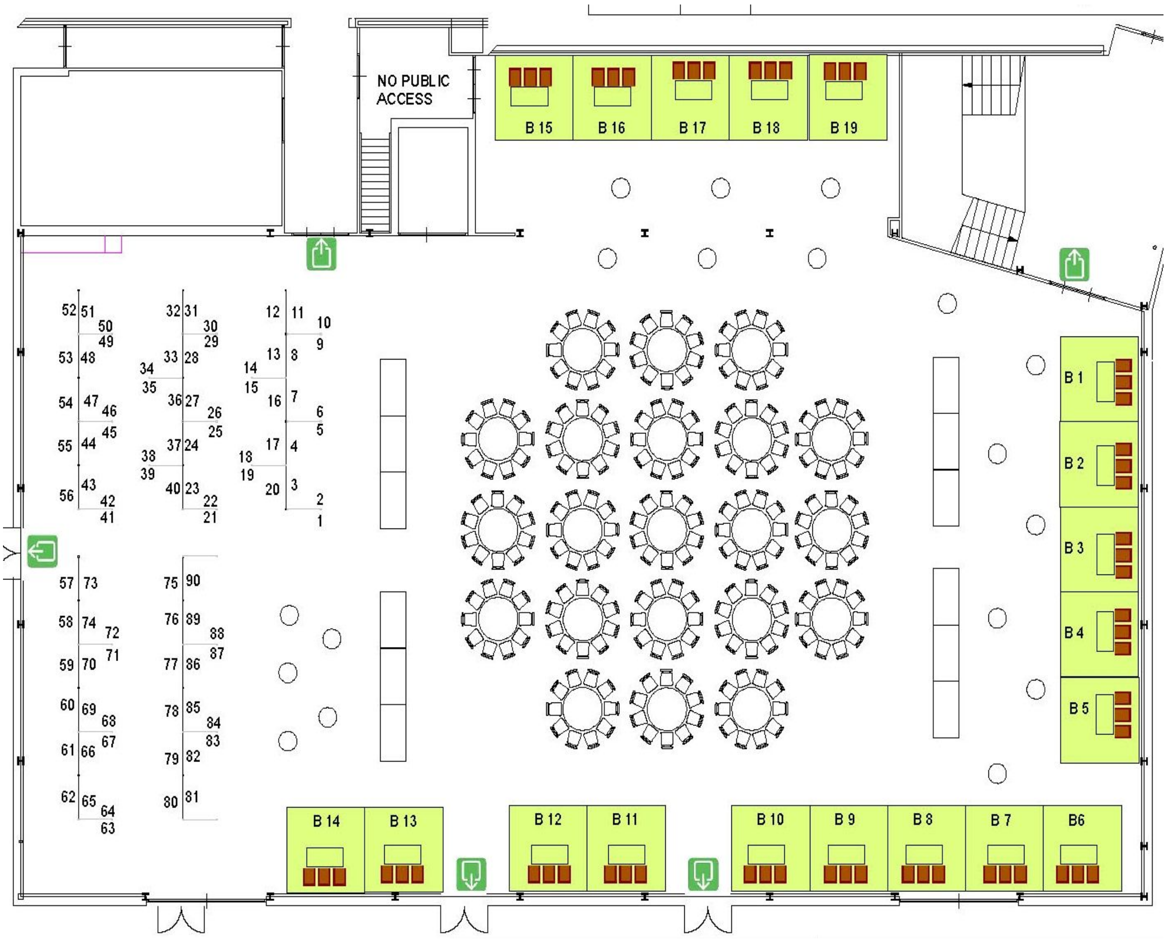
Xemistry GmbH

TS TECHNOLOGY

evolvus

The New
Tripos™

Exhibition Layout (Atrium)



B1 : Chemical Abstracts Service (CAS)
B2 : Molecular Networks GmbH
B3 : Digital Chemistry
B4 : Cidrux Pharminformatics B.V.
B5 : Transition State Technology Limited
B6 : OpenEye Scientific Software
B7 : ID Business Solutions Ltd. (IDBS)
B8 : ChemAxon
B9 : BioSolveIT
B10 : Cresset BioMolecular Discovery

B11 : CCDC
B12 : COSMOlogic
B13 : Evolvus Group
B14 : Chemical Computing Group (CCG)
B15 : Keymodule Ltd.
B16 : Xemistry GmbH
B17 : Simulations Plus
B18 : CambridgeSoft
B19 : Tripos International

Exhibition Hours

- Monday 15:00 – 19:30
- Tuesday 15:00 – 19:30

Group Excursion

Excursion to Amsterdam

Agenda:

13:00	Departure from the conference center, Noordwijkerhout
14:00	Arrival in Amsterdam ● possibility to visit the Van Gogh Museum (free ticket provided) ● explore Amsterdam
18:00	Boat cruise on the Amsterdam canals (Dutch: grachten) ● departure from the dock of Blue Boat Company
19:30	Arrival by boat at the restaurant d’Vijff Vlieghen, dinner & drinks
22:15	Departure from the restaurant d’Vijff Vlieghen by bus
23:00	Arrival at the conference center, Noordwijkerhout

Itinerary:

❶ **Van Gogh Museum**, Paulus Potterstraat 7 – the Van Gogh Museum houses the largest collection of Van Gogh's paintings and drawings in the world

❷ **Blue Boat Company**, Stadhouderskade 30

❸ **Restaurant d’Vijff Vlieghen**, Spuistraat 294-302, phone +31-20-5304060

Other Points of Interest:

❹ **Rijksmuseum** – the Rijksmuseum possesses the largest and most important collection of classical Dutch art, e.g. works from Vermeer, Hals, Rembrandt (The Nightwatch). Additionally, the collection consists of a large variety of decorative art, e.g. Delftware or 17th century dollhouses. Currently, only the most important works are on display, since the Rijksmuseum is undergoing reconstruction

❺ **Flower Market** – permanent market selling flowers, plants and tulip bulbs

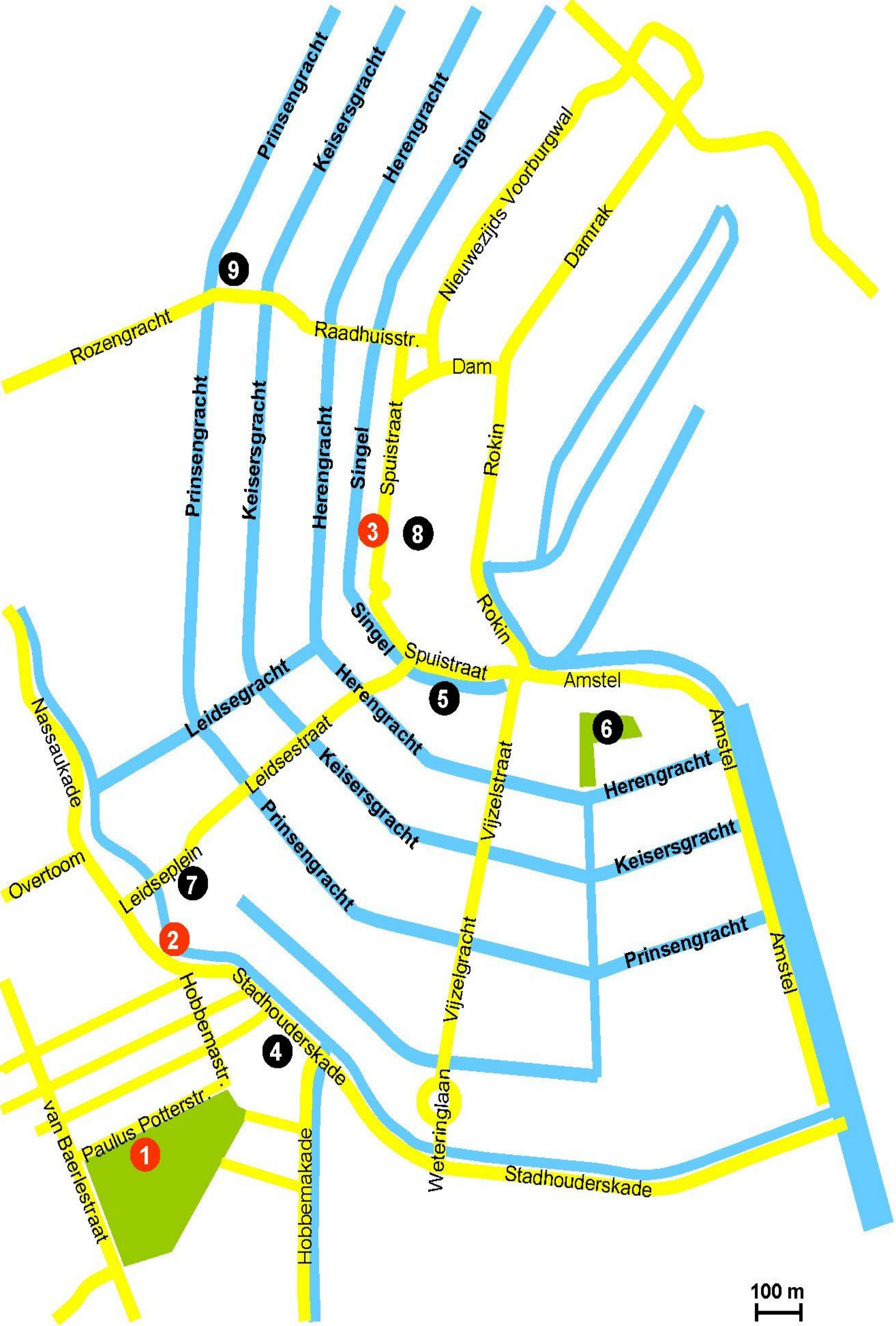
❻ **Rembrandtplein** – square with many pubs and outdoor cafes

❼ **Leidseplein** – square with many pubs and outdoor cafes

❽ **Begijnhof** – the Begijnhof is one of the oldest inner courts in the city which was used by Beguines, a lay sisterhood. Entrance from the Gedempte Begijnensloot

❾ **Anne Frank House** – The Anne Frank House is a museum dedicated to Jewish wartime diarist Anne Frank, who hid from Nazi persecution in hidden rooms at the rear of the building

Amsterdam Map



Technical Program

Technical Program

Plenary Session

Sunday, 1 June	
12:00 - 18:00	**Registration**
18:00 - 19:00	Welcome and Keynote Address
18:00 - 18:15	ORG-1 : *Welcome and Introduction* Bob Snyder, ICCS Program Chair
18:15 - 19:00	Opening Session - Keynote Address, CSA Trust Mike Lynch Award K-1 : *Challenges/opportunities for chemical structure databases in the 21st century* Alexander J. Lawson, Elsevier Information Systems, Frankfurt
19:00	**Welcoming Reception** – Atrium, courtesy of CCDC
20:00	**Rijsttafel Dinner** – Atrium, courtesy of Chemical Abstracts Service (CAS)

Monday, 2 June	
08:30 - 10:30	Informatics for Bridging Between Chemistry and Biology Kimito Funatsu, Presiding
08:30 - 09:00	A-1 : *Protein target prediction of toxic molecules identifies toxicological relationships between proteins* Florian Nigsch, University of Cambridge
09:00 - 09:30	A-2 : *Exploiting systems chemical biology to predict and understand (un)desired drug effects* Josef Scheiber, Novartis Institutes for Biomedical Research
09:30 - 10:00	A-3 : *Binding site similarity analysis for the functional classification of the protein kinase family* Richard M. Jackson, University of Leeds
10:00 - 10:30	A-4 : *Merging high-content screening and in silico approaches for compound profiling and mode-of-action analysis* Andreas Bender, Leiden / Amsterdam Center for Drug Research
10:30 - 11:00	**Break**
11:00 - 15:00	Cheminformatics Lothar Terfloth, Presiding
11:00 - 11:30	B-1 : *Performance of common similarity measures in virtual screening and lead-hopping* J Christian Baber, Wyeth Research
11:30 - 12:00	B-2 : *Maximum unbiased validation (MUV) of ligand based virtual screening* Knut Baumann, University of Technology Braunschweig
12:00 - 12:30	B-3 : *Molecular similarity by pattern recognition: Fast calculation of 3D pharmacophore resemblence* Gerhard Wolber, Inte:Ligand GmbH
12:30 - 13:30	**Lunch** – Atrium

Monday, 2 June	
13:30 - 14:00	B-4 : *Analysis of natural products: Lessons from nature inspiring the design of new drugs* Peter Ertl, Novartis Institutes for BioMedical Research
14:00 - 14:30	B-5 : *It's all in the bits: Improved database searching with better bits* Harold Helson, CambridgeSoft Inc.
14:30 - 15:00	B-6 : *Is there a general model for bioactivity?* Tudor I. Oprea, University of New Mexico
15:00 - 15:30	**Break**
15:00 - 19:30	**Exhibition & Posters** – Atrium
15:30 - 17:30	**Poster Presentations** (odd-numbered poster authors present)
18:30 - 19:30	**Reception** – Atrium
19:30 - 21:30	**Dinner** – Atrium

Tuesday, 3 June	
08:30 - 15:00	Structure-Based Drug Design and Virtual Screening Val Gillet & Bob Snyder, Presiding
08:30 - 09:00	C-1 : *SAMPL: Statistical assessment of the modeling of proteins and ligands* A. Geoffrey Skillman, OpenEye Scientific Software
09:00 - 09:30	C-2 : *HYDE: An integrated description of dehydration and H-bonding within protein ligand interfaces* Gudrun Lange, Bayer CropScience
09:30 - 10:00	C-3 : *Specificity scoring* Joannis Apostolakis, LMU Munich / MoDeST
10:00 - 10:30	C-4 : *Flexophore, a new versatile 3D pharmacophore descriptor* Modest von Korff, Actelion Ltd.
10:30 - 11:00	**Break**
11:00 - 11:30	C-5 : *A fragment-based computational protocol at PDB scale - Application to lead-optimization of DFG-out kinase inhibitors* Fabrice Moriaud, MEDIT
11:30 - 12:00	C-6 : *Can 3D ligand based virtual screening compete with docking? Application of molecular fields to virtual screening with the DUD dataset* Mark Mackey, Cresset BioMolecular Discovery Limited
12:00 - 12:30	C-7 : *Novel fragment-like PTR1 inhibitors discovered by virtual screening* Chido Mpamhanga, Dundee University (Drug Dicovery Unit)
12:30 - 12:45	**Group Photo**
12:30 - 13:30	**Lunch**
13:30 - 14:00	C-8 : *Fleksy: a flexible approach to induced fit docking* Sander B. Nabuurs, Radboud University Nijmegen
14:00 - 14:30	C-9 : *Index-driven structure-based virtual screening* Jochen Schlosser, Center for Bioinformatics Hamburg (ZBH)
14:30 - 15:00	C-10 : *Algorithmic design of ligand binding pockets on protein surfaces* Susanne Eyrisch, Center for Bioinformatics, Saarland University
15:00 - 15:30	**Break**
15:00 - 19:30	**Exhibition & Posters** – Atrium
15:30 - 17:30	**Poster Presentations** (even-numbered poster authors present)

Tuesday, 3 June	
18:30 - 19:30	**Reception** – Atrium
19:30 - 21:30	**Dinner** – Atrium

Wednesday, 4 June	
08:30 - 10:30	Virtual Chemistry Markus Wagener, Presiding
08:30 - 09:00	D-1 : *De novo drug design using multi-objective evolutionary graphs* Christos Nicolaou, University of Cyprus
09:00 - 09:30	D-2 : *Planning organic synthesis using reaction types derived from reaction databases* Christof H. Schwab, Molecular Networks GmbH
09:30 - 10:00	D-3 : *Knowledge-based de novo design using reaction vectors* Hina Patel, University of Sheffield
10:00 - 10:30	D-4 : *Recore: Instant 3D scaffold hopping using replacement fragments* Peter Richard Oledzki, BioSolveIT
10:30 - 11:00	**Break**
11:00 - 13:00	Analysis of Large Data Sets Christoph Steinbeck, Presiding
11:00 - 11:30	E-1 : *Turns revisited: Clustering turn structures using ESOMs leads to a uniform classification for open, normal and reverse turn families* Oliver Koch, The Cambridge Crystallographic Data Centre
11:30 - 12:00	E-2 : *Searching fragment spaces with feature trees* Uta Lessel, Boehringer Ingelheim Pharma GmbH & Co. KG
12:00 - 12:30	E-3 : *Three way comparison of chemical spaces avoiding structure exchange* Jens Loesel, Pfizer
12:30 - 13:00	E-4 : *Use of data mining to help identify compounds that are unstable in DMSO* Jameed Hussain, GlaxoSmithKline
13:00 - 13:00	**Box Lunch**
13:00 - 23:00	**Excursion**, dinner courtesy of Chemical Computing Group (CCG)

Thursday, 5 June	
07:30 - 08:30	**Hotel Check-out**
08:30 - 13:00	Structure-Activity and Structure-Property Prediction Matthias Rarey, Presiding
08:30 - 09:00	F-1 : *CypScore - in silico case studies on metabolic stability optimization* Andreas H. Göller, Bayer Healthcare AG
09:00 - 09:30	F-2 : *SyGMa: combining knowledge and empirical scoring in the prediction of metabolites* Lars Ridder, Organon, a part of Schering-Plough Corporation
09:30 - 10:00	F-3 : *TopoHERG – A highly selective pharmacophoric classifier for hERG-channel active compounds* Britta Nisius, Bayer Healthcare AG
10:00 - 10:30	F-4 : *Compound set optimization and sequential screening using emerging chemical patterns* Jens Auer, Bonn-Aachen International Center for Information Technology
10:30 - 11:00	**Break and Hotel Check-out**

Thursday, 5 June	
11:00 - 11:30	F-5 : *Interpretable Activity Models: exploring the limits of pharmacophore and 3D QSAR methods* David Anthony Evans, Eli Lilly
11:30 - 12:00	F-6 : *QSAR modeller seeks meaningful relationship* Craig L. Bruce, University of Nottingham
12:00 - 12:30	F-7 : *Rational design of M1-Muscarinic Antagonists using combinatorial transformation* Michael B. Bolger, Simulations Plus, Inc.
12:30 - 13:00	F-8 : *Structure-activity landscapes: a new way to study a structure-activity relationship* John H van Drie, John H Van Drie Research LLC
13:00 - 13:15	*Closing Remarks* Markus Wagener, ICCS Vice Chair
13:15 - 14:00	**Lunch or Box Lunch**
13:30 - 14:00	**Shuttle buses leave for Schiphol Airport**
14:30 - 15:00	**Shuttle buses leave for Schiphol Airport**

BioSolveIT Workshop

14:00 - 17:00	*Interactive Workshop on Virtual Screening and De Novo Design* BioSolveIT GmbH (registration with BioSolveIT required)
17:00 - 17:30	**Shuttle buses leave for Schiphol Airport**

Poster Session

P-1 : *Discovery Portal - a novel tool to increase productivity, efficiency and transparency across R&D organizations*
Jaroslaw Tomczak, Accelrys

P-2 : ~~*A simple language for conversing between diverse applications*~~
~~Omara Williams, Accelrys~~ *WITHDRAWN*

P-3 : *The use of stereo descriptors in the context of a structure validation workflow*
Pedro Gomez Fabre, Accelrys

P-4 : *OSIRIS, an entirely in-house developed drug discovery informatics system*
Thomas L Sander, Actelion Pharmaceuticals Ltd.

P-5 : *Scientific database application without borders: Empowering the scientists*
Man-Ling Lee, Aestel Scientific Information, LLC

P-6 : *Diversity oriented virtual compound selection strategy for high throughput screening of potential anticancer agents*
Gyorgy Dorman, AMRI

P-7 : *Investigating false predictions in mutagenicity QSAR models: What are we missing?*
Catrin Hasselgren, AstraZeneca

P-8 : *Selecting druglike pieces for the jigsaw puzzle: towards optimal fragment spaces*
Christof Gerlach, Bayer Schering Pharma

P-9 : *Going on SARfari in the protein kinase data jungle*
Judith Günther, Bayer Schering Pharma

P-10 : *A probabilistic approach to classifying metabolic stability*
Antonius ter Laak, Bayer Schering Pharma

P-11 : *MCS clustering - A hierarchical clustering approach for large data sets*
Alexander Böcker, Boehringer Ingelheim

P-12 : *Comparison of different approaches for cytochrome P450 modeling*
Paul Czodrowski, Boehringer Ingelheim

P-13 : *Mapping of activity class characteristic substructures extracted from random fragment populations*
Eugen Lounkine, Bonn-Aachen International Center for Information Technology

P-14 : *In silico prediction of efflux substrates classification*
Litai Zhang, Bristol Myers Squibb

P-15 : *Digging deep for GOLD - Using buriedness to improve discrimination between actives and inactives in docking*
Noel M. O'Boyle, Cambridge Crystallographic Data Centre

P-16 : *Representation, searching and enumeration of generic structures – From molecules towards patents*
Szabolcs Csepregi, ChemAxon Ltd.

P-17 : *Hierarchical clustering of chemical structures by learned scaffolds*
Miklos Vargyas, ChemAxon Ltd.

P-18 : *Molecular framework based analysis of large chemical spaces*
Anthony Joseph Trippe, Chemical Abstracts Service

P-19 : *Towards automated searching of data in Internet chemical databases*
Xiaoxia Li, Chinese Academy of Sciences

P-20 : *Chemotype bias in virtual screening: the elephant in the room*
Mark Mackey, Cresset BioMolecular Discovery Limited

P-21 : *Rapid property profiling and similarity calculations in large virtual libraries*
John Mordaunt Barnard, Digital Chemistry Ltd

P-22 : *Opportunities for integrating Markush patent searching with drug discovery*
John Mordaunt Barnard, Digital Chemistry Ltd

P-23 : *A mathematically more precise taxonomy and nomenclature for polymers*
Seymour Benjamin Elk, Elk Technical Associates

P-24 : *Indirect drug design using MD for flexible structure alignment application to HIV-1 protease inhibitors*
Alok Juneja, Freie Universität Berlin

P-25 : *Optimizing drug classification by feature selection: To bind or not to bind that is the question*
Ernst-Walter Knapp, Freie Universität Berlin

P-26 : *Understanding selective CDK4 inhibition through molecular dynamics*
Nahren Manuel Mascarenhas, Indian Institute of Chemical Biology

P-27 : *Extracting chemical CYP proteins interactions from literature using natural language processing methods*
Dazhi Jiao, Indiana University

P-28 : *An infrastructure for data mining public chemical & biological information*
David J Wild, Indiana University

P-29 : *Binding affinity prediction of distinct inhibitors of group-1 and group-2 Neuraminidases (NAs): ArgusLab4/AScore protocol*
Marija L. Mihajlovic, Institute for Multidisciplinary Research

P-30 : *Prediction of novel drug targets in the metazoan parasite schistosoma mansoni*
Frank Oellien, Intervet Innovation GmbH

P-31 : *Performance of different machine learning methods*
Uwe Koch, IRBM Merck Research Laboratories

P-32 : *Assessing and exploiting non-additivity in a structure-activity relationship*
John H van Drie, John H Van Drie Research LLC

P-33 : *CLiDE Pro: A chemical OCR tool*
Aniko Tunde Valko, Keymodule Ltd.

P-34 : *Molecular subgraph mining for analyzing ligand activity classes*
Julio E. Peironcely, LACDR

P-35 : *Frequent Substructure Mining of GPCR ligands*
Eelke van der Horst, Leiden University

P-36 : *Characterization of the inhibition of HIV-1 reverse transcriptase by non-nucleoside inhibitors and proteochemometric models which are able to predict compound activity against particular target mutants*
Gerard Jacob Pieter van Westen, Leiden University

P-37 : *Consensus modeling of chemical biodegradation pathways*
ML Patel, Lhasa Limited

P-38 : *Scaffold hunter: Exploiting holes in chemical space*
Stefan Wetzel, Max-Planck Institute for Molecular Physiology

P-39 : *Dynamic web application for drug design research*
John David MacCuish, Mesa Analytics & Computing

P-40 : *Parallel tiered clustering for large data sets using a modified Taylor's algorithm*
John David MacCuish, Mesa Analytics & Computing

P-41 : *Ligand-based models for the isoform specificity of Cytochrome P450 substrates*
Lothar Terfloth, Molecular Networks GmbH

P-42 : *Metabolomics approach for determining growth-specific metabolites based on Fourier transform ion cyclotron resonance mass spectrometry*
Hiroki Takahashi, Nara Institute of Science and Technology

P-43 : *Clustering peptidases emplying structural features of their inhibitors*
Mariusz Milik, Novartis Institutes for Biomedical Research

P-44 : *Prediction of cell permeability*
Paul Selzer, Novartis Institutes for Biomedical Research

P-45 : *Validation using the RCSB: Good idea or bad idea?*
Paul Charles Hawkins, OpenEye Scientific

P-46 : *Automated generation of fragment-based rules for mutagenicity prediction*
Olaf Günter Othersen, Radboud University Nijmegen

P-47 : *The detection of new active site conformations using molecular dynamics and rotamer assignments*
Gijs Schaftenaar, Radboud University Nijmegen

P-48 : *Automated extraction of kinase hinge-binding fragments from the protein data bank*
Dave John Wood, Radboud University Nijmegen

P-49 : *Get the best from substructure mining*
Jeroen Kazius, Research For Charity Foundation

P-50 : *The RSC's Project Prospect: identification and reuse of chemistry in publications*
Colin Batchelor, Royal Society of Chemistry

P-51 : *In silico studies on P63 as a new drug-target protein*
Anna Karawajczyk, RUMC

P-52 : *QSAR modelling of antineoplastic activities using NIH Roadmap Data*
Alexey Zakharov, Russian Academy of Medical Science

P-53 : *GUSAR: new approach for multiple QSAR*
Alexey Zakharov, Russian Academy of Medical Science

P-54 : *Fast empirical estimates of quantum mechanical descriptors for QSAR/QSPR modeling*
Robert Fraczkiewicz, Simulations Plus, Inc.

P-55 : *The representation, registration, and retrieval of substances with incompletely defined chemical structures*
Keith T Taylor, Symyx Technologies Inc

P-56 : *Exploring synthetically accessible chemical space*
Keith T Taylor, Symyx Technologies Inc

P-57 : *Development and visualization of the drug-likeness model*
Masamoto Arakawa, The University of Tokyo

P-58 : *Reverse analysis and multi-objective optimization of predictive models for polymer properties*
Shun Goto, The University of Tokyo

P-59 : *Development of a new regression analysis method using independent component analysis*
Hiromasa Kaneko, The University of Tokyo

P-60 : *Rule Induction of the site of metabolism by human Cytochromes P450 by data-mining*
Michio Koyama, The University of Tokyo

P-61 : *Dynamic interplay between chemotype, similarity and docking studies: Towards a virtual screening approach for protein kinase B inhibitors*
Jose L. Medina-Franco, Torrey Pines Institute for Molecular Studies

P-62 : *Multi-fusion similarity maps for comparing the chemical space of compound databases*
Jose L. Medina-Franco, Torrey Pines Institute for Molecular Studies

P-63 : *The effect of structural redundancy on virtual screen performance*
Robert D. Clark, Tripos International

P-64 : *Topomer CoMFA for rapid optimization*
Bernd Wendt, Tripos International

P-65 : *Development of an a priori ionic liquid design tool: Integration of a novel COSMO-RS molecular descriptor on neural networks*
Jose Palomar, Universidad Autónoma de Madrid

P-66 : *Radial scan of the electrostatic potential of RNA double helices: An application on tRNA acceptor stems*
Ray Marcel Marín, Universidad Nacional de Colombia

P-67 : *A graph theoretical approach to compare molecular electrostatic potentials*
Ray Marcel Marín, Universidad Nacional de Colombia

P-68 : *Engineering polymer informatics*
Nico Adams, University of Cambridge

P-69 : *Information extraction from the polymer literature*
Lezan Hawizy, University of Cambridge

P-70 : *MeFc and large chemical data sets*
Hamse Y. Mussa, University of Cambridge

P-71 : *Kernel based least squares and large data sets*
Hamse Y. Mussa, University of Cambridge

P-72 : *Molecular spam: Use of a modified spam filter for classification of bioactive molecules and drug target prediction*
Florian Nigsch, University of Cambridge

P-73 : *SPECTRa-T: Machine-based data extraction and semantic searching of chemistry e-theses*
Joseph Andrew Townsend, University of Cambridge

P-74 : *Creating chemo- & bioinformatics workflows: Further developments within the CDK-Taverna project*
Thomas Kuhn, University of Cologne

P-75 : *Protein-Protein interactions as targets for drugs: A view from the binding site*
Richard M. Jackson, University of Leeds

P-76 : *Determinants for selectivity in map kinase inhibitors by computational simulations*
Nikita Basant, University of Modena and Reggio Emilia

P-77 : *Fragment weighting schemes for similarity-based virtual screening: Use of occurrence weighting*
Shereen Arif, University of Sheffield

P-78 : *Effect of data standardization on the clustering of chemical structures*
Chia-Wei Chu, University of Sheffield

P-79 : *Multiobjective optimisation of pharmacophore hypotheses: Bias towards low-energy conformations*
Val Gillet, University of Sheffield

P-80 : *Weighted data fusion with turbo similarity searching to improve chemical similarity searching*
John Holliday, University of Sheffield

P-81 : *Using wavelets to represent GRID fields in virtual screening*
Richard Martin, University of Sheffield

P-82 : *A multiobjective approach to scoring functions for docking*
Iain Peter Mott, University of Sheffield

P-83 : *Neighbourhood behaviour studies for lead optimisation*
Georgios Papadatos, University of Sheffield

P-84 : *Maximum unbiased validation (MUV) datasets for virtual screening by PubChem based chemogenomics data mining*
Sebastian Georgios Rohrer, University of Technology Braunschweig

P-85 : *3D-Visualization of molecular conformations in the MOGADOC database*
Jürgen Vogt, University of Ulm

P-86 : *Similarity based correction for the predictions of compounds physicochemical properties*
Andrius Sazonovas, Vilnius University

P-87 : *Prediction of ionization constants for complex multicenter electrolytes utilizing proprietary ‘in house’ data*
Andrius Sazonovas, Vilnius University

P-88 : *A novel chemical database for sustainable development of synthesis routes - An instance of developing synthesis routes of quinolone derivatives*
Kenzi Hori, Yamaguchi University

P-89 : *Combinatrial chemistry using theoretical calculations: An application to boric acid catalyzed esterification of phenol*
Maki Shimeno, Yamaguchi University

P-90 : *Calculation of difference of free energy of solvations using the QM/MC/FEP method in chemical reactions*
Keita Uezu, Yamaguchi University

P-91 : *Toward in silico screening using transition state data base for developing new synthesis routes*
Toru Yamaguchi, Yamaguchi University

P-92 : *Tautomer generation. pKa based dominance conditions for generating the dominant tautomers*
Ferenc Csizmadia, ChemAxon Ltd.

P-93 : *Chemical terms – A language for cheminformatics*
Akos Papp, ChemAxon

Plenary Session Abstracts

Plenary Session Abstracts

K-1 : Challenges/Opportunities for chemical structure databases in the 21st century

Alexander J. Lawson, Elsevier Information Systems, Frankfurt, Germany

The 19th century witnessed the birth of large-scale structure-based data collections in chemistry as aids to harnessing the knowledge deposited in the primary literature. The developments of the 20th century enabled these collections (and the literature itself) to make the transition from print to electronic media. In the course of the 21st century further development will inevitably take place, shaping this synergy into new forms.

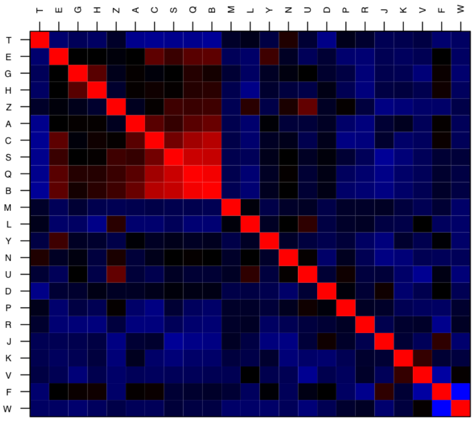
In this talk the challenges facing several possible scenarios will be explored.

A-1: Protein target prediction of toxic molecules identifies toxicological relationships between proteins

F. Nigsch, J.B.O. Mitchell, Unilever Centre for Molecular Science Informatics, Department of Chemistry, Cambridge, UK

Computational methods for protein target prediction can be used for the in silico identification of potential off-target activities. These off-target activities may often be the origin of clinically observed toxic effects or adverse drug reactions.

Based on a method that we recently applied successfully to a similar but easier classification problem, we built a model encompassing a larger dataset for protein target prediction of toxic molecules. [1] Our model uses the Winnow algorithm as underlying classification framework and circular fingerprints as molecular descriptors. A protein target dataset with 90,000 molecules spanning 233 activity classes was obtained by selecting all relevant classes from the MDL Drug Data Report (MDDR). Prior to the application of the protein target prediction model, we validated it using a 15-fold Monte Carlo cross-validation, each of which using a 50:50 split. We retained the 3 top-ranking predictions and found that in 82 percent of all cases the correct target was predicted within these three predictions. The first prediction was the correct one in almost 70 percent of cases.



This model was then applied to predict the protein targets of 150,000 molecules with experimentally determined toxicities contained in the MDL Toxicity Database. The resulting associations allowed us to determine proteins that are related with respect to their toxicities, as well as to cluster toxicities which are related with respect to the proteins likely to cause these toxicities. For both highly correlated protein clusters and also the top-ranking proteins for each toxicity class, we were able to confirm the significance of our results by independent evidence from published literature.

1. Nigsch, F.; Mitchell, J.B.O. How to Winnow Actives from Inactives: Introducing Molecular Orthogonal Sparse Bigrams (MOSBs) and Multiclass Winnow. *J. Chem. Inf. Model.* **2008**, ASAP article

A-2 : Exploiting systems chemical biology to predict and understand (un)desired drug effects

Josef Scheiber, Jeremy L. Jenkins, Dmitri Mikhailov, Meir Glick, John W. Davies, Novartis Institutes for Biomedical Research, Cambridge, MA, USA

Several drug withdrawals associated with adverse side effects –Vioxx® and Lipobay® being the most prominent ones – gained broad attention in recent years. To avoid such cases and thereby improve the life of patients it is without question highly desirable to identify and eliminate such problems in early research.

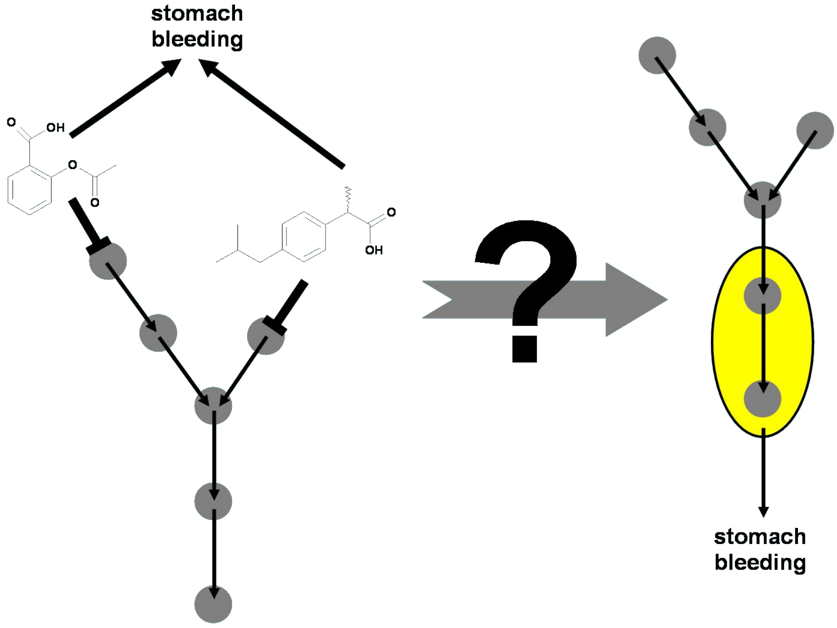


Figure 1: *This figure illustrates that the interaction of a compound with a single target may not always be the isolated cause for an undesired effect. Hitting two different targets can have the same outcome downstream in a pathway. This contribution will deal with approaches to address such cases through the combination of cheminformatics approaches with systems biology data.*

In some cases the reason for an undesired effect can be found in the interaction of the compound with a certain target, e.g. the prolonged QT-syndrome with the hERG-channel.[1] These cases can then be identified with well-established *in vitro*-methods,[2] as well as newly developed *in silico*-methods. [3-5] However, often chemically diverse compounds cause similar problems. This is the case when two different targets are hit in the same biological pathway, which is illustrated for stomach bleeding in Figure 1. In this case models can be established that predict certain adverse effects irrespective of target considerations, where the models are based on compound-adverse event pairings. After computing these models a link through chemical space can be made to compute correlations with different target prediction models.[3] Thereby it becomes possible to link certain phenotypic effects to the interaction between a molecule and a target.

This contribution will introduce an extension of these methods. On the one hand the predictive models for both adverse side effects and targets have been optimized, re-calculated and heavily validated using the MedDRA terminology for side effects[6] and sophisticated validation methods.[7] Further, we linked the predictions with biological network information to establish firm links between side effects and the interference of a compound with a certain pathway. The predictions can then be validated by analyzing the data from well-known pathway tools and databases like GeneGo's MetaCore and Ingenuitys' IPA. Also, new links between pathways and side effects can be established.

To summarize: The presentation will link Systems Chemical Biology[8] approaches to the field of adverse side effects of drugs to better understand the latter ones.

1. Curran, M. E.; Splawski, I.; Timothy, K. W.; Vincent, G. M.; Green, E. D.; Keating, M. T., A molecular basis for cardiac arrhythmia: HERG mutations cause long QT syndrome. *Cell* **1995,** 80, (5), 795-803.
2. Whitebread, S.; Hamon, J.; Bojanic, D.; Urban, L., In vitro safety pharmacology profiling: an essential tool for successful drug development. *Drug Discovery Today* **2005,** 10, (21-24),

1421-1433.

3. Bender, A.; Scheiber, J.; Glick, M.; Davies , J.; Azzaoui, K.; Hamon, J.; Urban, L.; Whitebread, S.; Jenkins, J., Analysis of Pharmacology Data and the Prediction of Adverse Drug Reactions and Off-Target Effects from Chemical Structure. *ChemMedChem* **2007,** 2, (6), 861-873.
4. 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.
5. Azzaoui, K.; Hamon, J.; Faller, B.; Whitebread, S.; Jacoby, E.; Bender, A.; Jenkins, J.; Urban, L., Modeling Promiscuity Based on in vitro Safety Pharmacology Profiling Data. *ChemMedChem* **2007,** 2, (6), 874-880.
6. ICH-MSSO Medical Dictionary for Regulatory Activities (MedDRA).
7. Breiman, L., Bagging Predictors. *Machine Learning* **1996,** 24, (2), 123-140.
8. Oprea, T. I.; Tropsha, A.; Faulon, J.-L.; Rintoul, M. D., Systems chemical biology. *Nat Chem Biol* **2007,** 3, (8), 447-450.

A-3 : Binding site similarity analysis for the functional classification of the protein kinase family

Richard Jackson, Sarah Kinnings, Institute of Molecular and Cellular Biology, Faculty of Biological Sciences, University of Leeds, Leeds, UK

Methods for analysing complete gene families in the drug discovery process are becoming of increasing importance, because similarities and differences within a family are often the key to understanding functional differences that can be exploited in drug design. Constituting around 1.7% of the human genome, the protein kinase family is one of the largest enzyme families and plays key roles in almost all signalling pathways. Since the deregulation of these signalling pathways often leads to disease, the control of protein kinase activity is a principle focus of pharmaceutical research. The vast majority of kinase inhibitors target the ATP-binding site. However, the high degree of sequence and structural conservation amongst the protein kinases means that the design of potent, selective kinase inhibitors is a significant challenge.

We have developed a large online database for the retrieval of ligand binding site similarities [1]. These are extracted automatically from the Macromolecular Structure Database using a geometric hashing algorithm [2]. We have undertaken a large-scale comparison of protein kinase ATP-binding sites. This has allowed us to discover binding site similarity in different sub-families of protein kinase that are not evident from sequence similarity alone. It has also enabled us to quantify the effect of how different drug molecules bind to the same binding site and influence the local binding site conformation. We propose a relevant classification of the protein kinase family based on the similarity of their binding sites. Not only does this classification highlight features that are important for the potency and selectivity of kinase inhibitors, but it also allows us to predict possible cross-reactivity among the protein kinases.

1. Gold, N.D., Jackson, R.M. SitesBase: A database for structure-based protein-ligand binding site comparisons. *Nucleic Acids Res.* **2006**, 34, D231-234.
2. Gold N.D., Deville K., Jackson, R.M. New opportunities for protease ligand-binding site comparisons using SitesBase. *Biochem Soc Trans.* **2007**, 35: 561-565.
3. Brakoulias A., Jackson RM. Towards a Structural Classification of Phosphate binding sites in protein-nucleotide complexes: an automated all-against-all structural comparison using geometric matching. *Proteins; structure function and bioinformatics*, **2004**, 56, 250-260.

A-4 : Merging high-content screening and in silico approaches for compound profiling and mode-of-action analysis

A. Bender [1,2,] *D. W. Young* [2], *J. L. Jenkins* [2], *Y. Feng* [2]

[1] *Leiden / Amsterdam Center for Drug Research, Division for Medicinal Chemistry, Leiden University, Leiden, The Netherlands*

[2] *Novartis Institutes for Biomedical Research, Inc., Cambridge, MA, USA*

High-content screening observes the reaction of a cell to an administered compound by multidimensional microscopy and it provides a potentially more information-rich complement to single-readout conventional assays. On the other hand, microscopy-based screening can also be more 'opaque' in the way that no mechanistic explanation for the observed effect is provided per se.

The work we performed to improve our ability to handle and understand high-content screening data consists of two parts[1]. Firstly, in order to reduce the vast amount of information obtained from microscopy based data, we performed factor analysis to reduce the amount of data to analyze, while at the same time retaining most of the information. We were able to define a six-dimensional factor space that defines cell state variables such as nuclear size and DNA replication, as depicted in Figure 1.

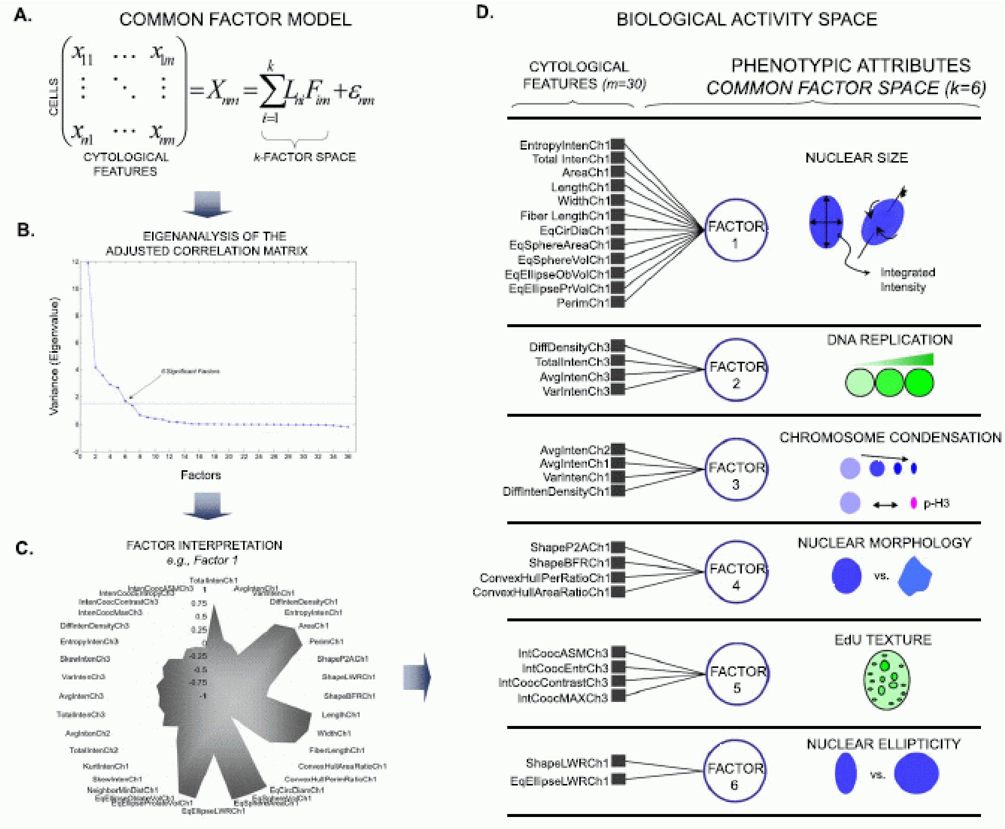


Figure 1. Factor analysis employed to project high-dimensional HCS readout space into low-dimensional space, using 6 variables to describe a cell state.

Next, by merging high-content screening with in silico target prediction, we merge both phenotypic and mechanistic approaches: by high-content screening we are able to observe the systems response, while at the same time providing hypotheses for the observed effects via the predicted targets of compounds.

We screened more than 6k compounds in high-content screenings and discuss cases where the phenotypic response and the predicted targets agree with each other, but also the even more interesting 'atypical' cases where similar phenotypes are observed by very different predicted targets (which might for example be located in the same pathway; Figure 2).

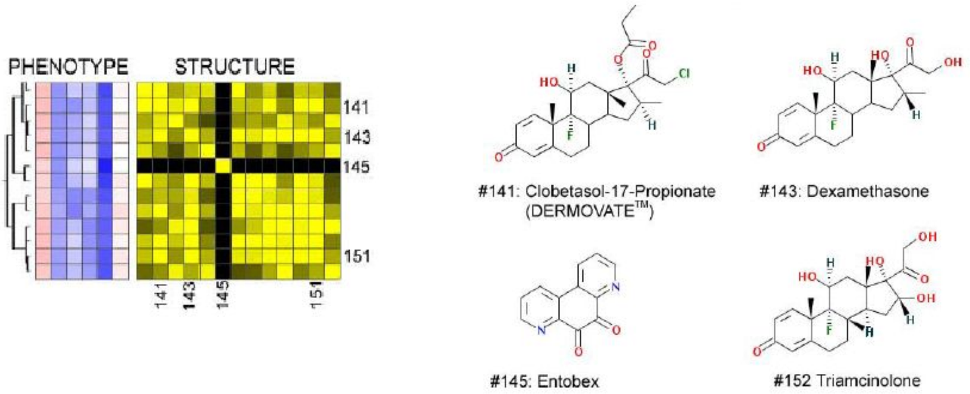


Figure 2. Four compounds, three of which steroids, which generate very similar phenotypes despite two distinct scaffolds present. Shown to the left are phenotypic readouts in six dimensions as well as a ligand structure similarity matrix.

Looking at the full compound set screened, one now has the opportunity to compare phenotypic similarity (‘systems response’) to structural similarity on a larger scale. This is shown in the similarity matrices in Figure 3 – while overall similar structures give similar readouts, also clear deviations from the rule are present. This analysis gives us the opportunity to compare compounds by not only using single or a defined set of targets, but the complete systems response instead.

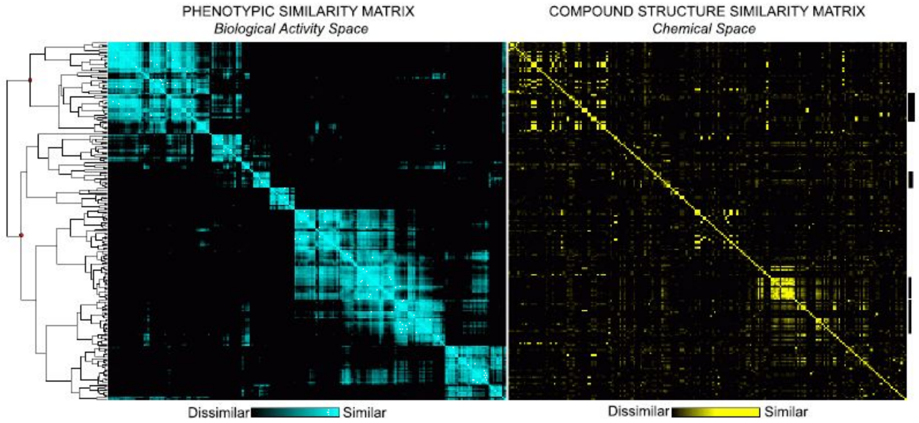


Figure 3. Comparison of phenotypic and ligand structure similarities of all ligand pairs. While both matrices show similarities, also clear differences are present, giving the possibility for example to define ‘phenotypic’ compound similarities.

1. Young, D.W.; Bender, A.; Hoyt, J. McWhinnie, E.; Chirn, G.W.; Tao, C.Y.; Tallarico, J.A.; Labow, M.; Jenkins, J.L.; Mitchison, T.J.; Feng, Y. Integrating high-content screening and ligand-target prediction to identify mechanism of action. *Nature Chem. Biol.* **2007**, 4, 59 – 68.

B-1 : Performance of common similarity measures in virtual screening and lead-hopping

J.C. Baber[1,] *G.. Tawa*[2], *R. Nilakantan*[3], *D. Mobilio*[3], *L. Greenblatt*[2], *K. Fan*[2], *C. Humblet*[2]

[1] *Chemical & Screening Sciences, Wyeth Research, Cambridge, USA*

[2] *Chemical & Screening Sciences, Wyeth Research, Princeton, USA*

[3] *Chemical & Screening Sciences, Wyeth Research, Pearl River, USA*

A large number of molecular similarity measures are now available to computational chemists and are routinely used in virtual screening exercises – particularly when no structural information is available. Virtual screening may be divided into two general types of problem: the follow up existing hits and lead-hopping to obtain new, structurally distinct, series. In the former compounds that are structurally similar to the query compound are generally sought whereas for lead-hopping compounds that are structurally dissimilar but have similar activity are required.

