ICCS

International Conference on Chemical Structures

11th International Conference on Chemical Structures

May 27 – 31, 2018 ♦ Noordwijkerhout ♦ The Netherlands

Program & Abstracts

www.int-conf-chem-structures.org

Attend our Pre-Conference Workshop on Sunday, May 27th 2018 from 15:00-17:00

"Orion - CADD on the Cloud"

- Orion is OpenEye's reimagining of all the elements of CADD conducted entirely within a cloud service, in our case Amazon Web Services (AWS), delivered as either a hosted system or an in-house VPC solution
- As a 'cloud native' platform Orion completely automates and manages access to large scale AWS processing and storage
- In-cloud facilities include molecular design, 3D visualization, data analysis, results/method sharing and project organization
- All of OpenEye science is included, enabling users to construct innovative workflows with Floe, our pipelining tool
- As an open platform Orion allows for straightforward integration of third-party code (customer, academic, vendor)
- Interaction with Orion is via a simple webpage, deliverable on any internet-enabled device

Interact with Orion via web browser on any internet-enabled device.

OpenEye has built a solid reputation as a scientific leader in the field of molecular design based on two decades of delivering useful applications and programming toolkits. Our scientific approach has focussed on the power of molecular 3D structure to inform and guide, in particular via the concept of shape similarity. We have changed industry perception of what is possible with the speed, robustness and scalability of our tools and have recently built these into a ground-up, cloud-native platform, Orion. Combining unlimited computation and storage with powerful tools for data sharing, visualization and analysis in an open development platform, Orion offers unprecedented capabilities for drug discovery and optimization.

OpenEye SCIENTIFIC

To learn more about OpenEye and Orion, please stop by our booth, or visit us at: www.eyesopen.com

Preface

Welcome to the 11th International Conference on Chemical Structures (ICCS). The organizers decided to deviate from the triennial ICCS schedule by one year, the event, as a result, being organized in 2018. The conference builds on a long and successful history, which started with a NATO Advanced Study Workshop in 1973[1] and with the previous edition being jointly organized with the German Conference on Cheminformatics. The ICCS meeting is among the most important events in this area of science and gives an accurate picture of the state-of-the-art in the computer handling and manipulation of chemical structures.

We have received 145 abstract submissions from over 24 different countries from 4 continents. All submissions were subject to a review process carried out by a Scientific Advisory Board of 20 international reviewers from academia and industry. This allowed us to compile an outstanding scientific program of 34 plenary and 78 poster presentations. Additionally, the conference hosts an exhibition which allows a sizable number of scientific institutions and vendors to present their latest applications, content and software. And most importantly, sufficient time is provided for scientific exchange and discussion among the attending scientist, both at the conference and also during the sailing excursion across the IJsselmeer to visit the Bataviawerf with a reconstruction of the Batavia, a 17th-century VOC ship, on Wednesday afternoon.

Once again, the conference was chosen as the venue to present the triennial CSA Trust Mike Lynch Award. This year, it is granted to Dr. Rudy Potenzone[2] in recognition of his outstanding accomplishments in the field of cheminformatics. Rudy Potenzone will open the conference by receiving the award and delivering the keynote address titled From Teletype Structure Input to Biology and Chemistry Intelligent Knowledge Graphs: My 45 Years in Cheminformatics on Sunday evening.

On Thursday, furthermore, we dedicate the cheminformatics session to Prof. Dr. Peter Willett, because of his important contribution to the ICCS and to the field of cheminformatics in general.

After the conference, you are encouraged to submit your presentation or poster for publication in a special ICCS article collection of the Journal of Cheminformatics, guest edited by Gerard van Westen and Markus Wagener. Papers can be submitted at any date up to the 1st of October 2018, and authors should mention in their cover letter that the manuscript is intended to be included in the ICCS 2018 article collection. Of course, all manuscripts will be subject to a peer review following the journal's guidelines.

This book of abstracts is intended to inform you about the scientific program of the conference and to help you to plan your attendance. Moreover, we also hope that the abstracts in this volume will serve you as a reminder of the presentations and posters as well as provide a snapshot of the current research in the area of cheminformatics and molecular modeling in 2018. Note that in the online program ORCID identifiers are provided where available, allowing you to learn more about past research by presenters. The ORCID identifiers are also used to create an online webapp.[3]

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

We hope that you enjoy the conference!