We will present the result of a recent examination of the virtual screening performance of over 60 different similarity measures in terms of both enrichment and overall performance as defined by the area under a ROC curve. Each method has been tested against the 40 different targets in the Directory of Useful Decoys (DUD) set[1] which provides a diverse range of drug-like classes of compounds. Tests were carried out both exhaustively using each known active in turn as an exemplar and by repeatedly choosing a random training set consisting of approximately 10% of the known actives. In order to test lead-hopping ability the known actives for each of the targets in the DUD set were classified into chemical series by experienced medicinal chemists. The compounds in each series were then used as known exemplars when screening the remainder of the set and the enrichment, overall performance and number of other series identified calculated.

Analysis of these results shows large differences in performance between methods and across targets. As would be expected, increasing the diversity of the test set generally results in a reduction in performance although some similarity methods appear to be more affected by structural diversity than others. This analysis provides a useful benchmark to assess the performance of new similarity methods and may also assist in selecting the most appropriate method, or methods, to use in order to achieve a given set of virtual screening goals.

1. Nuang, N.; Shoichet, B. K.; Irwin, J. J.; Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, 49, 6789-6801

B-2 : Maximum unbiased validation (MUV) of ligand based virtual screening

K. Baumann, S.G. Rohrer, Institute of Pharmaceutical Chemistry, Braunschweig University of Technology, Braunschweig, Germany

A common finding of many reports evaluating ligand-based virtual screening methods is that validation results vary considerably with changing benchmark datasets. Such effects are caused by the redundancy and self-similarity inherent to those datasets. These phenomena manifest themselves in the datasets' representation in descriptor space, which is termed the dataset *topology.* Three key findings that allow the design of MUV datasets are presented.

1. A methodology for the characterization of dataset topology based on spatial statistics is introduced. The method is non-parametric and can deal with arbitrary distributions of descriptor values. It utilizes two cumulative distribution functions of distances in chemical space, called the "*nearest-neighbor function*" *G(t)* and the "*empty space function*" *F(t),* which reflect the distributions of active-active and decoy-active distances, respectively. With this methodology it is possible to associate differences in virtual screening performance on different datasets with differences in dataset topology (correlation coefficient: 0.92, n = 234). Moreover, the better virtual screening performance of certain descriptors can be explained by their ability of representing the benchmark datasets by a more favorable topology (correlation coefficient: 0.91, n = 234).
2. It is shown, that the topologies of certain benchmark datasets cause over-optimistic validation results. Spatial statistics analysis as proposed here allows the detection of such biased datasets.
3. *G(t)* and *F(t)* can effectively be used as objective functions in the design of unbiased benchmark datasets. In order to design benchmark datasets with minimum bias, datasets should exhibit the lowest possible level of self-similarity, which can be monitored by *G(t).* Conversely, the set of decoys should be selected as similar to the benchmark set as possible. This process can efficiently be guided by *F(t).* Here, we apply this design strategy to a collection of datasets carefully selected from the bio-activity data available in PubChem.

The resulting maximum unbiased benchmark datasets are validated by retrospective virtual screening simulations and spatial statistics analysis.

The presented benchmark datasets will be available for download on our web-page. (http://www.pharmchem.tu-bs.de/forschung/baumann/)

B-3 : Molecular similarity by pattern recognition: Fast calculation of 3D pharmacophore resemblence

Gerhard Wolber [1,2,] Johannes Kirchmair [2], Thomas Seidel [2] and Fabian Bendix [2]

[1] Inte:Ligand Softwareentwicklungs- und Consulting GmbH, Vienna, Austria

[2] University of Innsbruck, Institute of Pharmacy, Innsbruck, Austria

Chemical-feature based pharmacophore models have been established as state-of-the-art technique describing interactions of small molecules with macromolecules and virtual screening [1, 2]. While there are already many approaches for molecular similarity in general, available similarity measures are commonly based on topological resemblance, atom pair coding or on n-point pharmacophores derived from subsets of chemical feature distances. If used for pharmacophoric similarity, these methods suffer from a topological bias to specified structure classes or by the combinatorial explosion occurring in the comparison algorithms if distance multiplets are involved.

We present a novel approach describing molecular similarity by multi-conformationally overlaying all their possible pharmacophoric features using pattern recognition and parts of our 3D alignment algorithm presented earlier [3]. The new similarity calculation method is based on a rotationally and translationally independent, but conformation-dependent representation of a molecule consisting of all chemical feature locations. Distance shells with feature proximity counts are derived from pharmacophore point locations, which are then paired and subtracted using a bipartite matching algorithm. The matching approach using distance shells, a method that has its original application in pattern recognition, bears the advantage of polynomial computational complexity and therefore allows for fast similarity measure calculation for conformational ensembles.

The use of pharmacophore points allows for a broader application scope such as measuring similarities between pharmacophore models. Applications for clustering and ligand-based pharmacophore creation are discussed.

References:

1. H. Kubinyi. In Search for New Leads, EFMC - Yearbook 2003, 14-28.
2. T. Langer, R. Hofmann. Pharmacophores and Pharmacophore Searches, R. Mannhold, H. Kubinyi, G. Folkers, series editors, Methods and principles in medicinal chemistry, pp. 131-148, Wiley-VCH, Weinheim, Germany, 2006.
3. G. Wolber, A. Dornhofer, T. Langer. Efficient overlay of small organic molecules using 3D pharmacophores. *J. Comput. Aided Mol. Des.;* **2007**; 20(12); 773-788.

B-4 : Analysis of natural products lessons from nature inspiring the design of new drugs

Peter Ertl, Novartis Institutes for BioMedical Research, Basel, Switzerland

Natural products (NPs) have evolved over a very long natural selection process to form optimal interactions with biological macromolecules. NPs are therefore a valuable source of inspiration for the design of new drugs. As illustrated in this study, application of cheminformatics techniques can provide useful help in this endeavor. First the physicochemical properties of NPs and their typical structural features are compared to those of bioactive molecules and average organic molecules. Then the substructure analysis of NPs is performed, with particular focus on comparing NP scaffolds with those of common synthetic molecules. The relationship between the structure of NPs scaffolds and the type of organism from which they have come is also analyzed.

To provide a guide for the design of NP-like bioactive structures a novel method to calculate natural product-likeness score is described. This score, which allows to determine how molecules are similar to the structural space covered by natural products, is shown to efficiently separate NPs from synthetic molecules in a crossvalidation experiment. Possible applications of the NP-likeness score are discussed and illustrated on several examples including prioritization of compound libraries towards NP-likeness and design of building blocks for the synthesis of NP-like libraries.

Hopefully the results of this analysis help to eliminate the old myth about NPs as being “too complex” or having “bad properties”, as well as help us to focus on these areas of NP structural space which are essential for biological activity, taking advantage of the long process of natural selection to guide us to new and as yet unexplored areas of the chemical structure universe.

1. Ertl, P.; Schuffenhauer, A. Cheminformatics Analysis of Natural Products: Lessons from Nature Inspiring the Design of New Drugs. In *Natural Compounds as Drugs*, Vol. II, Petersen, F.; Amstutz, R., Eds.; Birkhäuser Verlag: Basel, 2008.
2. Ertl, P.; Roggo, S.; Schuffenhauer, A. Natural Product-likeness Score and Its Application for Prioritization of Compound Libraries. *J. Chem. Inf. Model.* **2008**, 48, 68-74.
3. Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M.; Waldmann, H. The Scaffold Tree - Visualization of the Scaffold Universe by Hierarchical Scaffold Classification. *J. Chem. Inf. Model*. **2007**, 47, 47-58.
4. Ertl, P.; Jelfs, S.; Muehlbacher, J.; Schuffenhauer, A.; Selzer, P. Quest for the Rings. In Silico Exploration of Ring Universe to Identify Novel Bioactive Scaffolds. *J. Med. Chem.* **2006**, 49, 4568-4573.

Reprint download: http://peter-ertl.com/publications.html

B-5 : It's all In the bits: Improved database searching with better bits

Harold E. Helson, Andrew Smellie, Cambridgesoft Inc., Cambridge, USA

Traditionally, substructure searches in databases have been performed by first reducing the topological representation of the molecule into an encoded representation in a bit string, where each bit in the string codes for the presence of one or more substructures. In its simplest form, a molecule is decomposed into fragments which are hashed into an enormous bitstring. Various techniques are used to reduce the length of the bitstring so that it is tractable to store it in a computer. In this paper, we describe a technique of generating a reduced length bitstring whilst attempting to preserve the maximal amount of information.

Additionally, we introduce a novel data structure that takes particular advantage of the way distances are computed in a bitstring space (i.e. the tanimoto coefficient) to greatly accelerate nearest neighbor searching and similarity calculations in those spaces.

Examples will be shown that demonstrate the screening effectiveness with the modified bitstring by comparison with traditional methods. Using the improved bitstring, it will be shown that search speeds are greatly enhanced.

B-6 : Is there a general model for bioactivity?

T.I. Oprea, O. Ursu, C.G. Bologa, and L.A. Sklar, New Mexico Molecular Libraries Screening Center, University of New Mexico, Albuquerque, NM, USA

The Molecular Libraries Screening Centers Network (MLSCN) uploads bioactivity screening data in PubChem (http://pubchem.ncbi.nlm.nih.gov/) based on the Molecular Libraries Small Molecules Repository (MLSMR). On 01-30-2008, we found 222,878 unique MLSMR compounds tested on 478 MLSCN bioassays. Not all compounds were tested on all assays.

The majority of MLSMR compounds are “inactive” in the above numbers of assays. A smaller subset of MLSMR were active (25.03%, active in >1 assay), or among inconclusives (30.56%, in > 1 assay). Overall, a large percentage of compounds are inactive in all assays to date. This dataset allows us to address the following question: Is there a general model for bioactivity?

We used chemical fingerprints (560 predefined keys) and a machine learning technique (support vector machines, SVM), to discriminate actives from inactives. For “inactives” we used compounds that were found inactive in at least 60 assays; all actives or inconclusives were removed from this dataset, to yield over 22,271 compounds. For “actives” we used compounds that were found active in at least one assay; all “inactives” were removed from this data set to yield over 55,782 compounds. To the “actives” set of compounds were added the compounds from WOMBAT 2007.1 database ((http://www.sunsetmolecular.com)) to yield over 209,451 unique compounds.

From a pool of 209,451 actives and 22,271 inactives, two sets of 6000 randomly selected compounds were used to build/validate 100 different active/inactive models using a Radial Basis Function SVM kernel. Each SVM model was build using 300 random “actives”/”inactives” compounds from 6000 subset and was validated using the rest of the subset.

External prediction yields ~67% accuracy in the active class (135,467 out of 203,307 actives), and ~83% for the inactive class (13,350 out of 16,127 compounds).

An analysis based on molecular weight shows a small shift towards higher molecular weight for the "actives" set compared to the "inactives" set. Analyses based on estimated aqueous solubility and octanol/water partioning did not indicate significant differences between "actives" and "inactives". Furthermore, a large number of chemical scaffolds are present in both the active and inactive class. Artifactual results, such as florescent compounds, and aggregators, were not individually examined. However, in one MLSMR/MLSCN assay, only ~1100 compounds out of ~70,000 were considered potential aggregators.

Taken together, these results appear to indicate that bioactivity, as captured in the MLSMR and WOMBAT "actives", can be discriminated from the MLSMR "inactives". If validated by additional data, such models could be used to enrich screening libraries with compounds that are more likely to belong to the "active" class.

C-1 : SAMPL: Statistical assessment of the modeling of proteins and ligands

A. Skillman, G. Warren, P. Hawkins, A. Nicholls, OpenEye Scientific Software, Inc, Santa Fe, USA

Opportunities for prospective or blinded validation of computational models in drug discovery are rare yet valuable. SAMPL provides the computational chemistry community an opportunity to evaluate a variety of methods on previously unpublished or difficult to discover data.

We will report on the second annual SAMPL evaluation, a blinded trial of a variety of computational tasks. We received three unpublished data collections for use in the study. Abbott Laboratories provided twenty-seven Urokinase inhibitors with measured affinities and co-crystal structures. Vertex Pharmaceuticals provided fifty-two JNK-3 Kinase inhibitors with measured affinities and co-crystal structures. Finally, Peter Guthrie provided sixty-three water-vacuum transfer energies, calculated from data collected from obscure sources.

We used these data to generate eleven different blinded experiments in which participants could make predictions. All experiments were open to the public for approximately four months. Four experiments examined virtual screening, four experiments examined pose-prediction, two experiments examined affinity prediction, and one experiment examined vacuum-water transfer energies. We encouraged professional modelers to participate using third-party software in addition to both academic and industrial software developers. SAMPL attracted approximately fifty participants from North America and Europe and we accepted over one hundred public and anonymous predicted data sets.

The virtual screening experiments were designed to address important questions in evaluation design as well as generate feedback for individual algorithms. These four experiments allowed evaluation of both ligand-based and structure-based design programs. Pairs of experiments with the same ligands were used to measure the change in performance with improved knowledge of the protein structure. They also included simultaneous comparison of six different decoy sets including DUD-like decoys (1), Drug-like decoys (2) and the Rognan decoys (3) allowing evaluation of each method's performance on these well-known decoy collections in the same system. In addition, the predictions were evaluated on all the ligands provided as well as on a subset of low-potency chemically independent ligands, to simulate a prospective approach to a new target.

Pose prediction experiments included an initial cross-docking phase followed by a self-docking phase. Each method is evaluated based how well they perform in each experiment and how much they improve going from cross-docking to self-docking experiment. We will discuss important issues regarding crystal structure quality and its impact on structure reproduction evaluation.

At the last stage, the co-crystal models were provided to participants to predict binding affinities. This portion of SAMPL saw the most varied methods, including QSAR models, Docking algorithms, implicit solvent molecular mechanics calculations, explicit solvent molecular mechanics as well as Monte Carlo perturbation approaches. We will present an analysis of each model's ability to predict the absolute affinity, the relative affinity, and the rank of the inhibitors, including an assessment of the experimental error.

The vacuum-water transfer energy prediction data set included highly flexible multifunctional molecules that are much more challenging than typical transfer energy data sets. This data set also included some large drug-like molecules. This valuable data set allowed new insights into the effects of partial charge models upon salvation energy.

We will present a summary of all the results. It will include an assessment of the field in general, the effects of increasing information content on our ability to make predictions, trends in virtual screening predictions with different decoys and a discussion of affinity prediction and transfer energy. Finally, we will briefly discuss plans for the next SAMPL challenge in 2009.

1. Haung, N; Shoichet, B.K.; Irwin, J.J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, 49, 6789-6801.
2. Ajay; Walters, W.P.; Murcko, M.A. Can We Learn To Distinguish between “Drug-like” and “Nondrug-like” Molecules? *J. Med. Chem.* **1998**, 41, 3314-3324.
3. Bissantz, C.; Folkers, G.; Rognan, D. Protein-Based Virtual Screening of Chemical Databases. 1. Evaluation of Different Docking/Scoring Combinations. *J. Med. Chem.* **2000**, 43, 4759-4767.

C-2 : HYDE: An integrated description of dehydration and H-bonding within protein ligand interfaces

Gudrun Lange[1], Ingo Reulecke[2], Matthias Rarey[2], Robert Klein [1]

[1] Bayer CropScience, Frankfurt, Germany

[2] University of Hamburg, Hamburg, Germany

Scoring functions describe the interaction between molecules such as the binding of ligands to their target protein. They are used to identify the correct pose of a ligand with known inhibition in structure-based drug design. More importantly, they are used in a more automatic approach to score poses of thousands of putative ligands positioned into a target by docking programs and subsequently select those ligands which bind to the target (Virtual Screening). However, this goal was not always achieved and alternative scoring functions are needed (1). A comparison with experimental observations suggests that, for instance, the calculated contributions of interfacial H-bonds seem to be often overestimated while the hydrophobic effect is underestimated in many cases (2). In addition, comparing the size of the experimental ΔG with the size of the contribution attributed to the formation of an interfacial H-bond or the burial of an apolar surface, it becomes obvious that in addition to stabilizing contributions, there must exist a fair amount of counterbalancing destabilizing contributions to ΔG which are in most scoring functions not taken into account.

We believe that the underlying reason for the insufficient understanding of the interaction between molecules in aqueous solution lies in the imperfect description of water and its interaction with functional groups. Thus, we derived new dehydration terms for polar and apolar functions solely based on structural features of the water network and experimental logP values. These dehydration terms contribute stabilizing for apolar atoms (hydrophobic effect) or destabilizing in case of polar atoms and compare very well with experimental values. Our scoring function HYDE combines these dehydration terms with a term for H-bond energies and thus represents a very simple empirical approach describing the physics of protein ligand interactions (3). The balance between the hydrophobic effect and the contribution of H-bonds agrees well with experimental observations. In addition, significant destabilizing contribution to ΔG of individual atoms become apparent which lead to $\Delta G > 0$ if either the pose of a binder is incorrect or the ligand does not bind. The size of these destabilizing contribution explains why a single atom exchange within the binding site can lead to a significant altered affinity. This will be illustrated based on examples taken from the DUD data set (4). The examples show that HYDE is able to distinguish (a) between correct and wrong poses of known binders, (b) between protomers and tautomers of a binder and (c) between binders (Figure 1) and non-binders (Figure 2). This gives rise to drastically improved enrichments (Figure 3). In addition, the target-independ cut-off score allows a much more confident selection of compounds from huge libraries which is particularly important if not many binders to this particular target are known.

1. Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. Prediction of protein-ligand interactions. Docking and scoring successes and gaps. *J. Med. Chem*. **2006**, 49, 5851-5851.
2. Davies, A., D.; Teague, S. J. Hydrogen bonding, hydrophobic interactions, and failure of the rigid

receptor hypothesis. *Angew. Chem. International Edition* **1999**, 38, 736.

3. Reulecke, I.; Lange, G.; Albrecht, J.; Klein, R.; Rarey, M. Towards an integrated description of hydrogen bonding and dehydration: Reducing false positives in virtual screening using the HYDE scoring function. *ChemMedChem* accepted
4. Huang, N., Shoichet, B.K., Irwin, J.J. Benchmarking sets for molecular docking *J. Med. Chem.* **2006**, *49*, 6789-6801.

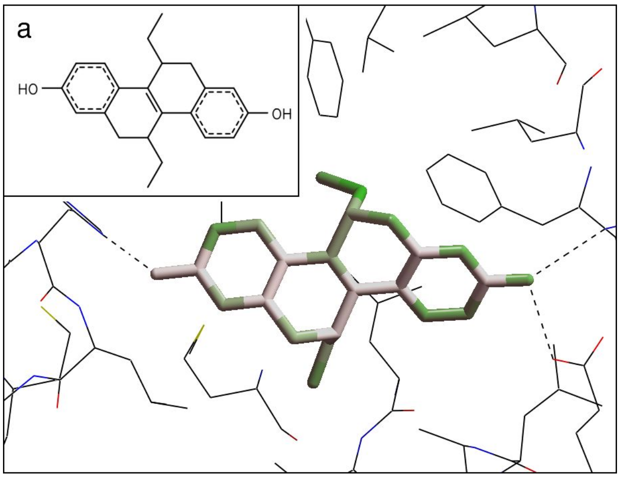


Figure 1: Contribution of individual atoms of the inhibitor in the crystal structure of the estrogen receptor (1l2i). Green coloured atoms contribute favourable to ΔG..

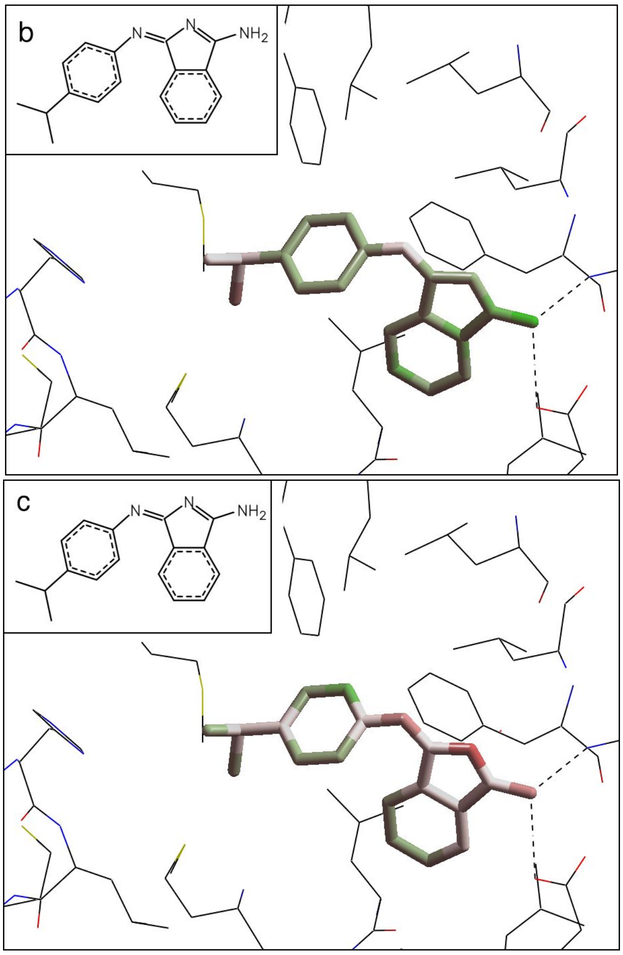


Figure 2: Contribution of individual atoms of decoy ZINC03977652 positioned in the estrogen receptor (1l2i) according to (a) the FlexX scoring function and (b) the HYDE scoring function. Green coloured atoms contribute favourable and read atoms unfavourable to ΔG..

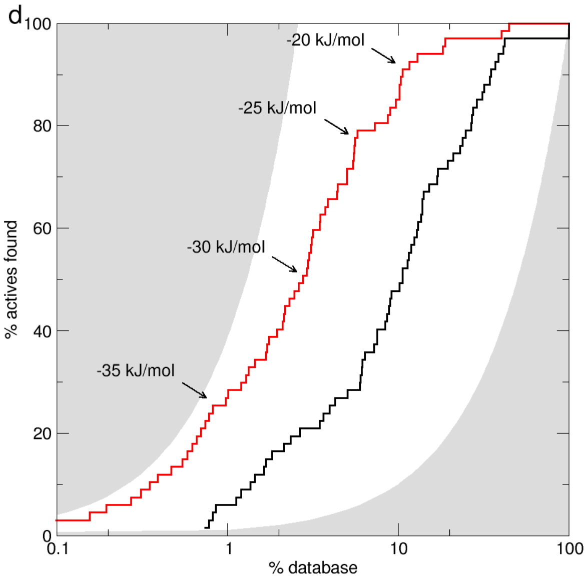


Figure 3: Enrichment plot using the estrogen agonist data set (4) and a random compound library as non-binders.

C-3 : Specificity scoring

S. Tietze, J. Apostolakis , Ludwig Maximilian University, Munich, Germany

Structure based virtual screening (SVS) is now a common tool in the study of molecular recognition, specificity, and the development of novel molecules for pharmaceutical and technological purposes. For a given target molecule the structure of its complex with different candidate ligands (or receptors) is predicted, and from the predicted structure the stability of the complex (binding affinity) is estimated. For the estimation, simple scoring functions are used, which are parametrized either by fitting to binding affinities or using simple statistical approaches. It has recently been shown that the most common of these scoring functions do not predict binding affinity significantly better than simple features of the ligands, such as molecular weight. It is therefore not particularly surprising that score based rankings in SVS often yield poor enrichments in screening benchmarks and applications.

Here we suggest two independent approaches for obtaining specific empirical scoring functions. In the first approach the parameters of the scoring function are optimized solely with respect to performance in screening, without using binding affinity. The potential is trained on a so called cross screening benchmark set, where 100 different ligands are screened against 100 different proteins. The accuracy of the trained potential is evaluated on a complementary cross screening set, which has been selected as to have no overlap with the training set proteins. Significant improvement over our previously validated regression based parameterization of the same functional form (ChillScore[1]) is observed, with the average Area Under ROC (AUC) over all 100 targets improving from 0.65 to 0.77.

The second approach is based on a simple two-step affinity regression method, where however, unspecific (based on ligand properties alone) contributions to the affinity are removed from the potential. A detailed comparison between the two obtained score parameterizations, among each other, and with a number of the generally used scoring functions was performed to highlight the peculiarities of both approaches. One of the interesting results of this study is that the two methods lead to similar parameterizations, even though the first is purely qualitative (no use of binding affinities and X-ray structures is ever made), while the second is based only on crystal structures and experimentally measured binding affinities (no use of predicted structures is made). To further validate the method we show results on standard screening benchmarks (taken from Jain et al. 2005[2]), where a significant improvement over the standard regression potential can be demonstrated: average AUC increases from 0.65 for our previous standard regression based potential to 0.87 for the potential trained on independent cross-screening data, and 0.9 for the two-step specificity fit. We finally discuss one of the most interesting results of this study, namely that scoring functions that show practically no correlation with binding affinity (R=0.32 and 0.26), are significantly better at screening than empirical scoring functions parametrized according to affinity (R=0.53).

1. Tietze S, Apostolakis J, GlamDock: development and validation of a new docking tool on several thousand protein-ligand complexes. J. Chem. Inf. Model. 2007, 47(4):1657-72.
2. Pham TA, Jain AN, Parameter Estimation for Scoring Protein-Ligand Interactions Using Negative Training Data. J. Med. Chem. 2005, 49(20): 5856-5868

C-4 : Flexophore, a new versatile 3D pharmacophore descriptor

M. Korff, T. Sander, J. Freyss, Actelion Ltd., Allschwil, Switzerland

A new molecular descriptor encoding three-dimensional pharmacophore information is described and evaluated. The encoding of a molecule starts by generating a reduced graph. Its nodes, which are called pharmacophore points, are classified by determining their enhanced atom types from a predefined list that was derived from an analysis of the protein data bank. This analysis also yielded a similarity matrix between these atom types characterizing the similarity of any two atom types concerning their coordination behavior to different protein atoms. The pharmacophore descriptor consists of the complete graph of these pharmacophore points. Any of the graph's edges is represented by a histogram of the distances through space between the corresponding nodes considering a representative set of conformers. These conformers are generated by a self-organization based algorithm for conformation sampling.

The pharmacophore similarity of two molecules is then determined by a sub-graph matching procedure considering node and edge similarities. Node similarities are taken from the enhanced atom type similarity matrix described above and edge similarities are calculated from the overlapping areas of the distance histograms.

To evaluate our descriptor's capability to model similarities of protein binding affinities we compiled a data set from the free available DUD dataset.[1] Because the DUD dataset contains proteins and ligands as well as decoys it is not only possible to compare ligand based and structure based screening, this dataset also enabled us to compare our results with other groups using the same dataset. The DUD database was designed to evaluate docking programs and contains 2,950 active compounds against 40 target proteins. Additionally the database contains 36 decoys for each ligand with similar physicochemical properties. We extracted the ligands from the target proteins to extend the applicability of the dataset to ligand based virtual screening. From the 40 target proteins 37 contained ligands which we used as query molecules for virtual screening evaluation. The query molecules were used to screen the test datasets consisting of ligands and decoys. In a large virtual screening experiment the Flexophore descriptor was challenged with five other descriptors and with our in-house docking tool. Four descriptors were chemical fingerprints, all encoding the molecular structure in a different way; the fifth descriptor was a topological pharmacophore histogram. Our experiments showed that the Flexophore descriptor outperformed the chemical descriptors as well as the topological pharmacophore descriptors considering the ability to detect structurally different actives while still being competitive concerning enrichment rates. Thus, it is well suited to find new chemical entities via "scaffold hopping". The Flexophore descriptor can be explored with a Java applet at http://www.cheminformatics.ch/flexophore. Its usage is free of charge and doesn't need any registration.

1. Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, 49, 6789-6801.

C-5 : A fragment-based computational protocol at PDB scale - Application to lead-optimization of DFG-out kinase inhibitors

F. Moriaud[1], T. Henry[1], S.A. Adcock[1], A.M. Vorotynsev[1], L. Martin[1], O. Doppelt[1,2], A.G. De Brevern[2], F. Delfaud[1]

[1] MEDIT, Palaiseau, France

2 EBGM University Paris, Paris, France

Fragment-based drug design has emerged in the last decade and has become an established paradigm at many pharmaceutical companies. This exciting field has been recently reviewed [1]. Obtaining structural information on the fragment complexed to the protein target is a key step and also a major limitation to the number and types of target that are amenable to fragment-based approaches. Therefore, computational methods are needed to mine efficiently all the available 3D structures of ligands complexed to proteins, both treated as a whole and as smaller fragments to increase the likelihood of fragment hopping from one target to another.

MED-SuMo [2,3], a target based drug design tool, offers a procedure to adequately characterize the protein binding site. This tool is based on the identification of local shape and 3D Surface Chemical Features similarities in the target binding site with other proteins (with their co-crystallized ligands). MED-SuMo uses the binding site of the target as a query to search either the whole Protein Data Bank (or any corporate protein structure databank) for all the binding sites that display a local match with the query. This valuable information can then be used to identify which residues of the binding site are potentially important for ligand binding affinity and selectivity. As the similar binding sites are overlaid, the co-crystallized ligands are aligned and are therefore a starting material for ligand hybridization. Among the hundreds of overlaid binding sites generated by MED-SuMo, we found the protein-ligand complexes overlaid by Pierce et al. [4] as their starting material for ligand hybridization in the BREED method. Interestingly, they found relevant combinations of ligands starting from only a few protein-ligand complexes structures and we believe that the output of MED-SuMo is a very promising input for automatic methods like BREED.

In this work, we've worked on a fragment database derived from the PDB: each pdb file is converted into one or more pdb files containing a single fragment as ligand. Fragments are converted to MED-Portions which are fragments annotated with protein 3D environnement and dummy bonds. We've used MED-SuMo to query and mine the Protein's Surface Chemical Functions surrounding MED-portions, seeking similarities with the kinase of interest (i.e. Vegfr DFGout, pdb code 2oh4, ligand code GIG) and collecting a library of 1129 unique MED-Portions positioned in the vegfr's active site and annotated with the counts of contacts and h-bonds. MED-Portions can be used to design novel ligand scaffolds (lead generation) or to optimize attachments on a fixed scaffold (lead optimization). Here we present the optimization of a substructure (i.e. phenylamide) of the GIG ligand to find others DFGout ligands. The 3D hybridisation in 5 iterations of the phenylamide moiety with 1129 fragments suggested by our MED-Hybridiser protocol leads to 22824 molecules. In this list, we identified 3585 different scaffolds, 298 are in PubChem, 46 in the PDB attesting of the diversity and quality of those generated molecules. 25 are marked as active on protein kinase in PubChem bioassay.

1. Hajduk PJ, Greer J. "A decade of fragment-based drug design: strategic advances and lessons learned" *Nat Rev Drug Discov.* **2007** Mar; 6(3):211-9.
2. Jambon M, Imberty A, Deléage G, Geourjon C "A new bioinformatic approach to detect common 3D sites in protein structures" PROTEINS **2003** Structure, Function, and Genetics 52:137-145
3. Jambon M, Andrieu O, Combet C, Deléage G, Delfaud F, Geourjon C « The SuMo server : 3D search for protein functional sites" *Bioinformatics.* **2005** Vol 21, n°20, 3929-3930
4. Pierce AC, Bemis GW, "BREED: generating novel inhibitors through hybridization of known ligands. Application to CDK2, p38, and HIV protease" *J. Med. Chem*. **2004** May 20;47(11):2768

C-6 : Can 3D ligand based virtual screening compete with docking? Application of molecular fields to virtual screening with the DUD dataset.

M Mackey, T Cheeseright, J Melville, S Rose, A Vinter, Cresset BMD, Welwyn Garden City, UK

The use of virtual screening to find new hits and leads has become common place within the pharmaceutical industry. However, the majority of examples and methods in the literature are based around docking to a protein active site or use 2D ligand similarity methods. 3D ligand methods are rarely referenced. Moreover, there is a common perception that ligand based methods are inferior to docking, in that the search ligands do not necessarily encode all the information necessary to find new active molecules, particularly those with completely novel chemotypes. Herein we will present the application of 3D molecular fields[1] to virtual screening using the "directory of useful decoys"[2] as modified by Good for testing of chemotype retrieval rates.[3]

The FieldScreen virtual screening method uses the similarity of molecular electrostatic, steric and hydrophobic fields to rank molecules according to their similarity to a known active (Figure 1). If a protein structure is available, then it can be used as an excluded volume to further focus the search.

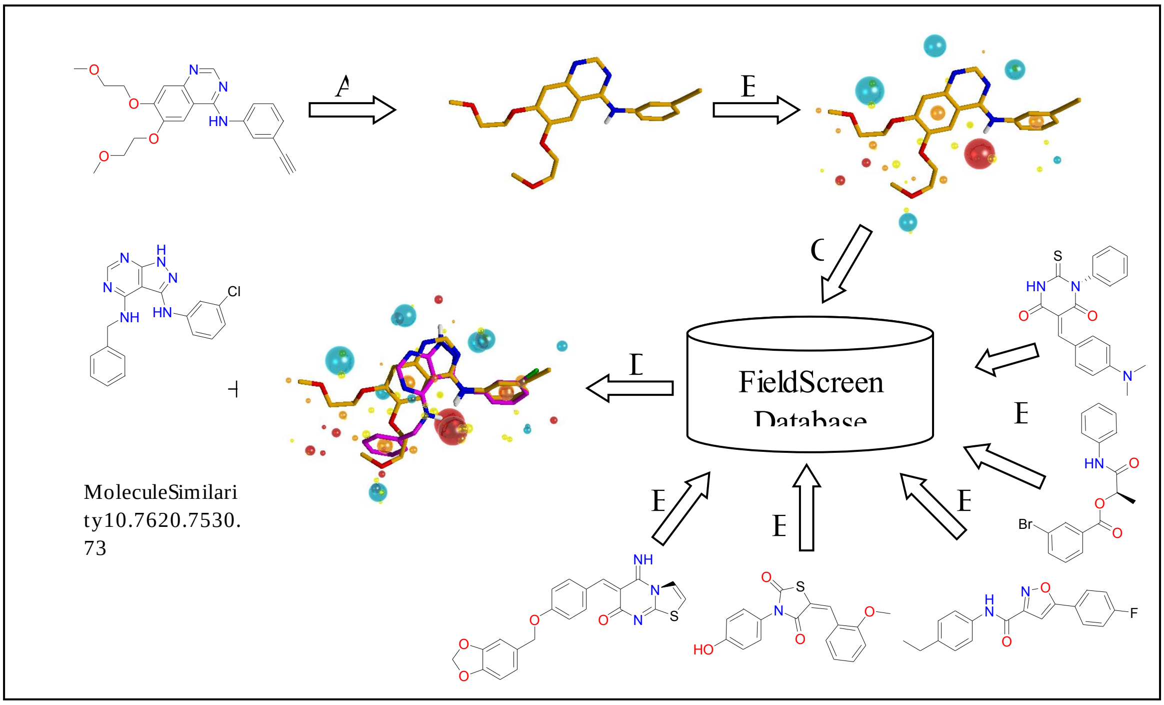


Figure 1. Schematic representation of the steps involved in searching the FieldScreen database.

Field similarity searching with FieldScreen is found to significantly outperform DOCK on almost all of the targets tested, both in terms of raw enrichment rates and in terms of enrichments of novel chemotypes. To allow fair comparison with the "fully automated" nature of the DOCK results[2], the FieldScreen searches were run where possible using the native ligands from the proteins used in the DOCK study with no optmisation or manual tweaking. The inclusion of protein information into the ligand-based screening protocol as an "excluded volume" is shown to further enhance enrichment rates (Figure 2). Moreover, FieldScreen preferentially retrieves small actives, which are more likely to be useful as leads (Figure 3).

We conclude that field similarity searching should be included either as a replacement for or in conjunction with docking in all 3D virtual screening situations.

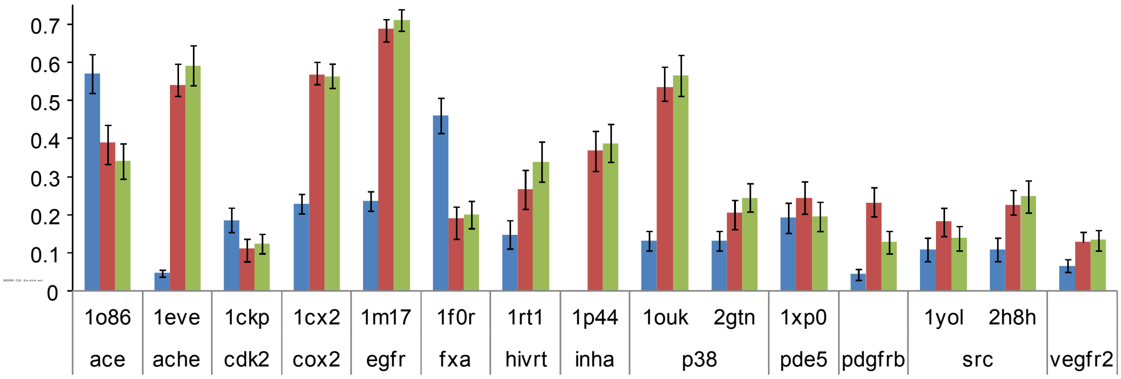


Figure 2. BEDROC enrichments for DOCK (Blue), FieldScreen (Red) and FieldScreen including excluded volume data (Green) for each target.

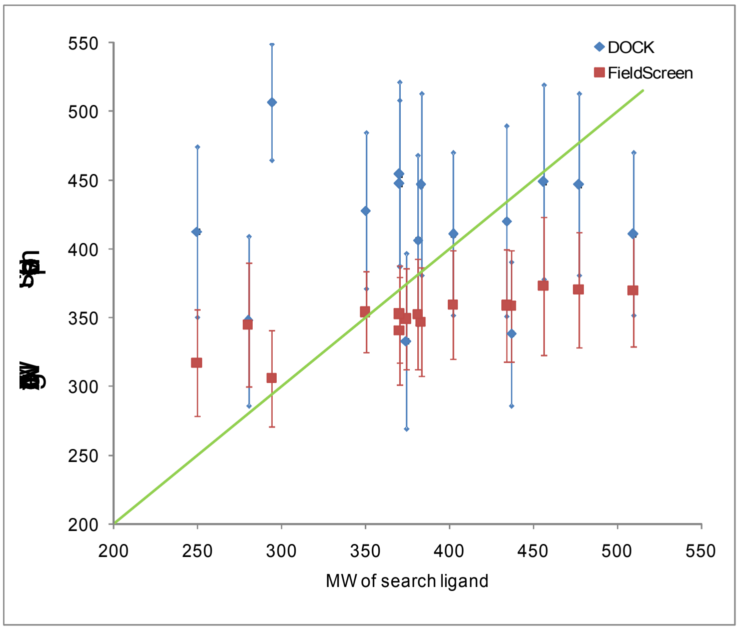


Figure 3 Average Molecular weight (MW) of the top scoring 500 compounds in the DUD all decoys dataset plotted against the MW of the search query. Error bars are 1 standard deviation.

1. T. Cheeseright, M. Mackey, S. Rose, A. Vinter; *J. Chem. Inf. Model*. **2006**; 46, 665-676
2. N. Huang, B. Shoichet, J.J. Irwin; *J. Med. Chem*. **2006**, 49, 6789-6801
3. A. C. Good, T. I. Oprea, *J. Comput. Aided Mol. Des*., in press, DOI 10.1007/s10822-007-9167-2

C-7 : Novel fragment-like PTR1 inhibitors discovered by virtual screening

C. P. Mpamhanga, L. Tulloch, E. Shanks, D. Robinson, W.N. Hunter P.W. Wyatt, R. Brenk, Biological Chemistry and Drug Discovery, College of Life Sciences, University of Dundee, Dundee, U.K.

It is a not uncommon in the drug discovery process to find projects that have been ensnared into intractable 'chemical cul-de-sacs'. This is often due to inherent poor physicochemical and ADMET properties of the existing hits or lead compounds. An increasingly popular method used to avoid this problem is fragment screening. The rationale being that identification of new fragments could provide new starting points for chemistry this process is now popularly referred to as 'scaffold hopping'. Typically these methods use particularly sensitive biophysical methods such as NMR and X-ray which are able to identify even fragments with low binding affinities; however they come with one major disadvantage that of limited throughput. We therefore explored if virtual screening can be used as an initial step to screen rapidly vast fragment libraries for novel scaffolds.

To validate our strategy we chose the enzyme pteridine reductase 1 (PTR1), a short-chain dehydrogenize responsible for the salvage of pterins in, *Trypanosoma brucei,* a protozoan parasite. This parasite is the causative agent of sleeping sickness or Human African Trypanosomiasis (HAT), a human epidemic affecting large numbers in the Sub-Saharan region of Africa. The parasites are auxotrophic for foliates thus making PTR1 a desirable potential drug target.**(1)** So far all known PTR1 scaffolds retain high PSA and as a result may suffer from poor blood brain permeability, a crucial property required for effective HAT therapeutic agents.

Our strategy involved use of DOCK 3.5 **(2)**; DrugScore **(3)**, Interaction fingerprints **(4)** and the MAB force field (**5**), to identify fragment-like compounds. This was followed by hit verification using appropriate biological-assays and X-ray crystallography to confirm the binding modes of the novel scaffolds. From an initial library of 25,000 commercially available fragments only 56 compounds were chosen for testing leading to the discovery of 15 compounds containing eight new scaffolds. One of these hits was subjected to crystal structure analysis and the predicted binding mode was confirmed. However, crystal structure analysis of two analogous revealed two distinct alternative binding modes. In these complexes, previously not observed protein movements and water-mediated protein-ligand contacts occur which prohibit prediction of the binding modes. This study demonstrates the power and pitfalls of using molecular docking for the discovery of fragment-like inhibitors.

1. Bello, A.R.; Nare, R.; Freedman, D.; Hardy, L.; Beverley, S.M. Proc. Natl. Acad. Sci. 1994, 91, 11442-11446.
2. Lorber, D.M.; Shoichet, B.K, Protein Sci. 1998 7 :938-50
3. Gohlke, H.; Hendlich, M.; Klebe, G. J. Mol. Biol. 2000, 295, 337-356.
4. Mpamhanga, C.P.; Chen, B.; McLay, I.M.; Willett, P. J. Chem. Inf. Model. 2006, 46, 2, 686 – 698
5. Gerber PR. J Comput. Aided Mol Des. 1998 12(1):37-51.

C-8 : Fleksy: a flexible approach to induced fit docking.

Sander B. Nabuurs [1,] Markus Wagener [2], Jacob de Vlieg [1,2]

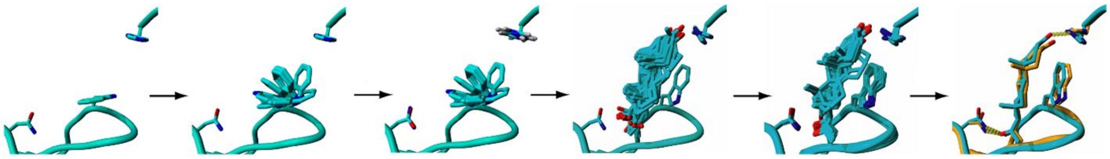
[1] Computational Drug Discovery, CMBI, Radboud University, Nijmegen, The Netherlands

[2] N.V. Organon, Oss, The Netherlands

Protein receptor rearrangements upon ligand binding are a major complicating factor in structure-based drug design. An accurate prediction of these so-called induced fit phenomena calls for ligand docking and virtual screening approaches capable of considering receptor flexibility.

We present Fleksy,[1] a flexible approach aimed at accurately positioning small molecule ligands into a protein receptor, while taking both ligand and receptor flexibility into account. Our method consists of an ensemble docking stage in which the ligand of interest is docked into a structural ensemble of receptor conformations, followed by a complex optimization stage during which both ligand and protein are allowed to move.

Pivotal to our method is the use of receptor ensembles to describe protein flexibility. To construct these ensembles we use a backbone dependent rotamer library and implement the concept of interaction sampling. The latter allows for the evaluation of different orientations and, when relevant, different tautomers of ambivalent interaction partners in the binding site such as asparagine, glutamine and histidine side chains. The docking stage comprises an ensemble-based soft-docking experiment using FlexX-Ensemble,[2] followed by an effective flexible receptor-ligand complex optimization using Yasara.[3] Ultimately Fleksy results in a set of receptor-ligand complexes ranked using a consensus scoring function which combines both docking scores and force field energies.



Averaged over three cross-docking datasets, in total containing 35 different pharmaceutically relevant receptor-ligand complexes, Fleksy reproduces the observed binding mode within 2.0 Å for 78% of the complexes. This compares favorably to the rigid receptor FlexX program[4] which on average reaches a success rate of 44% for these datasets.

1. Nabuurs, S.B.; Wagener, M.; de Vlieg, J. A flexible approach to induced fit docking. *J. Med. Chem.* **2007**, 50, 6507-6518.
2. Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, 261, 470-489.
3. http://www.yasara.org/
4. Claussen, H.; Buning, C.; Rarey, M.; Lengauer, T. FlexE: efficient molecular docking considering protein structure variations. *J. Mol. Biol.* **2001**, 308, 377-395.

C-9 : Index-driven structure-based virtual screening

Jochen Schlosser, Matthias Rarey, Center for Bioinformatics (ZBH), Hamburg, Germany

The standard approach to structure based high-throughput virtual screening nowadays is a sequential procedure. Each molecule of a given library is individually docked into the target protein in order to produce a ranked hit list. With the TrixX approach, we introduce a new paradigm avoiding the iterative process of virtual screening. The non-sequential character of our workflow allows a substantial speedup while yielding comparable re-docking results and enrichment rates.

In order to avoid the iterative processing of molecules, the TrixX method is based on a novel descriptor capable to cover pharmacophoric as well as shape information. Applying standard database technology, TrixX is able to retrieve active compounds in sets of up to several ten thousand ligands. Here, we introduce the next generation of our index-driven virtual screening technology named TrixX BMI with multiple new developments. We replaced the placing and linking procedure of small molecular fragments by rigid body docking of pre-processed conformational ensembles of small molecules or molecular fragments with up to ten rotatable bonds. Furthermore we extended the descriptor significantly by introducing an 80 dimensional steric bulk vector in addition to interaction types, directions and triangle side lengths (Fig. 1). We kept the promising idea of splitting virtual screening into disjoint phases. In the *Data Pre-Processing* phase, descriptors are computed based on conformational ensembles and stored in a database. This is a one-time effort. In the *Virtual High-Throughput Screening* phase, a given protein active site is used to generate complementary descriptors as query templates which are used to identify potential hit candidates within the database. Query matches are then translated into initial fragment placements which are then extended, optimized and scored. Because of the enormous amount of descriptors and their high-dimensional content there is the need for an efficient and overhead-free decision support system. This functionality is realized using compressed bitmap indices supplied by Fastbit.

Re-docking experiments on 115 protein-ligand complexes show that TrixX BMI correctly predicts the pose of the bioactive conformation within a RMSD of less than 2.5 Å of the co-crystallized ligand in 100 cases, thus achieving typical values for current docking tools (Table 1) and improving the runtime by about one order of magnitude. In addition to that several enrichment experiments demonstrate that TrixX BMI is competitive to current methods. TrixX BMI is especially suited for structure-based virtual screening under pharmacophoric constraints (Table 2). To show the systems scalability, a large test set consisting of 1.2 million random lead-like compounds, distributed over a 94-node computing cluster, is used. Four different targets (CDK2, DHFR, ER(agonists), ER(antagonists)) from the DUD, together with pharmacophores from the literature are used as benchmark set. TrixX BMI is able to finish the VHTS runs on all four targets in less than 20 minutes, whereas the average time is below 12 minutes with comparable enrichment rates (see Table 2). Due to its speed, index-based docking opens a new route for modelling protein flexibility in structure-based virtual screening.

1. Schellhammer I, Rarey M. TrixX. Structure-Based Molecule Indexing for Large-Scale Virtual Screening in Sublinear Time. *J. Comp. Aided Mol. Design,* **2007**, 1573-4951
2. Wu K, Otoo E, Shoshani A. An Efficient Compression Scheme for Bitmap Indices. *ACM Transactions on Database Systems,* **2006**, 31, 1-38.
3. Huang, Shoichet, Irwin. Benchmarking Sets for Molecular Docking. *J. Med. Chem.*, **2006**, 49(23), 6789 – 6801.

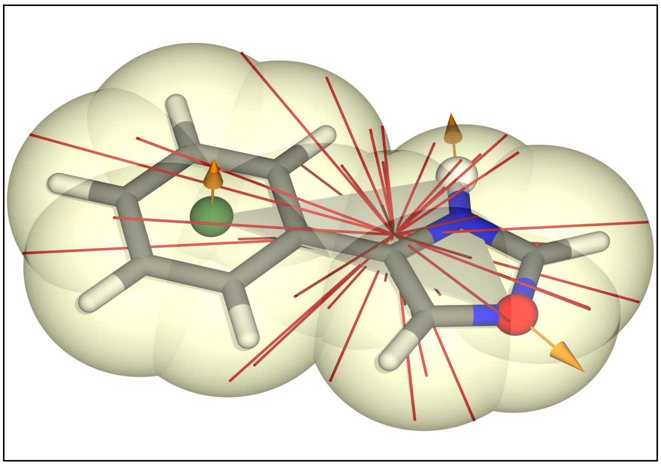


Figure 1: Example of a ligand descriptor

Table 1: Root-mean-square-deviation histogram of the re-docking experiments and the average runtime using all targets against a lead-like dataset of 12600 compounds.

RMSD [Å] ≤	1.0	1.5	2.0	2.5	Avg. runtime
FlexX 2.2.0	76	88	94	96	9.55 [sec/lig]
TrixX BMI	48	77	92	100	0.29 [sec/lig]

Table 2: Enrichment factors at 2% [with | without] pharmacophore constraints. In addition [TrixX BMI | FlexX] runtimes using pharmacophore constraints on the target specific dataset from the DUD and on a random lead-like dataset of 12600 compounds.