Markus Wagener (ICCS Chair), Frank Oellien (ICCS Co-Chair), Chris de Graaf, Lars Ridder, Egon Willighagen, and Gerard van Westen

1. https://tools.wmflabs.org/scholia/event-series/Q47501052 for an overview of all ICCS meetings.
2. https://tools.wmflabs.org/scholia/author/Q51614233
3. https://tools.wmflabs.org/scholia/event/Q47501229

Contents

The Conference

Organizing Committee

- Chris de Graaf, VU University Amsterdam, Amsterdam, The Netherlands
- Frank Oellien, AbbVie, Ludwigshafen, Germany
- Lars Ridder, Netherlands eScience Center, Amsterdam, The Netherlands
- Markus Wagener, Grünenthal, Aachen, Germany
- Gerard JP van Westen, University of Leiden, Leiden, The Netherlands
- Egon Willighagen, Maastricht University, Maastricht, The Netherlands

Scientific Advisory Board

- Andreas Bender, University of Cambridge, UK
- Peter Ertl, Novartis, CH
- Kimito Funatsu, University of Tokyo, JP
- Val Gillet, University of Sheffield, UK
- Chris de Graaf, Free University of Amsterdam, NL
- Rajarshi Guha, National Institutes of Health, US
- Esther Kellenberger, University of Strasbourg, FR
- Michael Lajiness, Eli Lilly and Company, US
- Frank Oellien, AbbVie, DE
- John Overington, Catapult Medicines Discovery, UK
- Matthias Rarey, University of Hamburg, DE
- Lars Ridder, Netherlands eScience Center, NL
- Christoph Steinbeck, University of Jena, DE
- Herman van Vlijmen, Janssen Pharmaceutica NV, BE
- Andrea Volkamer, Charité - Universitätsmedizin Berlin, DE
- Markus Wagener, Grünenthal, DE
- Pat Walters, Relay Therapeutics, US
- Gerard van Westen, University of Leiden, NL
- Antony Williams, EPA, US
- Egon Willighagen, University of Maastricht, NL

Supporting Societies

Chemical Information and Computer Applications Group of the Royal Society of Chemistry (RSC)	
Chemical Structure Association Trust (CSA Trust)	
Chemistry-Information-Computer Division of the German Chemical Society (GDCh)	
Division of Chemical Information of the American Chemical Society (ACS)	
Division of Chemical Information and Computer Science of the Chemical Society of Japan (CSJ)	
Royal Netherlands Chemical Society (KNCV)	
Swiss Chemical Society (SCS)	
European Association for Chemical and Molecular Sciences (EuCheMS)	

Sponsors

Premier Sponsor

OpenEye

Platinum Sponsors

Chemical Computing Group

Collaborative Drug Design

Gold Sponsors

AbbVie

Bayer

Chemical Abstract Service

Dotmatics

Discngine

KNIME

Schrödinger

Silver Sponsors

NextMove Software

Conference Bag Sponsors

inte:ligand

Notepad Sponsors

Schrödinger

KNIME

Poster Awards Sponsor

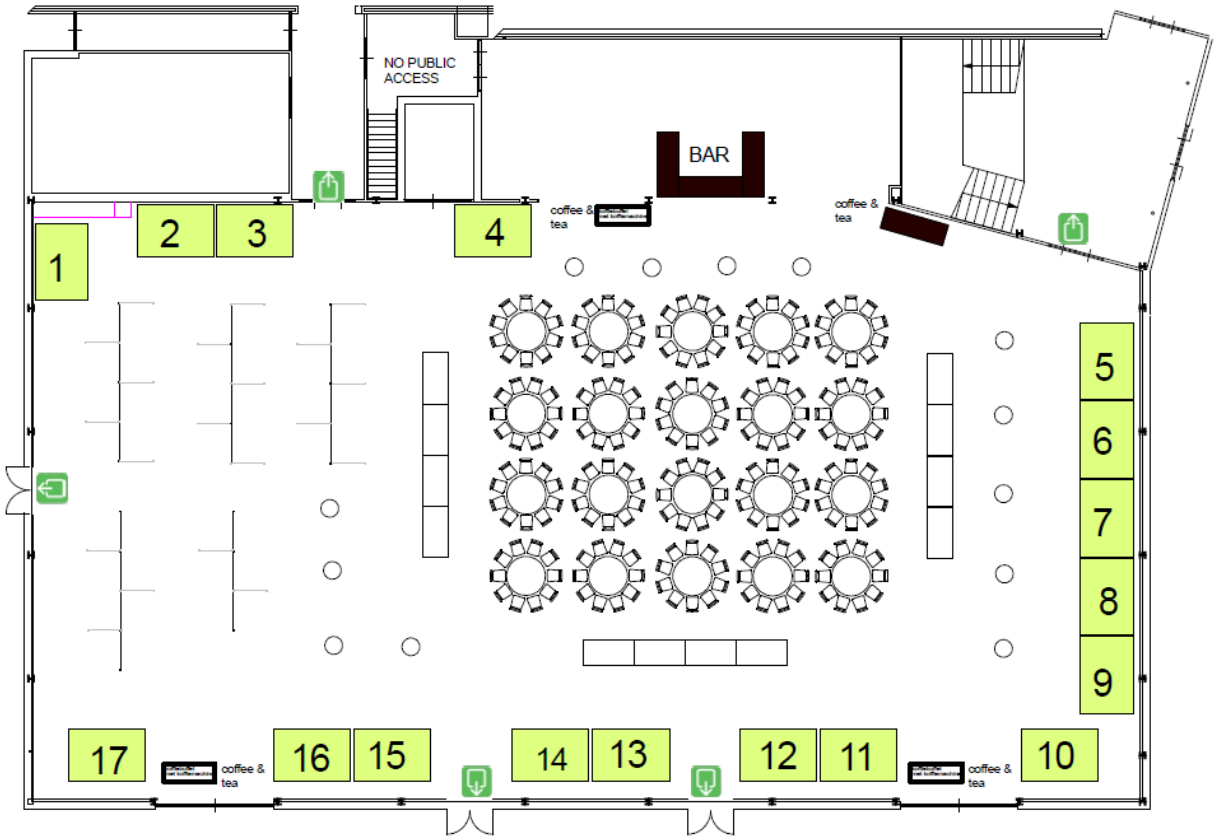
ChemMedChem

Other Sponsors

We would like to thank the Royal Netherlands Chemical Society (KNCV) for supporting the conference at the front desk. We would like to thank CCL.NET and Jan Labanowski for adding the conference to the CCL Conferences webpage. We would also like to thank the Center of Bioinformatics of the University of Hamburg for hosting the conference webpage.

Exhibition

Exhibition Layout



Exhibitor	Booth	Exhibitor	Booth
Acellera	B1	Cresset	B10
Discngine	B2	CCDC	B11
SilcsBio	B3	inte:ligand	B12
Schrödinger	B4	NextMove Software	B13
Collaborative Drug Design	B5	Dotmatics	B14
Chemical Abstract Service	B6	Xemistry	B15
Certara	B7	Chemical Computing Group	B16
Knime	B8	OpenEye	B17
Culgi	B9		

Exhibition Hours

- Monday, May 28th, 2018, 14:30 – 19:30
- Tuesday, May 29th, 2018, 14:30 – 19:30

Exhibitors

Chemical Computing Group

OpenEye

Xemistry

Dotmatics

NextMove Software

Collaborative Drug Design

KNIME

Inte:ligand

Schrödinger

Cresset

Chemical Abstract Service

Certara

CCDC

SilcsBio

Discngine

Acellera

Culgi

Workshops Sunday, May 27th

Chemical Computing Group Workshop: Application of Matched Molecular Pairs to Interactive SAR Exploration

Sunday May 27th 2018, 15:00-17:00, NH Conference Hotel Noordwijkerhout, Room: Boston 13

Managing and analyzing structure activity/property relationship data in medicinal chemistry projects is becoming ever more challenging, with larger data sets and parallel development of different structural series. Tools and methods for the efficient visualization, analysis and profiling of structures therefore remain of deep interest.

The workshop will start with a presentation about the use of interactive MMP analysis and R-group profiling to enhance typical medicinal chemistry workflows by interrogating the SAR data, thereby guiding a medicinal chemistry campaign in its development.