Target	$E_{[2\%]}$ TrixX BMI	$E_{[2\%]}$ FlexX 2.2.0	$E_{[2\%]}$ DOCK 3.5.54 (estimated)	Runtime [DUD]	Runtime [rand.]
ER(agonists)	39.39 | 7.57	21.21 | 15.15	n.a. | 8	0.05 | 6.18	0.04 | 4.65
ER(antagonists)	20.51 | 2.56	33.33 | 17.94	n.a. | 12	0.13 | 21.72	0.04 | 6.41
DHFR	44.48 | 40.13	50.16 | 44.81	n.a. | 35	0.03 | 14.39	0.01 | 2.66
CDK2	28.13 | 26.56	15.62 | 21.87	n.a. | 15	0.04 | 5.52	0.01 | 4.07

C-10 : Algorithmic design of ligand binding pockets on protein surfaces

S. Eyrisch, V. Helms, Center for Bioinformatics, Saarland University, Saarbruecken, Germany

In the last few years, the modulation of protein-protein interactions and, in particular, the discovery of so-called small molecule protein-protein interaction inhibitors (SMPPIIs) has become a very active field of research. So far, SMPPIIs have been identified for several protein complexes [1]. However, structure-based drug design of such inhibitors is still in its infancy. In contrast to the well-defined binding pockets in enzymes, most unbound structures of proteins involved in protein-protein interactions lack deep clefts or clearly shaped binding pockets. We therefore developed a pocket detection protocol that provides a starting point for *in-silico* drug design for such cases. This method was validated on three protein-protein interaction systems for which small molecule inhibitors are known, namely MDM2:p53, BCL-X_L:Bak, and IL-2:IL-2Rα. We found that large pockets not detectable in the unbound structure opened frequently on the protein surface during a 10 ns molecular dynamics (MD) simulation in explicit water at room temperature. These transient pockets represent potential binding sites of new inhibitors. At the native binding site, pockets of similar size as with a known inhibitor bound could be observed for all three systems, although these pockets were not - or only partly - present in the starting structure. Docking known inhibitors into these transient pockets resulted in docking poses with less than 2 Å RMS deviation from the crystal structures [2]. Unfortunately, the underlying MD simulations make this protocol quite time-consuming. However, if the potential binding site of a SMPPII (e. g. the protein-protein interaction interface) is known, conformational sampling focused on this region appears more promising than scanning the whole protein surface for transient pockets. Therefore, we present here an efficient method for generating putative binding pockets on protein surfaces algorithmically. After defining the starting structure(s), the approximate location and volume of the desired binding pockets, and a radius dictating which surrounding residues are treated as flexible, the algorithm calculates a pre-defined number of energetically favorable conformations containing these pockets. Internally, the algorithm represents the pockets by dummy atoms and scores their volume via van der Waals energies with the flexible part of the protein. The actual generation of conformations consists of two stages: In the first stage, all flexible residues are mutated to glycine and all rotameric states of their real side chains are pre-calculated. In the second stage, a tree is built up in which each node represents a (partial) conformation of the flexible part of the protein. Each node is scored according to the energy of this conformation and the pocket contribution. The top-scoring leaf nodes then represent the energetically most favorable conformations containing putative ligand binding pockets. For the three proteins mentioned before, the algorithm could generate promising low-energy pockets with realistic predetermined volumes within a few CPU minutes on a standard desktop PC.

1. Wells, J. A.; McClendon, C. L. Reaching for High-Hanging Fruit in Drug Discovery at Protein-Protein Interfaces. *Nature* **2007**, 450, 1001-1009.
2. Eyrisch, S.; Helms, V. Transient Pockets on Protein Surfaces Involved in Protein-Protein Interaction. *J. Med. Chem.* **2007,** 50, 3457-3464.

D-1 : De novo drug design using multi-objective evolutionary graphs

C. A. Nicolaou [1, 2,] C. S. Pattichis [1], J. Apostolakis [3]

[1] Computer Science Dept, University of Cyprus, Nicosia, Cyprus

[2] Noesis Chemoinformatics, Nicosia, Cyprus

[3] Ludwig Maximilian University, Munich, Germany

Drug discovery and development is a complex, lengthy process and failure of a candidate molecule can occur as a result of a combination of reasons, such as poor pharmacokinetics, lack of efficacy or toxicity. Successful drug candidates necessarily represent a compromise between the numerous, sometimes competing objective so that the benefits to patients outweigh potential drawbacks and risks[1]. De novo drug design involves searching an immense space of feasible, drug-like molecules to select those with the highest chances of becoming drugs using computational technology[2]. Traditionally, de novo design has focused on designing molecules satisfying a single objective, such as similarity to a known ligand or an interaction score, and ignored the presence of the multiple objectives required for drug-like behavior. Recently, methods have appeared in the literature that attempt to design molecules satisfying multiple predefined objectives[3] and thereby produce candidate solutions with a higher chance of serving as viable drug leads.

In the first section of this presentation we briefly describe the Multi-objective Evolutionary Graph Algorithm (MEGA), a new multi-objective optimization de novo design algorithm that can be used to design structurally diverse molecules satisfying one or more objectives. The algorithm combines evolutionary techniques with graph-theory to directly manipulate graphs and perform an efficient global search for promising solutions. In the experimental section we present results from the application of MEGA for designing molecules that selectively bind to a known pharmaceutical target using the ChillScore interaction score family[4]. The primary constraints applied to the design are based on the identified structure of the protein target and a known ligand currently marketed as a drug. A detailed explanation of the key elements of the specific implementation of the algorithm is given, including the methods for obtaining molecular building blocks, evolving the chemical graphs, and scoring the designed molecules. Our findings demonstrate that MEGA can produce several structurally diverse candidate molecules representing a wide range of compromises of the supplied constraints and thus, can be used as an “idea generator” to support expert chemists assigned with the task of molecular design.

1. Evans, D. A.; Fitch, D. M.; Smith, T. E.; Cee, V. J. Application of Complex Aldol Reactions to the Total Synthesis of Phorboxazole B. *J. Am. Chem. Soc.* **2000**, 122, 10033-10046.
2. Nicolaou, C. A.; Brown, N.; Pattichis, C. Molecular optimization using computational multi-objective methods. Curr. Opin. Drug Discov. Dev. **2007**, 10(3), 316-24.
3. Schneider, G.; Fechner, U. Computer-based *de novo d*esign of druglike molecules. Nat. Rev. Drug Discov. **2005**, 4(8), 649-663.
4. Brown, N.; McKay, B.; Gilardoni, F.; Gasteiger, J. A graph-based genetic algorithm and its application to the multiobjective evolution of median molecules. J. Chem. Inf. Comput. Sci. **2004**, 44(3), 1079-1087.
5. Tietze, S; Apostolakis, J. GlamDock: development and validation of a new docking tool on several thousand protein-ligand complexes. J. Chem. Inf. Model. **2007**, 47(4), 1657-1672.

D-2 : Planning organic synthesis using reaction types derived from reaction databases

C.H. Schwab [1,] B. Bienfait [1], J. Gasteiger [1,2]

[1] Molecular Networks GmbH, Erlangen, Germany

[2] Computer-Chemie-Centrum and Institute of Organic Chemistry, University of Erlangen-Nuremberg, Erlangen, Germany

A novel, reaction database-driven approach for a stepwise retrosynthetic analysis of a given target compound will be presented. The method uses a knowledge base of different reaction types that has been automatically derived from either a commercially available or in-house reaction database. New synthetic routes are suggested by applying appropriate reaction types to the target compound in a retrosynthetic and automated manner. At each step, the proposed precursors are automatically searched in integrated catalogs of available starting materials for their commercial availability. The rather general definition of chemical reactivity provides the user with new ideas for organic synthesis and deals with a broad range and diverse chemistry, including, *e.g.*, formation of heterocycles, pericyclic reactions, rearrangements and metathesis.

The method has been implemented in the web-based, easy-to-use program system THERESA (THE REtroSynthesis Analyser) and contains tools to browse the suggested syntheses and the corresponding published reaction data, such as literature data and experimental conditions, to build and display interactive synthesis trees and to generate reports.

The presentation will provide insights into the general algorithms of the approach and demonstrate the application of THERESA to some medicinally relevant synthetic targets.

D-3 : Knowledge-based de novo design using reaction vectors

H Patel[1,] *V Gillet*[1], *B Chen*[2], *M Bodkin*[3]

[1] *University of Sheffield, Department of Information Studies, Sheffield, UK*

[2] *University of Sheffield, Department of Chemistry, Sheffield, UK*

[3] *Eli Lilly UK, Windlesham, UK*

A number of *de novo* design tools have been described with the aim of generating novel molecules for drug design, however, they are limited in their ability to propose molecules which are synthetically feasible. Here we describe a novel method that utilises reaction vectors from databases of known reactions to generate structures of interest. The method has been implemented using the pipelining environment KNIME *(1)*.

The reaction vector captures the changes that take place at the reaction centre, without the need for complex reaction mapping procedures *(2)*. By first describing the individual components of a reaction using descriptors such as atom pairs, the overall reaction vector is generated using:

Reaction Vector = [Sum of product vectors] – [Sum of reactant vectors]

We show how reaction vectors can be used to generate novel molecules for synthesis based on simple transformations involving, for example, a simple functional group substitution, to more complex multi-component reactions of the form (R1 + R2 → P1 + P2). We demonstrate the application of the method to the design of known drugs from simple starting materials and a 'cleaned' reaction dataset, via mixing and matching of reaction transforms and reactants.

We also describe the how the method can be developed into an automated multi-objective application for *de novo* design.

References:

1. Konstanz Information Miner. www.knime.org
2. Broughton, H. B. et al. Methods for Classifying and Searching Chemical Reactions. United States Patent Application 367550, 25 Sept, 2000.

D-4 : Recore: Instant 3D scaffold hopping using replacement fragments

P Oledzki, C Detering, T Zuhl, M Gasterich , C Lemmen, BioSolveIT GmbH, Sankt Augustin, Germany

Recore is a 3D ligand scaffold replacement tool [1] which allows the user to generate new ligand cores within a few seconds. The method is an ideal solution for deriving new lead structures from existing ones, which may be patent protected or otherwise unusable.

Given a user-defined central part of a molecule (the 'core'), Recore identifies the geometrically best possible replacement from a 3D fragment database containing millions of moieties. This is achieved within seconds using an ultra-fast indexing mechanism, based on the exit vectors which connect the core with the side-chains.

We used Recore to define multiple 3D fragment sets based on either 3D crystal structures or generated conformers of drug-like molecules. In this step Recore identifies suitable fragments by shredding according to RECAP-type rules [2] for high likelihood of synthetic accessibility. We then applied the search engine on a number of pharmaceutically relevant targets using the bioactive conformations of known binders.

We present results showing that Recore is able to replace a central unit such as to jump from one chemical series to another, while preserving the position of the side-chains. In the hit-lists of Recore we identified other known actives in their bioactive conformation. We show that the use of additional pharmacophore-constraints help further guide the search towards relevant solution sets.

1. Maass, P.; Schulz-Gasch, T.; Stahl, M.; Rarey, M. Recore: A Fast and Versatile Method for Scaffold Hopping Based on Small Molecule Crystal Structure Conformations. *J. Chem. Inf. Model.* **2007**, 47(2), 390-399.
2. Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP—retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Model*. **1998**, 38(3), 511-522.

E-1 : Turns revisited: Clustering turn structures using ESOMs leads to a uniform classification for open, normal and reverse turn families

O. Koch [1/2]*, G.. Klebe* [2]

[1] *The Cambridge Crystallographic Data Centre, Cambridge, CB2 1EZ, UK*

[2] *Philipps-University Marburg, Institute of Pharmaceutical Chemistry, 35032 Marburg, Germany*

In contrast to helices and β-sheets, turns are irregular secondary structure elements. They are up to six residues in length and contain a hydrogen bond or a specific Cα-Cα distance between the first and last residue. Because of their irregularity and a lack of data in the past, previous classifications do not accommodate possible new turn-types in the current protein structures. Additionally, there is a lack of an overall classification for all turn families. Therefore, a new classification was done from scratch. Based on a non-redundant dataset of 1903 protein chains, all possible turn structures were retrieved from Relibase[1] using Reliscript (the Python-based interface) and clustered using Emergent SOMs[2]. The backbone torsion angles describing a turn, including the ω torsion angle, were used as the feature vector.

In general, a hydrogen bond between CO_i – NH_{i+n} is expected[3] within hydrogen bonded turns (‘*normal*’ turns, Figure 1b), but this analysis shows that NH_i – CO_{i+n} hydrogen bonded turns also exists (‘*reverse*’ turns, Figure 1a). They have been theoretically described[4], but never reported previously in proteins. Furthermore, a Cα-Cα distance cut-off of 10 Å was chosen for structures lacking a hydrogen bond followed by a visual inspection of the retrieved clusters to identify turn structures (‘*open*’ turns, Figure 1c).

An analysis of these turn families reveals that, based on the amino acid propensities, the differentiation into normal, open and reverse turn families seems reasonable. Additionally, a large fraction of open turn-types would be ignored using a shorter distance cut-off. Finally, this survey describes 3 *open* turn, 4 *normal* and 5 *reverse* turn families with several turn-types that have not been previously described (Table 1). Protein sequence-based turn prediction with high accuracy confirmed this new categorization based on machine learning methods as consistent and well-defined[5].

In addition to the information about helices and β-sheets retrieved from the PDB, this new uniform classification of turn families is integrated into Secbase, a new extension of Relibase. Relibase is an object-oriented data management system and stores the three dimensional structural information of protein-ligand complexes deposited in the PDB. Both tools provide integrated access for the analysis of secondary structure elements within proteins, protein-protein interfaces and ligand binding.

1. Hendlich, M.; Bergner, A.; Günther, J.; Klebe, G. Relibase: design and development of a database for comprehensive analysis of protein-ligand interactions. *J. Mol. Biol.* **2003**, 326(2), 607-620.
2. Ultsch, A. Maps for the Visualization of high-dimensional Data Spaces. In *Proceedings of Workshop on Self-Organizing Maps (WSOM);* Kyushu, Japan, **2003**, 225-230.
3. Chou, K.C. Prediction of tight turns and their types in proteins. *Anal. Biochem.* **2000**, 286, 1-16.
4. Toniolo, C. Intramolecularly hydrogen-bonded peptide conformations. *CRC Crit. Rev. Biochem.* **1980,** 9(1), 1-44.
5. Meissner, M.; Koch, O.; Klebe, G.; Schneider, G. Prediction of Turns in Protein Structure by Kernel-Based Machine Learning Methods, *submitted*

Figure 1: Different turn conformations: a) reverse ε-turn (3 residues), b) normal β-turn (4 residues), c) open β-turn (4 residues)

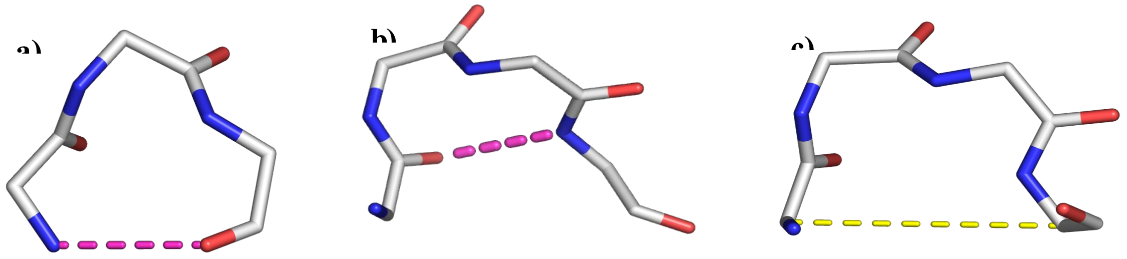


Table 1: Description of used datasets for clustering with number of retrieved turn-types (open turns show structure cluster that looks more like: a) a kink or b) a hook)

	designation	number of		
		residues	structures	turn-types
open	β	4	137101	11 + 6[a]
	α	5	19607	21 + 1[a]
	π	6	21204	22 + 6[b]
normal	γ	3	20198	2
	β	4	28718	6
	α	5	91726	9
	π	6	3994	8
reverse	δ	2	210	9
	ε	3	134	5
	-	4	1957	18
	-	5	1340	21
	-	6	954	13

E-2 : Searching fragment spaces with feature trees

Uta Lessel, Bernd Wellenzohn, Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach an der Riss, Germany

Virtual combinatorial chemistry easily produces billions of compounds, which can not be screened in a conventional manner even with the fastest methods available. An efficient solution for such a scenario is the generation of Fragment Spaces which encode huge numbers of virtual compounds by their fragments/reagents and rules of how to combine them. Fragment Spaces can be screened with so-called Fragment Space searches.

Rarey and Stahl [1] published a method for such searches based on the Feature Tree descriptor [2]. The Feature Tree descriptor is frequently used for virtual screening and has a potential for scaffold hopping [e.g. 3]. The Fragment Space searches are performed without ever fully enumerating all virtual products.

In this presentation we show the preparation of Fragment Spaces based on combinatorial chemistry and share our experiences with Fragment Space searches based on the Feature Tree descriptor in a possible workflow to use this methodology in a pharmaceutical setup.

1. Rarey, M.; Stahl, M. Similarity searching in large combinatorial chemistry spaces. *J. Comp.-Aided Molecular Design* **2001**, 15, 497-520.
2. Rarey, M.; Dixon, J.S. Feature trees: A new similarity measure based on tree matching. *J. Comp.-Aided Molecular Design* **1998**, 12, 471-490.

3. Good, A.C.; Hermsmeier, M.A.; Hindle, S.A. Measuring CAMD technique performance: A virtual screening case study in the design of validation experiments. *J. Comp.-Aided Molecular Design* **2004**, 18, 529-536.

E-3 : Three way comparison of chemical spaces avoiding structure exchange

Jens Loesel, Pfizer, Sandwich, United Kingdom

The comparison of chemical space between two compound collections tends to involve knowledge about the individual structures forming the collections. Unfortunately the exchange of structure data between pharmaceutical companies can pose such big administrative hurdles that the scientific effort of a comparison sometimes seems not worthwhile. In addition a generic comparison between two compound collections tends to outline the difference and novelty between them. To investigate the quality of an unknown set more labor intense work is often necessary.

We tried to address both these issues by developing a generic three way comparison between chemical spaces using a fixed six dimensional BCUT space as a frame of reference. Into this space we map a background set, a reference set as well as an unknown set of compounds.

The background and reference set it used to generate scores for individual bins between zero and one – based on enrichment or lack of it in the reference set compared to the background set. The unknown set is then mapped into the same bins and an overall score between zero and one is generated which is based on the occupancy figures of these set in the individual bins.
This work flow allows a comparison of compound collections by exchanging the frame of reference and the binning operation in one way and by getting bin occupancy values back. Multiple reference sets can be mapped into the same space. These reference sets represent the space of known drugs, active compounds in specific target classes or property spaces like high MW or ClogP areas. Each set will generate an individual score for the unknown set – giving an indication for the suitability of the unknown set to be fit for a set of properties or targets.

Once multiple reference sets are generated the process can be automated – allowing the scoring and evaluation of a new collection in mere minutes without manual intervention.

One often cited disadvantage of BCUT descriptor is the inability to use them for a targeted design. For the purpose of this work the obfuscation of chemical structure into Eigenvalues is seen as an advantage. The exchange of raw descriptors like fingerprints always bears the danger of re-engineering of the descriptor via a GA approach into a (similar) structure. This problem doesn't exit in the outlined work flow.

Some examples of scores for Gene classes and properties are shown in the presentation. A discussion will focus on limitations – as for some analysis you still need structure information. The sensibility of the score will be demonstrated using two sets with (known) different Molecular Weight distributions.

E-4 : Use of data mining to help identify compounds that are unstable in DMSO

J. Hussain, G Harper, Z Blaxill, I Areri, F Saremi-Yarahmadi, S Pickett, P Sidebottom, GlaxoSmithKline Research & Development Limited, Stevenage, UK

Dimethyl sulfoxide (DMSO) has the ability to dissolve a wide range of organic compounds hence it is not surprising that it is the solvent of choice for the solution storage of large compounds sets for use in High Throughput Screening (HTS). One of the issues that need to be addressed when new compounds are added to a HTS compound collection is the likely stability of the compounds in DMSO. GSK and other companies [1,2], employ a set of *in silico* substructural filters to remove known unstable compounds from their HTS compound collections. However, the diversity of these compound sets is large hence it is likely that many compounds that degrade in DMSO are not captured by these filters. The presentation will explain some of the work done within GSK to identify further compounds and substructures that are unstable within DMSO.

GSK has done a significant amount of compound quality assurance (QA) work to make sure its legacy HTS compound collection does not contain any impure compounds [1]. Additionally any new compounds that are purchased by GSK go through the QA process before entering the HTS collection. Because of this, GSK now has a large volume of data on compounds that do not pass the QA process. This data is inherently noisy as compound degradation is just one of the many reasons why a compound can fail QA. The talk will describe the use of a data-driven algorithm [3] to mine this noisy data and retrieve substructures that may be unstable in DMSO.

The data driven algorithm will be explained, along with the range of molecular descriptors used in the algorithm, from simple substructures to more complex pharmacophoric representations. The talk will present the results of the data driven analysis and describe subsequent experimental work which shows that a selection of the substructures selected by the data driven algorithm do indeed degrade in DMSO. The future direction of the work will also be discussed; describing how the data-driven technique has potential to highlight those compounds that need to be re-checked for purity to help maintain a high quality compound collection.

1. Lane, S. J.; Eggleston, D. S.; Brinded, K. A.; Hollerton, J. C.; Taylor, N. L.; Readshaw, S. R. Defining and maintaining a high quality screening collection: the GSK experience. *Drug Discovery Today.* **2006**, 11, 267-272.
2. Schopfer, U.; Engeloch, J.; Stanek, J.; Girod, M.; Schuffenhauer, A.; Jacoby, E.; Acklin, P. The Novartis Compound Archive – From Concept to Reality. *Combinatorial Chemistry & High Throughput Screening*. **2005**, 8, 513-519.
3. Harper, G.; Bravi, G. S.; Pickett, S. D.; Hussain, J.; Green, D. V. S.; The Reduced Graph Descriptor in Virtual Screening and Data-Driven Clustering of High-Throughput Screening Data. *J. Chem. Inf. Comput. Sci.* **2004**, 44, 2145-2156.

F-1 : CypScore - *in silico* case studies on metabolic stability optimization

A. H. Göller[1]*, M. Hennemann*[2]*, T. Clark*[2]

[1] *Bayer Healthcare AG, Bayer Schering Pharma Global Drug Discovery, Wuppertal, Germany*

[2] *Computer Chemie Centrum, Friedrich-Alexander University Erlangen-Nürnberg, Erlangen, Germany*

Metabolism (the M in ADMET) via first-pass clearance leads to low bioavailabilities and is one of the unfavorable ADMET properties leading to the termination of lead optimization and development projects. Additionally, toxic metabolites and metabolites altering the overall metabolism via inhibition or induction of CYP enzymes cause severe side effects.

It is therefore highly desirable to have a tool to predict the lability of specific atomic positions and the metabolites of any compound in silico, since (i) experimental metabolite determinations can often not done be for each interesting project compound due to resource limitations, (ii) even state-of-the-art experiments often provide only larger fragments but not the exact atomic position metabolized, (iii) an in silico method allows to calculate labilities for compounds not yet synthesized and the metabolism of the metabolites, i.e. to simulate multi-step reactions.

CypScore is an in silico prediction software for small molecule metabolic oxidations [1] mediated by cytochrome P450s by applying distinct models for the most important types of oxidation reactions. The models were established based on an in-house created literature compound database comprised of about 850 compounds with about 20000 non-hydrogen atoms labile to about 2400 metabolic transformations. The models were created by a combination of data-mining and linear regression techniques to yield optimized predictions of the labilities of the compounds. The models are based on Parasurf atomic reactivity descriptors [2] from VAMP AM1 quantum chemistry [3] calculated electron density distributions fitted against positionally defined labile molecular positions. The different reaction models in CypScore are appropriately weighted and allow for direct semi-quantitative comparison of the labile positions in one molecule (e.g. aliphatic oxidation vs. hydroxylation), in congeneric series, and in heterogenous datasets. Due to the quantum-chemistry base of the reactivity descriptors, CypScore does predictions ab initio and is not just a learned method. Thus, it is able to extrapolate outside the dataset used for calibrating the models.

CypScore was carefully validated against literature and in-house data sets where the software was able to find the major metabolite(s) for about 90 % of the compounds if considering the 3 weakest positions in each molecule. It is regularly applied to in-house projects for parallel SAR and metabolism optimization.

In this paper we present several literature examples of optimization strategies to reduce first-pass metabolism in CCR5 antagonists, endotehelin antagonists, farnesyltransferase inhibitors, Cox-2 inhibitors, thrombin inhibitors and mGlu5 allosteric antagonists and one generic in-house example which illustrate the capabilities of the CypScore approach.

1. (a) GRC Computational Chemistry, 2006, Les Diablerets, Switzerland, Talk; (b) SMI in silico ADMET conference, 2007, London, UK, Talk; (c) 4th Joint Sheffield Conference on Chemoinformatics, 2007, Sheffield, UK, Poster; (d) publication in preparation.
2. Parasurf, CEPOS InSilico Ltd., 26 Brookfield Gardens Ryde, Isle of Wight PO33 3NP.
3. VAMP 10.0, Clark, T.; et al., Erlangen 2007.

F-2 : SyGMa: combining expert knowledge and empirical scoring in the prediction of metabolites

L. Ridder, M. Wagener, Molecular Design & Informatics Department, Organon, part of Schering-Plough, The Netherlands

Predictions of potential metabolites, based on chemical structure, are becoming increasingly important in drug discovery to guide medicinal chemistry addressing metabolic issues and to support experimental metabolite screening and identification. We present a novel rule-based method, SyGMa (Systematic Generation of Metabolites), to predict potential metabolites of a given parent structure. A set of reaction rules covering a broad range of phase 1 and phase 2 metabolism has been derived from metabolic reactions reported in the Metabolite database to occur in man. An empirical probability score is assigned to each rule representing the fraction of correctly predicted metabolites in the training database. This score is used to refine the rules and to rank predicted metabolites. To obtain a better overview of which metabolic reactions are reproduced / not reproduced by SyGMa, and to support ongoing efforts to extend the rules, a similarity analysis of the reactions present in the database was performed and mapped with the SyGMa results. The current rule set of SyGMa covers approximately 70% of metabolic reactions observed in man.

Evaluation of the rule based predictions demonstrated a significant enrichment of true metabolites in the top of the ranking list: while, in total, 68% of all observed metabolites in an independent test set were reproduced by SyGMa, a large part, 30% of the observed metabolites, were identified among the top 3 predictions. From a subset of cytochrome P450 specific metabolites, overall 84% were reproduced, with 66% in the top 3 predicted phase 1 metabolites.

Specific examples are given to demonstrate the usage of SyGMa in experimental metabolite identification as well as the application of SyGMa to suggest chemical modifications improving the metabolic stability of compounds.

Finally, a method will be presented to supplement each prediction with the most similar experimental examples in a metabolite database. This enables the user to quickly assess with more detail the value of individual SyGMa predictions.

F-3 : TopoHERG – A highly selective pharmacophoric classifier for hERG-channel active compounds

B. Nisius [1], A. H. Göller [2]

[1] B-IT Life Science Informatics, University of Bonn, Germany, and Bayer Healthcare AG, Bayer Schering Pharma Global Drug Discovery, Wuppertal, Germany

[2] Bayer Healthcare AG, Bayer Schering Pharma, Global Drug Discovery, Wuppertal, Germany

Cardiac arrhythmia as a side-effect of many drugs has become a major pharmacological safety concern and led to the withdrawal of many drugs through recent years. Virtually every case of prolonged duration of cardiac action potential related to drug exposure (long QT syndrome) can be traced to one specific mechanism: blockade of I_{kr} current in the heart, conducted by an ion channel encoded by the human ether-a-gogo-related gene (hERG).

Experimental and modelling results provide clear evidence that ligand binding is rather unspecific and inhomogenous, and there will exist multiple binding SAR’s. It is therefore wise to combine multiple models based on different structure and property descriptors.

Our highly specific TopoHERG approach uses the Tripos Topomer Search [1] to compare any query molecule via topomer similarity to a database of known actives and inactives [2] using the fact that similar compounds tend to show similar biological effects. The classifier with strict topological distance classifies about 54 % of the compounds with a specificity of 95 % and a sensitivity of 76 % on the classified compounds.

The unclassified compounds were then run through a decision tree-based classifier based on clogP, CMR and the availability of a ionizable nitrogen we adapted from literature [3]. Since the classifier was created on 490 Johnson&Johnson in-house compounds, our data are a real validation set. The outcome of the combined TopoHERG and decision-tree classifiers on the complete dataset is a sensitivity of 54 %, a specificity of 94 % and overall 87 % of the compounds predicted correctly.

TopoHERG, has a high predictivity for any compound it is applicable to due to the fact that regardless of the binding mode of any chemotype the model will classify by inherent pharmacophoric similarity. By design, the model will improve with any new compound added to the database which can grow on a daily basis without re-training. Additionally, no split into training and test set is necessary, broadening the domain of applicability of the method. Combined with one or multiple orthogonal classifiers the overall performance of the model can fulfill the needs of pharmaceutical industry.

1. Cramer, R. D.; Jilek, R. J.; Andrews, K. M. dbtop: Topomer similarity searching of conventional structure databases, J. Molec. Graph. Mod., **2002**, 20, 447-462.
2. Database: The dataset consists of 475 compounds. 232 compounds were collected from publicly available hERG blockade IC_{50} data, the other compounds are in-house patch clamp measurements on HEK and CHO cells. 276 compounds in the dataset don't exhibit hERG blockade ($pIC_{50}<5$), 118 compounds show medium hERG activity ($5<=pIC_{50}<6$) and 82 compounds are highly hERG active ($pIC_{50}>=6$).
3. (a) Buyck, C.; Tollenaere, J.; Engels, M.; Clerck, F. D.; An in silico model for detecting potential hERG blocking; *Poster presentation, Euro-QSAR 2002*, *Bournemout*h, **2002**; (b) Aronov, A. M.; Goldman, B. B.; A model for identifying HERG K+ channel blockers; Bioorg. & Med. Chem. **2004**, 12, 2307-2315.

F-4 : Compound set optimization and sequential screening using Emerging Chemical Patterns

J. Auer , J. Bajorath, University of Bonn, Bonn, Germany

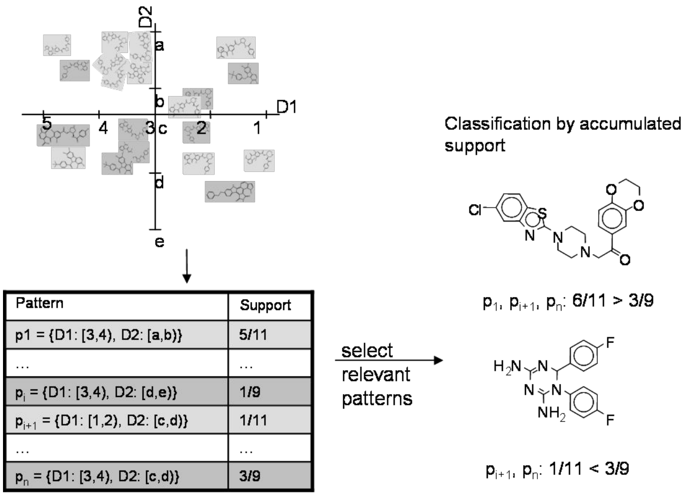
A method called "Emerging Chemical Patterns" (ECP) has recently been introduced as a novel approach to binary molecular classification[1]. The underlying pattern recognition algorithm was first introduced in computer science and then adopted for applications in medicinal chemistry and compound screening. The methodology makes it possible to extract key molecular features from very few known active compounds and classify molecules according to different potency levels. The approach was developed in light of the situation often faced during the early stages of lead optimization efforts: too few active reference molecules are available to build computational models for the prediction of potent compounds. The ECP method generates high-resolution signatures of active compounds by selecting class-specific combinations of 2D descriptor value ranges. These signatures can then be used to build highly accurate classifiers (Figure 1).

A special feature of ECP is its ability to accurately classify molecules on the basis of very small training sets containing only a few compounds. This feature is highly relevant for virtual compound screening when only very few experimental hits are available as templates. We designed an experiment based on four classes from literature sources (benzodiazepines, dihydrofolate reductase inhibitors, glycogen synthase kinase-3 inhibitors and HIV protease inhibitors), comparing ECP to a decision tree approach and a binary QSAR implementation. The analysis showed that ECP produced predictive models on the basis of training sets consisting of only three compounds.1

In addition to individual compound predictions, an iterative ECP scheme has been designed which optimizes a compound set's potency in a sequential manner. In each iteration, small compound sets are selected as training sets and used to remove weakly potent compounds. We could show that this iterative ECP classification produced compound selection sets with increases in average potency of up to 3 orders of magnitude (Figure 2).

The ability of ECP to produce highly accurate classifiers based on small training sets can also be used to reduce the experimental effort in high-throughput screening campaigns by combining experimental screening and ECP classification in a sequential screening methodology[2]. We simulated sequential screening using an experimental high-throughput screening (HTS) data set containing inhibitors of dihydrofolate reductase. We focused on minimizing the number of database compounds that need to be evaluated in order to identify a substantial fraction of available hits. Iterative ECP calculations recovered on average between 19% and 39% of available hits in the data set while dramatically reducing the number of compounds that need to be tested to 0.002% - 9% of the screening database.[2]

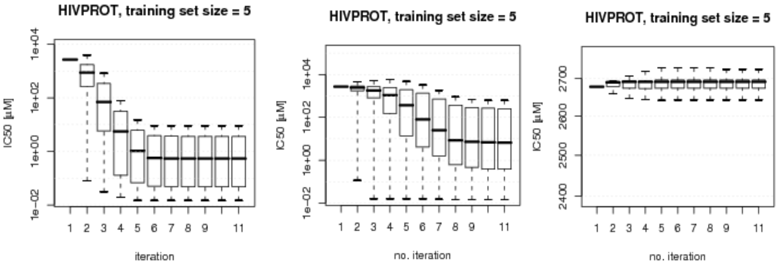
1. Auer, J. and Bajorath, J. Emerging Chemical Patterns: A New Methodology for Molecular Classification and Compound Selection *J. Chem. Inf. Model.* **2006**, 46, 6, 2502 – 2514.
2. Auer, J. and Bajorath, J. Simulation of Sequential Screening Experiments Using Emerging Chemical Patterns, *Medicinal Chemistry* **2008**, 4, 1, 80 – 90.



Pattern	Support
p1 = {D1: [3,4), D2: [a,b)}	5/11
…	…
p_i = {D1: [3,4), D2: [d,e)}	1/9
p_{i+1} = {D1: [1,2), D2: [c,d)}	1/11
…	…
p_n = {D1: [3,4), D2: [c,d)}	3/9

D2
a
b
c
d
e
D1
5
4
3
2
1
select relevant patterns
Classification by accumulated support
p_1, p_{i+1}, p_n: 6/11 > 3/9
p_{i+1}, p_n: 1/11 < 3/9

Figure 1: Classification of compounds based in ECPs. Class-specific descriptor combinations (patterns) are computed from a training set. For classification, the supports (fraction of training compounds that match a pattern) of all patterns matching the test compounds are accumulated and the class with the highest accumulated support is the predicted class.



a) ECP b) Binary QSAR c) Decision tree

Figure 2: Simulated lead optimization. ECP, binary QSAR, and a decision tree are used to select highly potent sets of active compounds in an iterative procedure. During each step, compounds are classified as highly active (<= 1µM) or weakly active (> 1µM) and weakly active compounds are iteratively removed.

F-5 : Interpretable activity models: Exploring the limits of pharmacophores and 3D QSAR methods

D.A. Evans , D.A. Thorner, M.J. Bodkin, Eli Lilly & Co. Ltd., Lilly Research Centre, Windlesham, UK

QSAR modeling in *pharma* is often focused on the rapid production and systematic updating of models based upon large numbers of easy to calculate molecular descriptors or fingerprints whose interpretation can be difficult. Such methods typically output either a predicted activity class or a numerical value with a significant uncertainty. This makes them most useful in the early stage of a drug discovery project to screen existing compounds, prioritize library design options or check for potential off-target activities. In the later lead optimization phase significant activity data on a particular compound series already exist and medicinal chemists tend to be searching for a systematic understanding of the SAR. Activity models which are interpretable by a medicinal chemist are more useful in deciding ‘what to make next?’

Recent validation studies on QSAR methods have demonstrated that equivalent or superior performance can be obtained with more interpretable 3D QSAR methods such as CoMFA compared to selected 2D methods.[1] The atom based grid method Phase was also demonstrated to be generally superior in performance to the pharmacophore alignment method Catalyst Hypogen.[2] The work to be reported aims to establish the viability of using pharmacophore-based alignment to produce 3D QSAR models from a project SAR of potentially thousands of compounds in an automated fashion. We compare the performance of Phase and CoMFA and the quality of QSAR models generated from manual and pharmacophore based alignments and also rank these against the now commonly used fingerprint-based support vector machine (SVM) models. Interestingly, we present results on the variation of predictive QSAR performance with training set size and investigate the observed performance plateau.

The use of automated pharmacophore overlays become computationally demanding when hundreds of molecules need to be considered simultaneously. We address this problem by pre-clustering the data set using 2D methods before building pharmacophore models within and across each cluster and combining the derived pharmacophores into a minimal set. This approach aims to answer the common question of whether compounds of two chemical series have the same binding mode and hence transferable SAR, by testing both whether they have the same pharmacophore and whether the combined series can produce a good quality QSAR model.[3] We will discuss the delivery of the approach as a desktop tool for medicinal chemists

1. Sutherland, J.J.; O'Brien, L.A; Weaver, D.F. A Comparison of Methods for Modeling Quantitative Structure-Activity Relationships. *J. Med. Chem.,* **2004,** 47, 5541 -5554,
2. Evans, D.A.; Doman, T.N.; Thorner, D.A.; Bodkin, M.J. 3D QSAR methods: Phase and Catalyst Compared. *J. Chem. Inf. Model.,* **2007**, 47, 1248 -1257**.**
3. Iyer, M.; Hopfinger, A. J. Treating Chemical Diversity in QSAR Analysis: Modeling Diverse HIV-1 Integrase Inhibitors Using 4D Fingerprints. *J. Chem. Inf. Model.,* **2007**, 47, 1945 -1960.

F-6 : QSAR modeller seeks meaningful relationship

C.L. Bruce[1], S.D. Pickett[2], J.D. Hirst[1]

[1] School of Chemistry, University of Nottingham, Nottingham, U.K.

[2] GlaxoSmithKline, Stevenage, U.K.

Disappointment with QSAR has been articulated recently[1] and although the technique is an important tool in the drug discovery process, improvements perhaps have not been as forthcoming as in other areas. A good model comprises several components. Predictive accuracy is paramount, but it is not the only important aspect. In addition, one should apply robust and appropriate statistical tests to the models to assess their significance or the significance of any apparent improvements. The real impact of a QSAR, however, perhaps lies in its chemical insight and interpretation, an aspect which is often overlooked.

Any insight into the relationship between descriptors and structure can be used to further our understanding, but obtaining this insight is not always as straightforward as calculating predictive accuracy. Interpretation is dependent on the classifier. For example, a decision tree is simple to interpret, but does not produce the most predictive models. Similarly, support vector machines offer excellent predictive capability, but generate a model that is difficult to interpret.

Previously, we have shown random forests predict with accuracies comparable to support vector machines.[2] A decision tree is easier to interpret than a random forest; Breiman gave them an ‘A+’ and ‘F’ for interpretability, respectively.[3] It is the different tree construction and number of trees present in a forest that makes their interpretation complicated. One cannot simply glance through the forest and readily see the model, whereas one can with a decision tree.

Therefore, to obtain useful interpretation from a random forest we have employed a selection of tools. This includes alternative representations of the trees using SMILES and SMARTS. Using existing methods we can compare and cluster the trees in this representation. Descriptor analysis and importance can be measured at the tree and forest level. Pathways in the trees can be compared and frequently occurring sub graphs identified. The ability to distinguish multiple modes of action in a data set is tested. In terms of model assessment, all test data can be assigned a level of confidence, reflecting the extent to which the prediction is an extrapolation from the model. These tools have been built around the Weka machine learning workbench[4] and are designed to allow further additions of new functionality.

References:

1. Johnson, S. R. The Trouble with QSAR (or How I Learned To Stop Worrying and Embrace Fallacy). *J. Chem. Inf. Model.* **2008**, 48, 25-26.
2. Bruce, C. L.; Melville, J. L.; Pickett, S. D.; Hirst, J. D. Contemporary QSAR Classifiers Compared. *J. Chem. Inf. Model.* **2007**, 47, 219-227.
3. Breiman, L. Statistical Modeling: The Two Cultures. *Statist. Sci.* **2001**, 16, 199-231.
4. Witten, I. H.; Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed.; Morgan Kaufmann: San Francisco, CA, 2005.

F-7 : Rational design of M1-Muscarinic Antagonists using combinatorial transformation

M.B. Bolger, R. Fraczkiewicz, D. Miller, J. Crison, W.S. Woltosz, Simulations Plus, Inc., Lancaster, CA, U.S.A.

Purpose. To develop an in silico process for de-novo design from high throughput screening (HTS) data.

Methods. HTS data for 61,044 M1-muscarinic antagonists (PubChem .AID 628) were filtered and classified by ClassPharmer™ (Simulations Plus, Inc., Lancaster, CA). A categorical Support Vector Machine Ensemble (SVME) model for an N-phenylpiperazine chemical class with 168 members was generated using ADMET Predictor™ (AP). SVME performance accuracy: 91% TP, 88% TN, with only 5% FP, and 5% FN. We automatically generated 10,000 bioisosteres of the lowest molecular weight active molecule in this class using the Combinatorial Transformation feature in ClassPharmer. The new compounds were filtered using AP rules for ADMET properties and for M1-antagonist activity. Finally, the resulting data was exported and loaded into ClassPharmer for final class generation.

Results. 10,000 potential antagonists produced 6 molecules predicted to be active and with only one adverse ADMET characteristic. The most active class (phenyl-perhydrodiazepines) had 226 members with 68% of the molecules predicted to be active based on the SVM categorical model. Three of the molecules with the lowest ADMET Risk score had no hits in a substructure search of Chemical Abstracts registry indicating novel composition of matter. These molecules scored three hits in the ADMET Risk scale indicating potential interaction with the estrogen receptor, potential toxicity on the fat-head minnow LD50 scale, and potential inhibitory activity at the hERG potassium channel.

Conclusions. Classification, activity modeling, and ADMET property estimation for HTS data, combined with Combinatorial Transformation feature in ClassPharmer, are a powerful set of tools for the rational design of novel M1-muscarinic receptor antagonists

1. May, L.T. and A. Christopoulos, Allosteric modulators of G-protein-coupled receptors. Curr Opin Pharmacol, 2003. 3(5): p. 551-6.
2. Gasparini, F., R. Kuhn, and J.P. Pin, Allosteric modulators of group1 metabotropic glutamate receptors: novel subtype-selective ligands and therapeutic perspectives. Curr Opin Pharmacol, 2002. 2(1): p. 43-9.
3. Spalding, T.A., et al., Discovery of an ectopic activation site on the M(1) muscarinic receptor. Mol Pharmacol, 2002. 61(6): p. 1297-302.

F-8 : Structure-activity landscapes: a new way to study a structure-activity relationship

J. van Drie[1], *R. Guha*[2]

[1] *John H Van Drie Research LLC, Andover, MA, USA,*

[2]*Indiana University, Bloomington, IN, USA*

When first confronted with a new SAR (structure-activity relationship), it is a challenge to identify the salient aspects of that SAR. We present a new method[1], Structure-Activity Landscapes, that facilitates the identification of those salient features, by studying the SAR pairwise, and ranking all pairs by a simple index

$$\mathrm{SALI} = \Delta A_{ij} / (1 - S_{ij})$$

which highlights those pairs of molecules most similar, with the largest change in activity. The similarity of a pair of molecules is denoted S_{ij}; their activity differences is denoted Δa_{ij}.

In addition to simply looking at a list of pairs sorted by the SALI, one can use the SALI to convert an SAR into a graph representation, which further facilitates obtaining an overall perspective on the SAR.

Finally, the SALI index leads to a novel metric for assessing the performance of a computational model of that SAR, and can even be used as a metric for discovering novel models.

Examples of each of these applications will be highlighted using literature SAR.

1. Guha, R.; van Drie, J.H., The Structure-Activity Landscape Index: Identifying and Quantifying Activity Cliffs. *J. Chem. Inf. Model.* **2008**, accepted for publication.

Poster Session Abstracts

Poster Session Abstracts

P-1 : Discovery Portal – a novel tool to increase productivity, efficiency and transparency across R&D organizations

J. Tomczak, Accelrys, Cambridge, United Kingdom

Data integration and data standardization is a major challenge and an important area of investment for most pharmaceutical companies. There are a number of commercial, in-house and even open source tools and solutions for bringing together and analyzing chemical and biological data. However, the data are usually disconnected from the processes that generate the data and internal drug discovery workflows. Furthermore, R&D scientists have to spend significant amount of time on learning and using generic tools for data extraction, mining and reporting and adapting them to their daily practices. Recognizing this bottleneck, a higher-level decision support and tracking tool can be designed on top of an existing, integrated infrastructure that fully implements internal discovery processes. The resulting Discovery Portal would be a Web-based resource where scientists and managers could not only share, mine and visualize all of their data and knowledge about drug discovery programs, but also define the programs in terms of experiments to be performed, which in turn could be requested and tracked on-line. The system developed by Accelrys combines Accelrys' products (in particular Pipeline Pilot), modern Web-based enterprise technologies and open source frameworks. The technology choice makes the system highly expandable and allows adding new capabilities to a running system. The portal demonstrates also that limitations of Web-based solutions are not as restrictive as they might seem. (Concepts will be illustrated with specific implementation examples).

P-2 : ~~A simple language for conversing between diverse applications~~

~~O Williams, A Westley, R Brown, Accelrys, Cambridge, UK~~

WITHDRAWN

P-3 : The use of stereo descriptors in the context of a structure validation workflow

Pedro Gomez Fabre, R Brown Accelrys, Cambridge, UK

The proper interpretation of a chemical structure drawing within a chemical information system is essential for storage and retrieval as well as for the analysis of chemical structures. A Structure Validation workflow will go through several stages of structure examination in order to verify consistency with proprietary drawing rules and to identify stereochemical features. This presentation describes these stages with particular focus on the stereochemical analysis. The latter relies on a robust, accurate, and high performing CIP calculator, which was developed in a joint collaboration of scientists and software engineers. The range of stereochemical descriptors provided by this calculator includes regular Cahn-Ingold-Prelog stereodescriptors as well as descriptors for identifying stereocentres which are not explicitly specified (hidden) in a given structure drawing or which are incorrectly drawn.

P-4 : OSIRIS, an entirely in-house developed drug discovery informatics system

T. Sander, J. Freyss, M. Korff, J. Reich, C. Rufener, Actelion Pharmaceuticals Ltd., Allschwil, Switzerland

We present OSIRIS, a drug discovery informatics system entirely self-developed at Actelion Pharmaceuticals Ltd. OSIRIS covers all information handling aspects from compound synthesis via biological testing to preclinical development. Its design principles were platform and vendor independence, a consistent look and feel, and complete coverage of the drug discovery process by custom tailored applications. These include electronic laboratory notebook applications for biology and chemistry, tools for high-throughput- and secondary screening evaluation, state-of-the-art chemistry-aware data visualization, physicochemical property prediction, 3D-pharmacophore comparisons, interactive modeling and computing grid based ligand-protein docking. Most applications were written in Java and built on top of a Java library layer that provides reusable functionality and GUI components such as structure canonicalization, combinatorial enumeration, chemical editors, etc.

We use an Oracle database as a scalable persistence engine for data and documents. Its search capabilities were extended by writing an Oracle Cartridge to allow for SQL based chemical substructure and similarity searches. We also employ the Oracle ConText Option to allow document full text searches.

Exemplarily we will demonstrate OSIRIS DataWarrior, a chemistry aware data analysis and visualization tool. It provides simultaneous tabular and graphical views on the same data and allows interactive record filtering on numerical, textual or chemical criteria. It opens various file formats or may directly query the research database. DataWarrior is used for diverse tasks such as combinatorial library planning, microarray data analysis, structure activity correlations on multidimensional project data, or HTS data analysis and lead selection.

Concluding we consider our experience a valid prove that a high degree of in-house software development may be a valid alternative even for smaller drug discovery departments versus the supposedly less risky approach of licensing best-of-bread tools and investing substantially into their integration.

Actelion is a 10 years old biopharmaceutical company with 1700 employees of which about 250 work in drug discovery. We have three drugs on the marked, which exceeded CHF 1.3 billion net revenues in 2007.

P-5 : Scientific database application without borders: Empowering the scientists

Man-Ling Lee, Ignacio Aliagas, Alberto Gobbi, Genentech Inc., South San Francisco, USA

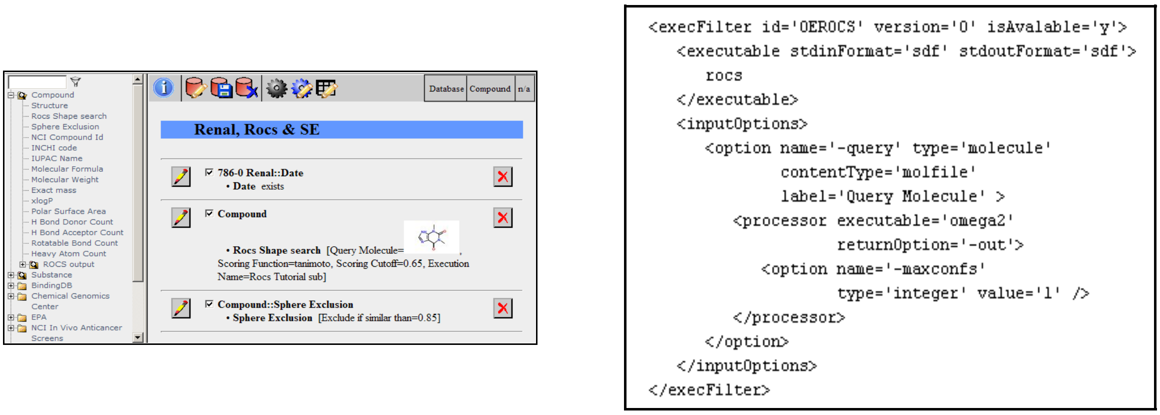
Traditionally, database applications are geared towards a scientific domain and their search capabilities are specific to that domain. For example, chemical database applications typically have only 2D structure search functionality. It is not possible to search by 3D shape similarity or other computational methods. For data such as sequence or crystal data, the users have to use other database applications. This requires users to familiarize themselves with multiple applications.

The goal of this presentation is to demonstrate that a database application is an ideal platform for end-users to explore scientific methods, applications and services offered on the intranet and internet. Scientists can operate in a familiar environment while still having easy access to external applications and data from different scientific domains.

AEREA is a web-based database application with interfaces that enable the straightforward integration of external programs and services allowing users to perform cross-border investigations. External programs can be integrated as filters in the query builder. A “filter" could be a 3D molecule shape comparison program, enabling scientists to perform other types of structure searches in the same application. Other programs can be integrated as hyperlinks or service requests. This enables post-processing of the retrieved data.

AEREA follows the paradigm of the architecture described by Hewitt et. al.[1] XML files are used to store the meta data describing the relationship of the data in the database and the access to the external programs. In addition, the meta data configures the graphical user interface. The xml-file is a flexible and human readable way to configure the application. It enables the integration of external programs with minimal programming knowledge. It also enables to use AEREA for searching in databases other than small molecule databases.
We will describe the plattform focusing on the interfaces that allow for the flexibility and extensibility. We will illustrate the interfaces and explain the configuration by the xml definition. Examples based on the data in the PubChem and WOMBAT databases will be presented.

1. Hewitt, R.; Gobbi, A.; Lee, M.; A Searching and Reporting System for Relational Databases Using a Graph-based Metadata Representation. *J. Chem. Inf. Comput. Sci.* **2005**, 45, 863-869.



Renal, Rocs & SE

☑ 786-0 Renal::Date
- **Date** exists

☑ **Compound**
- **Rocs Shape search** [Query Molecule= , Scoring Function=tanimoto, Scoring Cutoff=0.65, Execution Name=Rocs Tutorial sub]

☑ **Compound::Sphere Exclusion**
- **Sphere Exclusion** [Exclude if similar than=0.85]

```
<execFilter id='OEROCS' version='0' isAvalable='y'>
   <executable stdinFormat='sdf' stdoutFormat='sdf'>
      rocs
   </executable>
   <inputOptions>
      <option name='-query' type='molecule'
            contentType='molfile'
            label='Query Molecule' >
        <processor executable='omega2'
                returnOption='-out'>
          <option name='-maxconfs'
                type='integer' value='1' />
        </processor>
      </option>
   </inputOptions>
</execFilter>
```

Fig. 1 On the right is the query builder with a query containing a shape search. For the user, the ROCS Shape search appears like other search constraints. As shown in the xml configuration on the left, OMEGA is used to generate a 3D conformation from the 2D structure input and ROCS is used to perform the 3D shape comparisons. (OMEGA and ROCS are products of OpenEyes Scientific Software).

P-6 : Diversity oriented virtual compound selection strategy for high throughput screening of potential anticancer agents

György Dormán[1], Miklós J. Szabó[1], Angelo Carotti[6], Simona Distinto[2], Amiram Goldblum[4], Anna Gulyás-Forró[1], Johannes Kirchmair[3], Thierry Langer[3], David Marcus[4], Jordi Mestres[5], Orazio Nicolotti[6], Ferran Sanz[5] and Ismael Zamora[5]

[1]AMRI, Záhony u. 7, 1031 Budapest, Hungary

[2]Dipartimento Farmaco Chimico Tecnologico, Universita` degli Studi di Cagliari, Cagliari, Italy

[3]Inte:Ligand Software-Entwicklungs- und Consulting GmbH, Maria Enzersdorf, Austria

[4]Hebrew University of Jerusalem; Jerusalem, Israel

[5]University Pompeu Fabra; Barcelona, Spain

[6]University of Bari, Bari, Italy

The CancerGrid consortium was formed by ten life sciences companies and academic centers in 2007 to carry out a three-year multidisciplinary research program funded by the European Commission (www.cancergrid.eu). The consortium members work together to develop novel methods to increase the chance of finding potential anticancer agents. Grid-based computing technology is applied to the virtual screening of huge discovery libraries in order to identify promising lead compounds. According to the project plan 30,000 small molecules are selected by various state-of-the-art computational methods, and are then screened in cell-based and target-based assays. This stage will be followed by model development and validation based on the large number of screening data.

In order to discover novel chemotypes for anticancer agents, a multi-step virtual screening procedure was developed and carried out on the initial compound set which includes merged collections from repositories of University of Bari (1,500) and AMRI (199,100) leading to a diverse library (30,000) for biological screening. Forty percent of the compounds were selected against specific cancer targets (HSP90, RET, HDAC and MMP) , or their known, biologically active ligands by using *in silico* similarity and 2D/3D target-based methods [1-4]. Another 50% of the compounds were selected using *Drug Like Index* (DLI) [5] and strict ADME filters [6]. In order to support future works of HTS as well as QSAR model building, a reference set was selected randomly (5%) and a "Trojan horse"-type of counter set (5%) having poor *Drug Like Index* and ADME properties was also included. We present here the generation of the discovery screening library carried out by the various research groups.

CancerGrid is a multinational research project supported by the European Commission under the Framework Program 6 (#LSHC-CT-2006-037559, www.cancergrid.eu).

1. Nicolotti, O.; Miscioscia, T. F.; Leonetti, F.; Muncipinto, G.; Carotti, A. *J. Chem. Inf. Mod.*, **2007,** 47, 2439-48;
2. Langer, T.; Hoffmann, R. D., *Expert Opinion on Drug Discovery* **2006**, 1(3), 261-267;
3. Tovar, A.; Eckert, H.; Bajorath, J. *ChemMedChem*. **2007**, 2, 208-217;

4. Mestres, J.; Martín-Couce, L.; Gregori-Puigjané, E.; Cases, M.; Boyer S. *J. Chem. Inf. Model* **2006**, 46: 2725-2736;
5. Rayan, A.; Marcus, D.; Givaty, O.; Barasch, D.; Goldblum, A.; *Abstracts of Papers of the American Chemical Society* **2005**, 230, U1013 .
6. Fontaine, F., Pastor M., Zamora I., Sanz F. *J Med Chem*. **2005,** 48(7):2687-94.

P-7 : Investigating false predictions in mutagenicity QSAR models: What are we missing?

C. Hasselgren, S. Boyer, AstraZeneca, Mölndal, Sweden

Mutagenicity is the ability of a compound to induce permanent, heritable alterations in DNA sequence. The most common experimental test is the Ames[1] assay which is used as a predictor of carcinogenicity. The use of QSAR models to predict mutagenicity is current practice not only in the pharmaceutical industry but also in the manufacturing of industrial chemical and food additives etc. Numerous QSAR models have been published and report predictivities ranging from 60 to 85% depending on the dataset.

We have previously reported a rules-based system for risk assessment of mutagenicity. This comprises QSAR results, experimentally tested structural near neighbours and the presence of substructural alerts.[2] The overall predictivity of this system based on the QSAR alone was reported to be around 80-85% with sensitivity being slightly lower than specificity.

In this study, we report the temporal validation of this system based on data generated within AstraZeneca after the system was built. The aim is to assess true model predictivity on external data. In addition, work directed at understanding why some compounds are not correctly predicted by our QSAR models is presented. Emphasis was placed on experimentally active compounds which were falsely predicted as inactive, as these are of critical importance to guide experimental testing. Poor sensitivity has previously been discussed in terms of non-covalent interactions and poor structural coverage.[3]

The work includes various methodologies, such as assessing local structural environments in the dataset around the classified compounds, using a variable kNN approach. This was done to identify poor/variable coverage in the dataset. Also, separately modelling subgroups of the dataset, based on the presumed mechanism of interaction with DNA, was explored to investigate if non-covalently interacting agents could be modelled using electrostatic surfaces or pharmacophore/shape based methods.
The temporal validation shows that the models do not perform as well on an external dataset for active compounds but inactive predictions have high confidence. A number of public compounds were also included in the analysis and will be used to illustrate modelling results.

1. Ames, B. N.; McCann, J.; Yamasaki, E., Methods for detecting carcinogens and mutagens with the salmonella/mammalian-microsome mutagenicity test. *Mutat. Res.* **1975,** 31, (6), 347-63.
2. Hasselgren, C.; Carlsson, L.; Boyer, S., A rule-based method for comprehensive risk assessment of the mutagenic potential of drugs. *Manuscript*.
3. Snyder, R. D.; Smith, M. D., Computational prediction of genotoxicity: room for improvement. *Drug Discovery Today* **2005,** 10, (16), 1119-1124.

P-8 : Selecting druglike pieces for the virtual chemistry jigsaw puzzle: towards optimal fragment spaces

Christof Gerlach [1], Jörg Degen [2], Matthias Rarey [2] and Andrea Zaliani [2]

[1] Bayer-Schering Pharma AG, Med Chem VII-Comp. Chemistry, Berlin, Germany

[2] Universität Hamburg, Zentrum für Bioinformatik, Hamburg, Germany

Fragment-based approaches have become very popular within the lead finding phase of a drug design project. Different experimental techniques such as X-ray and NMR-supported protocols have been developed to detect and applied to successfully novel lead structures [1]. In addition, *in silico* approaches considering either descriptor- [2], ligand- [3] or structure-based [4] information for navigating within chemical fragment spaces have been established.

Still the question remains, how does a typical ‘druglike’ fragment looks like and from which source it should be derived?

We present our results from a comparison of two retrosynthetic sets of rules for the generation of fragment spaces. RECAP [5] was checked against our newly developed procedure BRICS (Breaking of Retrosynthetically Interesting Chemical Substructures). Within BRICS the shredding of molecules to fragments tries to mirror retrosynthetical concepts in a more elaborate way considering, for instance, bioisosteric replacements for cyclic and acyclic systems separately and also differentiates between activated and inactivated heterocyclic systems.

For a detailed analysis, fragment spaces from WDI [6] and from the ZINC database [7] were derived by both sets of rules. These datasets were characterized with respect to the number of generated fragments, connection points and size of the fragments. Finally, identical fragments which occur in both datasets have been compiled to generate an optimal fragment space consisting of approximately 5000 fragments.

The performance of these sets (generated by RECAP and BRICS) were evaluated by means of multiple FTree-FS searches using very large and diverse query sets. The BRICS-generated fragment space was able to exactly rebuild more than the double amount of query molecules in comparison with the RECAP-generated fragment space. Thus, the better performing BRICS-generated fragment space have been further enriched with fragments from ZINC having a reasonably high similarity to the WDI fragments. This led to two larger fragment spaces showing further improvements with respect to exact rebuilding of the query

In conclusion, our analysis underlines that the performance of a fragment space derived from ‘druglike’ molecules can be improved, using fragments which are originally derived from vendor catalogues. Thus, it seems that a high-performance set of fragments does not have to be derived solely from databases of drug molecules. Based on our findings three new fragment sets have been compiled, with different optimized performances in retrieving random sets of queries from different sources We also plan to make them publicly available in the near future. These can be used for further fragment-based searches to identify chemical probes for a given protein binding assay.

1. Hajduk, P. J.; Greer, J., A decade of fragment-based drug design: strategic advances and lessons learned. *Nat Rev Drug Discov* **2007,** 6, (3), 211-9.
2. Pärn, J.; Degen, J.; Rarey, M., Exploring fragment spaces under multiple physicochemical constraints. *J Comput Aided Mol Des* **2007,** 21, (6), 327-340.
3. Rarey, M.; Stahl, M., Similarity searching in large combinatorial chemistry spaces. *J Comput Aided Mol Des* **2001,** 15, (6), 497-520.
4. Degen, J.; Rarey, M., FlexNovo: structure-based searching in large fragment spaces. *ChemMedChem* **2006,** 1, (8), 854-68.
5. Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M., RECAP--retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J Chem Inf Comput Sci* **1998,** 38, (3), 511-22.
6. *World Drug Index, Version 2004,* Thomson: Philadelphia PA, 2004.
7. Irwin, J. J.; Shoichet, B. K., ZINC--a free database of commercially available compounds for virtual screening. *J Chem Inf Model* **2005,** 45, (1), 177-82.

P-9 : Going on SARfari in the protein kinase data jungle

J. Günther, Bayer Schering Pharma, Berlin, Germany

Focusing drug discovery on protein families can enable quite a number of synergies between individual projects, provided that appropriate tools are in place to analyse all relevant data in the context of the target family. In particular, interpretation of SAR data in view of the similarities and differences between the individual targets can guide the way to obtaining the right compounds with the desired selectivity profiles. At BSP, we developed an integrated chemogenomics workbench focused on protein kinases, called Kinase SARfari. It incorporates and links kinase sequence, structure, compounds and screening data. Both in-house data and data from the literature, as found in Biofocus DPI’s StARLITe database, are included.

Apart from giving a general overview on the data organisation in the database and highlighting some of the applications, the talk will introduce a novel methodology for picking nearest neighbours to a given target in the family. Identification of the targets with which cross-reactivities are most likely to be expected allows for the early set-up of relevant counter-screens. Moreover, drug-like ligands known to hit a close neighbour provide interesting starting points for a new lead-optimisation program. Accordingly, the approach we chose focuses on similarities in the recognition of small molecule inhibitor structures rather than on phylogenetic relationships between targets. To this end, we exploited the wealth of 3D data that is available for protein kinase targets, while making sure the method is applicable to any set of kinases, independent of whether their 3D structures have been determined or not. The physicochemical distance measure we developed to describe the binding site similarity of two protein kinases has been shown to correlate with the average difference in IC50 values obtained for these targets in our in-house kinase selectivity panel.

1. Chan, E.; StARLITe – A Chemogenomics Knowledge Base. *Abstract book of 7th International Conference on Chemical Structures*; Noordwijkerhout, 2005.

P-10 : A probabilistic approach to classifying metabolic stability

A. ter Laak[1], T Schroeter[2], A. Schwaighofer[2], S. Mika[2], P Lienau[1], A. Reichel[1], Müller, K-R[2], N. Heinrich[1]

[1]Bayer Schering Pharma AG, Müllerstrasse 178, 13342 Berlin, Germany

2Fraunhofer FIRST, Intelligent Data Analysis (IDA), Kekuléstraße 7, 12489 Berlin, Germany

Metabolic stability is an important property for drug candidates. Optimally,it should be taken into account already in the in silico phase of the drug design process. Yet, general purpose predictive tools for this endpoint are inherently difficult to obtain.

We present a machine learning approach to predicting metabolic stability, that is tailored to compounds from the drug development process at Bayer Schering Pharma AG. Our modelling is based on existing measurements of the percentage of each compound remaining after incubation with liver microsomes for 30 minutes. We built independent modelsfor4different species (human, male mouse, female mouse, male rat), with 1000 to 2100 measurements per species. From this data, we develop Bayesian classification models to predict the probability of a compound being metabolically stable.

A particular advantage of the chosen approach is that it implicitly takes the “domain of applicability” into account. For compounds outside the domain of applicability, the model output is 0.5, indicating that it is not possible to tell whether the compound is stable or not.

The developed models were validated on recent project data (200 to 700 compounds, depending on species), showing that the predictions are highly accurate. In particular, we could show that the accuracy of predictions increases when excluding compounds with a predicted probability around 0.5, that is, those that are outside the domain of applicability.

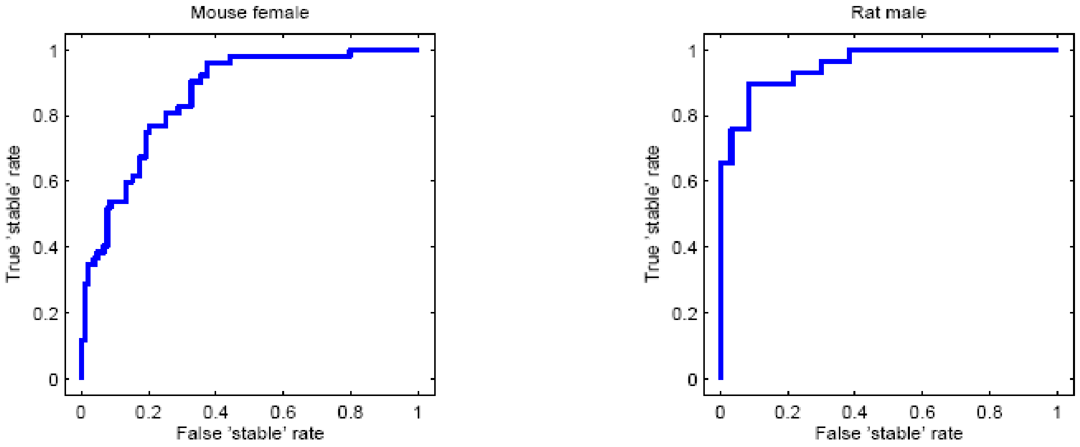


Figure 1. ROC-Curves for validating the developed models on data from recent projects that was not used for model building. Left: Model for female mouse. 258 unambiguous validation measurements, 156 of which are in the domain of applicability. Right: Model for male rat. 183 unambiguous validation measurement, 89 of which are in the domain of applicability.

P-11 : MCS clustering - A hierarchical clustering approach for large data sets

Alexander Böcker , Boehringer Ingelheim (Canada) Ltd., Laval, Canada

A new clustering algorithm has been implemented that is able to group large data sets with more than 500,000 molecules according to their chemotypes. The algorithm pre-clusters a data set using a fingerprint version of the hierarchical *k*-means algorithm.[1,2] Chemotypes are extracted from the terminal clusters using a maximum common substructure (MCS) approach.[3] Molecules forming a chemotype have to share a pre-defined number of rings, atoms and non-carbon heavy atoms. Each chemotype is represented by a MCS. Similar chemotypes and singletons are then fused to larger chemotypes. Singletons that cannot be assigned to any chemotype are finally grouped based on the proportion of overlap between the molecules. Representatives from each chemotype and singletons are used in a second round of the hierarchical *k*-means algorithm to provide a final hierarchical grouping. Results are reported to an interactive graphical user interface which allows preliminary conclusions to be drawn about the structure activity relationship of the molecules. An example application is shown for reverse transcriptase inhibitors in the MDDR database.[4] The algorithm allows the analysis of uHTS and virtual screening results with improved efficiency and quality.

1. Böcker, A.; Derksen, S.; Schmidt, E.; Teckentrup, A.; Schneider, G. A Hierarchical Clustering Approach for Large Compound Libraries. J. Chem. Inf. Model. **2005,** *45*, 807-815.
2. Böcker, A.; Schneider G,; Teckentrup, A.; Schneider, G. NIPALSTREE A new Hierarchical Clustering Approach for Large Compound Libraries and Its Application to Virtual Screening. J. Chem. Inf. Model. **2006,** *46*, 2220-2229.
3. Stahl, M.; Mauser, H.; Database Clustering with a Combination of Fingerprint and Maximum Common Substructure Methods. J. Chem. Inf. Comput. Sci. **2005,** *45*, 542-548.
4. MDL Information System Inc., San Leandro, USA. http://www.mdl.com/

P-12 : Comparison of different approaches for cytochrome P450 modeling

P Czodrowski, C Tautermann, T Fox, Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach an der Riss, Germany

The superfamily of cytochrome P450s (CYP) is involved in the metabolism of drugs in the human body. Understanding this complex process is of utmost relevance for the drug discovery process. Satisfactory results can be obtained by established ligand-based models. However, structure-based models have the advantage that they allow to visually explore crucial interactions to a certain CYP and to support the drug design process. Structure-based approaches have become accessible recently due to the availability of CYP crystal structures. We will show results from ligand-based approaches and compare these with structure-based models and elucidate the differences.

P-13 : Mapping of activity class characteristic substructures extracted from random fragment populations

E. Lounkine, J. Batista, J. Bajorath, University of Bonn, Bonn, Germany

Random molecular fragment populations generated by a novel method termed MolBlaster have been found to contain activity class related information suitable for molecular similarity assessment and virtual screening.[1,2] Based on fragment co-occurrence in different molecules and activity classes, these populations can be organized in hierarchies that revealed subsets characteristic for defined activity classes[3].

These findings led to the question of whether a chemically intuitive, structure-based rationalization of the origin of these activity class characteristic substructures (ACCS) could be found. A simplified approach to extract ACCS from activity class fragment populations was designed based on filtering of activity class fragment populations on the background of inactive molecules.

ACCS were mapped onto active compounds and an atom-based match rate metric was defined yielding a molecular map of ACCS overlap for each individual active molecule. Systematic analysis of more than 1.000 compounds spanning 45 activity classes revealed the formation of molecular core regions of high ACCS overlap. Cores of ten different overlap stringency levels were defined based on atom match rate binning (Figure 1).

Molecular cores were found to delineate well-defined, limited molecular regions that correspond to areas of ACCS origin. Furthermore, using ACCS as structural keys for fingerprinting, molecules belonging to an activity class could be clustered and stringent cores were found to be representative of intra-activity class clusters (Figure 2).

Our results suggest that molecular cores accurately define molecular regions rich in activity class characteristic information[4].

1. Batista, J; Godden, J. W.; Bajorath, J. Assessment of molecular similarity from the analysis of randomly generated structural fragment populations. *J. Chem. Inf. Model.* **2006**, *46*, 1937-1944.
2. Batista, J.; Bajorath, J. Chemical database mining through entropy-based molecular similarity assessment of randomly generated structural fragment populations. *J. Chem. Inf. Model.* **2007**, *47*, 59-68.
3. Batista, J.; Bajorath, J. Mining of randomly generated molecular fragment populations uncovers activity-specific fragment hierarchies. *J. Chem. Inf. Model.* **2007**, *47*, 1405-1413.
4. Lounkine, E.; Batista, J.; Bajorath, J. Mapping of activity-specific fragment pathways isolated from random fragment populations reveals the formation of coherent molecular cores. *J. Chem. Inf. Model.* **2007**, *47*, 2113-2119

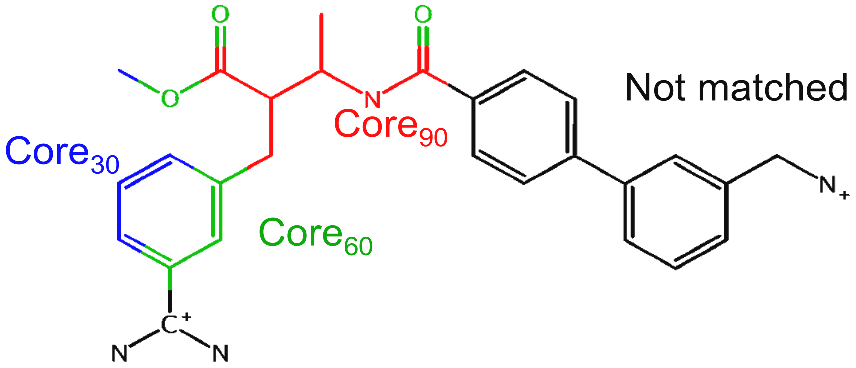


Figure 1. Core definition example. Different core levels can be distinguished based on atom match rates, which reflect ACCS overlap. Core$_{90}$ (red) spans atoms with a match rate >90%, core$_{60}$ (red and green) spans atoms with a match rate >60% and Core$_{30}$ (red, green, blue) atoms with a match rate >30%. Regions that were not matched by ACCS are shown in black.

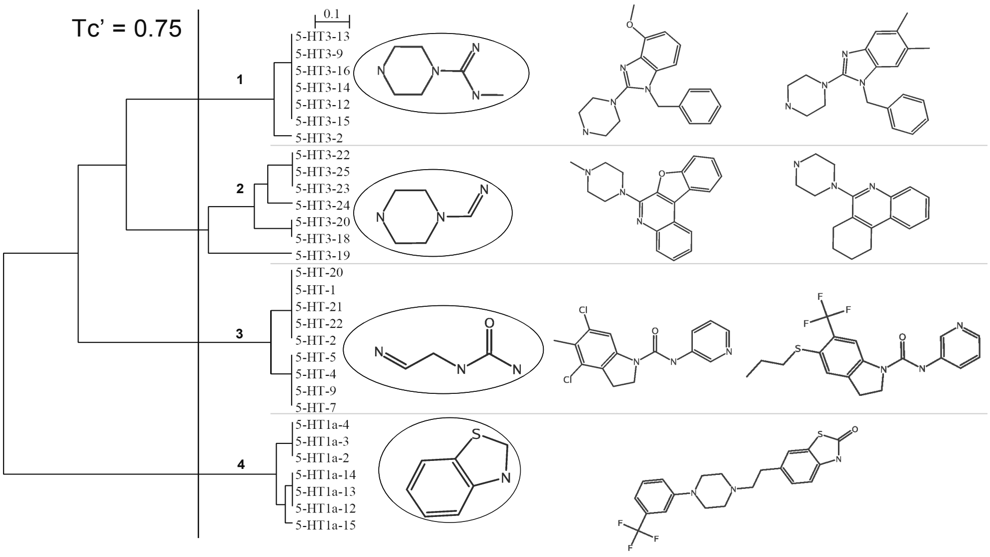


Figure 2. ACCS fingerprint and clustering. For 30 serotonin receptor ligands, ACCS fingerprint based clustering is shown together with representative cores (encircled) and exemplary molecules from each cluster.

P-14 : *In silico* prediction of efflux substrates classification

L. Zhang, P. V. Balimane, S. R. Johnson, S. Chong, Bristol-Myers Squibb, New Jersey, USA.

An in silico Efflux substrate classification model has been developed based on in vitro bidirectional Caco-2 cell permeability measured at 3mM concentration of test compounds, and seven physicochemical descriptors. The model suggests that efflux substrates tend to contain electron deficient aromatic rings, are highly branched, and most contain tertiary nitrogen. This model demonstrated ~80 % predictability of non-substrates and ~84% substrates from a training set of 172 compounds. For a validation set of 70 compounds the predictability was ~75% for non-substrates and ~76% for substrates. The model has the potential to be used as a filter for library designs to identify potential efflux substrates in early discovery.

P-15 : Digging deep for GOLD – How buriedness may be used to discriminate between actives and inactives in docking

N. M. O'Boyle [1], S. C. Brewerton [2], R. Taylor [1]

[1] Cambridge Crystallographic Data Centre, Cambridge, U.K.

[2] Astex Therapeutics, Ltd., Cambridge, U.K.

When docking software is used for virtual screening, the practical problem is to correctly rank the true ligands (actives) with respect to non-binders (inactives). A recent large-scale docking study by Warren et al.[1] has shown that this is still a difficult problem. One reason for the difficulty in ranking actives with respect to inactives is the scoring functions used in docking are typically trained only with positive data, that is, information relating to known binders. Negative data, or information relating to inactive molecules or inactive poses, is rarely included.

We describe the introduction and training of a new term in the scoring function used by the docking software GOLD.[2,3] The new term scales hydrogen bond, metal and lipophilic interactions based on the depth in the active site at which they occur. Depth was measured using the receptor density, the number of protein heavy atoms within 8Å of the point.

Parameters in the new term were optimised using negative data (poses of 99 inactive molecules) derived for the 85 proteins in the Astex Diverse Set (Hartshorn et al.[4]). The resulting scoring function gave a substantial improvement in the average rank of the active molecules (Figure 1).

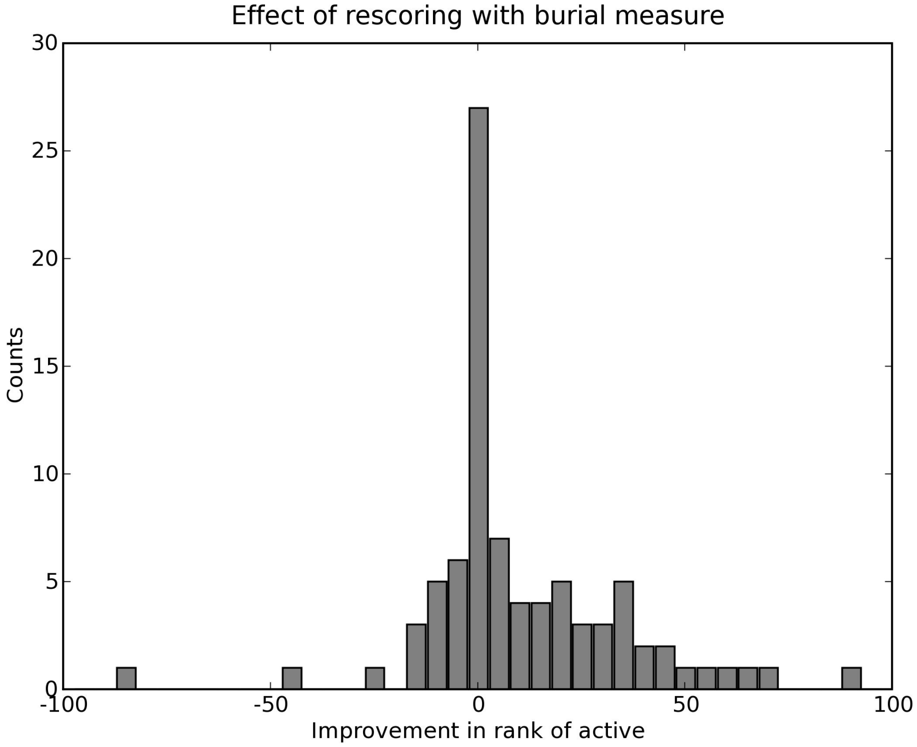


Figure 1 – A histogram of the improvement in the rank of each active when the new term in scoring

function is used. The ranks of each active in the training set were measured relative to 99 inactives.

1. Warren, G. L.; Andrews, C. W.; Capelli, A.-M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A critical assessment of docking programs and scoring functions. *J. Med. Chem.* **2006**, 49, 5912-5931.
2. Jones, G.; Willett, P.; Glen, R. C. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.* **1995**, 245, 43-53.
3. Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, 267, 727-748.
4. Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T. M.; Mortenson, P. N.; Murray, C. W. Diverse, high-quality test set for the validation of protein-ligand docking performance. *J. Med. Chem.* **2007**, 50, 726-741.

P-16 : Representation, searching and enumeration of generic structures – from molecules towards patents

S Csepregi, N Mate, S Dorant, E Biro, T Csizmazia, F Csizmadia, ChemAxon Ltd., Budapest, Hungary

Cheminformatics systems usually focus primarily on handling specific molecules and reactions. However, generic (Markush) structures are also indispensable in various areas, like combinatorial library design or 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) are extended to handle generic features (R-group definitions, atom and bond lists and link nodes). Markush structures can be drawn and visualized in the Marvin sketcher and viewer, registered in the database and their library space is searchable without the enumeration of library members. Different enumeration methods 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 with arbitrary precision.

Patent documents often involve further generic features, for example position and homology variation and bridged R-group definitions. The representation of these features will be discussed, as well as a future extension of the system towards full patent handling.

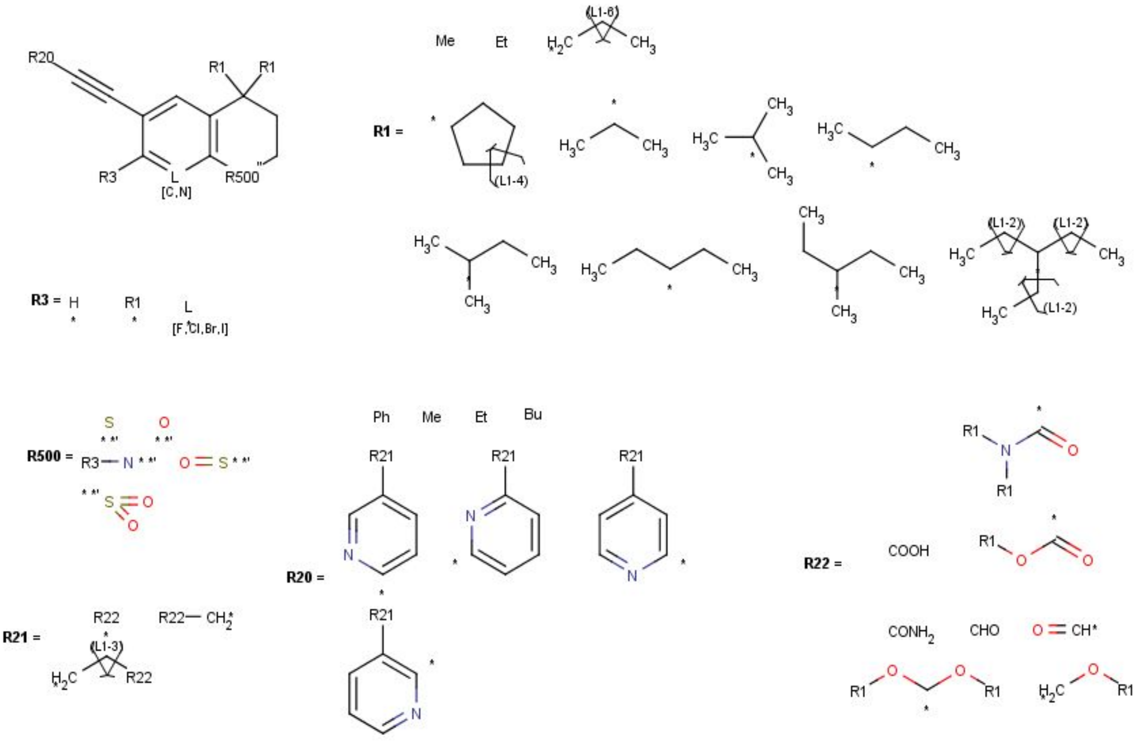


Figure 1. An example Markush structure containing atom lists, R-groups and link nodes. This generic

structure represents $4.1x10^{11}$molecules.

P-17 : Hierarchical clustering of chemical structures by maximum common substructures

M. Vargyas, F. Csizmadia, ChemAxon Ltd., Budapest, Hungary

Cluster analysis has been shown to be successful in the categorization of physico-chemical and biological properties of compounds. However, conventional approaches to clustering molecular structures, where chemical graphs are transformed into sequences of numbers, seldom meet chemists' expectations.

Graph based techniques that cluster compounds with respect to common structural motifs are gaining in popularity as these can better mimic human categorization. One such graph based method, called LibraryMCS, which clusters compounds according to their maximum common substructures (MCS) in a hierarchical manner is presented. Unlike some other graph based clustering methods, LibraryMCS neither involves a similarity based pre-clustering step nor relies on predefined fragments.

Recent evaluation by different research groups indicated that LibraryMCS was capable of producing high quality clusters agreeing with human categorization within practicable time (approximately 1000 structures).

The presentation will recount and demonstrate typical usages of LibraryMCS: virtual HTS hit set profiling, R-group decomposition by learned scaffolds, perception of novel scaffolds, reverse engineering of combinatorial libraries, diversity assessment of large chemical library and compound acquisition.

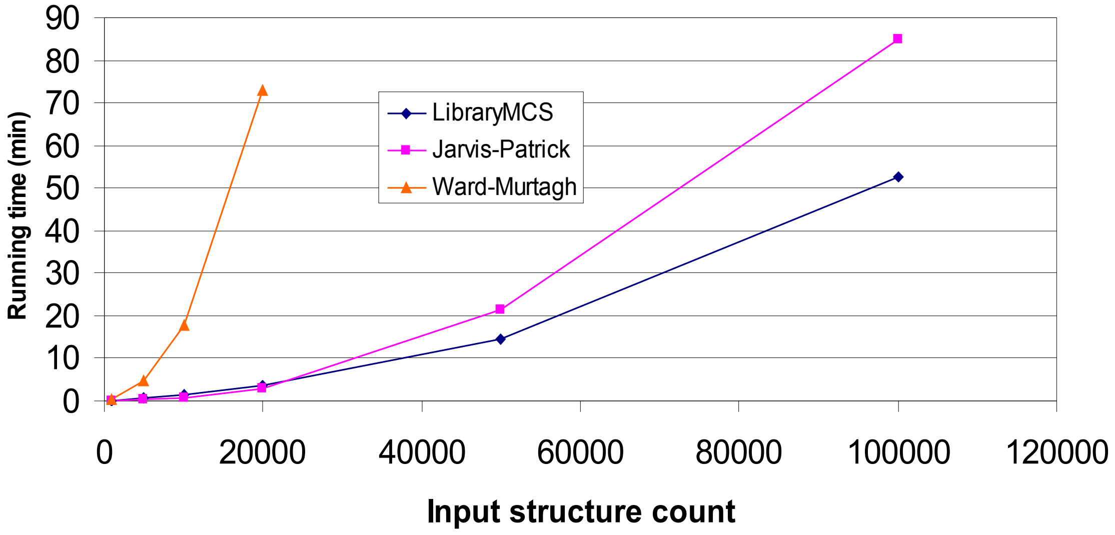


Figure 1. Performance comparison of three clustering methods. Randomly selected subset of druglike molecules from the Zinc database were clustered.

P-18 : Molecular framework based analysis of large chemical spaces

Anthony J. Trippe, Alan H. Lipkus, Chemical Abstracts Service, Columbus, USA

Since first introduced by Bemis and Murcko in 1996[1], molecular frameworks have proven to be a useful method for systematically organizing chemical structures and for analyzing the diversity of large substance collections. This presentation will focus on three analytical studies utilizing molecular frameworks. An analysis of the CAS Registry file has been conducted and a Power Law distribution amongst ringed organic substances has been discovered. A new method for visualizing chemical space utilizing molecular frameworks and a force-directed placement algorithm has been developed and finally a system for interactively exploring chemical space in conjunction with molecular properties including therapeutic bioactivity and biomolecule target association will be presented.

1. Bemis, G. W.; Murcko, M. A.; The Properties of Known Drugs. 1. Molecular Frameworks *J. Med. Chem.*. **1996**, 39, 2887-2893.

P-19 : Towards Automated Searching of data in Internet Chemical Databases

Xiaoxia Li, Xiaolong Yuan, Zengcai Liu, Li Guo, State Key Lab of Multiphase Complex System, Institute of Process Engineering, Chinese Academy of Sciences, Beijing, China

Public accessible chemical databases are valuable resources on Internet. To find the data on Web, first you should be aware if such data is available and where to get it. To search for possible data sources, general search engines like Google as well as comprehensive web directories of chemistry resources with list of chemical databases such as ChIN[1] that indexes more than 200 freely accessible chemical databases can be used.

Because of the diversity of chemicals and their properties, the coverage of compounds and property items varies in different chemical databases. To search data for a chemical, it is very often that one needs to search all possible database web sites one by one manually. The data in chemical databases cannot be indexed and searched by traditional search engines based on hyperlink analysis, because the Web pages containing the targeted data are dynamically generated by the database severs to respond to a query, which does not exist before the query and won't be kept on the server after the query, so cannot be crawled by crawlers of search engines following hyperlinks. Thus the Web databases are collectively called Deep Web,[2] the data collection as a whole in varies chemical databases is called Chemistry Deep Web herein accordingly. To create a searching tool for Chemistry Deep Web may not only overcome the limitation of current search engines in searching data for chemicals on Internet but also to make it possible for data integration from different sources that may be further used in computational applications.

To our knowledge, the ChemFinder of Cambridgesoft[3] is probably the only useful tool that helps searching the Chemistry Deep Web by automatically submitting a data query to different chemical databases.

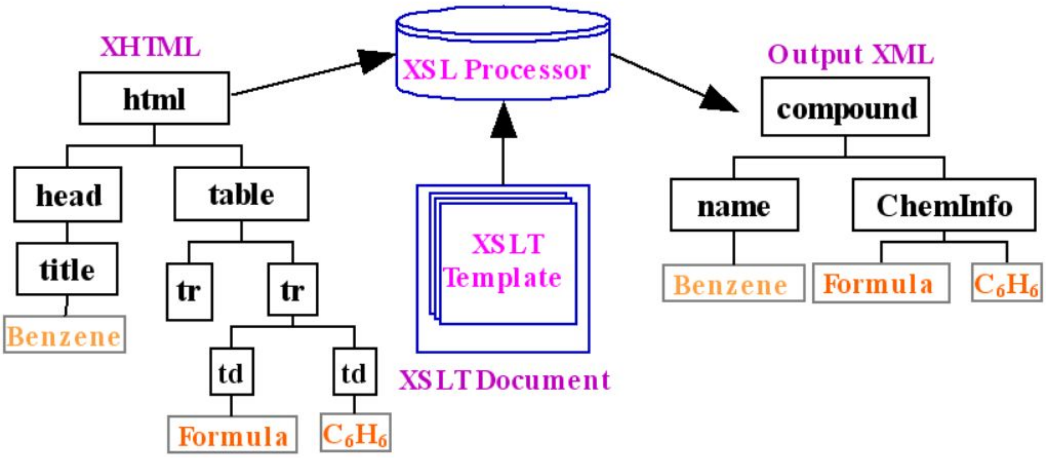


Figure. 1 Chemical data extraction based on XML language in ChemDB Portal

This presentation will report an approach in developing ChemDB Portal that aims at searching the data in various Web-based chemistry databases by one query. ChemDB Portal is implemented by combining HTTP, Java and XML technology.[4,5] In ChemDB Portal, a query is created and submitted to different web based chemical databases on Internet, the HTML documents with the target data returned from these sites are first transformed into XHTML by Tidy, then the target data can be extracted by a data extraction template in XSLT document into a XML document, which can be further mapped into database for XML based retrieval (see Figure 1).

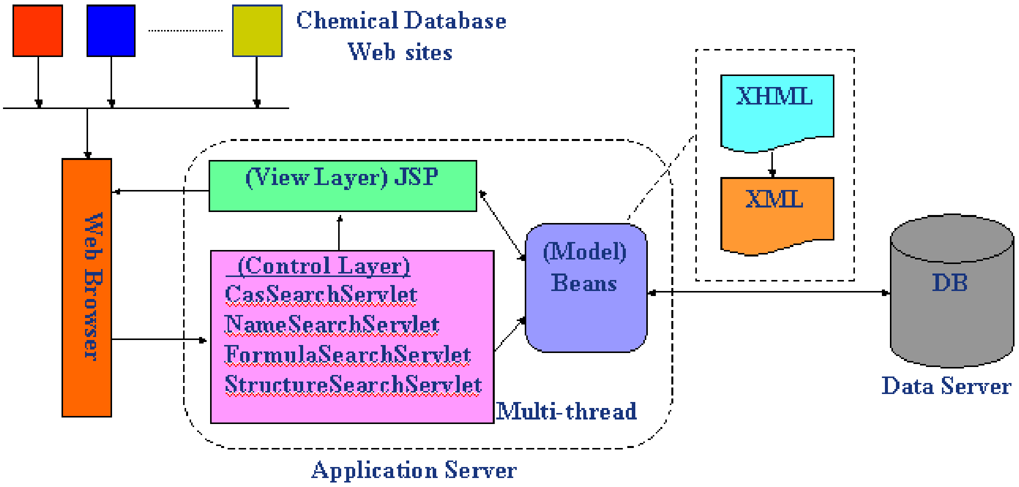


Figure 2. Flow chart of ChemDB Portal system with JSP Model2

How to create a data extraction template for the target data is the key to this XML based approach, which is not only tedious but also a challenging job to create it manually. A semi-automated tool called XE_ChemD that helps create data extraction templates for the chemistry Deep Web has been created. XE_ChemD gets a HTML document by given URL and normalize it to XHTML which is parsed to a XML tree at the same time. After the target data in the source tree are chosen, the candidate XPath expressions that forms the template can be automatically generated based on the context of the target data in terms of their dependence on content, structural, or formatting features.

The data in chemical databases indexed in ChemDB Portal can be searched by a query with identifications of a compound such as CAS registration number, formula, names or structure (see Figure 2). Searching 8 databases simultaneously by one query to ChemDB Portal is now possible that demonstrates its potential to be a search engine dedicated to chemistry data in Deep Web in future.

Acknowledgement: Thanks for the support of National Science Foundation of China (NSFC20673119, NSFC90612015, NSFC20221603).

References

1. ChIN, the Chemistry Portal, Chinese National Science Digital Library (CSDL), http://chin.csdl.ac.cn/ (accessed Feb 4, 2008)
2. Chang, K. C.; He, B.; Li, C.; Patel, M.; Zhang, Z. Structured databases on the web: Observations and implications. *ACM SIGMOD Record*, **2004**, 33, 61-70
3. 3. ChemFinder, CambridgeSoft Corporation, http://www.chemfinder.com/ (accessed Feb 4, 2008)
4. 4. Chu, C.; Li, X.; Guo L. Directed Query Engine Applications in the Integrated Retrieval of Chemical Web Databases, Computers & Applied Chemistry (CN), 2005, 22, 659-666
5. 5. Zhuo, L.; Li, X.; Guo, L. Chemical deep Web data extraction with XML-based technology, Computers & Applied Chemistry (CN), 2**006**, 23, 1137-1141

P-20 : Chemotype bias in virtual screening: the elephant in the room

M Mackey, T Cheeseright, J Melville, S Rose, A Vinter, Cresset BMD, Welwyn Garden City,UK

The analysis of virtual screening methods is complicated by the “chemotype problem”: the sets of actives used are often present as clusters of highly similar molecules. Failure to correct for this can lead to spuriously high enrichment rates being performed. For example, if a large proportion of the actives come from a congeneric series, then a 2D fingerprint-based search using a molecule from that series will appear to perform exceptionally well, while in reality providing little useful information.

Very few literature assessments of virtual screening studies have attempted such a correction. These all err either in the direction of downweighting clustered actives too much[1], or by inadvertently applying a bias which can make random results appear to be highly significant.[1,2,3] In addition, the commonly used metrics used to assess virtual screens (enrichment factors, ROC, BEDROC etc) all have flaws: either they are insensitive to performance on the early part of the retrieval curve or they are oversensitive to the precise number of actives and inactives used.[2] We present a modification of the BEDROC metric called BAROC which is intuitive, concentrates on early performance, is interpretable, and is insensitive to the size of the dataset, and suggest its use as the standard metric for measuring early enrichment in VS studies.

The application of the chemotype correction to BAROC leads to the CC-BAROC metric (Figure 1).

$$CC - BAROC(\alpha) = \sum_{i=1}^{m} w_i \sum_{j=1}^{k_i} e^{-\alpha \beta_{ij}}$$

Figure 1. m is the number of chemotypes, km is number of actives assigned to chemotype i, □ij is the false positive of the jth active in the ith chemotype, and wi is a weighting factor depending on the correlation between the VS scoring method and the clustering method

Applying CC-BAROC to literature data sets[4,5] reveals that ignoring chemotype bias often leads to erroneous conclusions about the efficacy of VS methods, and suggests that many published VS studies greatly overstate the effectiveness of VS techniques.

1. Clark, R. D., Webster-Clark, D. J., *J. Comput. Aided Mol. Des*., in press, DOI 10.1007/s10822-008-9181-z
2. Good, A. C. *et al.*, *J. Mol. Graph. Model.*, **2003**, 22, 31-40
3. Warren, G. L. *et al*, *J. Med. Chem.*, **2006**, 49, 5912-5931
4. N. Huang, B. Shoichet, J.J. Irwin; *J. Med. Chem*. **2006**, 49, 6789-6801
5. Jain, A. N., *J. Comput. Aided Mol. Des*., in press, DOI 10.1007/s10822-007-9151-x

P-21 : Rapid property profiling and similarity calculations in large virtual libraries

J. Barnard, G. Downs, Digital Chemistry Ltd., Leeds, UK

Extremely large virtual compound libraries, containing up to 10^{12} or even 10^{15} different molecules, may be used in drug discovery, and challenge even the fastest virtual high-throughput screening (VHTS) analyses. Even using Markush structure-based property-calculation techniques[1], production of property profiles for simple drug-like characteristics such as the Lipinksi properties can be prohibitively slow. Appropriate sampling of library members can permit much faster analyses, and factors affecting the accuracy of property distribution profiles based on such sampling are discussed. Direct analysis of the Markush structure can also be used to calculate upper and lower bounds on the range of values for a particular property, without the need to enumerate individual values.

Exhaustive comparisons of individual library members against a target molecule, to identify the most similar, are also too time-consuming to be used on very large libraries. An approximate similarity search algorithm has been developed, which allows selection of molecules from a library that are highly similar to a specified target, though not guaranteed to be the most similar. The sets of molecules selected by this algorithm are compared to those identified by exhaustive similarity search.

1. Barnard, J. M.; Downs, G. M.; von Scholley-Pfab, A.; Brown, R.D. Use of Markush structure analysis techniques for descriptor generation and clustering of large combinatorial libraries. J. Mol. Graph. Modelling, 2000, 18, 452-463.

P-22 : Opportunities for Integrating Markush Patent Searching with Drug Discovery

J. Barnard, M. Wright, Digital Chemistry Ltd., Leeds, UK

Information about existing chemical compounds is regularly used in designing chemical libraries for new drug discovery. The information is normally derived from a diverse range of specific structure databases, including internal, bought-in, and public ones, which are managed using Oracle RDBMS data cartridges. Integrated systems incorporate similarity searching, physicochemical property profiling, structure-based cluster analysis etc. across a variety of databases with a common user interface. There is clear potential for utilizing information about the chemical structures in patents in the drug discovery process. This could allow the avoidance of areas of chemistry already covered by patents, identification of un-patented areas of chemical space, and also analysis of the different classes of chemical structure being patented in particular therapeutic areas.1

Though some limited use is made of databases of specific structures from patent claims, searching of the Markush structures currently remains the province of dedicated on-line systems with proprietary databases. Some Oracle cartridges have recently been extended to handle Markush representations of un-enumerated combinatorial libraries alongside discrete molecules. These provide a possible basis for extension to handle the more complex Markush structures found in patents. The issues involved in this extension, both technical and commercial, are discussed.