The presentation will be followed by working through some real examples of the use of the new MOEsaic application, and some complementary capabilities in the MOE (Molecular Operating Environment) software system;

R-Group Profiles and Analysis / MOEsaic / MMP Analysis / Template-Forced Docking / Scaffold Replacement / MedChem Transformations

Trial copies of MOE can be provided; see www.chemcomp.com/Product-Free_Trial.htm

OpenEye Workshop: *Orion - CADD on the Cloud*

Sunday May 27th 2018, 15:00-17:00, NH Conference Hotel Noordwijkerhout, Room: Boston 15

The cloud will increasingly become the destination for a wide variety of tasks, in computational chemistry and elsewhere. In this workshop we will introduce Orion, OpenEye's new cloud-native CADD platform. By seamlessly integrating almost limitless computing capacity with well validated workflows and powerful analysis tools Orion substantially increases the scale of problems that can be addressed and makes finding solutions to those problems easy for anyone.

In this workshop we will use Orion to address a frequent problem in medicinal chemistry – using protein structural knowledge to find new lead compounds from a large number of molecules and understanding how these active compounds interact with the protein binding site. To solve this problem effectively we will use a variety of approaches; docking at various levels of accuracy, re-scoring and pose refinement using higher levels of theory. This workflow will proceed from a pool of millions of molecules to produce a few 10's of high probability candidates for experimental validation.

The ability to set up and monitor a large-scale calculation on the cloud, analyse its results, share that analysis and make decisions based on it, all through the same interface, a standard web browser, is extremely powerful. We will illustrate all these capabilities in the course of the workshop.

Workshops Thursday, May 31st

Schrödinger Workshop: Maximizing the impact of Computational Modelling on Drug Design

Thursday May 31st 2018, 14:00-16:00, NH Conference Hotel Noordwijkerhout, Room: Boston 13

LiveDesign is a novel platform delivering cheminformatics and expert computational models side by side in a highly collaborative and intuitive web-based tool. By presenting experimental data alongside predictive data and models, a broad range of scientists can drive new ideas by asking the key questions and easily exploring chemical space.

In this workshop we will introduce LiveDesign in the context of real-world medicinal chemistry workflows. This will range from rapid querying of the existing SAR, through to graphical exploration of experimental and predictive data to aid profiling and prioritization of new ideas. Embedded 3D docking and pharmacophore model visualization is a key component of the LiveDesign platform and we will show how to make the most of this information. We will also show how the administration interface allows modelers to publish validated Glide1 docking models, for use in a selectivity study of COX1 and COX22. Finally we will show how new ideas can easily be pushed and pulled into Maestro for deeper analysis with more complex computational methods, for a truly cyclic workflow.

1. Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren,T. A.; Sanschagrin, P. C.; Mainz, D. T., "Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein-Ligand Complexes" J. Med. Chem., 2006, 49, 6177–6196
2. Plount-Price, M. L.; Jorgensen W. L., “Analysis of Binding Affinities for Celecoxib Analogues with COX-1 and COX-2 from Combined Docking and Monte Carlo Simulations and Insight into the COX-2/COX-1 Selectivity” J. Am. Chem. Soc., 2000, 122 (39), pp 9455–9466

Joint Xemistry & KNIME Workshop: Chemistry Data Workflows – Leveraging Explorative Native KNIME Technology and Xemistry Custom Nodes

Thursday May 31st 2018, 14:00-16:00, NH Conference Hotel Noordwijkerhout, Room: Boston 15

The KNIME software has quickly become a prime player in the chemistry data processing arena. Additional chemistry capabilities are continuously added – as built-in support features, packaged standard nodes, and third-party vendor offerings.

Xemistry and KNIME will present a joint workshop highlighting new chemistry-related developments in and around the KNIME software.

In the first part, Daria Goldmann of KNIME will explain and demonstrate new core chemistry features and interactive analysis and exploration capabilities which support the implementation of reproducible KNIME workflows for chemistry data.

In the second part, Wolf Ihlenfeldt of Xemistry introduces the CACTVS KNIME node builder environment – for those occasions where you need a custom chemistry data processing node which is not available as a turnkey solution, and you really do not want to dig into the intricacies of native KNIME Java development.