1. Calcagno, M. An investigation into analyzing patents by chemical structure using Thomson's Derwent World Patent Index codes. World Patent Information 2008, in press: doi: 0.1016/j.wpi.2007.10.007

P-23 : A mathematically more precise taxonomy and nomenclature for polymers: (1) Replacing the heuristics embodied by the words "polymer" and "monomer" with mathematically delineatable terms, (2) Canonically segregating that aggregation of atoms and bonds which forms "monomers" in 3-space

Seymour Benjamin Elk, Elk Technical Associates; New Milford, New Jersey; USA

In order to formulate a useful system of taxonomy and nomenclature that includes all presently known aggregations, as well as allowing for the inclusion of newly discovered chemical entities, one needs a much higher degree of precision than is presently associated with either the word "polymer" or its building block, "monomer". Such a canon must consistently, and unambiguously, prescribe precisely which aggregation of atoms and bonds have the mathematical ideal associated with an infinitely repeating entity in a finite world ─ including the ability to canonically mark the beginning and end of that repeating aggregation. In other words, in lieu of the present heuristics that is conjured up by the use of these two "words", a mathematically more precise taxonomy and nomenclature is presented. Special attention is focused on the premise that the "monomer" be limited to the ideal of congruence of modules in an n-dimensional domain, that is infinite in some or all of these n dimensions. This is augmented by creating an expanded line of demarcation for that nebulous concept of "few" in such a domain, which leads to a concomitant differentiation between refined ideals referred to "oligimers" vs. "oligomers". (In place of these two words, which historically have been absolutely synonymous in common parlance, a precise, different, denotation for each of these terms is promulgated.) A distinguishing feature of the newly proposed definitions is the ability to formulate a consistent system of taxonomy, based on dimension, that, not only canonically assigns an ordered sequence of atoms and bonds which specifies which atoms and bonds (or parts thereof) comprise the monomer, but which also incorporates the orientation of these moieties. The nomenclature developed in a previous monograph ("A New Unifying Biparametric Nomenclature that Spans All of Chemistry", Elsevier 2004) that was used in order to canonically "nomenclate" finite length molecules is expanded and, in selected instances, reformulated so as to retain consistency when applied to unlimited repetition of the monomer. Additionally, the canonical assignment of nomenclature to that aggregation of atoms and bonds which lack the "regularity" to meet the proposed limitation to the definition of "polymer" (therein called "multimer") is likewise expanded from its original definition. The implications of this set of re-definitions is illustrated for important classes of newly formulated "polymers" such as a recently formulated cobalt-polypyrrole composite catalyst, in which polymers that are inherently three-dimensional are depicted as though they were intrinsically coplanar, and for that class of "PIMs" (polymers of intrinsic microporisity).

P-24 : Indirect drug design using MD for flexible structure alignment application to HIV-1 protease inhibitors

Alok Juneja[1], Milan Hodozcek[2], Henning Riedesel[1], Ernst Walter Knapp[1]

[1]*Institute of Chemistry and Biochemistry, Freie Universität Berlin, Berlin, Germany*

[2]*National Institute of Chemistry, Ljubljana, Slovenia*

The rational drug design is one of the major challenges in structural and computational biology. Most of the known theoretical approaches on drug design are based on knowledge of the structures of the biological targets and its active sites where the drug binds. Such approaches use conventional force fields from molecular dynamics simulations as well as empirical force fields derived from data bases containing structures of different drugs bound to different targets where by complex modelling procedures a drug molecule is built into the binding site of the target molecule. The concept of indirect drug design tries to circumvent these hurdles. This computational approach is suitable to model drugs, if the knowledge on their binding sites is absent, assuming that drugs binding in the same pocket have common properties, which can be elucidated by appropriate similarity measures. Hence, by exploring the similarities between drugs that bind to the same target, we may also obtain information on the conformation of the drug in the binding pocket.

We used the maximum OVERLAP method (implemented in CHARMM-version 29a1) that can perform flexible structure alignment (FSA) of drug molecules on the basis of volume, charge and electrostatic potential. As a test case, we have considered 44 co-crystallized structures of HIV-1 protease (HPR) inhibitors that are available in the PDB. FSA has been carried out between randomly generated structures of HPR inhibitors, to superimpose their structures optimally. During this structure optimization, the inhibitors exchange information they have about the binding site and assist each other to attain the conformations that show maximum similarity amongst them. The conformations of inhibitors obtained after FSA are then compared with their respective crystal structures. Apparently, the drug conformations modelled by FSA are close to the conformations that the drugs acquire to fit in the binding site of the crystal structure. By an extensive search of structure alignments we succeeded to assign the inhibitors to 4 different clusters, which is a hallmark of different binding modes in the same binding site. Our clustering approach based on the similarity measure was related to the structure data, where clustering has been done looking at hydrogen bonding pattern between inhibitor and target. Also, here four clusters are observed.

According to our results for the test case of HPR inhibitors, the introduced procedure of indirect drug design looks promising. We demonstrated that it can solve the problem to find the most probable conformation of a drug in the binding pocket without knowledge on the binding site. In a nutshell, this indirect drug design approach will not only assist in designing new drugs but may also be helpful to optimise lead compounds in an efficient way.

P-25 : Optimizing drug classification by feature selection: To bind or not to bind that is the question

Ernst-Walter Knapp, Henning Riedese, Freie Universität Berlin, Berlin, Germany

Typical scenarios in drug design are to start from a database of molecules characterized as being active or inactive with respect to a specific target. The task is to use this information to determine the activity for a set of potential drug molecules. If drug activities are quantified (for instance in terms of association constants), a regression scheme interpolating the activities can be applied. Otherwise classification methods apply. In 2006, *COEPRA* [Comparative Evaluation of Prediction Algorithms] a contest for regression and classification of drugs was created [http://www.coepra.org/]. *COEPRA* works similar as *CASP*, the contest for protein structure prediction. This initiative turned out to be most useful for the research community that develops and applies drug classification algorithms. For the first time a wide spectrum of prediction algorithms was put to a test on equal footing. It allowed learning from success and failures of all participating groups world wide.

In this contribution, we will consider the four classification samples from *COEPRA* and go beyond what we have applied there. Each of the four samples consists of a training (learning) and a test (prediction) set of nonapeptides, which are described by two different types of feature vectors: 180 component sequence based vectors and 5787 component chemoinformatic fingerprint based vectors. We use an approach based on a scoring function, which can be employed for classification and regression. The scoring function is linear in parameter space but can have linear and quadratic terms in feature space. Thus, it defines a quadratic hyper-surface in feature space, which for classification serves as separatrix and for regression as interpolating regression surface.

Since the scoring function is linear in the unknown parameters the support vector machine, or a least square optimization approach can be used to determine the parameters. The later leads to a linear equation system, whose coefficient matrix elements are constituted by pair correlations of the different features. In preliminary work, this approach was used to classify peptide binding to the major histocompatibility complex [1].

Often the feature space generated by chemoinformatic fingerprint methods is much too large. Some negative consequences are (i) too many parameters leading to over-fitting, (ii) poor performance using suboptimal or even conflicting features, and (iii) unnecessary long computation times. Therefore, we performed feature selection with a genetic algorithm resulting in very small optimized feature sets. For the four *COEPRA* examples there are considerably less then ten optimized features necessary yielding performances, which are equal or better than achieved by the best contributor.

To optimize the performance of prediction furthermore, we investigated the similarity of each individual molecule from the test set with the molecules of the training set and derived weights for each pair of molecules, which are used for the learning procedure. In this way, we use a prediction scheme, which is tuned for each molecule of the prediction set individually resulting in additional improvements of the prediction performance.

1. Riedesel, H.; B. Kolbeck, B.; Schmetzer, O.; Knapp, E.W. Peptide Binding at Class I Major Histocompatibility Complex Scored with Linear Functions and Support Vector Machines, *Genome Informatics* **2004**, 15, 198–212.

P-26 : Understanding Selective CDK4 Inhibition Through Molecular Dynamics

N.M. Mascarenhas, N. Ghoshal, Indian Institute of Chemical Biology, Kolkata, India

Objective: CDK4 is a bonafide cancer target [1] and understanding the critical protein-ligand interactions for developing potential inhibitors is of paramount importance. We were interested in understanding the difference between the modes of inhibition of two very similar ligands and bring to light with the aid of molecular dynamics the differences in protein-ligand interactions responsible for selectivity.

Scheme: Two ligands, lig16, displaying equal potency towards CDK2 and CDK4, and lig17 exhibiting selective inhibition towards CDK4 [2] were considered (Figure 1, Table 1). A homology model of CDK4 was constructed and both ligands were docked into the active site using GOLD (Table 2). The binding site was then immersed in a 30Å water sphere from the center of the ligand to mimic the aqueous environment. Amino acid residues within 15 Å away from the center of the ligand were kept flexible during the entire simulation. The simulations were performed using CFF91forcefield of Discover in InsightII [3] for a duration of 1 ns.

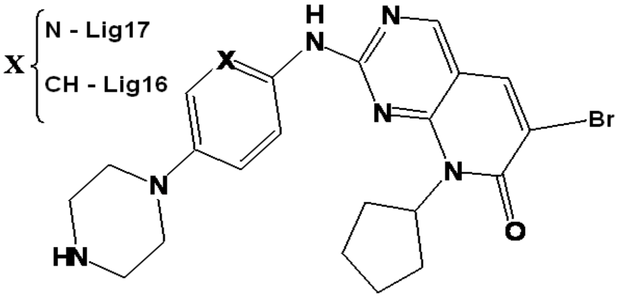


Figure 1. Structures of lig16 and lig17

Table 1. Biological activity of lig16 and lig17.

	IC$_{50}$ (nM)	
	CDK4	CDK2
Lig16	2	43
Lig17	16	>5000

Results: (i) A good homology model of CDK4 was constructed (90.1 % residues in core region Figure 2) (ii) Lig16 and lig17 showed different mode of inhibition within ATP pocket of CDK4 and CDK2. (iii) Residue His95 was found to be critical in offering selectivity. (iv) Both ligands (Lig16 and Lig17) were stabilized by strong Columbic interactions with CDK4.

Table 2. Results of GOLD docking

	CDK4	
	Lig16	**Lig17**
Chemscore	33.939	33.549
H-Bonding **[Residue (no of H-bonds)]**	V96(2), T102(1)	H95(1),V96(2) ,

	T102(1)	

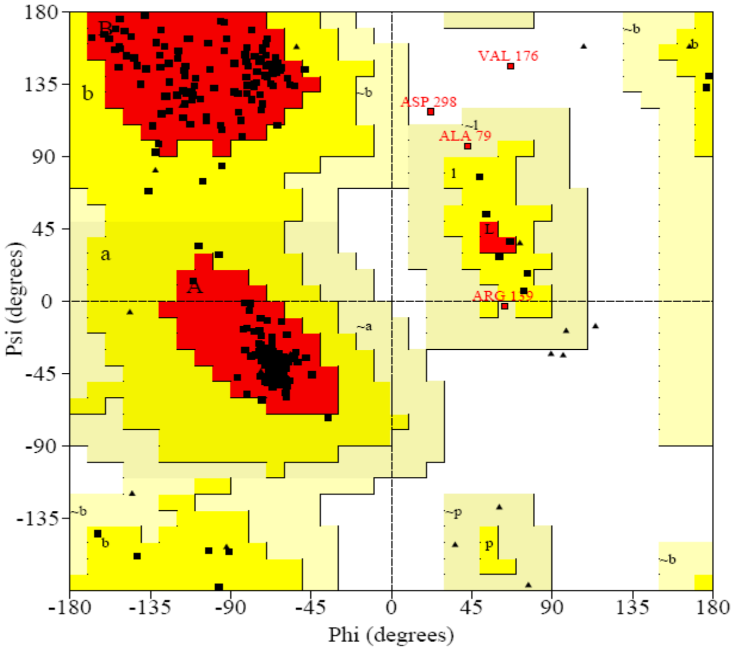


Figure 2. Phi-Psi plots of the modeled CDK4 obtained by Procheck. Two residues in the generously allowed regions Arg159 & Ala79 and two in the disallowed regions Asp298 & Val176. These residues lie far away from the active site.

References

1. Landis, M.W., Pawlyk, B.S., Li, T., Sicinski, P. and Hinds, P.W., Cyclin D1-dependent kinase activity in murine development and mammary tumorigenesis. Cancer Cell, 2006, 9, 13-22.
2. Toogood, P.L, Harvey, P.J., Harvey, P.J., Repine, J.T., Sheehan, D.J., VanderWel, S.N., Zhou, Hairong, Keller, P.R., McNamara, D.J., Sherry, D., Zhu, T., Brodfuehrer, J., Choi, C., Barvian, M.R., Fry, D.W., Discovery of a potent and selective inhibitor of cyclin-dependent kinase 4/6. J. Med. Chem., 2005, 48, 2388-2406.
3. A product of Accelrys Inc.; http://www.accelrys.com.

P-27 : Extracting chemical CYP proteins interactions from literature using natural language processing methods

D Jiao, D Wild, School of Informatics, Indiana University at Bloomington, Bloomington USA

This poster describes the development of an information extraction system which maps interactions between chemicals and CYP proteins from existing literature, using machine learning and natural language processing methods. The interaction between CYP proteins and chemicals is important in drug discovery and development. In this system, abstracts from articles related to CYP and chemical interactions are preprocessed using named entity recognition methods to identify chemical names and CYP names, with the help of dictionaries generated from biological and chemical ontologies. Chemical structures are also attached to chemical names for future processing. The texts are then parsed by a syntactic parser to create a dependency graph in which grammatical relationships between constituents of the sentences are generated. Then interactions between CYP and chemicals are extracted by identifying certain keywords, together with the protein and chemical names based on the dependency graph. The extracted information, including the chemical compounds, their structures, the proteins, and the interactions between chemicals and proteins are stored in a database for retrieval and further analysis. In this poster, the training process to build certain components of the system, problems encountered during the system creation, and the evaluation of the system will be discussed in detail.

References:

1. Corbett, P.; Murray-Rust, P. High-Throughput Identification of Chemistry in Life Science Texts.

Computational Life Sciences II. **2006**, 107-118.

2. Clegg, A. B.; Shepherd, A. J. Benchmarking natural-language parsers for biological applications using dependency graphs. *BMC Bioinformatics* **2007** 8, 24+.
3. Feng, C.; Yamashita, F.; Hashida, M. Automated Extraction of Information from the Literature on Chemical-CYP3A4 Interactions. *J Chem Inf Model* **2007** 47 (6), 2449 -2455.
4. Kulick, S.; Bies, A.; Liberman, M.; Mandel, M. ; McDonald, R.; Palmer, M.; Schein, A.; Ungar, L.. *Integrated Annotation for Biomedical Information Extraction.* HLT/NAACL, 2004.

P-28 : An infrastructure for data mining public chemical & biological information

David J. Wild, Rajarshi Guha, Indiana University School of Informatics, Bloomington, Indiana, USA

At Indiana University School of Informatics, we are exploring several projects aimed at increasing the accessibility and usability of the increasing volumes of chemical and biological information that are readily available on the web, in public databases, and in accessible documents. In this presentation, we will describe our progress with and results for each of the projects and we will detail our vision for how these projects might together enable better data mining of public information in drug discovery projects. The projects that will be described include: public information access using aggregate compound information web services; automatic generation of workflows of web services using semantic web techniques; calculation of chemical similarity between web documents; mining of information in chemistry journal articles; and network models to integrate compound, protein and assay information.

P-29 : Binding affinity prediction of distinct inhibitors of group-1 and group-2 Neuraminidases (NAs): ArgusLab4/AScore protocol

*M. L. Mihajlovic [1, 2], P. M. Mitrasinovic [2, *]*

[1] Faculty of Physical Chemistry, University of Belgrade, Belgrade, Serbia

[2] Institute for Multidisciplinary Research, Belgrade, Serbia

** Correspondence to: petar.mitrasinovic@cms.bg.ac.yu*

Due to a recent pandemic threat by the worldwide spread of H5N1 avian influenza, the World Health Organization has shown its profound concerns regarding the possibility of having the virus spread among humans soon. Reports about the virus resistance to two approved anti-influenza drugs, oseltamivir (Tamiflu) and zanamivir (Relenza), as well as the lack of adequate vaccine have raised the urgent question of developing new anti-viral drugs (1).

The design of new NA inhibitors is a dynamic field of research. Because of often insufficient structural infromation that may be exploited for structure-based design, computational methods are currently a viable partner with experiment (2).

By using the crystal structures of known inhibitors bound to group-1 and group-2 NAs, the ArgusLab4/AScore/ShapeDock (GaDock)-docking protocol (3) was shown to indentify the correct binding modes of all inhibitors in their own protein/ligand crystal structures. In order to study the dependence of binding mode prediction on small changes in protein crystal structure, 132 experiments (11 inhibitors docked in 12 protein structures) for group-2 NAs and 88 docking experiments (8 inhibitors docked in 11 protein structures) for group-1 NAs were performed. In a total of 132 docking experiments, ArgusLab4/AScore/ShapeDock (GaDock) identified the corect binding modes of 116 protein/ligand complexes of group-2 NAs. This outcome shows that ArgusLab4/AScore/ShapeDock (GaDock) significantly outperforms the Dock4/PMF approach (4). In addition, 80 binding modes of 88 protein/ligand complexes of group-1 NAs were correctly identified.

Our study suggests that the ArgusLab4/AScore/ShapeDock docking protocol, as a very consistent and reproducible algorithm, can be employed for the determination of reliable binding affinities of NA inhibitors, thus providing a promising base for the future design of novel anti-viral drugs.

1. Russell, R. J.; Haire, L. F.; Stevens, D. J.; Collins, P. J.; Lin, Y. P.; Blackburn, G. M.; Hay, A. J.; Gamblin, S. J.; Skehel J. J. The Structure of H5N1 Avian Influenza Neuraminidase Suggests New Opportunities for Drug Design *Nature* **2006**, 443, 45-49.
2. Sangma, C.; Hannongbua, S. Structural Information and Computational Methods Used in Design

of Neuraminidase Inhibitors *Current Computer-Aided Drug Design* **2007**, 2, 113-132.

3. Thompson, M. A. ArgusLab 4.0.1. Planaria Software LLC, Seattle, WA, http://www.arguslab.com.
4. Muegge, I. The Effect of Small Changes in Protein Structure on Predicted Binding Modes of known Inhibitors of Influenza Virus Neuraminidase: PMF Scoring in Dock4 *Med. Chem. Res.* **1999**, 9, 490-500.

P-30 : Prediction of novel drug targets in the metazoan parasite schistosoma mansoni

F. Oellien[1], C. R. Caffrey[2], A. Rohwer[1], R. J. Marhöfer[1], S. Braschi[2], G. Oliveira[3], J. H. McKerrow[2], P. M. Selzer[1]

[1]Intervet Innovation GmbH, Schwabenheim, Germany

[2]Sandler Center for Basic Research in Parasitics, University of California San Francisco, San Francisco, USA

[3]Laboratory of Cellular and Molecular Parasitology,Centro de Pesquisas René Rachou, Belo Horizonte, Brazil

Schistosomiasis is a prevalent and chronic helmintic disease in tropical regions. Treatment and control relies on chemotherapy with just one drug, praziquantel and this reliance is of concern should clinically relevant drug resistance emerge and spread. Therefore, to identify potential target proteins for new avenues of drug discovery we have taken a comparative chemogenomics approach utilizing the putative proteome of *Schistosoma mansoni* as compared to the proteomes of two model organisms, the nematode, *Caenorhabditis elegans* and the fruitfly, *Drosophila melanogaster.* Using the genome comparison software Genlight, two separate *in silico* workflows were implemented to derive a set of parasite proteins for which gene disruption of the orthologs in both the model organisms yielded deleterious phenotypes (e.g., lethal, impairment of motility), *i.e.,* are essential genes/proteins.

Finally 35 *S. mansoni* proteins were identified for which druggable protein homologs exist in the literature and 18 of these were homologous to proteins with 3D structures including co-crystallized ligands with which structure-based drug design approaches can be prosecuted.

P-31 : Performance of different machine learning methods

U. Koch[1], S.A.H. Spieser[1], K. Eitner[1,2], D. Plewczynski[3]

[1] IRBM (Merck Research Laboratories Rome), Rome, Italy

[2] Adam Mickiewicz University, Faculty of Chemistry, Poznań, Poland

[3] Interdisciplinary Centre for Mathematical and Computational Modeling, University of Warsaw, Warsaw, Poland

The performance of different classification methods in the context of ligand based virtual screening is evaluated[1,2]. Five different biological targets were chosen for this comparison. It will be shown that some methods perform particularly well in avoiding false negatives, others in avoiding false positives. Application to the analysis of patents and ligand based virtual screening will be demonstrated. The performance of these methods in terms of finding new scaffolds and the effect of combining lists obtained from different methods will be discussed.

1. Plewczynski, D.; Spieser, S.A.; Koch, U. Assessing different classification methods for virtual screening *J. Chem. Inf. Model.* **2006,** 46,1098-106.
2. Plewczynski, D.; von Grotthuss, M.; Spieser, S.A.; Rychlewski, L.; Wyrwicz, L.S.; Ginalski, K.; Koch U. Target specific compound identification using a support vector machine *Comb. Chem. High Throughput Screen.* **2007**,10, 189-96

P-32 : Assessing and exploiting non-additivity in a structure-activity relationship

J. van Drie, John H Van Drie Research LLC, Andover, MA, USA

Virtually all of our analyses of structure-activity-relationships make an implicit assumption of additivity, e.g. if adding an acetyl to template X produces a 10-fold improvement in affinity, then adding an acetyl to template Y should produce a similar 10-fold improvement in affinity. This additivity assumption is not only deeply buried in the intuitive approaches of medicinal chemists, it is also implicit in many of the modeling approaches we use, most prominently in the calculation of statistical potentials, and in docking and scoring.

Ken Dill1 has noted that the assumption of additivity is sometimes called the “4th law of thermodynamics”, and suggests that this assumption may underlie our inability to accurately predict binding affinities, etc.

We present a simple method for assessing the degree of additivity in an SAR, and show that some SAR’s are highly additive, some are not. Our computational methods tend to work quite well when things are additive, but non-additivity poses special challenges to drug design. We will show how to exploit non-additivity, and will explore the implications of non-additivity to all aspects of computer-aided drug discovery. The SAR of hepatitis C viral protease will be highlighted as a prominent example of one displaying a high degree of non-additivity; how non-additivity was exploited in the discovery of clinical candidates targeting HCVP will be described2.

1. Dill, K.A. Additivity principles in biochemistry *J Biol. Chem.,* **1997**, 272:701.
2. Perni, R.B. *et al*., Inhibitors of hepatitis C virus NS3.4A protease. Part 3: P2 proline variants. *Bioorg Med Chem Lett*. **2004** 14:1939.

P-33 : CLiDE Pro: A chemical OCR tool

A. Valko [1], P. Johnson [2]

[1] Keymodule Ltd., Leeds, UK

[2] University of Leeds, Leeds, UK

Depictions of two-dimensional chemical structures published in the literature are stored as bitmap images in most electronic sources of chemical information such as reports, journals and patents. Although the original chemical structures are usually created using chemical drawing programs which generate complete structural information, this information is lost during the publication process and if required, is normally regenerated by redrawing the structure with a computer program, which is time-consuming and prone to errors.

CLiDE Pro is a chemical OCR software tool aimed at automatic extraction of chemical information from either the printed chemistry literature, or from the equivalent electronic pdf version. CLiDE Pro is the latest incarnation of software to emerge from the long-term CLiDE (Chemical Literature Data Extraction) project [1-3]. Chemical OCR involves three main problems: (a) identification of chemical images within a document, (b) compilation of chemical graphs of individual molecules from chemical images, and (c) interpretation of complex objects such as generic molecules and reaction schemes using the retrieved chemical graphs. The structure recognition methods implemented in CLiDE Pro will be presented. Structure features which frequently cause problems such as crossing bonds, lines found in various chemical entities such as single bonds attached to triple bonds, dashed bonds and parts of atom labels commonly misclassified as lines (e.g. I and Cl) will be discussed together with our solutions to these problems. A key component of the presentation will be CLiDE Pro's approach to the interpretation of generic structures.

The chemical OCR tool which has 100% accuracy in all situations has yet to be developed, and indeed is unlikely to be developed in the foreseeable future. This exactly parallels the situation for text OCR, where despite decades of research, accuracy of recognition still falls a little short of 100% and requires some manual editing, but is still very useful. If chemical OCR can reach similar levels of accuracy then automated mining of the chemical literature will become a powerful and cost-effective process.

1. Ibison, P.; Jacquot, M.; Kam, F.; Neville, A. G.; Simpson, R. W.; Tonnelier, C.; Venczel, T; Johnson, A. P. Chemical Literature Data Extraction: The CLiDE Project. *J. Chem. Inf. Comput. Sci.* **1993**, 33(3), 338-344.
2. Ibison, P.; Kam, F.; Simpson, R. W.; Tonnelier, C.; Venczel, T; Johnson, A. P. Chemical Structure Recognition and Generic Text in the CLiDE Project. In *Proceedings on Online Information 92*; London, England, 1992.
3. Simon, A.; Johnson, A. P. Recent Advances in the CLiDE Project: Logical Layout Analysis of Chemical Documents. *J. Chem. Inf. Comput. Sci.* **1997**, 37(1), 109-116.

P-34 : Molecular subgraph mining for analyzing ligand activity classes

Julio E. Peironcely[1], *Eelke van der Horst*[1], *Adriaan P. IJzerman*[1], *Michael Emmerich*[2], *Andreas Bender*[1]

[1] *Medicinal Chemistry, Leiden/Amsterdam Center for Drug Research, Leiden, The Netherlands*

[2] *Leiden Institute for Advanced Computer Science, University of Leiden, The Netherlands*

Protein sequence or overall structural similarities are often employed to categorize the similarity of receptors, but this approach might not be ideal from the chemist's perspective. It can happen that chemically similar ligands interact with proteins without any obvious sequence similarity. Relating receptors by the similarity of their ligands can provide relationships that may be missed if we only study the sequence of the targets. In this study we grouped targets by finding frequent substructures in their ligands employing different graph mining approaches. Knowing these frequencies allowed us to discover substructures that are useful to effectively separate two families, i.e. when they are frequent in one family and infrequent in the others. We used these frequencies to build a phylogenetic tree to visualize the distance at which the target families are related according to the similarities of their ligands. We analyzed activity classes that are similar from the ligand-side, despite having a small sequence similarity, and assume these similarities to be relevant in the context of drug side effect predictions. On the other hand, our study provides tools to detect which fragments increase the specificity of a ligand, reducing promiscuity and off-target interactions.

P-35 : Frequent substructure mining of GPCR ligands

E. van der Horst, A. Bender, A. Ijzerman, Division of Medicinal Chemistry, Leiden-Amsterdam Center for Drug Research, Leiden University, Leiden, The Netherlands

In this study, we conducted frequent substructure mining to find the structural features that discriminate between ligands that either do or do not bind to G protein-coupled receptors (GPCRs). Finding which substructures are rare and which are common in GPCR ligands will help in the design of new ligands and for prioritizing compounds for screening. Besides the normal 2D structure notation, three other chemical representations were used. The first 'elaborate' representation used a special type for aromatic bonds, the second also added a special type for any aromatic atom, and the third representation used a special notation for planar, not necessarily aromatic, structures. In all but the normal representation, wildcards were used for halogens and aliphatic heteroatoms with an extra label indicating the atom type. A set of 16k GPCR ligands was compared against a roughly equal number from a screening set of compounds (Chembridge). For analysis of the results, two decision trees were constructed, one for the most-common substructure for GPCR ligands and one for the most-common substructure in the screening set. The alkylamine substructures were most discriminating for GPCR ligands as compared to the Chembridge set. This reflects the presence of aminergic receptor ligands in the GPCR dataset. Carboxamide substructures were most common in the Chembridge dataset. This is probably due to particular reaction types used to construct the screening library. The 'normal'representation mode led to the most significant substructure for GPCR ligands; the aromatic bonds representation yielded the most significant substructure for the screening compounds. In conclusion, frequent substructure mining is a useful approach for characterizing heterogeneous ligand datasets.

P-36 : Characterization of the inhibition of HIV-1 reverse transcriptase by non-nucleoside inhibitors and proteochemometric models which are able to predict compound activity against particular target mutants

Gerard van Westen[a], *J. Wegner*[b], *A. Bender*[a], *A. IJzerman*[a], *H van Vlijmen*[b]

[a] *Leiden / Amsterdam Center for Drug Research, Leiden University, Netherlands*

[b] *Tibotec BVBA, Mechelen, Belgium*

HIV-RT has traditionally been a valuable target in anti-HIV drug design. NNRTIs are a class of inhibitors very specific for HIV-1 RT, however (cross) resistance forms an increasing problem in the treatment of HIV patients with these drugs.[1] In order to minimize the onset of resistance and design drugs that retain activity even in the presence of several mutations, a detailed understanding of the interaction mechanism of NNRTIs and the mutations leading to resistance is required.

Presented here is a characterization of the NNRTI binding pocket on RT and the mechanism leading to inhibition of the polymerase protein. It was found that the disturbed catalytic site theory [2] is the most likely cause of polymerase function inhibition. Furthermore, using several approaches, the entry channel for NNRTIs into the binding pocket was considered. It was found that there is a high probability that the entry channel for NNRTIs is not formed by the solvent accessible channel[3] but by the movement of one of the two sheets making up the binding pocket.[2, 3]

Using a group of 36 crystal structures, a field-based 5D QSAR process was implemented including modeling of mutant structures, conformational analysis, induced fit minimization, and calculation of interaction energies and consensus interaction fields. This approach allowed characterization of NNRTI resistance and prediction of the effect of point mutations on the interaction between NNRTIs and RT. The interaction energies and consensus interaction fields support lead-optimization chemistry. Furthermore it allows rationalization and confirmation of resistance hypotheses in literature. In total, 432 different structures divided over 18 drug classes were analyzed.

Finally a simple proteochemometric model was created based on an internal data set containing interaction data between a series of 455 compounds and 6 different mutants of HIV-1 RT. Using several classification techniques, the performance varied between 60 and 80 % correct prediction. Two models were built, predicting interaction between the compounds and the entire mutant panel and predicting interaction of all compounds within each sequence.

We conclude that the applied SAR approaches confirm to a large degree the known resistance patterns in literature, but are less effective for mutations causing backbone movements. Especially these cases will be refined in future research to help understand NNRTI resistance mechanisms.

1. Sallie, R., Replicative homeostasis: a fundamental mechanism mediating selective viral replication and escape mutation. *Virol J* **2005,** 2, 10.
2. Balzarini, J., Current status of the non-nucleoside reverse transcriptase inhibitors of human immunodeficiency virus type 1. *Curr Top Med Chem* **2004,** 4, (9), 921-944.
3. Rodriguez-Barrios, F.; Balzarini, J.; Gago, F., The molecular basis of resilience to the effect of the Lys103Asn mutation in non-nucleoside HIV-1 reverse transcriptase inhibitors studied by targeted molecular dynamics simulations. *J Am Chem Soc* **2005,** 127, (20), 7570-7578.

P-37 : Consensus modeling of chemical biodegradation pathways

ML Patel, MD Hobbs, PN Judson, MA Ott, M Ulyatt, JD Vessey, Lhasa Limited, Leeds, United Kingdom

As part of the NoMiracle integrated European research project exploring novel methods and tools to better evaluate chemical risks, we investigated whether a consensus model approach could be developed to advise on biodegradation pathways by reasoning about the results from different *in silico* prediction systems and databases.

A working prototype ‘Mira’ has been created that is able to query two biodegradation prediction systems (CATABOL and MEPPS) and a database of known biodegradation pathways (included in CATABOL) about a compound of interest and reason about the results returned. Mira is able to provide an assessment of both the biodegradability of a chemical and the structure of its biodegradants. Each prediction is associated with an overall level of likelihood and is also assigned a level of confidence based on the user’s confidence in each of the underlying systems.

The reasoning methodology used within Mira addresses a number of challenges common to consensus modeling, including approaches to the combination of qualitative and quantitative outcomes. In addition, consideration is also given to varying confidence in, and the extent of concordance between, different prediction systems.

P-38 : Scaffold Hunter: Exploiting holes in chemical space

S. Wetzel[1], K. Klein[2], A. Schuffenhauer[3], P. Ertl[3], P. Mutzel[2], H. Waldmann[1]

[1]Max-Planck-Institute for Molecular Physiology, Dortmund and Chemical Biology, Technical University of Dortmund, Germany

[2] Chair of Algorithm Engineering (Ls11), Department of Computer Science, Technical University Dortmund, Germany

[3] Novartis Institute for Biomedical Research, Basel, Switzerland

“Space”, as Douglas Adams famously said “*is big. You just won't believe how vastly, hugely, mind-*

bogglingly big it is."[1] So navigating chemical space is quite an effort and still everyone seeks "*To boldly go where no man has gone before.*"[2] How can one know where the promising parts of chemcial space are where no man has gone before?

We developed a hierarchical scaffold classification strategy3,4 to chart chemical spaces. Our approach is based on Murcko scaffolds and a rule set which will create a scaffold tree. The resulting tree diagrams where the nodes represent scaffold structures are very intuitive to chemists. In general they allow a quick orientation in the structure space depicted. Features like colour coding according to properties and the like further increase the information content without reducing clarity.

The Scaffold Hunter program interactively displays the tree diagrams generated by the scaffold tree algorithm and thus facilitates navigation in and exploitation of chemical space. It allows to quickly identify "holes" in the structure space analyzed and to link them to bioactivity. Thus promising starting points for library design can be easily identified.

Scaffold Hunter will enable chemists to directly work with the results of hierarchical classifications in an intuitive and easy way - independent of the underlying algorithm. It is in use for hitlist triaging and compound management but can also serve as a tool to navigate and analyze chemical structure space.

1. Adams, D. Hitchhikers Guide to the Galaxy, Del Rey, Reissue Edition **1995**.
2. From the title theme of the original StarTrek science fiction television series
3. Koch, M.A.; Schuffenhauer, A.; Scheck, M.; Wetzel, S.; Casaulta, M.; Odermatt, A.; Ertl, P.; Waldmann, H. Charting Biologically relevant chemical space: A structural classification of natural products (SCONP), *PNAS* **2005**, 102, 17272-17277.
4. Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M.A.; Waldmann, H. The Scaffold Tree - Visualization of the Scaffold Universe by Hierarchical Scaffold Classification. *J. Chem. Inf. Model.* **2007**, 47, 47-58.

P-39 : Dynamic web application for drug design research

J. MacCuish[1], M. Chapman[1], N. MacCuish[1], J. Bradley[2], J. Blankley[3]

[1] Mesa Analytics & Computing, LLC, Santa Fe, NM, USA

[2] Drexel University, Philadelphia, PA, USA

[3]Pfizer (retired), Ann Arbor, MI, USA

WebFlowNets, a development framework for the construction of dynamic web applications, is presented in the context of eLearning and research. The dynamic Web application, WebFlowDD, built with WebFlowNets, will be described. WebFlowDD provides researchers with workflows in drug design utilizing a variety of freeware and Web-services. WebFlowDD demonstrates the broader implications of the WebFlowNets framework for molecular modeling, drug design, and QSAR.

P-40 : Parallel tiered clustering for large data sets using a modified Taylor's algorithm

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

Clustering large sets has many applications in drug discovery, among them compound acquisition decisions and combinatorial library diversification. Molecular fingerprints (2D) and molecular shape conformers (3D) from PubChem are the basic descriptors comprising the large sets utilized in this study. A parallel tiered clustering algorithm, implementing a modified Taylor's algorithm, will be described as an efficient method for analyzing datasets of such large scale. Results will be presented in SAESAR (Shape And Electrostatics Structure Activity Relationships).

P-41 : Ligand-based models for the isoform specificity of Cytochrome P450 substrates

L. Terfloth[1], B. Bienfait[1], J. Gasteiger[1,2]

[1] Molecular Networks GmbH, Erlangen,Germany

[2] Computer-Chemie-Centrum and Institute of Organic Chemistry, University of Erlangen-Nuremberg, Erlangen, Germany

In silico prediction of ADMET (absorption, distribution, metabolism, elimination, and toxicity) properties is expected to detect and eliminate compounds with inappropriate pharmacokinetic properties at an early stage of the drug discovery process. A central step in the ADMET process is drug metabolism. Metabolic stability, drug toxicity, and drug-drug interactions have to be considered.

Oxidation reactions mediated by cytochrome P450 isoforms play a crucial role in phase I of the human metabolism of xenobiotics. Here, we report on the isoform specificity for CYP3A4, CYP2D6, and CYP2C9 substrates.[1] The influence of the descriptors used for structure representation and the impact of the modeling method on the predictability of the models will be discussed. A thorough CV (cross-validation) scheme is presented to assess the reliability of the models. Furthermore, the prediction of a more diverse and larger external validation data set with an accuracy of up to 83% underlines the validity of the models. It will be shown that the random selection of a test set can be rather misleading to assess the predictability of a classification model (Figure 1).

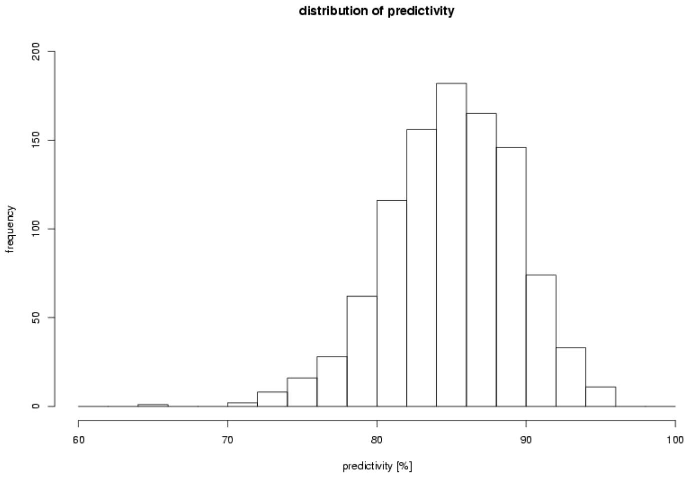


Figure 1: Distribution of the predictability for 1000 randomly selected test data sets.

A classification model for the isoform specificity is implemented in the application isoCYP.[2,3]

1. 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.
2. The software package isoCYP is available from Molecular Networks GmbH, Erlangen, Germany. http://www.molecular-networks.com (accessed Feb 24, 2008).
3. A Web service of isoCYP is available from Molecular Networks GmbH, Erlangen, Germany. http://www.molecular-networks.com/online_demos/cyp450 (accessed Feb 24, 2008).

P-42 : Metabolomics approach for determining growth-specific metabolites based on Fourier transform ion cyclotron resonance mass spectrometry

Hiroki Takahashi, Yoko Shinbo, Md.Altaf-Ul-Amin, Ken Kurokawa, Shigehiko Kanaya, Graduate School of Information Sciences, Nara Institute of Science and Technology, Nara, Japan

Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR/MS) is the best MS technology for obtaining exact mass measurements owing to its great resolution and accuracy, and several outstanding FT-ICR/MS-based metabolomics have been reported. Development of a general scheme of FT-ICR/MS-based metabolic profiling, with the aid of its potential for the high-resolution measuring power together with ion-signal intensity information, should thus make a significant contribution to metabolomics studies. To attain the purpose and to understand the cell system based on the components of metabolites, we apply chemometrics and bioinformatics approach to FT-ICR/MS data. Among a variety of metabolomics strategies, FT-ICR/MS offers a unique opportunity in non-targeted metabolomics studies owing to its extreme accuracy (below 1 ppm) in the mass measurement. Thus, chemical formulas and molecular identities of metabolites can be predicted with the aid of high precision mass spectrometry (MS) data and can also be easily linked to reported metabolites.

Metabolomics researches currently confront a problem that high-through put data-acquisition technologies including chromatography-coupled mass spectrometry (MS) and FT-ICR/MS have facilitated simultaneous detection and quantification of a large number of metabolite-derived peaks without metabolite assignment. The problem in annotation of metabolites is that there is only a piece of information about peaks corresponding to precise molecular weight for metabolite-derived ions in MS. Annotation process of ions as metabolites is most important for interpreting cellular condition by metabolite composition. To attain this, we propose a procedure of metabolite annotation using the data obtained from FT-ICR/MS by grouping ions originated from identical metabolites by a correlation analysis. The procedure mainly consists of three steps; data acquisition and constructing data matrix (Step 1), classification of ions into metabolite-derivative groups (Step 2), and annotation of ions as metabolites using chemical and biochemical knowledge (Step 3). In Step 2, ions can be classified into metabolite-derivative groups by a graph clustering algorithm combining with a correlation analysis. This procedure makes it possible to annotate ions as metabolites in high-quality, and to interpret the cellular condition of *Escherichia coli* by metabolite composition of cells. We searched 174 ions using KNApSAcK (http://kanaya.naist.jp/KNApSAcK/), and obtained 163 metabolite candidates from the search of the entire metabolite inventory in the database. Based on the species-metabolite relationship and MS/MS analyises, we were finally able to assign 33% of 220 detected ions to candidate metabolites. In this study, the percentage of ions annotated to metabolite candidates is much higher than that in case of plant reported by Nakamura et al (10% of peaks in *Arabidopsis thaliana*).

Using PLS regression we constructed a linear relation between OD_{600} values and metabolite profiles. High correlation between predicted and observed OD_{600} values certifies the correctness of the linear model. Our analyses reveal that global cyclopropane fatty acid formation of phosphatidylglycerols occurs as *E. coli* enters stationary phase from exponential phase. The results indicate that non-targeted metabolomics based on direct-infusion FT-ICR/MS is useful for analyzing the responses of biological systems to a variety of changes. Our integrated methodology is applicable to metabolic studies involving other organisms.

1. Nakamura, Y.; Kimura, A.; Saga, H.; Oikawa, A.; Shinbo, Y.; Kai, K.; Sakurai, N.; Suzuki, H.; Kitayama, M.; Shibata, D.; Kanaya, S.; Ohta, D., Differential metabolomics unraveling light/dark regulation of metabolic activities in Arabidopsis cell culture. *Planta* **2007**, *227*, (1), 57-66

P-43 : Clustering peptidases employing structural features of their inhibitors

M. Milik[1], *A. Bender*[1,2], *M. Glick*[1]

[1] *Novartis Institutes for Biomedical Research, Cambridge, USA*

[2] *current address: Division of Medicinal Chemistry, Leiden University, Leiden, The Netherlands*

Presented is a method for classification of enzyme proteins basing on structures of their small-molecule inhibitors. The goal is to provide a protein target classification method which is orthogonal in its concept to the classification based on protein sequence similarity, and which may provide additional input for focused molecular library design and lead finding procedures.

The feasibility of the method was tested on the subset of data from the WOMBAT database [1]. Compounds defined in this database as “active” against peptidase targets was extracted and used to define an activity-based fingerprint. The compounds were clustered according to their 2D structure similarity to remove redundancy and generate more robust classification. In the follow, the activity-based binary fingerprint was created for every selected peptidase target. The length of the fingerprint vector was equal to the number of structural clusters. The value on the position ‘n’ in the fingerprint vector was set to ‘0’ if no compound from the structural cluster number ‘n’ inhibits the given peptidase; in the opposite case the value was set to ‘1’.

This fingerprint was then used to build a phylogenetic-like tree for peptidases, as it is presented in the Figure 1 (left side). For comparison, on the same figure we show the phylogenic tree for the same set of proteins built on the basis of their sequence similarity (Figure 1, right side). Table 1 gives more detail description of the branches identified in the activity-based tree. The presented results show that our method may give an additional input into early analysis of candidate lead molecules, for example, by hinting of their putative off-target activity and toxicity problems. These kinds of problems are difficult to evaluate based only on protein target sequence or substrate analysis.

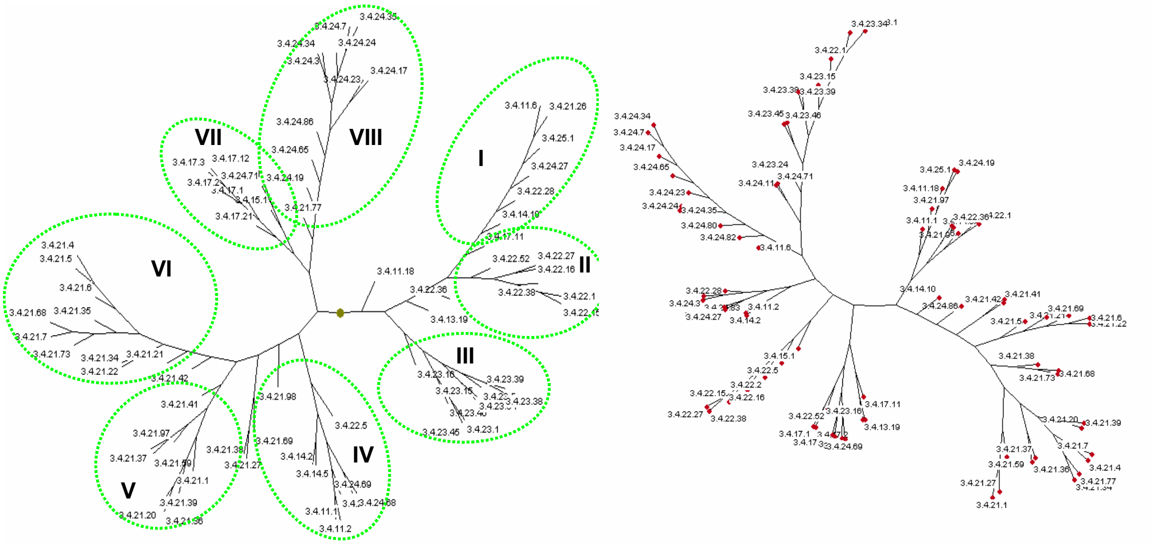


Figure 1. Phylogenetic-like trees for selected peptidase enzymes. The left one is based on the activity data, as it is defined in the WOMBAT database; the right tree is based on CLUSTALW [2] multiple alignment of the protein sequences. The peptidases are labeled by their Enzyme Classification codes. The Roman numbers enumerate activity-based clusters presented with more details in Table 1. The figure was prepared using PhyloDraw package [3].

Table 1. List of the peptidases used in the presented study, with their activity-based clustering deduced from the branching of the activity fingerprint tree from Figure 1 (left tree). The selected peptidases are provided with their Enzyme Commission number and peptidase family name – both extracted from SwissProt [4] database. While some of the clusters follow the EC classification up to the sub-subclass level, the other ones are more diversified and contain peptidases with different action mechanisms. More detail analysis of structural and chemical meaning of these clusters will be presented.

EC #	Enzyme Name	Pept.[1] Family	EC #	Enzyme Name	Pept.[1] Family
	Branch I			**Branch II**	
3.4.11.6	Aminopeptidase B	M1	3.4.22.1	Cathepsin B	C1
3.4.14.10	Tripeptidyl peptidase II	S8	3.4.22.15	Cathepsin L	C1
3.4.17.11	Glutamate carboxypeptidase	M20A	3.4.22.16	Cathepsin H	C1
3.4.21.26	Prolyl oligopeptidase	S9A	3.4.22.27	Cathepsin S	C1
3.4.22.28	Picornain 3C	C3	3.4.22.38	Cathepsin K	C1
3.4.24.27	Thermolysin	M4	3.4.22.52	Calpain-1	C2
3.4.25.1					
	Branch III			**Branch IV**	
3.4.23.1	Pepsin A	A1	3.4.11.1	Leucyl aminopeptidase	M17
3.4.23.5	Cathepsin D	A1	3.4.11.2	Membrane alanyl aminopeptidase	M1
3.4.23.15	Renin	A1	3.4.14.2	Dipeptidyl peptidase II	S28
3.4.23.16	HIV-1 retropepsin	A2	3.4.14.5	Dipeptidyl peptidase IV	S9B
3.4.23.38	Plasmepsin I	A1	3.4.22.5	Fruit bromelain (transfered entry: 3.4.22.33)	C1
3.4.23.39	Plasmepsin II	A1	3.4.24.68	Tentoxilysin	M27
3.4.23.45	Memapsin 1	A1	3.4.24.69	Bontoxilysin	M27
3.4.23.	Mamepsin 2	A1	3.4.24.	Anthrax lethal factor en-	M34

46			83	dopept.	
Branch V			**Branch VI**		
3.4.21.1	Chymotrypsin	S1	3.4.21.4	Trypsin	S1
3.4.21.20	Cathepsin C	S1	3.4.21.5	Thrombin	S1
3.4.21.36	Pancreatic elastase	S1	3.4.21.6	Coagulation factor Xa	S1
3.4.21.37	Leukocyte elastase	S1	3.4.21.7	Plasmin	S1
3.4.21.39	Chymase	S1	3.4.21.21	Coagulation factor VIIa	S1
3.4.21.59	Tryptase	S1	3.4.21.22	Coagulation factor IXa	S1
3.4.21.97	Assemblin	S21	3.4.21.34	Plasma kallikrein	S1
			3.4.21.35	Tissue kallikrein	S1
			3.4.21.68	T-plasminogen activator	S1
			3.4.21.73	U-plasminogen activator	S1
Branch VII			**Branch VIII**		
3.4.15.1	Peptidyl dipeptidase A	M2	3.4.24.3	Microbial collagenase	M9
3.4.17.1	Carboxypeptidase A	M14	3.4.24.7	Interstitial collagenase	M10B
3.4.17.2	Carboxypeptidase B	M14	3.4.24.17	Stromelysin 1	M10B
3.4.17.3	Lysine carboxypeptidase	M14	3.4.24.23	Matrilysin	M10B
3.4.17.12	Carboxypeptidase M	M14	3.4.24.24	Gelatinase A	M10B
3.4.17.21	Glutamate carboxypeptidase	M28	3.4.24.34	Neutrophil collagenase	M10B
3.4.21.77	Semenogelase	S1	3.4.24.35	Gelatinase B	M10B
3.4.24.11	Neprilysin	M13			
3.4.24.71	Endothelin-converting enz.	M13			

1. Peptidase family according to SwissProt [4] database

1. Olah M.; Mracec, M.; Ostopovici, L.; Rad, R.; Bora, A.; Hadaruga, N.; Olah, I.; Banda, M.; Simon, Z.; Mracec, M.; Oprea, T. I. WOMBAT: World of Molecular Bioactivity, in Chemoinformatics. In *Drug Discovery*; Oprea. T. I. (Ed), Wiley-VCH, New York, **2004**, pp. 223-239.
2. Chenna, R.; Sugawara, H.; Koike,T.; Lopez, R.; Gibson, T. J.; Higgins, D. G.; Thompson, J. D.; Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.* **2003**, 31, 3497-500
3. Jeong-Hyeon Choi, Ho-Youl Jung, Hye-Sun Kim, Hwan-Gue Cho. PhyloDraw: A Phylogenetic Tree Drawing System, *Bioinformatics*, **2000**, 16, 1054-1058.
4. Bairoch A.; Apweiler R.; Wu C.H., Barker W.C.; Boeckmann B.; Ferro S.; Gasteiger E.; Huang H.; Lopez R.; Magrane M.; Martin M.J.; Natale D.A.; O'Donovan C.; Redaschi N.; Yeh L.S. The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **2005,** 33:D, 154-159.

P-44 : Prediction of cell permeability

Paul Selzer[1], Peter Ertl [1], Daniela Gabriel [1], Christian N. Parker[1], Meir Glick [2], Mariusz Milik [2]

[1] Novartis Institutes for BioMedical Research, Basel, Switzerland

[2] Novartis Institutes for BioMedical Research Inc., Cambridge MA, USA

High content imaging allows one to analyze biological processes at the sub-cellular level, providing valuable contributions to hit finding. However, this assay format can be limited in terms of throughput due to difficulties in assay protocol and reagent cost. Therefore, there is a need to screen focused sets of compounds instead of randomly screening the full set of available compounds . This highlights the need for reliable and robust methods to predict cell permeability that help to concentrate the screening resources on those compounds that have a higher probability of being cell permeable.

To support this goal an in-house expert system for the prediction of the cell permeability of small molecules has been developed. The system was developed by the application of several machine learning techniques to this problem leading to a combination of two complementary approaches:

1. Bayesian Model
   The model has been developed with data resulting from cell based screens providing a binary classification (cell permeable vs. non cell permeable) as output. The chemical structures were encoded as Pipeline Pilot fingerprints. The experimental data have been divided into training and test set (4:1). Applying the model to the test set yielded a 9.5 fold enrichment (compared to random) cell permeable compounds in a virtual screen.

2. Random Forest Model
   This model was built on Caco2 assay data. The output of the model is the permeation coefficient P which is then binned to provide a binary classification (highly permeable vs. medium and low permeable). The model was validated in a pseudo prospective study – meaning that it has been built with compounds that have been screened earlier in time and tested against compounds that have been screened later. This experiment has been repeated several times with different time-point thresholds yielding ROC curve AUCs of 0.8.

Both methods were combined to provide a consensus score to rank compounds according to their cell permeability potential. Currently the system is being validated to analyze it's applicability for the productive screening set selection process.

P-45 : Validation using the RCSB: Good idea or bad idea?

Paul C. Hawkins , Gregory L. Warren, A. Geoffrey Skillman, OpenEye Scientific Software, Santa Fe, USA

Protein-ligand co-complexes from the RCSB database have been used in many studies on the quality of docking and conformer generation. However, due to the poor quality of some of the structures, many of their conclusions are invalid. This paper will discuss pitfalls associated with using structures from the RCSB for comparison or validation purposes. These pitfalls include local problems, such as poor quality fits to electron density (of ligand or protein), highly strained ligand structures and global issues such as lack of consideration of experimental error in the structural data. While nominal resolution has been frequently used for identifying good quality structures from the RCSB, much better assessments of quality can be obtained from global measures such as the diffraction-component precision index (DPI) and local measures including the real-space correlation coefficient. Consideration of these measures is mandatory when assembling a reliable set of structures for validation. Many of the problems associated with using ligand structures from the RCSB are eliminated when using small molecule crystal structures from the CCSD, as there is a much greater degree of precision in these structures.

With these issues in mind, datasets for validation of conformer generation applications derived from both the RCSB and the CCSD will be presented and the performance of a selection of methods on these datasets will be discussed using a number of different metrics.

P-46 : Automated generation of fragment-based rules for mutagenicity prediction

O.G. Othersen[1], R. Wehrens[1], L. Buydens[1], L. Ridder[2], M. Wagener[2]

[1] Analytical Chemistry, Institute for Molecules and Materials (IMM), Radboud University Nijmegen, Nijmegen, The Netherlands

[2] Molecular Design & Informatics Department, Organon, part of Schering-Plough, Oss, The Netherlands

Prediction of mutagenicity is an area of interest in pharmaceutical research as genotoxicity of a drug candidate is detrimental in development. Since mutagenicity is directly related to the chemical properties of a compound, i.e. its reactivity and affinity towards DNA, it is generally believed that prediction of mutagenicity on the basis of chemical structure should be feasible at a reasonable accuracy. However, several recent reviews[1-3] indicate that actual commercial packages for the prediction of mutagenicity perform not very well on drug-like compounds as the programs are not trained on these compound classes and the difficulty of describing non-covalent DNA interactions.

An automated and systematic method to derive molecular fragments related to mutagenicity has been developed. This approach allows to merge related specific mutagenic fragments into more general alerts, to incorporate the chemical context into the alert description (i.e. R-groups) and to refine alerts through logical combination with more specific fragments. The graphical representation of fragments and their interrelationships facilitates the analysis of the structural data and allows to assess the quality and relevance of the derived alerts. Furthermore, this method is not limited to mutagenicity and may be used to find alerts and structural dependencies within other binary classified data.

The performance of the method is illustrated using sets of structures with experimentally determined genotoxic potential and compared to other state-of-the art mutagenicity prediction methods.

1. Snyder, R.D.; Ewing, D.; Hendry, L.B. DNA intercalative potential of marketed drugs testing positive in in vitro cytogenetics assays. *Mutation Research*, **2006**, 609, 47–59.
2. Snyder, R.D.; Pearl, G.S.; Mandakas, G.; Choy, W.N.; Goodsaid, F.; Rosenblum, I.Y. Evaluation of DNA intercalation potential of pharmaceuticals and other chemicals by cell-based and three-dimensional computational approaches. *Environ. Mol. Mutagen*., **2004**, 43, 143-158.
3. Cariello, N.F. ; Wilson, J.D.; Britt, B.H.; Wedd, D.J.; Burlinson, B.; Gombar, V. Comparison of the computer programs DEREK and TOPKAT to predict bacterial mutagenicity *Mutagenesis*, **2002**, 17(4), 321-329.

P-47 : The detection of new active site conformations using molecular dynamics and rotamer assignments.

G. Schaftenaar, B. Vroling, Computational Drug Discovery Group, Radboud University Nijmegen, Nijmegen, The Netherlands

Protein flexibility is an important, but often neglected aspect of the drug development process. The flexibility of the binding pocket residues of the peroxisome proliferator-activated receptor was investigated using a novel molecular dynamics (MD) protocol. A comparison is made between a standard MD protocol, using water probes in the binding pocket, and a novel protocol, using hydrophobic probes. The ligand-complexed protein is simulated and used as a reference by comparing it to the obtained results of the hydrophobic and water probes trajectories. It is hypothesized that due to the hydrophobic nature of the natural ligands, the MD simulation using hydrophobic probes would result in a more accurate description of the binding pocket dynamics compared to the water probes trajectory.

The raw MD data is described by rotameric conformations, which results in sequence-like descriptions of active site conformations. Clustering and multi-dimensional scaling were used to validate the hypothesis. It was found that the use of a hydrophobic water model increases the sampling of ligand-like pocket conformations. The conversion of MD data to rotameric sequences allows for easy data analysis and intuitive visualizations.

1. Vroling, B, Schaftenaar, G.. The detection of new active site conformations using molecular dynamics and rotamer assignments. , *to be submitted to J. Comput. Aided Mol. Des.*

P-48 : Automated extraction of kinase hinge-binding fragments from the protein data bank

D. Wood[1], *S. Nabuurs*[1], *J. de Vlieg*[1], *M. Wagener*[2]

[1] *Computational Drug Discovery Group, Radboud University, Nijmegen, Netherlands*

[2] *Department of Molecular Design and Informatics, Organon NV, Oss, Netherlands*

Protein kinases are key components in signal transduction pathways. As abnormal kinase activity is observed in a wide range of diseases, inhibition of protein kinases has become an area of significant interest to the pharmaceutical industry [1]. An important mechanism for kinase inhibition involves targeting the ATP binding pocket with small, ATP-competitive compounds. These compounds form a common interaction with the highly conserved hinge region of the protein kinase structures.

A method was developed to automatically extract the substructures that bind to the hinge residues from the protein kinase entries of the Protein Data Bank. These substructures are bioisosteric in nature [2] and can be used to make substructural replacements to obtain new screening candidates from lead compounds. The substructures were collated and presented as a catalogue to assist computational chemists at Organon in the development of new kinase inhibitors.

1. Scapin, G. Structural biology in drug design: selective protein kinase inhibitors. *Drug Discovery Today.* **2002***,* 7(11), 601-611
2. Wagener, M.; Lommerse, J. P. M. The quest for bioisosteric replacements. *J. Chem. Inf. Model.* **2006** 46, 677-685

P-49 : Get the best from substructure mining

J. Kazius, Research For Charity Foundation, Utrecht, the Netherlands

All cheminformatics approaches neglect most of the chemical information that is present in a set of compounds. As no descriptor set can capture all biologically important features, valuable chemical knowledge can thus stay hidden from the process of hypothesis-based drug design. One example of a straightforward structure-activity relationship (SAR) is a substructure that predisposes compounds towards reduced or enhanced biological activity.

Substructure miners provide rapid access to a substantial repertoire of chemical descriptors that otherwise remains hidden: *substructures*. Substructure mining is therefore at least complementary to other methods of chemical data mining. Simply put, substructure mining consists of a focussed, but exhaustive, series of substructure searches.

This study explains problems that can now be solved through substructure mining. AweSuM is the new *Awe*some *Su*bstructure *M*ining software from Curios-IT and it is designed to efficiently learn the most interesting substructures. Several chemical datasets were mined (including those from HTS) [1]: we examine the resulting pharmacophores and scaffolds along with their biological and statistical relevance. For instance, an automatically extracted pharmacophore for hERG channel blockade shows predictive power and validates published chemical knowledge. Other datasets were analysed to identify scaffolds that affect binding affinity toward G Protein-Coupled Receptors (GPCRs). These results demonstrate that AweSuM extracts useful SAR knowledge from the vast space of substructure descriptors. More specifically, AweSuM reveals scaffolds that summarise the chemical content of datasets and key substructures (such as toxicophores or pharmacophores) that predict biological activities.

1. Wheeler, D. L.; Barrett, T.; Benson, D.A.; Bryant, S. H.; Canese, K.; Chetvernin, V., et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. **2008**, 36, D13-21.

P-50 : The RSC's project prospect: Identification and reuse of chemistry in publications

Colin Batchelor, RSC Publishing, Royal Society of Chemistry, Cambridge, UK

The RSC's Project Prospect, launched in early 2007, incorporates chemical structures and ontology terms into journal articles, and provides additional information on the compounds found to help the reader better understand the article and find related material. Uniquely, this structural information is also published in our RSS feeds. These continuing developments aim to extend further the markup of compounds using open standards and to promote new ways of publishing and finding structural data to improve the accessibility of chemistry. We will share our thinking on the future of structural information within chemistry publications.

P-51 : *In silico* studies on p63 as a new drug-target protein

A. Karawajczyk , G. Schaftenaar, RUMC/CMBI, Nijmegen, The Netherlands

p63 protein is a key regulator of ectodermal, orofacial and limb development [1]. Mutants of this protein are involved in the development of rare skin diseases such as EEC syndrome (Ectrodactyly-ectodermal dysplasia–clefting), Hay-Wells syndrome (AEC), Limb-mammary syndrome (LMS) and ADULT syndrome (Fig. 1). Clinical distinction among these syndromes is sustained both by the degree of expressivity of each disorder and by the occurrence of unique characteristic. The mutations associated with related disorders have been characterized and they are located in different domains of the protein (Fig. 2) [2]. Five protein motifs can be distinguished: a transactivation domain (TA), a DNA binding domain (DBD), oligomerisation domain (ISO), a sterile alpha motif domain (SAM), and transinactivation domain (TI). Finding a way to restore the function of p63 is one of the greatest challenges faced by scientists today.

The structure of the p63 is not known yet, however it shares a high sequence homology with the tumor suppressor p53. Based on that fact the homology model of the DNA binding domain is build and optimized. Furthermore, it was found that small molecules like PRIMA, MIRA and STIMA are capable of reactivating a wide range of mutant forms of p53 [3]. Again by analogy it is believed that they can also interact with p63 with the same effect. The binding site, the binding forces and the mechanism of action are to be determinated.

We will present the results of the recent studies on the structural properties of p63 mutants that may contribute to rational drug design. Additionally, the docking study has indicated the possible binding places for MIRA and PRIMA that steers an identification of pharmacophores of potential drugs.

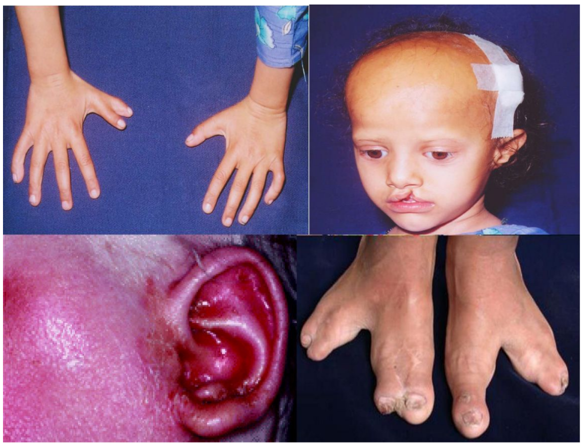


Figure 1 The illustration of EEC, SHFM, and ADULT syndromes.

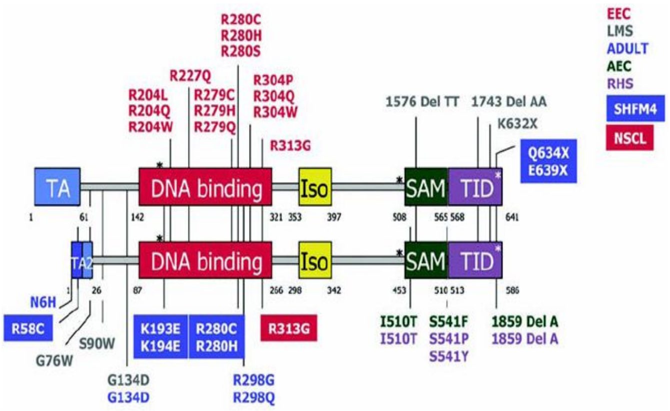


Figure 2 Mutation in p63 domains causing different diseases. The colors of indicated mutations correspond to the list of syndromes in the right top corner of the figure.

1. Mikkola, M. L. p63 in skin appendage development. *Cell Cycles* **2007**, 6, 285-290; Aberdam D., Gambaro K., Rostagno P., Aberdam E., Rouleau M. Key role of p63 in BMP-4-induced epidermal commitment of embryonic stem cells. *Cell Cycles* **2007**, 6, 291-294
2. Rinne T., Brunner H. G., Bokhoven van H. p63-associated disorders. *Cell Cycles* **2007**, 6, 262-268
3. Selivanova, G., Wiman K. G. Reactivation of mutant p53: molecular mechanism and therapeutic potential. *Oncogene.* **2007**, 26, 2243-2254.

P-52 : QSAR modelling of antineoplastic activities using NIH roadmap data

A. Zakharo, A. Lagunin, D. Filimonov and V. Poroikov, Institute of Biomedical Chemistry, Russian Academy of Medical Sciences, Moscow, Russia

A new database of diverse chemical structures and their biological activities is being developed by National Center for Biotechnology Information at NIH. The database, called PubChem, contains both structural information from scientific literature as well as screening and probe data from Molecular Libraries Screening Center Network. We selected more than 3200 compounds with more than 8 atineoplastic activities from PubChem for Quantitative Structure-Activity Relationships (QSAR) analysis by GUSAR program. QSAR models for several antineoplastic activities (Cytotoxicity of p53ts Cells at the Permissive Temperature, Activators of CRE-BLA: S1P2 Purchased Analogues, Agonists of the S1P2 Receptor of Purchased Analogues, Tumor Hsp90 Inhibitors, Inhibitors of HPGD (15-Hydroxyprostaglandin Dehydrogenase) and others) were obtained with reasonable accuracy (R^2 obtained by GUSAR was more then 0.60). They were validated by leave-one-out cross-validation procedure and on external test sets. Q^2 obtained by GUSAR exceeded 0.50. We calculated RMS_{test} for assessment of accuracy of test set prediction. RMS_{test} was less than 0.40 within the domain applicability of model for all activities.

Acknowledgements. The study is supported by FP6-grant LSHB-CT-2007-037590 (Net2Drug).

P-53 : GUSAR: new approach for multiple QSAR

A. Zakharov, A. Lagunin, D. Filimonov and V. Poroikov, Institute of Biomedical Chemistry, Russian Academy of Medical Sciences, Moscow, Russia

We proposed a new QSAR method based on QNA (Quantitative Neighbourhoods of Atoms) descriptors and self-consistent regression that was realized in computer program GUSAR. It predicts the quantitative values of biological activity of chemical compounds on the basis of their structural formulae. The method uses universal descriptors and does not require selecting the most relevant descriptors like in the classic 2D QSAR. The method does not require also selection of model on the basis of Q^2 values. GUSAR was compared with several widely used methods including CoMFA, CoMSIA, GRID, HQSAR, EVA and 2D QSAR. Nine evaluation sets with data on toxicity, metabolism, ligand-enzyme and ligand-receptor interactions were used for assessment of GUSAR predictive abilities. It was shown the GUSAR accuracy was comparable or better than the accuracy of other QSAR methods both on heterogeneous (CDK inhibitors, DHFR inhibitors, ACE inhibitors) and homogeneous (*Vibrio fischeri*, *Chlorella vulgaris*, *Tetrahymena pyriformis,* 5-HT1A serotonin receptor ligands, estrogen receptor ligands and CYP2A5 inhibitors) data sets. It was shown that Q^2 obtained by GUSAR for these sets varied from 0.67 to 0.87 and R^2_{test} varied from 0.53 to 0.95. GUSAR showed reasonable prediction ability and robustness in leave-20%-out cross-validation procedure. Average R^2 prediction accuracy for different test sets in leave-20%-out cross-validation procedure was 0.714 and varied from 0.53 to 0.89. Thus, GUSAR can be widely applied to different routine QSAR tasks, for building models and prediction different quantitative characteristics for many activities simultaneously.

Acknowledgements. This work is supported by European Commission under FP6-'Life sciences, genomics and biotechnology for health' contract LSHB-CT-2007-037590 'Net2Drug'.

P-54 : Fast empirical estimates of quantum mechanical descriptors for QSAR/QSPR modeling

R. Fraczkiewicz , M. Waldman, J. Crison, W.S. Woltosz, Simulations Plus, Inc., Lancaster, CA, U.S.A.

A sizable number of published studies suggest that molecular descriptors derived from quantum chemical calculations can yield effective QSAR/QSPR models, in particular related to chemical reactivity. For example, Gross, et al, have determined that partial atomic charges derived from Natural Population Analysis (NPA) of DFT wavefunctions have a superior predictive power of aqueous ionization constants of phenols and anilines [1]. Quantum mechanical descriptors have also been found useful in predicting toxicity [2], drug metabolism [3], and intestinal absorption [4] – these are mere examples of wide applications of these descriptors. Unfortunately, implementation of sufficient quality quantum mechanical descriptors in Ultra High Throughput (UHT, >200,000 compounds/hour) QSAR/QSPR models for drug candidate screening is seriously limited by time consuming *ab initio* calculations where a single molecule processing can take hours, or even days.

We decided to overcome this difficulty by creating ultra-fast empirical estimates of certain quantum mechanical descriptors (sigma and pi partial atomic charges, pi system HOMO/LUMO energies, chemical hardness and electronegativity, and Fukui reactivity indices) at the atomic and molecular level by fitting high quality *ab initio* electron densities calculated for a dataset of almost 700 organic molecules. This dataset, containing neutral as well as formally charged molecules, was composed with maximum diversity of individual atomic environments in mind. All molecular geometries were optimized at the

B3LYP/6-311G** level, followed by extraction of approximately 13000 sigma and pi partial atomic charges with the aid of the NPA and Natural Bond Orbital (NBO) schemes. One part of the data set (11000) was used to train a new empirical model for very fast estimation of the atomic charges; the remainder (2000) was sequestered as an external validation set. Two separate empirical models, both using only 2D molecular structures on input, were created: one for sigma, the other for pi subsystems. Sigma and pi electron densities on atoms were estimated with an excellent quality: for both models the root-mean-square-error (RMSE) on external test set was close to 0.05 electron units.

The importance and usefulness of thus derived estimated quantum mechanical descriptors were subsequently demonstrated on a wide array of QSPR models related to Absorption, Distribution, Metabolism, Excretion and Toxicity (ADMET) properties of molecules built by our group.

1. Gross, K. C.; Seybold, P. G.; Hadad, C. M. Comparison of Different Atomic Charge Schemes for Predicting pKa Variations in Substituted Anilines and Phenols. *Int. J. Quantum Chem.* **2002**, 90, 445-458.
2. Benigni, R.; Giuliani, A.; Franke, R.; Gruska, A. Quantitative Structure-Activity Relationships of Mutagenic and Carcinogenic Aromatic Amines. *Chem. Rev.* **2000**, 100, 3697-3714.
3. Singh, S. B.; Shen, L. Q.; Walker, M. J.; Sheridan, R. P. A Model for Predicting Likely Sites of CYP3A4-mediated Metabolism on Drug-like Molecules. *J. Med. Chem.* **2003**, 46, 1330.
4. Jones, R.; Connolly, P. C.; Klamt, A.; Diedenhofen, M. Use of Surface Charges from DFT Calculations to Predict Intestinal Absorption. *J. Chem. Inf. Model.*, **2005**, 45, 1337-1342.
5. Reed, A. E.; Weinstock, R. B.; Weinhold, F. A. Natural population analysis. *J. Chem. Phys.* **1985**, 83, 735-746.

P-55 : The representation, registration, and retrieval of substances with incompletely defined chemical structures

K Taylor, B Christie, D Grier, B Leland, J Nourse, Symyx Technologies Inc, San Ramon, USA

A new bond type is described that embodies a one to many relationship. This allows the description of structures that have indeterminate positional substitution, such structures are commonly encountered in metabolite determination studies, natural product chemistry, and analytical chemistry in general. It also allows the description of structures that involve pi-bonding, for example metallocenes. When coupled with a new extension to formula definition, this bond type allows the description of generic structures with a level of variability that exceeds the capabilities of existing technologies

Fully characterized chemical entities that have chemical structures whose atoms and bonds conform to the valence bond model, for example aspirin, present few difficulties for chemical registration systems. Frequently during the research process the structures of substances are incompletely characterized. This paper describes and illustrates an extension to Symyx's chemical representation; a new style of chemical bond that enables the registration and retrieval of structures with unknown positional substitution In addition, this bond type, allows a general structure to be defined for many industrial chemicals, and it has the characteristics required to define structures with pi-bonded ligands.

P-56 : Exploring synthetically accessible chemical space

K Taylor, J Durant, R Hillard, Symyx Technologies Inc, San Ramom, USA

A KNIME (www.knime.org) workflow is presented that takes a generic reaction, parses out the generic reactants and searches databases of known substances, for example Symyx Available Chemicals Directory, PubChem, or Symyx Compound Index. The potential reactants are filtered to remove those that are isotopically labeled, de-salted, and enumerated. The products are de-duplicated using Symyx's NEMA technology, and filtered for drug likeness, to provide a virtual library that may be further analyzed synthetic opportunity.

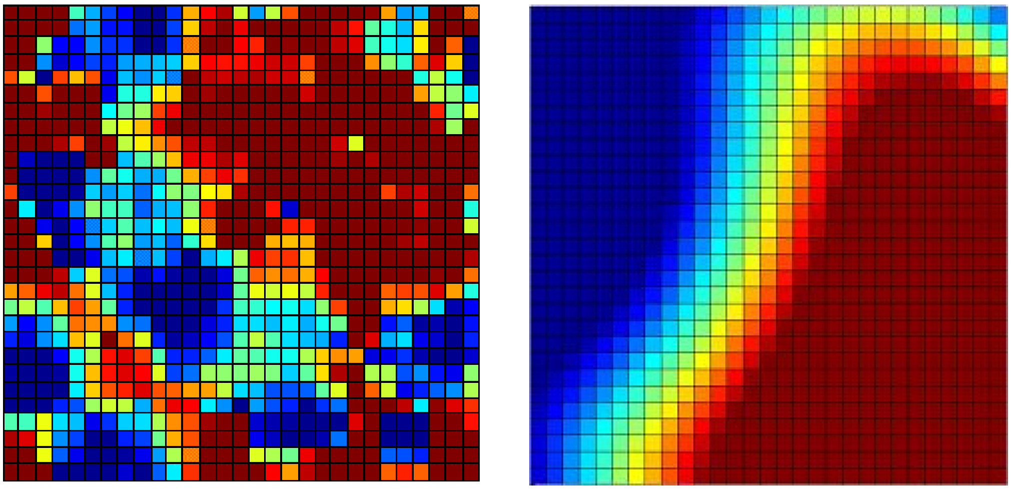
P-57 : Development and visualization of the drug-likeness model

M. Arakawa, T. Miyao, K. Funatsu, The University of Tokyo, Tokyo, Japan

In the field of drug design, virtual screening is widely used for discovering novel lead candidates. By exploring virtual library, compounds having certain activities for target enzyme would be found out efficiently. One of the most important factors for the success of virtual screening is quality of virtual libraries. Thus the reliable methodologies to construct high quality libraries have been expected. It is desired that chemical structures in virtual library exhibit drug-like properties in order to avoid ADME problems in the later phase of drug development process. Thus statistical model for estimating drug-likeness of chemical structures has been required in order to build high quality virtual library.

The first object of this study is to construct a model for estimating drug-likeness from chemical structures. Drug and non-drug molecules have been taken from the database of CMC (comprehensive medicinal chemistry) and ACD (available chemicals directory), respectively, and used to construct the drug-likeness model. Chemical descriptors were calculated with DRAGON 5.4, and the drug-likeness model was constructed by using SVR (support vector regression) method with Gaussian kernel.

Another object of this study is to visualize the drug-likeness model. The SOM (self organizing map) and GTM (generative topographic mapping) [1] methods have been adopted to map multi-dimensional descriptor space to two-dimensional map. In this visualization, smooth mapping is preferable because more complex 2D map is difficult to use in drug design process. However, estimation method of the smoothness of multidimensional mapping has not been established, so we propose novel criteria named RMS of midpoint (RMSM). RMSM is calculated as RMS of mapping error of all midpoints instead of the original data points. As a result of calculation of RMSM of SOM and GTM, we concluded that GTM is able to give slightly better nonlinear mapping than SOM. The 2D maps obtained by SOM and GTM are shown in Figure 1.



(A) SOM

(B) GTM

Figure 1 two-dimensional maps of drug-like (red) and non-drug-like (blue) compound.

1. C. M. Bishop, M. Svensén, C. K. I. Williams, Neural Computation, 10, 215-234, 1998.

P-58 : Reverse analysis and multi-objective optimization of predictive models for polymer properties

S. Goto, M. Arakawa, K. Funatsu, The University of Tokyo, Tokyo, Japan

The glass transition temperature (T_g) is one of the most important properties of polymers. The T_g determines the usage of polymers because it is strongly related to physical properties. In industrial field, a lot of polymer syntheses are tried under various conditions in order to accomplish target values of polymer properties. If we can accurately estimate the T_g, we can decrease the number of useless experimental trials. However, it is difficult to predict the T_g because it is affected by many factors, such as the amount and the kind of monomer and catalyst.

In general, there are two ways to estimate the T_g, theoretical calculation and statistical calculation. Theoretical calculation, such as molecular simulation [1], enables us to estimate the T_g with higher accuracy but takes a long time. On the other hand, statistical calculation, such as quantitative structure-property relationship (QSPR) analysis, is suitable for designing monomer composition and structure. C. Camacho-

Zuniga and F. A. Ruiz-Trevino proposed a model between the T_g and repeating units of polymers [2]. In this study, we built partial least squares (PLS) models between T_g's and molecular descriptors of monomers or values of monomer composition. PLS is one of the linear regression methods and PLS model is more chemically understandable than that of non-linear methods. In order to calculate molecular descriptors, we used general group contribution method and DRAGON [3]. By the same procedure, we built the predictive models concerning other properties.

As an application of our models, we can design not only monomer composition but also new monomer structure. When a functional polymer is developed, there are usually plural target values of polymer properties. In such a case, reverse analysis of the models by multi-objective optimization method is required. With the models using the values of monomer composition as explanatory variables, we can propose the candidates of monomer compositions of the polymers which might give the target properties. With the models using the kind and number of atomic groups as explanatory variables, we can get the sets of atomic groups. By constructing them, we can generate the candidate structures of monomers for the polymers which might give the target properties. New structures of the monomers might be included in them. At the end, with the models using DRAGON as explanatory variables, we can verify the validity of the above design for new polymers.

1. Nakanishi, K. Physical Property Prediction based on Molecular Simulation for Simple Model Fluids *J. Chem. Eng. Japan*, **1995**, 28, 1-7
2. Camacho-Zuniga, C.; Ruiz-Trevino, F. A. A New Group Contribution Scheme To Estimate the Glass Transition Temperature for Polymers and Diluents *Ind. Eng. Chem. Res.*, **2003**, 42, 1530-1534.
3. Talete srl. DRAGON for Windows (Software for Molecular Descriptor Calculations). Version 5.4 - 2006 - http://www.talete.mi.it/

P-59 : Development of a new regression analysis method using independent component analysis

Hiromasa Kaneko, Masamoto Arakawa, Kimito Funatsu, The University of Tokyo, Tokyo, Japan

Multivariate techniques such as multiple linear regression (MLR), principal component regression (PCR), and partial least squares (PLS) are powerful tools for handling several problems in chemoinformatics. It is possible to construct an accurate model by using these methods, for example, PLS, but is difficult to construct a predictive model. As for a problem we often face, there is the possibility of statistical problems occurring, such as overfitting.

It is important to indicate the magnitude of contribution of each variable to a model. However, it is dangerous to simply consider regression coefficients as important for each variable because there are correlations in explanatory variables. Thus, it is desirable that a prediction model be constructed that has high predictive power and that is easy to interpret in various fields of science.

In our study, independent component analysis (ICA) [1] and regression analysis are combined to extract significant components and construct a model that has high predictive power and that is easy to interpret. ICA is a method that extracts mutually independent components from explanatory variables, and is used in many fields such as signal processing. Through making full use of the high-order statistical characteristics of the source, that is, the fourth-order central moment, ICA can effectively resolve the independent components from the measured mixed signals without any additional information about the source signals. A relationship between the independent components and an objective variable is constructed by the least-squares method. This method is named ICA-MLR [2].

We verified the superiority of ICA-MLR over PLS with simulation data and tried to apply this method to a quantitative structure-property relationship (QSPR) analysis of aqueous solubility. We constructed models between aqueous solubilities based on the experimental data for 1290 molecules [3] and 173 molecular descriptors. PLS and genetic algorithm PLS (GAPLS) models were constructed for a comparison of ICA-MLR. R^2, Q^2, and R_{pred}^2 values of the PLS model are 0.836, 0.819, and 0.848, respectively. These values of the ICA-MLR model are 0.937, 0.868, and 0.894, respectively. ICA-MLR achieved higher predictive accuracy than PLS. ICA-MLR could extract effective components from explanatory variables and construct the regression model with high predictive accuracy. In addition, the information of regression coefficients $\mathbf{b}_{ICA\text{-}MLR}$ indicates the magnitude of contribution of each descriptor in the analysis of aqueous solubility.

1. Comon, P. Independent component analysis, A new concept? *Signal Process*. **1994**, 36, 287-314.
2. Kaneko H, Arakawa M, Funatsu K. Development of a New Regression Analysis Method Using Independent Component Analysis. *J. Chem. Inf. Model.* **2008**, in press.
3. Hou, T. J.; Xia, K.; Zhang, W.; Xu, X. J. ADME Evaluation in Drug Discovery. 4. Prediction of Aqueous Solubility Based on Atom Contribution Approach. *J. Chem. Inf. Comput. Sci*. **2004**, 44, 266-275.

P-60 : Rule induction of the site of metabolism by human cytochromes P450 by data-mining

M. Koyama, M. Arakawa, K. Funatsu, The University of Tokyo, Tokyo, Japan

It is indispensable to investigate ADMET(absorption, distribution, metabolism, elimination, and toxicity) properties in the drug discovery process. Generally, examinations of metabolism take a long time in ADMET test process. Therefore, the method for predicting metabolite of a drug is strongly needed. In order to predict it, we have to reveal several selectivities such as isoform specifity, regioselectivity, and so on. In this study, we focus on the regioselectivity, which is the selectivity of the position where metabolic reaction preferentially occurred in a molecule by enzymes.

R. P Sheridan reported QSAR-based regioselectivity models for cytochromes P450 3A4, 2D6 and 2C9, which had higher predictive power than their mechanistic model [1]. Descriptors used in their models were those which describe the local environment around each nonhydrogen atom in each molecule. In this study, we applied data-mining methods to their datasets of 3A4, 2D6, and 2C9. Data-mining is often defined as "non-trivial process of indentifying valid, novel, potentially useful, and ultimately understandable patterns in the data". With this method, we could not only build prediction models but also induce rules of regioselectivity. Chemically understandable rules give chemists insight not only "what" is the metabolite of a drug but also "why" the metabolite was produced from a drug. And thus data-mining is considered to be efficient approach to solve regioselectivity.

We used several data-mining methods, such as C4.5, Ripper, ensemble learning, and Inductive Logic Programing(ILP).C4.5 and Ripper are one of the most popular classifying methods and often give easily interpretable classification rules.

Ensemble learning is the method of combining several learning machines such as C4.5. It has higher predictive power than using single learning machine in many cases. But it also has a weak point that the results of it are often hard to translate into rules. So far, several ensemble learning methods have been developed such as bagging, boosting, stacking and so on, and we examined some of them.

Inductive Logic Programing(ILP) is also the methodlogy of data-mining but differs somewhat from the others. First, ILP uses a language much closer to one used normally by chemists and therefore is considered to more likely yield chemically understandable rules than other methods. Second, ILP can learn from the background knowledge and thus induce rules from datasets hard to express a single table. There have been developed various useful ILP systems such as Aleph, Progol, Foil and so on, and we used Progol system [2].

We used methods as above and compared each other. As for 2C9, the best predictive model was ensemble learning model, 86 % of atoms to be considered as metabolized sites were correct. But we couldn't extract any rules from the model. On the other hand, we obtained several chemically understandable rules by C4.5, Ripper, and ILP. In this poster we show the some of them, considering the chemical meanings.

1. Robert P. Sheridan; Kenneth R. Korzekwa Empirical Regioselectivity Models for Human Cytochromes P450 3A4, 2D6, 2C9. J. Med. Chem. **2007**, 50, 3173-3184.
2. S.Muggleton Inverse entailment and progol. New Generation Computing. 1995, 13, 245-286

P-61 : Dynamic interplay between chemotype, similarity and docking studies: Towards a virtual screening approach for protein kinase B inhibitors

J. L. Medina-Franco[1], K. Martínez-Mayorga,[1] M. Giulianotti,[1] T. Scior,[2] Y. Yu,[3] C. Pinilla[4], R. Houghten[1,4]

[1] Torrey Pines Institute for Molecular Studies, Fort Pierce, Florida, USA

[2] Departamento de Farmacia, Facultad de Ciencias Químicas, Benemérita Universidad Autónoma de Puebla, Puebla, México

[3] College of Pharmaceutical Science, Zijin Campus, Zhejiang University, Hangzhou, China

[4] Torrey Pines Institute for Molecular Studies, San Diego, California, USA

Protein kinase B (PKB), also named Akt, is a serine/threonine kinase that plays a key role in the regulation of a number of important events such as cell survival, proliferation and growth. Over expression of PKB is associated with several types of cancer.[1] In this work we present a systematic classification of known PKBβ (AKT2) inhibitors stored in a public database[2] using a number of chemoinformatics and docking methods including 2D- and 3D-similarity-based classification and a chemotype-based classification.[3] 2D- and 3D-similarity studies were performed using structural fingerprints and the Rapid Overlay of Chemical Structures (ROCS) approach, respectively. Docking-based classification of the known inhibitors using the Fast Rigid Exhaustive Docking (FRED) program and Genetic Optimization for Ligand Docking (GOLD) led to a promising fast virtual screening strategy for identifying PKBβ inhibitors in compound databases.

Using the similarity and docking approaches described above we screened a large collection with more than 100 thousand compounds. For docking, multiple crystallographic structures of PKBβ were used[1] and for similarity studies several known PKBβ inhibitors were employed as queries. Here we discuss the results of the virtual screening including an analysis of the relationship between 2D- and 3D-similarity measures and docking scores. A chemotype-based analysis of docking results revealed potential scaffolds for structural modification.

1. Davies, T. G. et. al. A Structural Comparison of Inhibitor Binding to PKB, PKA and PKA-PKB Cimera. *J. Mol. Biol.* **2007**, 367, 882-894. b) Saxty, G. et. al. Identification of Inhibitors of Protein Kinase B Using Fragment-Based Lead Discovery. *J. Med. Chem.* **2007**, 50, 2293-2296.
2. Scior, J. T.; Bernard, P.; Medina-Franco, J. L.; Maggiora, G. M. Large Compound Databases for Structure-Activity Relationships Studies in Drug Discovery. *Mini-Rev. Med. Chem.* **2007**, 7, 851-860.
3. Medina-Franco, J. L.; Petit, J.; Maggiora, G. M. Hierarchical Strategy for Identifying Active Chemotype Classes in Compound Databases. *Chem. Biol. Drug. Des.* **2006**, 67, 395-408.

P-62 : Multi-fusion similarity maps for comparing the chemical space of combinatorial libraries

J. Medina-Franco[1], G. Maggiora[2], M. Giulianotti[1], J. Rios[1], C. Pinilla[3], R. Houghten[1,3]

[1] Torrey Pines Institute for Molecular Studies, Fort Pierce, Florida , USA

[2] College of Pharmacy & BIO5 Institute, University of Arizona, Tucson, Arizona, USA

[3] Torrey Pines Institute for Molecular Studies, San Diego, California 92121, USA

A low-dimensional method for graphically depicting and characterizing relationships among molecules in high-dimensional chemical spaces is described.[1] The method is based on the use of multiple fusion-based similarity measures.[2] Specifically, max-fusion and mean-fusion similarity measures are used to construct multi-fusion similarity maps that characterize the relationship of a set of "test" molecules to a set of "reference" molecules, but other types of fusion-based similarity measures, such as median-fusion similarity, can also be used (Figure 1). The reference set is very general and can be made of molecules from, for example, the set of test molecules itself (the self-referencing case), molecules from a small library or large compound collection, or from molecules that are active in a given assay or group of assays. The use of multiple fusion similarity measures tends to provide more information than single fusion-based measures including, importantly, information on the nature of the chemical-space neighborhoods surrounding reference-set molecules.

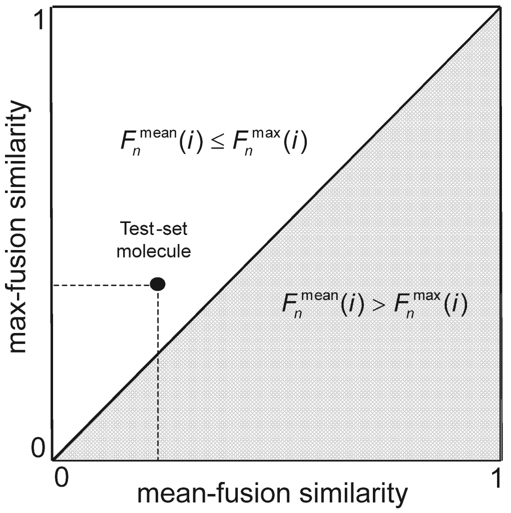


Figure 1. General form of a multi-fusion similarity map.

In this work a general discussion is presented on how to interpret multi-fusion similarity maps, and several examples are given that illustrate how these maps can be used to compare compound libraries or collections,[3] to select compounds for screening or acquisition, and to identify new active molecules using ligand-based virtual screening. Specific applications to visually characterize the chemical space of combinatorial libraries[4] are presented (Figure 2). Although the methodology described in this work is focused on applications to small molecules, it can be applied to any sets of objects (e.g., proteins) for which a similarity measure can be determined computationally or otherwise.

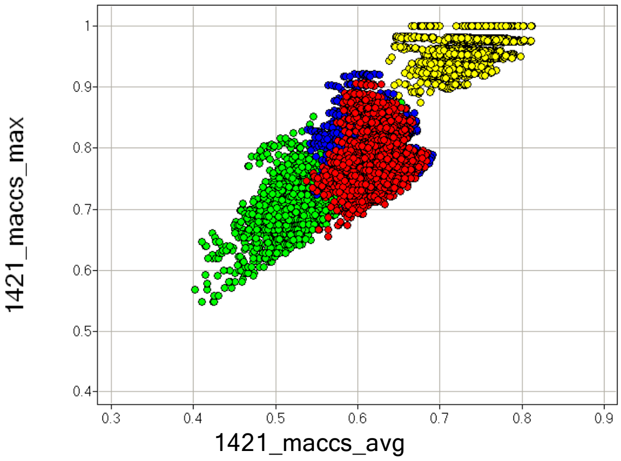


Figure 2. Multi-fusion similarity map comparing combinatorial libraries[4] represented in different colors.

1. Medina-Franco, J. L.; Maggiora, G. M.; Giulianotti, M. A.; Pinilla, C.; Houghten, R. A. A Similarity-based data-fusion approach to the visual characterization and comparison of compound databases. *Chem. Biol. Drug. Des.* **2007**, 70, 393-412.
2. Ginn, C. M. R., Willett, P., Bradshaw, J. Combination of molecular similarity measures using data fusion. *Perspect. Drug Discov. Design* **2000**, 20, 1-16.
3. Scior, J. T.; Bernard, P.; Medina-Franco, J. L.; Maggiora, G. M. Large Compound Databases for Structure-Activity Relationships Studies in Drug Discovery. *Mini-Rev. Med. Chem.* **2007**, 7, 851-860.
4. Houghten, R. A.; Pinilla, C.; Appel, J. R.; Giulianotti, M. A.; Nefzi, A.; Ostresh, J. M.; Dooley, C. T.; Maggiora, G. M.; Medina Franco, J. L.; Brunner, D.; Schneider, J. Mixture-based Synthetic Combinatorial Libraries: Direct in vivo Testing, Scaffold Ranking, and Enhanced Deconvolution Using Computational Approaches. *J. Comb. Chem.* **2008**, *10*, 3-9.

P-63 : The effect of structural redundancy on virtual screen performance

Robert D. Clark, Tripos International, St. Louis MO, USA

Large-scale combinatorial synthesis and high-throughput screening (HTS) once held out the promise of making prediction of biological activity on the basis of chemical structure a matter of purely historical interest, but this scenario has not been realized. Instead, in silico screens (virtual HTS, or vHTS) have become critically important for determining which possible focused libraries should be pursued, for subsequently identifying important lead series that may have been missed by biological screening, and for identifying the structural series necessary for establishing useful quantitative structure-activity relationships (QSARs). Several different kinds of virtual screen are now in use, each serving a somewhat different purpose in lead discovery and optimization. The primary use of substructural and topological similarity searching is to identify analogues for follow-on synthesis or purchase. Docking, pharmacophoric and shape similarity methods, in contrast, are used to identify potential lead- and scaffold hops, bringing in novel alternative chemistries to serve as a hedge against development issues with ADME, pharmacokinetic properties or ADME that can affect entire structural classes.

It is tempting to search for a "philosopher's stone" of vHTS that can solve all problems for all targets, but the reality is that different methods are suitable for different targets and will continue to be so for the foreseeable future. Hence researchers need to be able to compare the performance of different methods in different contexts and against different targets. The area under the receiver operating characteristic curve (ROC AUC) has many good statistical properties and has seen considerable use, particularly in connection with docking and scoring.

An ROC analysis entails plotting the recovery rate for true positives against that for false positives across a large set of reference compounds made up of compounds known to be active against the target in question as well as decoys that are known to be (or, more often, that are presumed to be) inactive. Unfortunately, the data sets used for carrying out such evaluations are almost always drawn from compound collections accumulated over many years. The composition of such collections reflects many historical influences, including incidental development series and bias due to particular offensive and defensive patent strategies employed over the years. They are more or less structurally "clumpy" as a result, especially when they have been compiled across diverse research programs (public databases) or as a result of merger and acquisition activities (pharmaceutical databases).

Such clumpiness can seriously skew retrospective ROC AUC analyses, misleading researchers as to which in silico screen is likely to give the best prospective performance [1]. Two methods have recently been proposed for addressing such biases: using the semilogarithmic integral rather than the linear one (pROC AUC) and down-weighting the true positive rate for "hits" that are overly similar to other actives [2]. This talk will center on various applications and how the relationship between the searching and clustering methods used affects the statistics produced.

1. Good, A.C.; Hermsmeier, M.A.; Hindle SA. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 529-536.
2. Clark, R.D.; Webster-Clark, D.J. *J. Comput.-Aided Mol. Des.* **2008***, in press*; Online First DOI 10.1007/s10822-008-9181-z.

P-64 : Topomer CoMFA for rapid optimization

Bernd Wendt, Tripos International, Munich, Germany

Recently the Topomer CoMFA method[1] was released as a 3D-QSAR tool that automates the creation of models for predicting the biological activity or property of compounds. Since its inception in 2001 the tool had been in productive use at Tripos' chemistry research center to drive collaborative drug discovery projects. Over the last 2 years the tool was evaluated in technological previews at several pharmaceutical companies. The presentation will cover the lessons learned from prospective as well as retrospective studies.

One of the critical parameters identified in the analysis of QSAR models is the composition of the series. Work on a set of 16 published QSAR datasets resulted in the development of a new procedure for 3D-QSAR analysis. Quantitative Series Enhancement Analysis (QSEA)[2] will be proposed for determination whether compounds belong to an emerging structure-activity relationship and which compounds can be predicted within reliable limits.

An important aspect of 3D-QSAR models is its application for virtual screening where the purpose is to identify compounds with superior properties through database searching. The combined use of Topomer Search and Topomer CoMFA will be presented on the basis of retrospective studies.

1. Cramer, R.D.. Topomer CoMFA: A Design Methodology for Rapid Lead Optimization. *J. Med. Chem.* **2003**; 46; p 374-388.
2. Wendt, B. ; Cramer R.D. Quantitative Series Enrichment Analysis: A novel procedure for 3D-QSAR Analysis. *J. Comp. Aid. Des.* **2008**, in press.

P-65 : Development of an a priori ionic liquid design tool. Integration of a novel COSMO-RS molecular descriptor on neural networks

J. Palomar[1], J. S. Torrecilla,[2] V. R. Ferro,[1] F. Rodríguez.[2]

[1]Sección de Ingeniería Química, Universidad Autónoma de Madrid, Spain

[2]Departamento de Ingeniería Química, Universidad Complutense de Madrid, Madrid, Spain

An innovative computational approach is proposed to design ionic liquids directly on the computer, based on a new a priori molecular descriptor of ionic liquids (ILs) derived from quantum chemical COSMO-RS methodology.1 Several previous studies have shown that COSMO-RS performs fast and accurate statistical thermodynamic calculations using only quantum chemical information of the molecular structures modeling the IL compounds.2 We recently showed the capability of COSMO-RS method to predict accurate specific density of representative series of imidazolium based ILs.3 In addition, we suggested the possibility of new interesting applications of COSMO-RS methodology. Thus, in this work, it is probed that the charge distribution on the polarity scale given by COSMO-RS can be used to characterize the chemical nature of both cation and anion of the IL structures, using simple molecular models in the calculations. As result, a novel a priori quantum-chemical parameter, Sσ-profile, is defined for forty five imidazolium based ILs, as a quantitative numerical indicator of their electronic structures and molecular sizes. Subsequently, neural networks (NNs) are successfully applied to establish relationship between the electronic information given by Sσ-profile molecular descriptor and a set of IL properties of pure and mixture fluids, including density, solubility and partition coefficients. As consequence, we develop here an a priori computational tool for screening ILs with required properties simultaneously, using COSMO-RS predictions to NN design and optimization. The performance of this computational approach was demonstrated following a classical quantitative structure-property relationship (QSPR) scheme, which is the main aim of this work. In this study, two very simple molecular models for obtaining Sσ-Profile were validated. When ion-paired structures (CA model) are used in calculations, a more reliable description of the intermolecular interactions in pure IL fluid is obtained. However, a model of independent ions (C+A model) presents the clear advantage of a reasonable reliability at much less computational cost.

1. Klamt, A. *COSMO-RS: From Quantum Chemistry to Fluid Phase Thermodynamics and Drug Design*, 1st Edition; Elsevier: Amsterdam, **2005**.
2. Diedenhofen, M.; Klamt, A.; Marsh, K.; Schäfere, A. Prediction of the vapor pressure and vaporization enthalpy of 1-n-alkyl-3-methylimidazolium-bis-(trifluoromethanesulfonyl) amide ionic liquids. *Phys. Chem. Chem. Phys.* **2007,** *9*, 4653-4656.
3. Palomar, J.; Ferro, V. R.; Torrecilla, J. S.; Rodríguez, F. Density and Molar Volume Predictions Using COSMO-RS for Ionic Liquids. An Approach to Solvent Design. *Ind. Eng. Chem. Res.* **2007**, *46,* 6041-6048.

P-66 : Radial scan of the electrostatic potential of RNA double helices. An application on tRNA acceptor stems

R. Marín, W. Agudelo, E. Daza, Theoretical Chemistry Group, National University of Colombia, Bogotá D. C, Colombia

We have developed a methodology to characterize and compare quantitatively the Molecular Electrostatic Potential (MEP) induced by the more exposed atoms in the minor and major grooves of the RNA double helices. This method is general and is also applicable to other RNA or DNA double helix systems where minor and/or major grooves must be characterized electrostatically. By means of a radial MEP scan, the MEP associated with each base pair in the minor and major grooves can be characterized by a set of MEP values organized in a *n*-tuple for each groove ($\mathbf{V}_{\mathbf{minor}} = [V_i,\ldots,V_n]$ and $\mathbf{V}_{\mathbf{major}} = [V_j,\ldots,V_m]$), that can be compared through an Euclidean distance (see Fig. 1). Our biochemical interests in developing such a method is to understand the highly specific recognition between tRNAs and aminoacyl tRNA synthetases (aaRS), needed for the genetic code translation. We believe that the discrimination among several tRNAs is mainly directed by electrostatic interactions with the enzyme. As application cases, we took the recognition base pairs present in the first base pairs of the tRNA^{Ala} and tRNA^{Thr} acceptor stems[1-4]. The MEP similarity measure computed from the Euclidean distance between *n*-tuples, allowed us to obtain classifications (using hierarchical clustering) that showed good correlation with the aminoacylation activity, i.e. base pairs classified as similar showed similar activities.

When the resulting *n*-tuples of the radial MEP scan are plotted as V vs. θ, the obtained graphics give a deeper understanding of the electrostatic patterns of the grooves. For the tRNA^{Ala} variants, the V vs. θ plots revealed clear correlations between the MEP profile and the aminoacylation activity for three different base pair positions. The results for the tRNA^{Thr} also showed a good correlation with the activity, but in addition, we compare our results with crystallographic data. As the ThrRS-tRNA^{Thr} complex geometry is available at the Protein Data Bank[5], we show how the possible electrostatic interactions with the enzyme, expected from the MEP profiles, are in agreement with the X-ray crystal structure. These findings agree with the notion of *electrostatic complementarity* or *electrostatic lock*, which suggests that a more complete key-and-lock model to recreate enzyme-substrate interaction has to consider the electrostatic fit as well as the geometrical fit, since the enzyme has to find its electrostatic counterpart in the binding site in order to allow maximum interaction between molecules[5]. They also support the idea that recognition elements present in some tRNA positions allow discrimination among tRNAs, acting as “electrostatic keys” that must fit in the “electrostatic lock” placed at the enzyme recognition sites. These results go beyond the purely geometric explanations commonly suggested when studying the tRNA-aaRS recognition problem.

1. Beuning, P. J.; Gulotta, M.; Musier-Forsyth, K. Atomic Group Mutagenesis Reveals Major Groove Fine Interactions of tRNA Synthetase with an RNA Helix. *J. Am. Chem. Soc.* **1997**, 119, 8397–8402.
2. Beuning, P. J.; Nagan, M. C.; Cramer, C. J.; Musier-Forsyth, K.; Gelpi, J.-L.; Bashford, D. Efficient Aminoacylation of the tRNA^{Ala} Acceptor Stem: Dependence on the 2:71 Base Pair. *RNA,* **2002**, 8, 659–670.
3. Nagan, M. C.; Kerimo, S. S.; Musier-Forsyth, K.; Cramer, C. J. Wild-Type RNA $\text{Microhelix}^{\text{Ala}}$ and 3:70 Variants: Molecular Dynamics Analysis of Local Helical Structure and Tightly Bound Water. *J. Am. Chem. Soc.* **1999**, 121, 7310–7317.
4. Hasegawa, T.; Miyano, M.; Himeno, H.; Sano, Y.; Kimura, K.; Shimizu, M. Identity Determinants of E. coli Threonine tRNA. *Biochem. Biophys. Res. Commun*. **1992**, 184, 478-484.
5. Sankaranarayanan, R.; Dock-Bregeon, A.-C.; Romby, P.; Caillet, J.; Springer, M.; Rees, B.; Ehresmann, C.; Ehresmann, B.; Moras, D. The Structure of Threonyl-tRNA Synthetase-tRNA^{Thr} Complex Enlightens Its Repressor Activity and Reveals an Essential Zinc Ion in the Active Site. *Cell*, **1999**, 97, 371-391.
6. Náray-Szabó, G. Quantum Chemical Calculation of the Enzime-Ligand Interaction Energy for Trypsin Inhibition by Benzamidines. *J. Am. Chem. Soc.* **1984**, 106, 4584–4589.

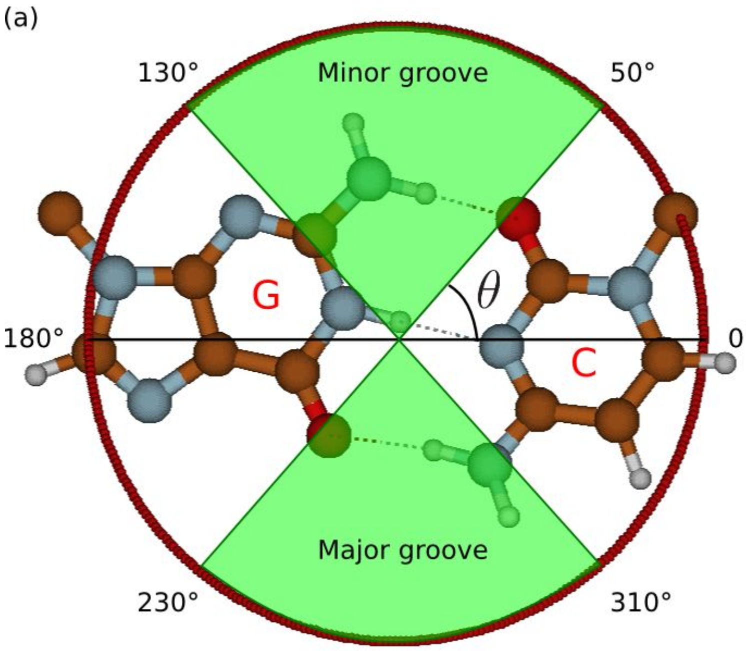


Figure 1. Radial MEP scan performed around a base pair. Potential values are computed every 1 degree according to the green interval.

P-67 : A graph theoretical approach to compare molecular electrostatic potentials

R. Marín, N. Aguirre, E. Daza, Theoretical Chemistry Group, National University of Colombia, Bogotá D. C, Colombia

In this paper we introduce a similarity measure based on negative molecular isopotential surfaces. We propose to represent the MEP through a graph, more precisely, by a weighted rooted tree that encodes some geometrical information and topological relations of successive isopotential surfaces. Trees are compared through an edit distance[1] from which a normalized similarity measure is derived. This alternative approach overcomes the difficulties of molecular alignment and avoids the definition of some particular descriptors to represent the MEP field or the molecule itself, being these its major advantages[2]. We have also implemented this method in the program *TARIS: Tree Analysis and Representation of Isopotential Surfaces,* which may be downloaded at http://taris.sourceforge.net.

As application examples we have performed the similarity study over two different sets of molecules. We took 46 small organic molecules, which represent eight different functional groups (acids, alcohols, aldehydes, amines, amides, ethers, esters and ketones). The classification by similarity using hierarchical clustering gave rise to a clear partition of molecules according to their chemical function, i.e. eight groups were obtained, each one of them corresponding to one functional group (Fig. 1a). Only three molecules were misclassified. In the second example, we studied the well known set of the 31 steroids commonly used as a reference point in several papers[3]. Structure-Activity Relationships (SAR), built by hierarchical clustering, showed a clear partition of the set in high, intermediate and low activities (Fig. 1b). Quantitative Structure-Activity Relationships (QSARs) were built by means of Partial Least Squares regressions. Similar or better results were obtained when compared with the most widely used methods in literature[2].

1. Zhang, K.; Shasha, D. Simple Fast Algorithms for the Editing Distance between Trees and Related Problems. *SIAM J. Comput.* **1989**, 18, 1245-1262.
2. Marín, R. M.; Aguirre, N. F.; Daza E. E. Graph Theoretical Similarity Approach to Compare Molecular Electrostatic Potentials. *J. Chem. Inf. Model*. **2008**, 48, 109-118.
3. Robert, D.; Amat, L.; Carbó-Dorca, R. Three-Dimensional Quantitative-Activity Relationships from Tuned Molecular Quantum Similarity Measures: Prediction of the Corticosteroid-Binding Globulin Binding Affinity for a Steroid Family. *J. Chem. Inf. Comput. Sci*. **1999**, 39, 333-344.

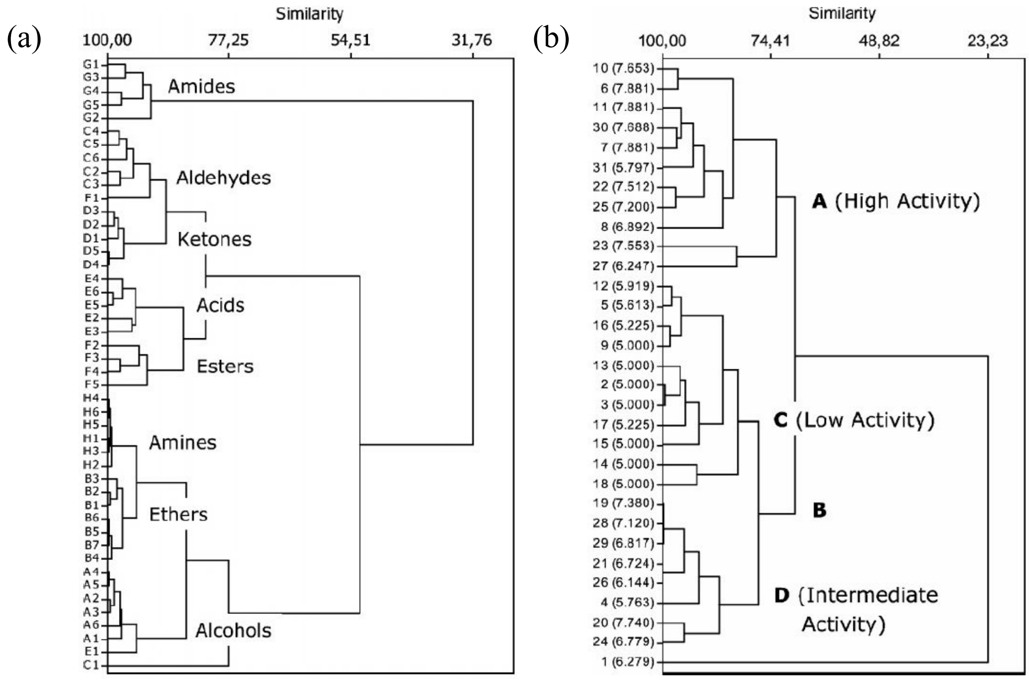


Figure 1. (a) Dendrogram obtained from the similarity matrix for the 46 organic molecules using average linkage. (b) Dendrogram obtained from the 31 steroids similarity matrix. The *logK* values corresponding to their activities are in parentheses.

P-68 : Engineering polymer informatics: Towards the computer-aided design of polymers

N. Adams, N. England, D. Jessop, P. Murray-Rust, Unilever Centre for Molecular Science Informatics, University Chemical Laboratory, Department of Chemistry, University of Cambridge, Cambridge, United Kingdom

Polymers are an important and ubiquitous class of materials and can be found in a wide variety of applications, ranging from home and personal care products to polymer pharmaceuticals. Due to changes in the way in which polymer science is being carried out (experimentation is increasingly driven by high-throughput and combinatorial approaches, combined with the development of novel synthesis techniques),1 it is becoming increasingly data-driven and is, therefore, in need of good informatics support. Unfortunately polymer informatics is almost non-existent, due to the intrinsic nature of polymers themselves: polymers are ensembles of macromolecules, all which have slightly different structures, which, in turn introduces a certain amount of fuzziness into the description of these systems and causes traditional metaphors, such as the connection table to break down.

To address this situation, we have developed Polymer Markup Language (PML)2,3 as an extension to Chemical Markup Language (CML).4 Polymer Markup Language is an extensible language, designed to support the (structural) representation of polymers and polymer-related information. It is semantically completely explicit and allows polymers to be represented at various levels of certainty. As an example, it is possible to represent an ill-defined system such as a phenol/formaldehyde resin in exactly the same way in which a well-defined polymer such as poly(styrene) could be represented. This is achieved through coarse-graining the description of the polymer, while preserving the possibility for mapping the coarse representation onto an atomistic description: in the case of poly(styrene), we are be able to expand the representation into a connection table, whereas this difficult for the phenol/formaldehyde system. At the level of PML, however, the descriptions are consistent, which, in turn allows for the comparison of polymers at different levels of certainty. As such, Polymer Markup Language also provides a level of normalization. Furthermore, the language allows a wide variety of annotations, such as group contributions, measures of reactivity and probability, which can be used to model competing reactive centres.

Furthermore, we are also currently engaged in the development of a domain ontology for polymers. Ontologies are data models representing a domain, in our case polymers, and are used to reason over objects in the domain and the relationships between them. Ontologies are valuable for structured comparative searching of knowledge, document classification, knowledge sharing between information systems and the development of machine generated hypotheses.

1. Meier, M. A. R., Schubert, U. S.; Selected Successful Approaches in Combinatorial Materials Research, Soft Matter, 2006, 2, 371-376
2. Adams, N., Murray-Rust, P.; Engineering Polymer Informatics: Towards the Computer-Aided Design of Polymers, Macromol Rapid Commun., 2008, in press
3. http://www.dspace.cam.ac.uk/handle/1810/194888
4. Murray-Rust, P., Rzepa, H.; Chemical Markup, XML and the World-Wide Web. 1. Basic Principles, J. Chem. Inf. Comp. Sci., 1999, 39, 928-942 :

P-69 : Information extraction from the polymer literature

L. Hawizy N. Adams , J. Downing, P. Murray-Rust, Unilever Centre for Molecular Science Informatics, University Chemical Laboratory, Department of Chemistry, University of Cambridge, Cambridge, United Kingdom

Polymers are an important and ubiquitous class of materials and can be found in a wide variety of applications, ranging from home and personal care products to polymer pharmaceuticals. Due to changes in the way in which polymer science is being carried out (experimentation is increasingly driven by high-throughput and combinatorial approaches, combined with the development of novel synthesis techniques),[1] it is becoming increasingly data-driven and is, therefore, in need of good informatics support. Unfortunately polymer informatics is almost non-existent, due to the intrinsic nature of polymers themselves: polymers are ensembles of macromolecules, all which have slightly different structures, which, in turn introduces a certain amount of fuzziness into the description of these systems and causes traditional metaphors, such as the connection table to break down.

To address this situation, we have previously developed Polymer Markup Language (PML)[2] as an extension to Chemical Markup Language (CML).[3] The combined use of PML and CML allows a wide variety of polymer information to be represented. Here we present an approach to the automated extraction of synthesis information from the polymer literature and the representation of this information using a combination of CML, PML and RDF,[4] with the main aim of developing repositories of polymer information. We show that chemical reaction information and sequences can be extracted from experimental sections of papers with good results using a number of different techniques. This can be as simple as “following the bold numbers” (i.e. numbers identifying reactants and products) in a reaction and creating “reactant-yields-product” graphs. Alternatively, we show how natural language processing techniques can be applied to extract not only reaction information, but also characterization and other materials data. Extracted information is then stored in a repository in a semantically rich way and can, in the future, be used for the development of novel representations of chemical reaction information, aspects of laboratory automation as well as polymer synthesis expert systems.

1. Meier, M. A. R., Schubert, U. S.; Selected Successful Approaches in Combinatorial Materials Research, *Soft Matter*, **2006**, 2, 371-376
2. Adams, N., Murray-Rust, P.; Engineering Polymer Informatics: Towards the Computer-Aided Design of Polymers, *Macromol Rapid Commun*., **2008**, in press
3. Murray-Rust, P., Rzepa, H.; Chemical Markup, XML and the World-Wide Web. 1. Basic Principles, *J. Chem. Inf. Comp. Sci*., **1999**, 39, 928-942
4. World, Wide Web Consortium, RDF Primer, **2004**, *http://www.w3.org/TR/rdf-primer/*

P-70 : MeFc and large chemical data sets

H. Y. Mussa, R.C. Glen, Unilever Centre for Molecular Informatics, Department of Chemistry, University of Cambridge, Cambridge, U.K

The recognition of patterns is fundamental to many analyses of chemical data. Recognition of patterns in the use of molecular similarity in drug discovery is a classic example. Neural Networks (NN) are typical examples of machine learning algorithms where the pattern recognition problem can be reduced to finding/approximating a mapping function $y_k = h(x_k; w_k)$ that might be able to relate x_k (the input) with

y_k (the output) in a given sample $\{(x_k, y_k)\}_{k=1}^{N}$ by estimating w (parameters) through learning/training. Over the years NN have been an invaluable tool for data analysis. The algorithm possesses excellent modelling and approximating capabilities provided that the appropriate training/learning technique is employed. Training NN with the Extended Kalman Filter (EKF) scheme gives excellent performance, but EKF requires the updating of large covariance matrices. This renders the EKF training approach computationally impractical even for networks of moderate size [1, 2, 3].

In order to address this problem Shah *et al.* [3], and Puskorieus and Feldkamap [4] proposed a family of decoupled forms of EKF. The essence of these authors' solution to the memory and CPU-time problem is to approximate the appropriate covariance matrix (with a block diagonal one) by dividing the synaptic weights into mutually exclusive groups, and retaining information between only weights within a group. However, owing to the *ad hoc* nature of the decoupling (discarding information heuristically), deterioration in the quality of resultant NN models is understandably inevitable as pointed out in Refs.[1,3]. Recently we have proposed an extended Kalman filter training technique, MeFc (Memory efficient Fully coupled Extended Kalman Filter Scheme) that a) leaves all the NN nodes connected b) has computational costs similar/cheaper to/than that of the decoupling approach and c) achieves accuracy similar to that of the EKF algorithm.

In this paper we would like to present both the method and its performance on realistic chemical data sets.

1. Haykin, S. Neural Networks: A Comprehensive Foundation (2nd edition), Prentice Hall, 1998.
2. L.Wu, L. in:Proc. IEEE Int.Conf. on Speech and Signal Processing 1, MIT Press, Glasgow 1989.
3. Shah, S.; Palmiere, F.; and Datum, M. Neural Networks 1992, 5, 779.
4. Puskorius, G. V.; and Feldkamp, L.A. in IJCNN-91-Seatle International Joint Conference on Neural Networks 1 pp.771, Seatle (1991).

P-71 : Kernel based least squares and large data sets

H. Y. Mussa· R.C. Glen, Unilever Centre for Molecular Informatics, Department of Chemistry, University of Cambridge, Cambridge, U.K

Kernel based methods (KBM) are arguably the best data analysis technique currently available. Unlike Neural Networks in which, besides a global minimum, several local minima exist, a Kernel based fitting/classifying problem is a convex optimization problem with a single minimum. However, finding this minimum (and in doing so yielding optimal parameters of a given observational model) in practice requires the manipulation, such as inversion, of large matrices.

This has been challenging even when the number of data points is just over a few thousands [1,2].

The well established direct methods for updating, or inverting huge matrices fail due to the expense of a large increase in core-memory storage and CPU-time cost, even for moderate size problems. The root of the problem is that direct methods have O(N^2) core memory storage requirement and the CPU-time scales as O(N^3), where N is the dimension of the matrix (the number of data points, here). Despite the advances in computer power, ``conventional" computers can only solve relatively small problems ($N \approx 10^4$ to 10^5).

In this paper we would like to present a computationally efficient training scheme for KBM for obtaining the global minimum. Some preliminary results on chemical data sets will also be presented.

1. Chua,K. S. *Pattern Recognition Letters* **2003**, 24,75.
2. Mangasarian, O. L.; and Musicant, D. R. *J. Mach. Learn. Res.* **2001**, 1, 161

P-72 : Molecular spam: Use of a modified spam filter for classification of bioactive molecules and drug target prediction

F. Nigsch, J.B.O. Mitchell, Unilever Centre for Molecular Science Informatics, Department of Chemistry, Cambridge, UK

Spam filtering on the internet is a challenging task in order to identify those messages that are legitimate and contain useful information, considering that 90% of all global email traffic may be spam.

Computational methods used to support drug discovery are faced with similar problems: Which out of tens of thousands of possible molecules will show a similar biological activity to a handful of known and promising examples?

To discriminate between biologically active and inactive molecules from large *in silico* virtual libraries, we have converted a highly accurate spam filter. The adaptations introduced allow us to use the method for the prediction of the likely protein targets of druglike organic molecules, and for the identification of the molecular fragments giving rise to each observed bioactivity.

Our results reveal an error rate of below 4% on a dataset of 8,500 molecules in 11 classes, including HIV protease inhibitors, protein kinase C inhibitors and GPCR-binding ligands. Addition of 95,000 inactive molecules as decoys ("molecular spam") results in a decrease in accuracy to just below 90%. We have followed this up with an even more challenging experiment, in which we show that our method can predict the correct protein target, from almost 200 possibilities, for 70% of 43,000 molecules.

A straightforward identification of important molecular fragments contributing to the bioactivity of each class is possible due to the nature of the underlying algorithm. [1] The method may therefore be employed to gain structural insights, and we show how this can be used to colour-code molecules in a chemically intuitive way, displaying the propensity of each fragment to confer a given bioactivity.

1. Littlestone, N. Learning Quickly When Irrelevant Attributes Abound: A New Linear-threshold Algorithm. *Machine Learning* **1988**, 2, 285–318.

P-73 : SPECTRa-T: Machine-based data extraction and semantic searching of chemistry e-theses

J Townsend[1], *J Downing*[1], *M Harvey*[2], *P Morgan*[3], *P Murray-Rust*[1], *H Rzepa*[2], *D Stewart*[1], *A Tonge*[1]

[1] *Unilever Centre for Molecular Science Informatics, Cambridge, UK*

[2] *Imperial College, London, UK*

[3] *Cambridge University Library, Cambridge, UK*

Chemical theses typically contain a wealth of experimental data which is currently untapped. Data such as chemical names, spectra and spectral assignments are not routinely captured and exposed to search tools, and are typically stored without being subjected to appropriate preservation techniques which would enable data re-use. The SPECTRa-T (Submission, Preservation & Exposure of Chemistry Teaching & Research Data from Theses) project has developed text- and data-mining tools to extract named chemical entities (NCEs) and chemical objects (COs) e.g. spectral assignments and physical chemistry properties, from electronic theses (e-theses); where appropriate COs are associated with a particular NCE (figure 1). Semantic Web standards for searching data have been developed by the W3C and are being increasingly adopted by research and development laboratories in the pharmaceutical industry [1]. The extracted information is deposited in a persistent Resource Description Framework [2] (RDF) triple-store which enables users to conduct semantic searches. The level of sophistication of such searches far exceeds that of a normal free-text search and RDF is more easily extensible then SQL databases.

SPECTRa-T uses OSCAR [3] to identify the NCEs and COs which requires the documents to be converted to SciXML [4]. Portable Document Format (PDF) is the *de facto* format for the majority of repository deposition and we initially developed software to extract NCEs from PDF e-theses. We show that the loss of formatting introduced by conversion to PDF makes it impossible for machines to construct semantically rich SciXML from e-theses. Although machines can identify NCEs and COs the loss of formatting prevents any associations between them. We also show that a significant number of false-positives and false-negatives may be identified as a result of data corruption.

Office Open XML [5] (OOXML) is a XML-based file format specification for electronic documents developed by Microsoft. We demonstrate that it is possible to construct semantically rich SciXML from OOXML which enables significant further processing – such as the automatic reconstruction of reaction pathways in RDF (figure 2) as well as associating COs to NCEs. The COs currently identified are: NMR (proton and carbon) and IR spectra, (high-resolution) mass spectrum, melting/boiling points, elemental analysis, optical rotation, R_f values, physical description and yield. All COs are parsed to CML [6] as which allows automated validation of the data and placed in a repository. OOXML also allows chemical diagrams and reaction pathways embedded in the document to be retrieved as binary objects which are converted to

CML and stored.

1. Mullin, R.; The Big Picture, *Chem. Eng. News.* **2007,** 40, 13-17.
2. Resource Description Framework (RDF), http://www.w3.org/RDF/
3. Batchelor, C. R.; Corbett, P. T.; Semantic enrichment of journal articles using chemical named entity recognition. *In Assoc. Comp. Linguistics Companion Volume* 2007 45-48.
4. Copestake, A.; Corbett, P.; Murray-Rust, P.; Rupp, C. J.; Siddharthan, A.; Teufel, S.; Waldron, B.; An Architecture for Language Processing for Scientific Texts, *Proc. UK e-Science. Prog. AHM 2006.*
5. Introducing the Office (2007) Open XML File Formats, http://msdn2.microsoft.com/en-us/library/aa338205.aspx
6. P. Murray-Rust, P.; Rzepa, H. S.; Wright, M.; Development of Chemical Markup Language (CML) as a System for Handling Complex Chemical Content, *New J. Chem.*, **2001**, 25, 618-634.

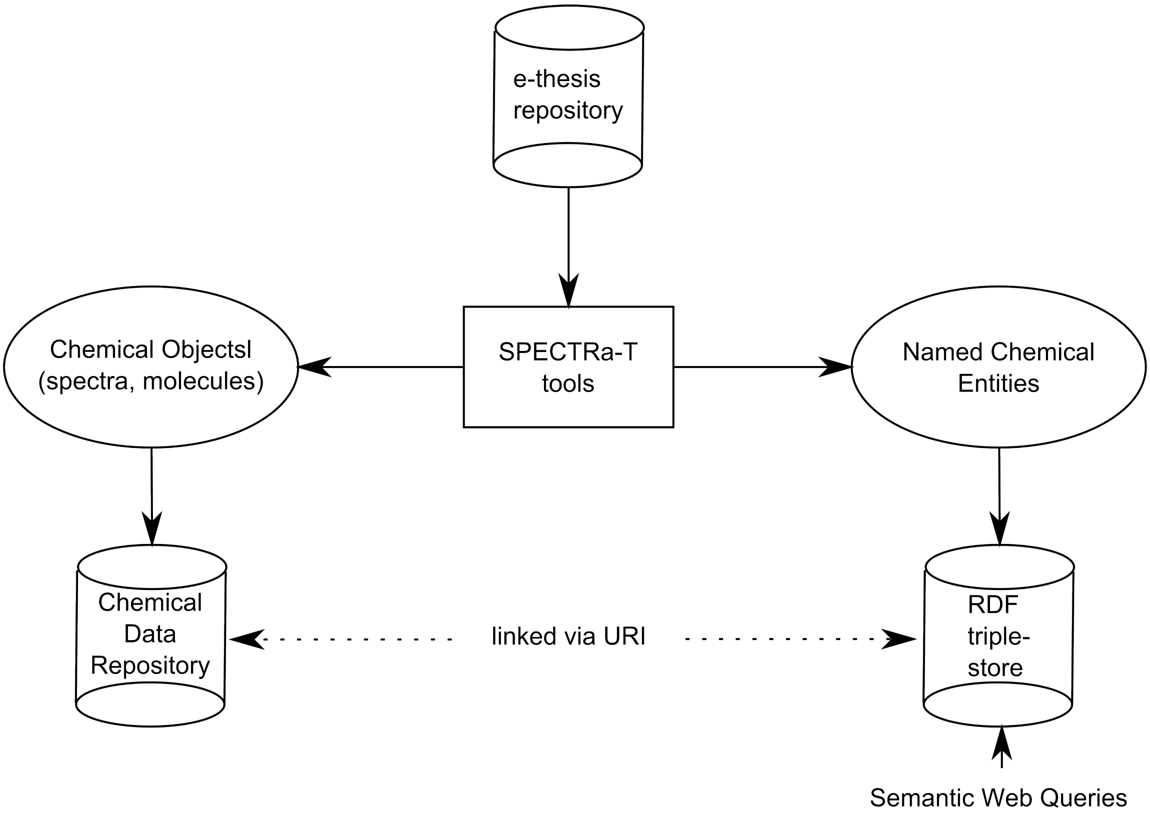


Figure 1: Overview of the SPECTRa-T data-mining architecture

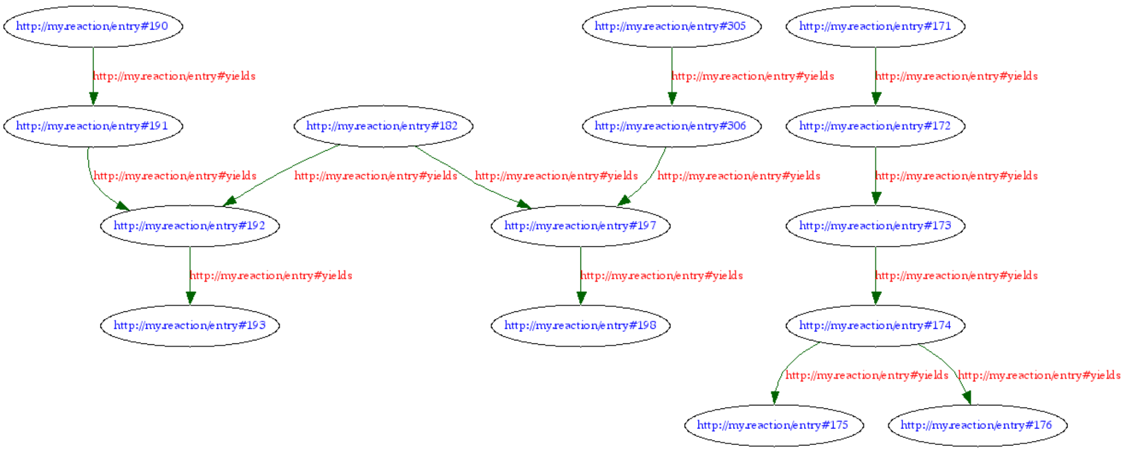

http://my.reaction/entry#190
http://my.reaction/entry#305
http://my.reaction/entry#171
http://my.reaction/entry#yields
http://my.reaction/entry#yields
http://my.reaction/entry#yields
http://my.reaction/entry#191
http://my.reaction/entry#182
http://my.reaction/entry#306
http://my.reaction/entry#172
http://my.reaction/entry#yields
http://my.reaction/entry#yields
http://my.reaction/entry#yields
http://my.reaction/entry#yields
http://my.reaction/entry#yields
http://my.reaction/entry#192
http://my.reaction/entry#197
http://my.reaction/entry#173
http://my.reaction/entry#yields
http://my.reaction/entry#yields
http://my.reaction/entry#yields
http://my.reaction/entry#193
http://my.reaction/entry#198
http://my.reaction/entry#174
http://my.reaction/entry#yields
http://my.reaction/entry#yields
http://my.reaction/entry#175
http://my.reaction/entry#176

Figure 2: A part of the automatically generated reaction pathway from an organic PhD thesis

P-74 : Creating chemo- & bioinformatics workflows: Further developments within the CDK-Taverna project

Thomas Kuhn [1,2], , Achim Zielesny [2] and Christoph Steinbeck [1]

[1] Cologne University Bioinformatics Center (CUBIC), Cologne, Germany

[2] University of Applied Sciences of Gelsenkirchen, Institute for Bioinformatics and Chemoinformatics, Recklinghausen, Germany

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] and Bioclipse[3].

Pipelining or workflow tools allow for the Lego™-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.
Current developments in CDK-Taverna focus on a soft computing framework which allows a flexible use of different methods from, for example, the WEKA[4] library. Here, properties of chemical substances may be calculated using descriptors from the QSAR / QSPR package of the Chemistry Development Kit (CDK).

Further, a reaction enumeration algorithm for combinatorial chemistry based on existing methods of the Chemistry Development Kit is being developed. This algorithm allows for the enumeration of a reaction given that reactants and products are provided as “Markush” structures.

1. Oinn T, Addis M, Ferris M, Marvin D, Senger M, Greenwood M, Carver T, Glover K, Pocock M, Wipat A and Li P. Taverna: A tool for the composition and enactment of bioinformatics workflows Bioinformatics Vol. 20(17) pp 3045-3054, 2004
2. Steinbeck C, Han YQ, Kuhn S, Horlacher O, Luttmann E, Willighagen E. The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics. J Chem Inf Comput Sci 2003; 43: 493-500
3. Spjuth, O., Helmus, T., Willighagen, E. L., Kuhn, S., Eklund, M. et al. Bioclipse: An open rich client workbench for chemo- and bioinformatics, BMC Bioinformatics 2007
4. Witten I. H. and Frank E. Data Mining: Practical machine learning tools and techniques. 2nd Edition, Morgen Kaufmann, San Francisco, 2005
5. Hassan, M., Brown, R. B., Varma-O`Brien, Rogers, D. Cheminformatics analysis and learning in a data pipelining environment, Molecular Diversity 2006; 10: 283-299

P-75 : Protein-protein interactions as targets for drugs: A view from the binding site

Richard Jackson, Jonathan Fuller, Nicholas Burgoyne, Institute of Molecular and Cellular Biology, Faculty of Biological Sciences, University of Leeds, Leeds, UK

The ability to control protein-protein interactions therapeutically is of great current interest due to the many important processes involving these interactions. We have taken a structural informatics perspective to analyse pockets on protein surfaces[1,2], likely to correspond to binding ‘hot-spots’ according to properties thought to be important in stabilizing the native complex. This includes, sequence conservation and measures of physical properties including hydrophobicity, desolvation, electrostatic and van der Waals potentials. The resulting differences between predicting binding-sites at protein–protein and protein–ligand interfaces are striking. Generally, the prediction accuracy for protein–protein interfaces is much lower[3].

We have further addressed the differences between proteins that bind marketed drugs and the proteins that are targeted by small molecule protein-protein interaction inhibitors. It is observed that the former bind deeper within the contact surface of the target protein, with higher ligand efficiencies. In addition, conformational change on ligand binding plays an important role in the druggability of specific protein-protein interaction targets. In line with some recent successes in this field our results suggest that drug discovery methods that target several pockets are likely to increase the chances of success in this new field of therapeutics.

1. Laurie, A.T., Jackson, R.M. Methods for the prediction of protein-ligand binding sites for structure-based drug design and virtual screening. *Current Protein and Peptide Science* **2006**, 7, 395-406.
2. Laurie, A.T., Jackson, R.M. Q-SiteFinder: an energy-based method for the prediction of protein-

ligand binding sites. *Bioinformatics* **2005**, 21, 1908-1916.

3. Burgoyne, N.J., Jackson, R.M. Predicting protein interaction sites: Binding hot-spots in protein-protein and protein-ligand interfaces. *Bioinformatics* **2006**, 22, 1335-1342.

P-76 : Determinants for selectivity in map kinase inhibitors by computational simulations

Nikita Basant, Maria Christina Menziani, Department of Chemistry, University of Modena

P38 MAPKs (Mitogen Activated Protein Kinase) are currently among the most important target classes for drug development as they regulate signal transduction by phosphorylating tyrosine, threonine, and serine residues in key proteins involved in signal pathways featuring relevance in many diseases.[1] The pathophysiological dysfunction of protein kinase signaling pathways underlies the molecular basis of many cancers and of several manifestations of cardiovascular disease, such as hypertrophy and other types of left ventricular remodeling, ischemia/reperfusion injury, angiogenesis, and atherogenesis.[2]

Over the past 10 years there has been a dramatic escalation in the development of kinase inhibitors for the treatment of many chronic disease and other conditions. Because all Protein Kinases use ATP as cofactor, they share a highly conserved ATP binding pocket, which is the molecular binding site of most inhibitors. Preferred interaction patterns of ligands with amino acid residues of proteins in and near the biding site cavity are of utmost importance for the prediction of binding modes and structural adaptations.

However different kinases regulate very closely related cellular events some of which may be vital for survival, hence, the selectivity of inhibitors for kinase families is very critical to avoid unwanted side effects. Here we compare and analyze "in silico" the binding interactions of the inhibitor in the binding pocket of the kinase proteins therefore determinants influencing selective binding of inhibitors for P38 and JNK3 classes of kinase proteins are being investigated. A computational characterization of their unique binding interactions is discussed.

References:

1. Kung, C.; Shokat, K. M. Small-molecule kinase-inhibitor target assessment. *ChemBioChem* **2005**, *6*, 523-526.
2. Cohen P. Protein kinases: the major drug targets of the twenty-first century? *Nat Rev Drug Disc*. **2002**, 1, 309–315.

P-77 : Fragment weighting schemes for similarity-based virtual screening: Use of occurrence weighting

S. Arif, J. Holliday, P. Willett, University of Sheffield, Sheffield, United Kingdom

The calculation of molecular similarity using 2D fingerprints is one of the most important methods currently available for ligand-based virtual screening [1]. In a similarity search, the similarity is computed between a reference structure of known biological activity and each of the structures in a database. The similarity is measured by identifying the substructures common to fingerprints of the two molecules that are being compared, then computing the value of a similarity coefficient, such as the Tanimoto coefficient, between the reference structure and the current database structure.

A 2D fingerprint has traditionally been a binary vector that encodes the presence or absence of topological substructures in a molecule, but there is no reason why this should necessarily be the case. Instead, it is possible to assign weights to fragments that describe their relative degree of importance in the molecules in which they occur. There are various ways in which weighting can be carried out [2]: here, we compare occurrence-based representations (encoding how many times a fragment substructure occurs in a molecule) with incidence-based representations (encoding merely the presence or absence of a fragment substructure). Previous work has suggested that the former are to be preferred, but the few studies that have been carried out have all been quite limited in scope. Here, we have used five different fragment weighting schemes: simple binary weighting (W1), the raw number of occurrences (W2), and then three variants of W2: the log (W3) and the square root (W4) of the occurrences and then a more complex variant that has been used to normalize the occurrence of words in text retrieval systems (W5) [3]. These weights have been applied to molecules from the MDL Drug Data Report (MDDR) database represented by molecular holograms, vectors in which each element contains the number of times that a specific bit has been set by a fragment-hashing scheme, and the resulting representations used in similarity searches for eleven MDDR activity classes. Future work that will have been completed by the time of the conference in June will have also used different types of fragment and different databases.

Each of the five different weighting schemes can be applied to the reference structure and to each of the database structures, giving a total of 25 possible similarity measures for the searches: here, we have considered all those 19 schemes that involve either W1 or W2. Ten representative molecules were chosen from each activity class to be the reference structures for searching, and a note made of the number of actives retrieved in the top-5% of the ranking resulting from the similarity search; the results were then averaged over the ten reference structures for each activity class. Initial results are shown in Table 1, where it will be seen that the best results are obtained when the reference structure and the database structures are both encoded using W2 or W3, a finding that is confirmed by Kendall's W test of statistical significance. Our results to date hence suggest that the use of fragment occurrence data can significantly enhance the effectiveness of similarity-based virtual screening systems.

1. Willett, P. Similarity-Based Virtual Screening Using 2D Fingerprints." *Drug Discov. Today*, **2006**, *11*, 1046-1053.
2. Willett, P.; Winterman, V. A Comparison of some Measures of Inter-Molecular Structural Similarity. *Quant. Struct.-Activ. Relat.*, **1986**, *5*, 18-25.
3. Salton, G.; Buckley, C. Term-Weighting Approaches in Automatic Text Retrieval. *Inf. Proc. Manag.*, **1988**, *24*, 513-523.

Table 1. Mean numbers (averaged over searches for ten different reference structures) of actives retrieved in the top-5% of
the ranked database in searches for eleven MDDR activity classes (denoted by abbreviated names in the top row of the table).
The right-hand columns give the mean numbers of actives retrieved, and the mean ranks when the different similarity measures
are ranked in decreasing order of numbers of actives retrieved.

Similarity measure[1]	Activity class											Mean	
	5HT3	5HT1	5HT	D2	REN	ANG	THR	SUBP	HIV	COX	PKC	Actives	Rank
M11	107.7	83.6	33.8	29.6	421.0	231.2	89.7	119.3	118.5	29.2	64.9	120.8	11.27
M12	83.0	48.4	24.4	19.6	316.3	270.0	73.1	122.4	92.7	21.5	90.8	105.7	13.8
M13	132.8	137.3	47.1	51.9	518.6	201.6	97.3	194.1	111.0	51.3	55.7	143.3	6.7
M14	102.3	70.3	31.3	25.7	368.4	253.2	83.2	118.1	104.4	24.0	80.1	114.6	12.6
M15	129.9	125.1	42.2	47.9	510.9	209.2	110.9	161.5	120.1	45.8	53.2	141.5	7.4
M21	145.9	95.8	36.9	33.2	111.3	84.0	88.2	19.4	37.7	45.1	20.6	65.2	12.7
M22	150.9	117.4	36.2	46.8	788.2	321.5	115.4	208.6	159.8	54.9	59.9	187.2	3.9
M23	133.0	123.6	37.1	45.1	338.7	120.2	89.2	97.3	55.1	58.2	37.9	103.0	10.6
M24	155.4	127.9	36.0	47.6	448.3	203.6	112.8	133.7	88.3	54.2	46.3	146.0	7.9
M25	133.1	85.4	35.4	31.0	78.7	49.0	65.4	13.2	23.6	43.8	17.4	52.4	14.8
M31	93.0	71.5	27.0	27.4	412.4	246.0	85.9	164.9	124.3	31.3	74.7	123.5	11.4
M32	69.9	53.7	21.0	23.7	244.1	237.6	75.7	198.9	105.2	16.3	86.3	103.2	13.6
M33	134.7	139.8	41.6	51.2	726.6	294.2	118.5	201.4	141.6	53.7	63.7	178.8	3.5
M41	144.4	117.1	38.7	40.9	494.2	254.3	116.5	124.2	118.1	50.4	41.4	140.0	7.4
M42	120.6	84.5	32.8	28.8	459.3	339.0	92.6	201.2	139.1	33.2	74.4	132.2	8.4
M44	139.2	107.0	39.6	40.2	659.5	330.0	111.4	204.0	144.9	45.5	56.7	170.7	5.4
M51	87.2	52.8	26.2	18.4	273.4	220.6	74.9	81.3	87.3	21.1	82.5	93.2	15.0
M52	94.6	46.7	20.9	17.0	269.1	251.4	73.9	152.1	84.4	21.0	102.6	103.1	14.5
M55	112.9	88.6	33.6	33.3	413.5	274.0	92.8	166.1	129.6	31.6	61.9	130.7	9.0

[1] Each similarity measure, denoted by M*ab*, describes the weight applied to the database structures' fingerprints (*a*) and the weight applied to the reference structure's fingerprint (*b*): thus M13, e.g., refers to the searches in which the database structures are coded using W1 (conventional binary weighting) and the reference structures are coded using W3 (the natural logarithm of the occurrence frequencies).

P-78 : Effect of data standardization on the clustering of chemical structures

C.-W. Chu· J. Holliday, P. Willett, University of Sheffield, Sheffield, United Kingdom

The clustering of chemical structures is of importance in several areas of chemoinformatics [1]. A little-discussed aspect of clustering is standardisation, the application of a mathematical transformation to each of the individual components of a multi-attribute descriptor so that all of the attributes make a comparable contribution to the measurement of similarity. Here, we report a detailed comparison of the effectiveness of different standardisation methods when applied to chemical datasets.

Our experiments used two datasets: 10191 molecules from the MDDR database and 11607 molecules from the IDAlert database, with molecules being noted as active or inactive in eleven bioactivity classes. The molecules in the two datasets were characterised in three ways: 12 physicochemical properties (PipelinePilot software); 53 topological indices (Molconn-Z software) and 998-element molecular holograms (Unity software). Each of the three characterisations was used to generate eight different representations, these being the raw data vectors and the vectors resulting from application of the seven different standardisation methods described in the review by Milligan and Cooper [2]. The eight resulting representations for each descriptor-type were then clustered using the K-Means and Ward's methods (Digital Chemistry clustering software), generating classifications with 25, 50 or 100 clusters.

The extent to which molecules with the same bioactivity occurred in the same clusters was determined in two ways: an entropy measure based on the distribution of actives across the clusters (with a good standardisation method being one that minimised the spread of actives); and a measure based on the numbers of inactive molecules occurring in clusters that contained active molecules (with a good standardisation method being one that minimised the numbers of inactives in such clusters). Each measure was calculated for each activity class and the results averaged over the eleven activity classes to quantify the effectiveness of a given standardisation method.

The resulting mean values of each of the two measures were then used to rank the eight standardisation methods in order of decreasing effectiveness, for each of the two datasets, two clustering methods and three numbers of clusters. The extent of the agreement between the rankings engendered by the three different characterisations was measured using the Kendall W test. A significant degree of agreement at the 0.05 level of statistical significance was obtained for none of the 12 analyses on the MDDR data, and for only 3 of the 12 analyses on the IDAlert data. This suggests that there is no consistent pattern when the various standardisation methods are ranked in order of decreasing effectiveness, and hence that there is no obvious performance benefit that is likely to be obtained from the use of any particular method.

1. Downs, G. M.; Barnard, J. M. Clustering Methods and their Uses in Computational Chemistry. Rev. Comp. Chem., 2002, 18, 1-40.
2. Milligan, G. W.; Cooper, M. C. A Study of Standardization of Variables in Cluster Analysis. J. Classif., 1988, 5, 181-204.

P-79 : Multiobjective optimisation of pharmacophore hypotheses: Bias towards low-energy conformations

V Gillet[1], *E Gardiner*[1], *D Cosgrove*[2], *R Taylor*[3]

[1] *University of Sheffield, Sheffield, UK*

[2] *AstraZeneca, Alderley Park, UK*

[3] *Cambridge Crystallographic Data Centre, Cambridge, UK*

Pharmacophore elucidation is a difficult problem involving the determination of the interactions between a small molecule and a protein without knowledge of the 3D structure of the protein. Given several small molecules that are known to bind to the protein, the aim is to generate an alignment such that their common features are overlaid. However, it the absence of the receptor site, it is unlikely that the true pharmacophore can be determined unambiguously. Thus we have developed a Multi-Objective Genetic Algorithm (MOGA) with the aim of generating multiple feasible pharmacophore hypotheses by exploring trade-offs in conformational energy, volume overlay and number and quality of pharmacophore points [1,2]. In previous work we have demonstrated the ability of the MOGA to identify pharmacophores that are consistent with existing knowledge together with alternative hypotheses that, in the absence of knowledge of the true pharmacophore, are equally feasible. However, the search space for multiple potential pharmacophores can become very large and generally increases with the number of molecules being considered. In this work we have combined a clique-detection algorithm with the MOGA in order to limit the MOGA exploration to a feasible region of solution space to increase both the efficiency and effectiveness of the program. In a further enhancement we bias the search towards low-energy conformations using torsion angles obtained from surveys of crystal structures [3]. We report the results of these enhancements in terms of speed, solution quality and solution diversity on datasets of up to ten molecules.

1. Cottrell, S.J.; Gillet, V.J. and Taylor, R. Incorporating Partial Matches within Multiobjective Pharmacophore Identification *J. Comput-Aided Mol. Design*, **2006**, 20, 735-749.

2. Cottrell, S.; Gillet, V.J.; Taylor R. and Wilton, D. Generation of Multiple Pharmacophore Hypotheses Using Multiobjective Optimisation Techniques *J. Comput-Aided Mol. Design*, **2004**, 18, 665-682.
3. Bruno, I.J.; Cole, J.C.; Kessler, M.; Luo, J.; Motherwell, S; Purkis, L.H.; Smith, B.R.; Taylor R. Retrieval of Crystallographically-Derived Molecular Geometry *J. Chem. Inf. Comput. Sci.,* **2004**, 44, 2133 -2144.

P-80 : Weighted data fusion with turbo similarity searching to improve chemical similarity searching

J. Holliday [1], J. Chen [1], J. Bradshaw [2]

[1] University of Sheffield, Sheffield, United Kingdom

[2] Daylight Chemical Information Systems, Inc., United Kingdom

Similarity searching is perhaps the simplest tool available for ligand-based virtual screening of chemical databases, requiring just a single known bioactive molecule, the reference or target structure, as the starting-point for a database search. The most common similarity search involves the use of a simple association coefficient, normally the Tanimoto coefficient, with a 2D fragment bit-string representation of molecular structure. More recently, data fusion in similarity searching has emerged which uses more than one coefficient to evaluate the similarity between the target structure and the database structures [1-3]. In addition, using multiple reference structures with group fusion (or turbo similarity searching) has also been applied [4, 5] with considerable success.

In this presentation, we first conclude that four coefficients: Forbes, Simple Matching, Tanimoto and Russell/Rao; are the most suitable coefficients to use in data fusion in the context of similarity searching due to the complementary nature of their individual performances. We then implement a systematic approach to optimising the weightings applied to the four coefficients in data fusion process. The approach uses the turbo similarity search methodology in the training and testing stages. All three fusion-rules are studied: MIN, MAX and SUM.

We divided the MDL Drug Data Report database into two parts; one for training and one for testing. Using targets from selected active classes, we conducted several turbo similarity searches of the training set, varying the weights systematically and scored the weighting scheme by the proportion of active nearest neighbours. We then applied these optimum weights, using different targets from the active classes, to turbo searches of the test set. We also carried out turbo similarity searches using the industry standard Tanimoto coefficient on its own, for comparative purposes. The improvement rates over Tanimoto for the testing stage are shown in Table 1. These figures show a clear improvement in almost all cases over the Tanimoto. Once training has been carried out for an active class, there is little extra computation cost in carrying out similarity searches of this type. In addition, the weights can be optimised further as more actives from each respective class are identified.

A selection of optimum weights for four of the classes tested is given in Table 2. This indicates that the targets from activity classes which comprise compounds which are generally larger in size than the database average show better performance when combinations involve the Tanimoto and Russell/Rao whilst, for targets from classes which comprise smaller compounds, the Forbes and Simple Match are more effective.

1. Holliday, J. D.; Hu, C.-Y.; Willett, P. (2002). Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bit-strings. Comb. Chem. High Throughput Screening, 5, 155-166.
2. Salim, N.; Holliday, J.D.; Willett, P. (2003). Combination of fingerprint-based similarity coefficients using data fusion. J. Chem. Inf. Comput. Sci. 43, 435-442.
3. Whittle, M.; Gillet, V. J.; Willett, P.; Alex, A.; Loesel, J. (2004). Enhancing the effectiveness of virtual screening by fusing nearest neighbor lists: A comparison of similarity coefficients. J. Chem. Inf. Comput. Sci. 44, 1840-1848.
4. Hert, J.; Willett, P.; Wilton, D.J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer A. (2005). Enhancing the effectiveness of similarity-based virtual screening using nearest-neighbour information. J. Med. Chem. 48, 7049-7054.
5. Hert, J., Willett, P., and Wilton, D. J. (2006). New methods for ligand-based virtual screening: use of data fusion and machine learning to enhance the effectiveness of similarity searching. J. Chem.

Inf. Model. 46, 462-470.

Table 1: Improvement rates over the Tanimoto coefficient at the testing stage

Fusion Method	**MIN**		**MAX**		**SUM**	
	Avg.	S.D.	**Avg.**	S.D.	**Avg.**	S.D.
Renin Inhibitor	**7.70**	4.72	**7.60**	4.75	**9.30**	3.34
Angiotensin II AT1 Antagonist	**33.65**	12.87	**39.10**	9.39	**40.40**	10.61
Thrombin Inhibitor	**111.60**	97.40	**42.40**	36.33	**113.15**	100.80
Substance P Antagonist	**5.15**	6.77	**3.25**	8.59	**4.00**	6.79
5HT3 Antagonist	**35.40**	68.06	**15.70**	13.64	**22.80**	21.02
5HT1A Agonist	**61.90**	50.02	**33.60**	19.29	**48.30**	31.77
5HT Reuptake Inhibitor	**12.40**	17.95	**-0.30**	14.38	**0.75**	27.07
HIV-1 Protease Inhibitor	**270.90**	436.05	**80.15**	96.34	**13.90**	15.53
D2 Antagonist	**0.00**	0.00	**40.10**	23.00	**29.40**	22.77
Coclooxygenase Inhibitor	**-2.40**	14.09	**23.00**	25.82	**7.70**	6.69
Protein Kinase C Inhibitor	**4.70**	4.84	**7.80**	14.14	**13.90**	9.52
Overall Average	**49.20**	**64.80**	**24.50**	**24.17**	**27.60**	**23.26**

Table 2: Selection of optimum weights for various active classes

Fusion method/Class/Compound #	**For Wt**	**SM Wt**	**Tan Wt**	**RusWt**
TMIN/Renin Inhibitor/157104	0.0	0.0	0.4	0.6
TMAX/Renin Inhibitor/157104	0.0	0.0	0.5	0.5
TSUM/Renin Inhibitor/157104	0.0	0.0	0.75	0.25
TMIN/Thrombin Inhibitor/245228	0.0	0.0	0.4	0.6
TMAX/Thrombin Inhibitor/245228	0.0	0.0	0.25	0.75
TSUM/Thrombin Inhibitor/245228	0.0	0.0	0.25	0.75
TMIN/5HT1A Agonist/156667	0.5	0.0	0.5	0.0
TMAX/5HT1A Agonist/156667	0.25	0.25	0.5	0.0
TSUM/5HT1A Agonist/156667	0.0	0.75	0.25	0.0
TMIN/5HT Reuptake Inhibitor/272569	0.0	0.4	0.2	0.4
TMIN/5HT Reuptake Inhibitor/272569	0.25	0.75	0.0	0.0
TMIN/5HT Reuptake Inhibitor/272569	1.0	0.0	0.0	0.0

P-81 : Using wavelets to represent GRID fields in virtual screening

R Martin [1], V Gillet [1], E Gardiner [1], S Senger [2]

[1] Dept. of Information Studies, University of Sheffield, Sheffield, UK

[2] GlaxoSmithKline, Molecular Discovery Research, Computational and Structural Chemistry, Medicines Research Centre, Stevenage, UK

One of the most revealing three-dimensional descriptors available for analysis of small molecules is the *molecular interaction potential* (MIP). Perhaps the most commonly used of these is the *GRID* field [1], which is comprised of a discrete grid placed over the ligand for which potential interaction energies between the molecule and a *probe group* (e.g. water) are calculated at each vertex. A disadvantage of such a field is its large size and hence the demanding nature of the computations required. One way in which this is overcome is to extract features into a linear fingerprint (e.g. GRIND descriptors). However, this results in a loss of information and this requires the selection of an appropriate set of parameters. Here we use wavelets to encode the entire field in a holistic manner.

We show that the nonstandard *Daubechies 4-tap wavelet transform* (D4 WT) can be exploited to represent finely sampled GRID maps requiring only 1% of the storage of the original fields. This reduced representation can be used without loss of accuracy in ligand-based similarity searching, compared with using the whole field. Nearly identical search results were observed when searching over sets of CDK2, ESR1 and HIV inhibitors also used by Chen et al [2]. We also describe the impact of wavelet approximation methods upon the retrieval of active compounds from amongst a large number of decoys procured from the *DUD* [3].

1. Goodford, P.J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *Journal of Medicinal Chemistry* **1985**, 28 (7), 849-857.
2. Chen, Q.; Higgs, R.E. & Vieth, M. Geometric Accuracy of Three-Dimensional Molecular Overlays. *Journal of Chemical Information and Modelling* **2006**, 46 (5), 1996-2002.
3. Huang, N.; Shoichet, B.K. & Irwin, J.J. Benchmarking Sets for Molecular Docking. *Journal of Medicinal Chemistry* **2006**, 49 (23), 6789-6801.

P-82 : A multiobjective approach to scoring functions for docking

I. Mott[1], P. Gedeck[2], V. Gillet[1]

[1] Department of Information Studies, University of Sheffield, Sheffield, UK

[2] Novartis Institutes for Biomedical Research, Novartis Horsham Research Centre, West Sussex, UK

We describe work undertaken to develop a multiobjective approach to scoring function optimisation for use with the docking problem.

The limitations of current docking and scoring protocols are well documented[1,3,4]. Namely, the failure to correctly prioritise compounds according to their known binding affinities. Central to this dysfunction is the inability of the scoring functions employed to work universally given the diverse range of protein systems studied. We therefore argue a more targeted approach to the scoring problem is justified.

In previous studies[2,5] negative training data has been employed in scoring function optimisation, in which a genetic algorithm optimises a function so that it ranks a known binding mode in preference to a set of decoys poses. This approach has the advantage of explicitly encoding information of disfavoured interactions, whilst not relying on potentially incomparable experimental affinity values, as per the typical regression-based scoring function optimisation. However, these studies aimed to produce functions with general applicability, which ultimately suffer the same shortcomings as many of the commonly used scoring functions.

Using a multiobjective evolutionary algorithm (MOEA), we demonstrate a scoring function optimisation protocol that extends the previously reported method. We utilise multiple proteins simultaneously, with the aim of producing ‘customised’ empirical scoring functions for different proteins and protein classes for use in high-throughput virtual screening applications.

1. Coupez, B. & Lewis, R.A. (2006). "Docking and scoring - Theoretically easy, practically impossible?" Current Medicinal Chemistry, 13, 2995-3003.
2. Fenu, L.A. (2007). The development of novel scoring methods for virtual screening. P.h.D., University of Southampton.
3. Kontoyianni, M., McClellan, L.M. & Sokol, G.S. (2004). "Evaluation of docking performance: comparative data on docking algorithms". Journal of Medicinal Chemistry, 47 (3), 558-565.
4. Leach, A.R., Shoichet, B.K. & Peishoff, C.E. (2006). "Prediction of protein-ligand interactions. Docking and scoring: successes and gaps". Journal of Medicinal Chemistry, 49 (20), 5851-5855.
5. Smith, R., Hubbard, R.E., Gschwend, D.A., Leach, A.R. & Good, A.C. (2003). "Analysis and optimization of structure-based virtual screening protocols: (3). New methods and old problems in scoring function design". Journal of Molecular Graphics and Modelling, 22 (1), 41-53.

P-83 : Neighbourhood behaviour studies for lead optimisation

G. Papadatos[1], V. Gillet[1], P. Willett[1], I. McLay[2], T. Cooper[2], S. Macdonald[2], S. Pickett[2]

[1]Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield, Sheffield, UK

[2]GlaxoSmithKline Medicines Research Centre, Stevenage, Hertfordshire, UK

The *similar property principle* is a well-established heuristic which has been the rationale for numerous Chemoinformatics and Medicinal Chemistry applications. It states that similar compounds tend to exhibit similar properties, and therefore similar chemical and biological activities.[1] Closely related is the notion of the *neighbourhood principle*. According to this principle, small structural dissimilarities, as defined *in silico* by a molecular descriptor, are likely to lead to small property differences.

The neighbourhood principle becomes very important during the Lead Optimisation phase when the main task for the medicinal chemists is to explore iteratively the chemical space in the vicinity of the lead structure with the aim to improve its property profile. It becomes fundamental then to identify a set of appropriate molecular descriptors which satisfy the neighbourhood principle, or in other words, molecular descriptors that exhibit the so-called *neighbourhood behaviour* (NB).[2, 3]

In this study, we have compared two existing computational methods to assess the neighbourhood behaviour of selected molecular descriptors, namely Patterson plots and Horvath Optimality criterion.[3, 4] To that end, we have employed a wide set of established 2D and 3D fingerprints, including dictionary-based, path, circular and pharmacophores, as well as multiple assay data from several GSK lead optimisation projects. By adapting the Optimality criterion method, we have generated graphs analogous to precision-recall plots which illustrate the extent to which the neighbourhood behaviour is valid for a given descriptor, as well as the appropriate Tanimoto similarity cut-off. Compared to the Patterson plots approach, this provides a faster, more robust and data-driven framework for the evaluation of similarity metrics in the context of NB.

Performance-wise, the results indicate that 2D circular substructure fingerprints performed consistently better especially among the bioactivity datasets, having an optimal Tanimoto similarity cut-off of approx. 0.65. This could provide an alternative to the well-established “Daylight & Tanimoto 0.85” practice for similarity as well as diversity studies.[5]

1. Johnson, M. A.; Maggiora, G. M., *Concepts and application of molecular similarity.* Wiley and sons: New York, 1990.
2. Dixon, S. L.; Merz, K. M., One-dimensional molecular representations and similarity calculations: methodology and validation. *Journal of Medicinal Chemistry* **2001,** 44, (23), 3795-3809.
3. Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E., Neighborhood behavior: a useful concept for validation of "molecular diversity" descriptors. *Journal of Medicinal Chemistry* **1996,** 39, (16), 3049-59.
4. Horvath, D.; Jeandenans, C., Neighborhood Behavior of in silico structural spaces with respect to in vitro activity spaces - A novel understanding of the molecular similarity principle in the context of multiple receptor binding profiles. *Journal of Chemical Information and Computer Sciences* **2003,** 43, (2), 680-690.
5. Martin, Y. C.; Kofron, J. L.; Traphagen, L. M., Do structurally similar molecules have similar biological activity? *Journal of Medicinal Chemistry* **2002,** 45, (19), 4350-4358.

P-84 : Maximum unbiased validation (MUV) datasets for virtual screening by pubchem based chemogenomics data mining

S.G. RohrerK. Baumann, Institute of Pharmaceutical Chemistry, Braunschweig University of Technology, Braunschweig, Germany

Recently, work from our lab provided a non-parametric methodology within the framework of spatial statistics to quantify the bias introduced by the composition of benchmark datasets into the validation of ligand based virtual screening methods.

Here, we use spatial statistics based design applied to a collection of bio-activity datasets selected from PubChem to generate maximum unbiased validation (*MUV*) datasets. Compared to other sources of bioactivity data, PubChem features several major advantages: (i) All data in PubChem, including structures, bio-assay conditions and experimental readouts are publicly accessible. (ii) The compound collections tested exhibit a remarkable level of diversity. (iii) The vast majority of tested compounds are “drug-like”. (iv) Compounds that were found *inactive* are listed in addition to those found *active*. This provides the unique opportunity to design decoy sets, for which the inactivity against the target is experimentally validated.

On the downside, most of the bio-activity data available from PubChem is based on High-Throughput Screening (*HTS*) experiments. HTS data is notoriously affected by experimental noise and artifacts. Thus, for the design of benchmark datasets it is essential to scrutinize PubChem bio-activity data with extreme care.

A workflow is presented that purges PubChem bio-activity datasets from unselective hits and enforces the selection of a representative set of decoys. In a first step, a collection of datasets of compounds active against pharmaceutically relevant targets was selected. A distance matrix was calculated, linking all PubChem bio-assays with protein target information by the sequence similarity of their respective targets. Based on a statistical analysis of all compounds tested in at least one bio-assay, those were flagged as unselective hits and removed from the datasets, that were found active in a significantly large number of assays with unrelated targets. Furthermore, active compounds located in regions of chemical space under-sampled by decoys were excluded from the datasets. Topological optimization using experimental design strategies monitored by spatial statistics functions was then used to generate corresponding datasets of actives and decoys that are unbiased with regard to analogue bias and artificial enrichment.

The presented datasets, the selectivity filter and a MATLAB toolbox for the spatial statistics analysis of chemical datasets will be available from our web-page.

(http://www.pharmchem.tu-bs.de/forschung/baumann/)

P-85 : 3D-Visualization of molecular conformations in the MOGADOC database

Jürgen Vogt[1], *Natalja Vogt*[1], *Evgeny Popov*[2]

[1] *Chemieinformationssysteme, University of Ulm, Ulm, Germany*

[2] *Nizhegorodsky State Architectural and Civil Engineering University, Nizhny Novogod, Russia*

The MOGADOC Database (Molecular Gasphase Documentation) has grown up to 32,800 bibliographic references for about 9,000 inorganic, organic and organometallic compounds, which were studied in the gas-phase by microwave spectroscopy, radio astronomy and electron diffraction. The database also contains 7,700 numerical datasets with internuclear distances, bond angles and dihedral angles.

The standard retrieval features of the WWW browser supported database have been described elsewhere in details [1]. By means of an implemented Java-based structure editor the user can retrieve molecular structures and their fragments [2]. Hereby the user has the choice to take into account or to ignore cis-trans-isomerism in cyclic compounds and (E)-(Z)-isomerism at double bonds [3]. Moreover, a Java-based applet has been developed, which enables the user to visualize three-dimensionally the molecular structures. The user can interactively rotate, shift, and scale the displayed 3D structures. Furthermore it is possible to allow or suppress the display of bond orders, atom labels (which are necessary to assign the corresponding internal coordinates in that entry) and the principal axis system.

The full set of Cartesian coordinates for all atoms of a molecule is necessary for the 3D visualization. However, in most of the original papers, the Cartesian coordinates are not given, not completely determined or cannot be determined (for example, for trigonometrically inconsistent thermal-average structures from the conventional gas electron diffraction analysis). In these cases the Cartesian coordinates have been derived by means of our own program supplemented by GaussView™ from available experimental internal coordinates using the symmetry of the equilibrium configuration and/or structural parameters of similar compounds as supplementary data.

Presently we are developing a Java based modul for the visualization of the energy hypersurface for multi-conformational molecular systems.

The project has been supported by the Dr. Barbara Mez-Starck Foundation, Freiburg/Br. (Germany).

1. J. Vogt and N. Vogt: Statistical Tools of the MOGADOC Database (Molecular Gasphase Documentation). Struct. Chem. 2003, 14,137-141
2. J. Vogt, N. Vogt, and R. Kramer: Visualzation and Substructure Retrieval Tools in the MOGADOC Database. J. Chem. Inform. Comput. Sci. 2003, 43, 357-361
3. J. Vogt and N. Vogt: Structure Searching in the MOGADOC Database. J. Mol. Struct. 2004, 695, 237-241

P-86 : Similarity based correction for the predictions of compounds physicochemical properties

Andrius Sazonovas[3] , P. Japertas [1,] R. Didziapetris [2], A. Petrauskas [4]

[1] Pharma Algorithms, Inc., A.Mickeviciaus g. 29, LT-08117 Vilnius, Lithuania; Faculty of Chemistry, Vilnius University, Naugarduko g. 24, LT-03225 Vilnius, Lithuania

[2] Pharma Algorithms, Inc., A.Mickeviciaus g. 29, LT-08117 Vilnius, Lithuania

[3] Faculty of Chemistry, Vilnius University, Naugarduko g. 24, LT-03225 Vilnius, Lithuania; Pharma Algorithms, Inc., A.Mickeviciaus g. 29, LT-08117 Vilnius, Lithuania

[4] Pharma Algorithms, Inc., A.Mickeviciaus g. 29, LT-08117 Vilnius, Lithuania

A number of problems have been long known to prevent the effective use of third-party predictive algorithms in the pharmaceutical industry. Among them is the training set not covering the specific part of the chemical space occupied by the compounds that a certain company is working with or a specific experimental protocol used to measure the property of interest that yields the results contrasting with the experimental values for the same compounds in the training set. Therefore the need arises for a method that would allow any company to tailor a third-party predictive algorithm to its specific needs using proprietary in-house data.

Here we present a novel similarity based methodology that provides a possibility for a user to expand the Applicability Domain of the existing Pharma Algorithms models with the help of a custom database of experimental values for the property of interest. A Reliability Index (RI) is also calculated as a measure of the quality of the particular prediction. The use of the method is illustrated with examples of its application in predicting logP, logD and solubility of the compounds. It is shown that a relatively small amount (5 to 10) of similar compounds has to be added to substantially improve the prediction for a group of problematic compounds that is not represented in the original training set. The Reliability Index is shown to be closely related to the overall quality of any given prediction that is represented by a clear correlation of the RI and RMSE values.

Given that the improvement of any Pharma Algorithms model in this way is instant as it occurs that very moment when new compounds with experimental values are added to the similarity database and there is no need to retrain the model, this method opens completely new possibilities for their use in the industry.

P-87 : Prediction of ionization constants for complex multicenter electrolytes utilizing proprietary ‘in house’ data

Andrius Sazonovas, P. Japertas [1], R. Kubilius [2], D. Simelevicius [3]

[1] Pharma Algorithms, Inc., A.Mickeviciaus g. 29, LT-08117 Vilnius, Lithuania; Faculty of Chemistry, Vilnius University, Naugarduko g. 24, LT-03225 Vilnius, Lithuania

[2] Pharma Algorithms, Inc., A.Mickeviciaus g. 29, LT-08117 Vilnius, Lithuania

[3] Faculty of Mathematics and Informatics, Vilnius University, Naugarduko g. 24, LT-03225 Vilnius, Lithuania; Pharma Algorithms, Inc., A.Mickeviciaus g. 29, LT-08117 Vilnius, Lithuania

Ionization is one of the key parameters that affects absolute majority of the physicochemical properties and biological activities that are of interest to the developers of any new marketed chemicals, regardless of their intended use. Therefore, estimation of pKa values has always been the field where prediction accuracy received special attention from the industry.

In this work we present the methodology of pKa prediction developed by Pharma Algorithms. It is a multi-step procedure involving estimation of pKa microconstants for all possible ionization centers in an uncharged molecule (“fundamental microconstants”), numerous corrections of these initial pKa values according to the surrounding of the reaction center and calculation of charge influences of ionized groups to the neighboring ionization centers. In total, algorithm utilizes a data set of >12,000 compounds with experimental pKa measurements, a database of 4,600 ionization centers, a set of ca. 500 various interaction constants and four interaction calculation methods for different types of interactions, producing a full range of microconstants from which pKa macroconstants are obtained. This allows for a simulation of complete distribution plot of all protonation states of the molecule at different pH conditions.

Finally the above described predictions are used as the baseline values in a novel similarity based routine, allowing estimation of reliability for each prediction (evaluation of the Model Applicability Domain). In addition, this methodology provides industrial users with a unique possibility to expand the Model Applicability Domain with the help of any user-defined proprietary ‘in house’ databases of experimental pKa values. As it will be shown, this functionality is capable of substantially increasing the accuracy of predictions for the compounds not represented in the initial training set, thus opening new possibilities of the model use in prediction of pKa values for specific classes of proprietary ‘in house’ compounds.

P-88 : A novel chemical database for sustainable development of synthesis routes: An instance of developing synthesis routes of quinolone derivatives

K. Hori, T. Yamaguchi, H. Sadatomi, M. Sumimoto, Graduate school of Science and Engineering, Yamaguchi University, Ube, Japan

Transition State Database (TSDB)[1] is a system storing information of TSs, reactants, intermediates and products derived from quantum chemical calculations. One of the usages of TSDB is to make initial geometries for new TSs using data in the TSDB and to reduce computational times of searching TSs. Another usage is to confirm whether or not a new route designed by a synthesis route designing system such as AIPHOS, KOSP[2] or TOSP is useful to synthesize target compounds. Figure 1 show the concept of TSDB. We have been developing TSDB and studying how to use it. The present TSDB has a web interface which shows 3D structures, IRC animations and so on and searches transition states in the TSDB by comparing chemical structures of reactants and products. Trial version is released at https://trial.tsdb.jp/. The present study describes how to use the TSDB in developing synthesis routes of target compounds.

Antibacterial spiro compounds are usefull as medicines and preservatives. Quinolone derivatives including 7-amino-5-azaspiro[2.4]heptyl substitutions offer super antibacterial activity and oral absorptive property. 6-metyl-6-azaspiro[2.4]heptane-4-one is one of manufactured intermediates for 7-amino-5-azaspiro[2.4]heptane. TOSP that is one of SRDSs proposed 24 synthesis routes for 6-metyl-6-azaspiro[2.4]heptane-4-one. Similarity reaction searching about these 24 routes in TSDB was executed and 4 similar reactions were found. We performed PM3 or B3LYP/6-31G* level of theory calculations to analyze these routes using found data. Accordingly, it was clarified that 6-metyl-6-azaspiro[2.4]heptane-4-one could be synthesized using 3-methyl-3-aza-bicyclo[3.2.0] heptane-1,5-diol, the starting compound of the pinacol rearrangement. We also generated next routes for synthesizing 3-methyl-3-aza-bicyclo[3.2.0]heptane-1,5-diol using TOSP. 84 routes were proposed and 3 similar reactions with these routes were found from TSDB. 3 routes were also analyzed to clarify that 3-methyl-3-aza-bicyclo[3.2.0]heptane-1,5-diol could be synthesized from 5-chloro-3-methyl-3-aza-bicyclo[3.2.0]heptan-1-ol using the substitution reaction.

It is possible to use other simple starting compounds with reiterating the present procedure shown in Figure 2. This procedure makes it possible to evaluate synthesis routes created by SRDS using theoretical calculations together with the TSDB for developing new synthesis routes.

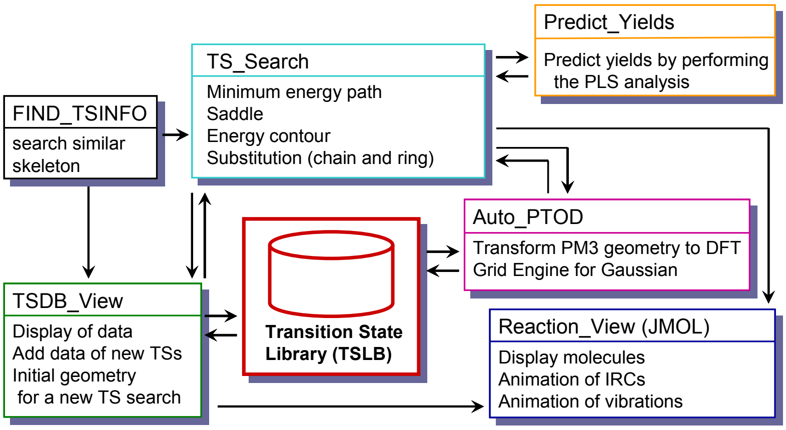


Figure 1. The concept of TSDB

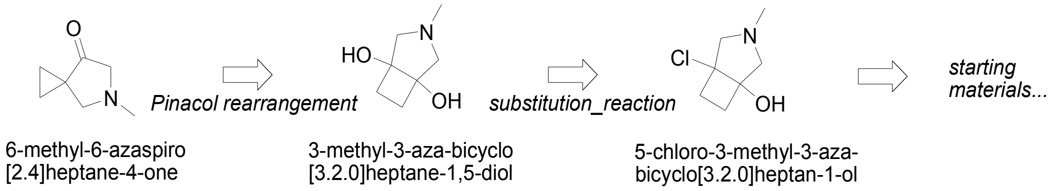

O
N
Pinacol rearrangement
HO
N
OH
substitution_reaction
Cl
N
OH
starting materials...
6-methyl-6-azaspiro [2.4]heptane-4-one
3-methyl-3-aza-bicyclo [3.2.0]heptane-1,5-diol
5-chloro-3-methyl-3-aza-bicyclo[3.2.0]heptan-1-ol

Figure 2. Abuilt synthesis route for 6-methyl-6-azaspiro[2.4]heptane-4-one

1. Hori, K.; Okano, K; Yoshimura, K.; Yamamoto, H. *J. Comp. Aided Chem.*, **2005**, 6, 30.
2. Satoh, K.; Funatsu, K., *J. Chem. Inf. Comput. Sci.,* **1999**, 39, 316. (b) AIPHOS/KOSP V1.0, Fujitsu Limited, Tokyo Japan, 2003

P-89 : Combinatrial chemistry using theoretical calculations. An application to boric acid catalyzed esterification of phenol

M. Shimeno, M. Sumimoto, K. Hori, Graduate school of Science and Engineering, Yamaguchi University, Ube, Japan

Combinatrial technique is rather difficult in organic synthetic chemistry since it is not so easy to synthesize all the compounds with desired substituents. However, the same technique in computational chemistry is very attractive since it is very easy to introduce complicated substituents in theoretical calculations and reactions are proceeding in the virtual world. This is also true when we try to investigate the role of catalysis, for example, boric acid catalyzed esterification of phenol.

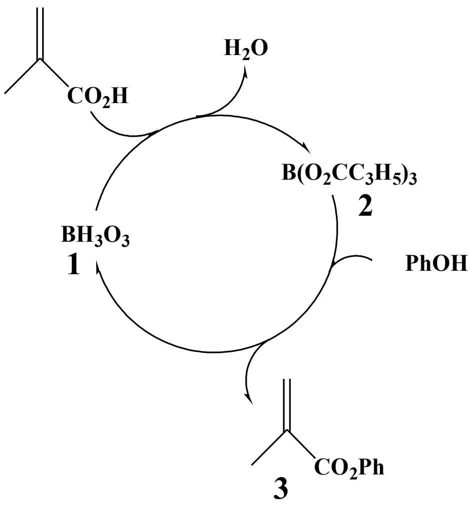
Boric acid / sulfuric acid catalyzed esterification has been using to synthesize phenyl methacrylate [1]. In the beginning of this reaction, boric acid reacts with methacrylic acid to form boric methacrylic anhydride and H2O. The anhydride then reacts with phenol to form phenyl methacrylate 3. While sulfuric acid has been using as the catalyst for the reaction, other protic acids such as phosphoric acid, trichloroacetic acid, and so on may be useful. In order to understand the role of acids and find other acids useful for the reaction, following calculations were performed.

1. Investigate the reaction mechanism without considering protic acids using Density Functional Theory (DFT) calculations.
2. Clarify the role of sulfuric acid in the esterification reaction.
3. Confirm whether or not other protic acids play a same role as sulfuric acid.

Table 1 Activation energies (Ea) and heat of reactions depending onprotic acids.

Fig. 1: boric acid catalyzed reaction cycle.

acid	2-3	Ea	Δ
H_2SO_4 (pKa=1.9)	1-step	28.0	-0.2
	2-step	27.6	-0.5
	3-step	24.9	-0.5
CCl_3COOH (pKa=0.77)	1-step	37.8	-4.0
	2-step	29.2	-4.8
	3-step	31.0	
PO_4H_3 (pKa=2.12)	1-step	37.8	-9.1
	2-step	24.5	-9.9
	3-step	20.8	-13.3
H_2O (pKa=15.7)	1-step	39.3	-4.8
	2-step	41.0	-6.7
	3-step	40.8	-3.9



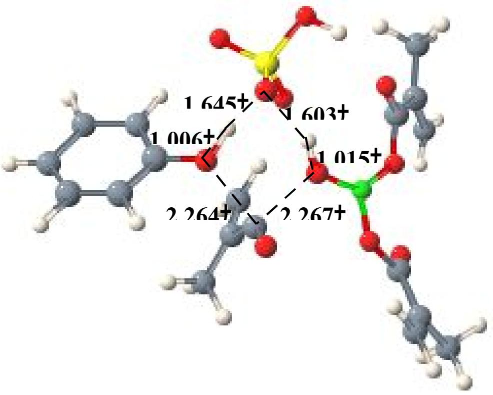


Fig. 2: 1-step transition state (TS) from sulfuric acid.

Figure 1 displays the mechanism investigated in the present study. As boric acid have three hydroxyl groups, all the groups react with methacrylic acid to form **2,** followed by the reaction with phenol to form **3**. It was confirmed that the reaction proceeds with help of one water molecule, i.e., a proton relay that the water molecule accepts a proton from phenol and donates its proton to regenerate boric acid [2]. The activation energies (Ea) of three similar reactions were calculated to be almost 40 kcal mol^{-1} as listed in Table 1. We tried to calculate that sulfuric acid can also play a similar role to that of the water molecule. Figure 2 displays the transition state geometry including a sulfuric acid for the reaction. The intrinsic reaction coordinate calculations showed that the obtained TS connect 2+PhOH with the product 3. The calculated Ea is 28.0 kcal mol-1, lower by 11.3 kcal mol-1 than that with a water molecule. This mechanism suggests that HCl cannot catalyze the esterification since the acid has no donor atom used for the proton relay.

In order to search other protic acids useful for the proton relay, the same mechanism applied to the esterification using trichloroaacetic acid and phosphoric acid. Although TS structures similar to sulfuric acid were obtained for the both acids, the Ea's of the first step reaction are larger by ca. 10 kcal mol-1 than that for sulfuric acid. The high barrier for phosphoric acid resulted in observation that the acid is useless for producing phenyl methacrylate using boric acids. The same conclusion is applicable to the reaction of trichloroacetic acid. We now try to find useful protic acids for the esterification using this combinatorial technique.

1. William, W., *Tetrahedron Letters*, **1971**, 37, 3453.
2. Hori, K.; Ikenaga, Y.; Arata, K.; Takahashi, T.; Kasai, K.; Noguchi, Y.; Sumimoto, M.; Yamamoto. H. *Tetrahedron*, **2007**, 63, 1264

P-90 : Calculation of difference of free energy of solvations using the QM/MC/FEP method in chemical reactions

K. Uezu, T. Yamaguchi, M. Sumimoto, K. Hori, Graduate school of Science and Engineering, Yamaguchi University, Ube, Japan

Monte Carlo (MC) simulations or molecular dynamic (MD) calculations using classical force fields have been used for many studies which calculate free energies of activation ($\Delta G^{\ddagger}$) in solutions. We have been developing the QM/MC method, Monte Carlo simulation using quantum mechanical (QM) calculations in order to investigate solvent effects for chemical reactions. The simulation uses the QM calculations to obtain energies of ensembles so that it is not necessary to determine charges and van der Waals parameters of solvent molecules. Although MC simulations using fully quantum mechanical calculations are free from classical parameters, it takes long CPU times to obtain energies with statistical meanings.

We developed a program which calculates difference of free energy of solvation, ΔΔG(solv), using the QM/MC method connected with free energy perturbation (FEP) calculations according to Eq. 1, i.e., the QM/MC/FEP simulation. $\Delta G^{\ddagger}$ in solution can be calculated by adding $\Delta G^{\ddagger}$ in the gas phase to ΔΔG(solv) between the TS and the reactant in solution. The QM/MC/FEP calculations use the thermodynamic cycle shown in Scheme 1. We adopted he droplet model that includes a solute molecule in its center and have solvent molecules around the solute within a radius depending on the sides of solvent molecules. The solute has structures optimized using quantum mechanical calculations. The QM/MC/FEP simulation was applied to calculate ΔΔG(solv) between the reactant and TS, between TS and the product by dividing differences between the structures by 50 points.

$$\Delta G_{1-0} = G_1 - G_0 = -k_B T \ln \left\langle \exp\left(-\frac{E_1^{sol} - E_0^{sol}}{k_B T} \right) \right\rangle_0 \qquad (1)$$

Scheme 1

$$\begin{tikzcd} \textit{In gas:} & A_0 \arrow[r, "\Delta G_{gas}(A_{0\to 1})"] \arrow[d, "G_0^{sol}"'] & A_1 \arrow[d, "G_1^{sol}"] \\ \textit{In solution:} & A_0 \arrow[r, "\Delta G_{solution}(A_{0\to 1})"'] & A_1 \end{tikzcd}$$

The QM/MC/FEP simulation was applied to Diels-Alder reaction between methylvinylketone and cyclopentadiene. The activation barrier was calculated to be 18.3 kcal mol^{-1} and the heat of reaction to be 12.5 kcal mol^{-1} in the gas phase at the B3LYP/6-311++G**//MP2/6-31G* level of theory. The QM/MC(B3LYP/6-311++G**//MP2/6-31G*,PM3,PM3) calculation gave averaged activation energies included solvent effects to be 12.7, 13.8 and 15.2 kcal mol^{-1} in aqueous, methanol and propane solutions, respectively. These results indicate that the rate of the reaction in aqueous solution is much faster than that in propane and is consistent with the experimental data.

Applications to other reactions such as acid-catalyzed hydrolysis, Cope Elimination and decarboxylation reactions will be shown in the poster.

1. R. Breslow, T.Guo, *J. Am. Chem. Soc*, **1988**, 110, 5615.
2. in isooctane

P-91 : Toward *in silico* screening using transition state data base for developing new synthesis routes

T. Yamaguchi, H. Sadatomi, M. Sumimoto, K. Hori, Graduate School of Science and Engineering, Yamaguchi University, Ube, Japan

We have been constructing a data base with information of transition states for chemical reactions[1]. The data base is called the transition state data base (TSDB) which helps to develop synthesis routes for target compounds and works together with the synthesis route design systems (SRDS) such as TOSP and KOSP[2]. The present study describes what is the TSDB and how to use it for "*in silico screaning*" of synthesis routes from chemoinformatic calculations.

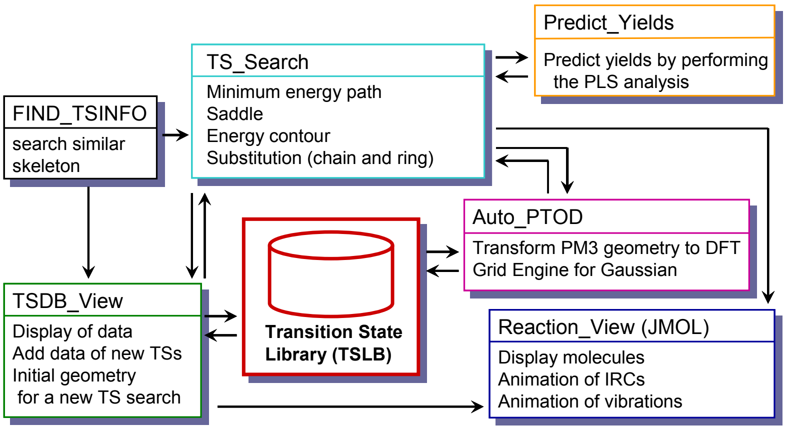
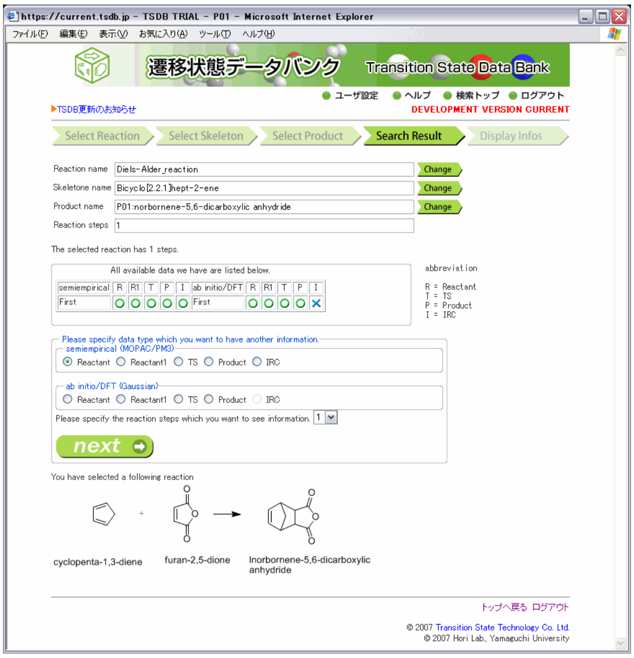


Figure 1. The concept of TSDB

The structure of the TSDB is shown in Figure 1. The TSDB has a library of TS information (TSLB) including more than 1000 TSs for 30 kinds of namely reactions at present. We are expecting to gather more than 3000 TS information for 100 namely reactions for two years. The library categorizes information depending on names of reactions, skeleton structures and kinds of products and contains coordinates, energies and results of intrinsic reaction coordinate (IRC) calculations. The data base has six programs with different aims. TSDB_View displays 3D structures of selected reactions as well as coordinates which will be used as initial geometries for locating new TSs. To build an initial geometry for the TS_Search program, FIND_TSINFO can perform searches based on similarities of chemical structures. This function is deliveried by PostgreSQL + pgchem::tigress programs[3], both of which are open source softwares. TS_Search is a program which is developed for trying several methods to search TSs semi-automatically. The Predict_Yield program predicts the trends of experimental yields of synthesis reactions using the calculated activation energies. For this purpose, PLS analysis was used to obtain a relationship between the experimental data and the activation energies obtained from the DFT calculations[4].

It is very important to access the TSDB via the internet. Figure 2a shows a window found data for Diels-Alder reaction of cyclopentadiene and furan-2,5-dione. Figure 2b displays results of the IRC calculations for the reaction as well as energy relations between reactants, TS and the product. The interface has been developed using PHP5 and is able to display the energy relations. You can use a trial version of the TSDB at http://traial.tsdb.jp.

It is also important to show how to use the TSDB for developing new synthesis routes for synthetic organic chemist. They can create synthesis routes of targets but cannot give answers for their possibility without experiments. Theoretical analyses of reaction mechanisms for the synthesis routes are considered to be one of the tools for examining possibilities of the synthesis routes. It means that the combination of computational chemistry and chmoinformatics offers a different way to develop new synthesis routes for compounds although real experiments have to be done and improved in order to create synthesis routes for industrial usages. These efforts are called "*in silico screaning*" of synthesis routes. We have been investigating this concept and found that information previously obtained is very useful for new calculations to locate TSs for similar reactions. This is why we have been building the TSDB for these years. In fact, *in silico screaning* for 2,6-dimethylchroman-4-one make it was possible to narrow down more than 20 routes to only four routes using the DFT calculations.


https://current.tsdb.jp - TSDB TRIAL - P01 - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
遷移状態データバンク Transition State Data Bank
ユーザ設定 ヘルプ 検索トップ ログアウト
▶TSDB更新のお知らせ
DEVELOPMENT VERSION CURRENT
Select Reaction
Select Skeleton
Select Product
Search Result
Display Infos
Change
next
cyclopenta-1,3-diene
furan-2,5-dione
Inorbornene-5,6-dicarboxylic anhydride
トップへ戻る ログアウト

Reaction name: Diels-Alder_reaction

Skeletone name: Bicyclo[2.2.1]hept-2-ene

Product name: P01:norbornene-5,6-dicarboxylic anhydride

Reaction steps: 1

The selected reaction has 1 steps.

All available data we have are listed below.

semiempirical	R	R1	T	P	I	ab initio/DFT	R	R1	T	P	I
First	○	○	○	○	○	First	○	○	○	○	✕

abbreviation

R = Reactant
T = TS
P = Product
I = IRC

Please specify data type which you want to have another information.

semiempirical (MOPAC/PM3)
● Reactant ○ Reactant1 ○ TS ○ Product ○ IRC

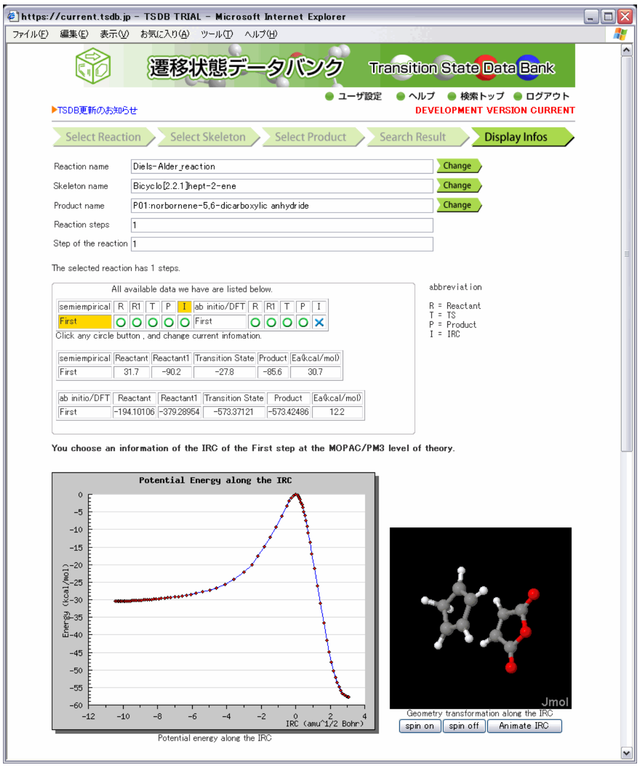
ab initio/DFT (Gaussian)
○ Reactant ○ Reactant1 ○ TS ○ Product ○ IRC

Please specify the reaction steps which you want to see information. 1

You have selected a following reaction

© 2007 Transition State Technology Co. Ltd.
© 2007 Hori Lab., Yamaguchi University

Figure 2a. Showing reaction schemes and sumaries of data


https://current.tsdb.jp - TSDB TRIAL - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
遷移状態データバンク Transition State Data Bank
ユーザ設定 ヘルプ 検索トップ ログアウト
▶TSDB更新のお知らせ
DEVELOPMENT VERSION CURRENT
Select Reaction
Select Skeleton
Select Product
Search Result
Display Infos
Change
Potential Energy along the IRC
0
-5
-10
-15
-20
-25
-30
-35
-40
-45
-50
-55
-60
Energy (kcal/mol)
-12
-10
-8
-6
-4
-2
0
2
4
IRC (amu^1/2 Bohr)
Potential energy along the IRC
Jmol
Geometry transformation along the IRC
spin on
spin off
Animate IRC

Reaction name: Diels-Alder_reaction

Skeleton name: Bicyclo[2.2.1]hept-2-ene

Product name: P01:norbornene-5,6-dicarboxylic anhydride

Reaction steps: 1

Step of the reaction: 1

The selected reaction has 1 steps.

All available data we have are listed below.

semiempirical	R	R1	T	P	I	ab initio/DFT	R	R1	T	P	I
First	○	○	○	○	○	First	○	○	○	○	✕

Click any circle button , and change current infomation.

semiempirical	Reactant	Reactant1	Transition State	Product	Ea(kcal/mol)
First	31.7	-90.2	-27.8	-85.6	30.7

ab initio/DFT	Reactant	Reactant1	Transition State	Product	Ea(kcal/mol)
First	-194.10106	-379.28954	-573.37121	-573.42486	12.2

abbreviation

R = Reactant
T = TS
P = Product
I = IRC

You choose an information of the IRC of the First step at the MOPAC/PM3 level of theory.

Figure 2b. Showing geometries of TSs and IRC graphs

1. Hori, K.; Okano, K; Yoshimura, K.; Yamamoto, H. *J. Comp. Aided Chem.*, **2005**, 6, 30.
2. Satoh, K.; Funatsu, K., *J. Chem. Inf. Comput. Sci.,* **1999**, 39, 316. (b) AIPHOS/KOSP V1.0, Fujitsu Limited, Tokyo Japan, 2003
3. Schmid, E.-G.., http://pgfoundry.org/projects/pgchem/
4. Hori, K.; Yamaguchi, T.; Sumimoto, M.; Okano, K.; Yamamoto, H., *Tetrahedron,* **2008**, 64, 1759.

P-92 : Tautomer generation. pKa based dominance conditions for generating the dominant tautomers

József Szegezdi, Ferenc Csizmadia, ChemAxon Ltd., 1037 Budapest, Máramaros köz 3/A, Hungary

We developed a Java based application for generating the tautotmeric forms (protomers) of a molecule. The generation of tautomers is performed after the identification of the hydrogen donor and acceptor groups of the submitted molecule. Hydrogen acceptors are atoms which have unsaturated bond. An atom in the molecule may be considered a donor if it is connected to an unsaturated bond at alpha position.

The numbers of the generated tautomers largely depend on the donor and the acceptor sites, which are allowed to interact with each other through the tautomerization path. The length of the tautomerization path is also an important parameter of the tautomer generation. In our method, no restriction is applied for the length of the tautomerization path.

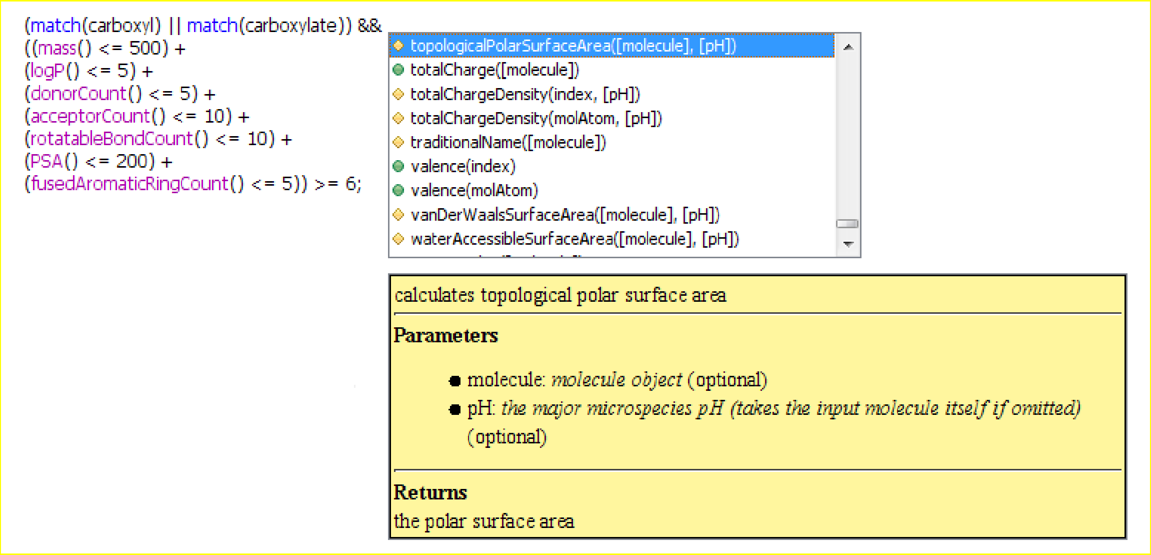
The calculation of the distribution of the different tautomeric forms in aqueous solution is particularly important both from theoretical and practical point of view. This is why we pay special attention for the generation of the dominant tautomeric forms of the submitted molecule. To generate the major tautomeric forms, we defined simple pKa conditions for the donor and acceptor groups. The ratio of the major and the minor tautomeric forms also depend on the solution's pH. The prediction of these ratios is also possible from the calculated pKa values. Predicted dominant tautomeric forms are in good agreement with the practical observations.

P-93 : Chemical terms – A language for cheminformatics

G. Pirok, Z. Mohacsi, N. Mate, J. Szegezdi, I. Cseh, A. Szabo, M. Vargyas, S. Csepregi, A. Papp, F. Csizmadia, ChemAxon, Budapest, Hungary

Pharmaceutical research is not just about molecules, it is about realizable molecules having interesting properties. The available set of computable properties is growing, functions usually calculate a specific physicochemical parameter. These functions like partial charge distribution, pKa, logD carry important chemical information, but the most interesting questions today are more complex. Many of those questions are related to ADMET. Will a planned specific compound be absorbed well, what are its major metabolites, what is its risk of being toxic? Chemical reactivity is another field where the problem is too complex to be described by a simple prediction function.

Scientists need an easy way to formulate new functions by the combination of property predictions, mathematical functions, and structural calculations. The Chemical Terms language was developed with this purpose in mind. More than a hundred functions are currently provided, and can be extended through a public plugin interface. The evaluator engine is an integratable component, and the Chemical Terms language has been implemented within other areas of ChemAxon's technology to add or tune chemical feasibility for cheminformatics tools in various areas such as database filtering, pharmacophore screening, drug design, virtual synthesis, and xenobiotic metabolic pathway prediction.



```
(match(carboxyl) || match(carboxylate)) &&
((mass() <= 500) +
(logP() <= 5) +
(donorCount() <= 5) +
(acceptorCount() <= 10) +
(rotatableBondCount() <= 10) +
(PSA() <= 200) +
(fusedAromaticRingCount() <= 5)) >= 6;
```

◇ topologicalPolarSurfaceArea([molecule], [pH])
● totalCharge([molecule])
◇ totalChargeDensity(index, [pH])
◇ totalChargeDensity(molAtom, [pH])
◇ traditionalName([molecule])
● valence(index)
● valence(molAtom)
◇ vanDerWaalsSurfaceArea([molecule], [pH])
◇ waterAccessibleSurfaceArea([molecule], [pH])

calculates topological polar surface area

Parameters

- molecule: *molecule object* (optional)
- pH: *the major microspecies pH (takes the input molecule itself if omitted)* (optional)

Returns
the polar surface area

Sponsoring Societies

- Division of Chemical Information (CINF), American Chemical Society (ACS)
- Royal Netherlands Chemical Society (KNCV)
- Chemistry-Information-Computer Division, Gesellschaft Deutscher Chemiker (Society of German Chemists) (GDCh)
- The Chemical Structure Association Trust (CSA Trust)
- Chemical Information Group, Royal Society of Chemistry (RSC)
- Division of Chemical Information and Computer Science of the Chemical Society of Japan (CSJ)
- Swiss Chemical Society (SCS)