Excursion: Sailing Cruise on the IJsselmeer (Lake IJsel) and visit to Batavia Yard

Schedule

13:00	Busses depart from the conference center, Noordwijkerhout
14:00	Arrive at the harbor of Volendam, board the sailing boats *Willem Barentsz* and *Abel Tasman*
16:30	Arrive at Batavia Yard (Lelystad) and join the guided tour
18:00	Return to the ships, dinner will be served on board
22:00	Disembark at Volendam, return to Noordwijkerhout by bus
23:00	Arrive at the conference center

Batavia Yard

The destination of the sailing cruise is the Batavia Yard. As described at their website:

“Batavia Yard is a shipyard with extraordinary ambitions, reconstructing ships from the Golden Age that were important to the Netherlands' maritime history. This heritage was demolished at the time because of its limited lifespan, or has sunk to the bottom of the sea. In April 1995, the Batavia, which is the most authentic reconstruction of a 17th-century VOC ship ever made, was launched after ten years in the making. The initiator was master shipbuilder Willem Vos. After this reconstruction was complete, a second project was started in the yard to reconstruct ‘De 7 Provinciën', a 17th-century battleship with which Michiel de Ruyter fought many sea battles.”

http://www.bataviawerf.nl/who-are-we.html

Scientific Program

Plenary Session

Sunday, May 27

12:00 - 18:00	**Registration** *Atrium Lounge*
15:00 - 17:00	**Pre-conference workshops**
	Application of Matched Molecular Pairs to Interactive SAR Exploration **Workshop** Chemical Computing Group *Boston 13*
	Orion - CADD on the Cloud **Workshop** OpenEye *Boston 15*
17:00 - 18:00	**Free Time**
18:00 - 18:15	**Welcome** *Rotonde*
18:15 - 19:00	**Keynote Address - CSA Trust Mike Lynch Award** **From Teletype Structure Input to Biology and Chemistry Intelligent Knowledge Graphs: My 45 Years in Cheminformatics** Awardee Dr. Rudy Potenzone
19:00 - 20:00	**Welcome Reception** *Atrium*
20:00 - 22:00	**Reception Dinner** *Atrium*

Monday, May 28

8:30 - 10:00	**Session A - Integration of Chemical Information** Herman van Vlijmen, Presiding *Rotonde*
8:30 - 9:00	**A-1: Synthetically Accessible Virtual Inventory (SAVI) – Reaction Generation and Handling at the One-Billion Compounds Scale** Hitesh Jayantilal Patel, National Cancer Institute, United States
9:00 - 9:30	**A-2: Fast Molecular Searching Tools and Their Extension at GSK** Peter Pogany, Glaxo Smith Kline, United Kingdom
9:30 - 10:00	**A-3: Analysis of the ToxCast & Tox21 Compound Set Using Regulator-derived GHS Toxicity Annotations and in silico-derived Protein-target Descriptors** Chad Henry George Allen, University of Cambridge, United Kingdom
10:00 - 10:30	**Coffee Break** *Atrium*
10:30 - 14:30	**Session B - Structure-Activity and Structure-Property Prediction** Andreas Bender, Presiding *Rotonde*
10:30 - 11:00	**B-1: How Do You Build and Validate 1500 Models and What Can You Learn from Them? An Automated and Reproducible System for Building Predictive Models for Bioassay Data** Greg Landrum, KNIME AG, Switzerland
11:00 - 11:30	**B-2: Machine Learning of Partial Charges From QM Calculations and the Application in Fixed-Charge Force Fields and Cheminformatics** Sereina Riniker, ETH Zurich, Switzerland
11:30 - 12:00	**B-3: Artificial Intelligence for Predicting Molecular Electrostatic Potentials (ESPs): A Step Towards Developing ESP-guided Knowledge-based Scoring Functions** Prakash Chandra Rathi, Astex Pharmaceuticals, United Kingdom
12:00 - 13:00	**Lunch** *Atrium*
13:00 - 13:30	**B-4: Next-Generation MD-QSAR Models of Dynamic Kinase-Inhibitor Interactions Based on Machine Learning and Molecular Dynamics** Denis Fourches, North Carolina State University, United States
13:30 - 14:00	**B-5: Automated Selectivity Inversion of Kinase Inhibitors** Simone Fulle, Novo Nordisk, Denmark
14:00 - 14:30	**B-6: Multivariate Regression with Left-censored Data – Efficient Use of Incompletely Measured Bioactivity Data for Predictive Modelling** Knut Baumann, TU Braunschweig, Germany
14:30 - 15:00	**Coffee Break** *Atrium*
15:00 - 19:30	**Poster Session & Exhibition** *Atrium*
15:00 - 17:00	**Poster Presentations Red** Egon Willighagen, Presiding *Atrium*
18:30 - 19:30	**Reception** *Atrium*
19:30 - 21:30	**Dinner** *Atrium*

Tuesday, May 29

08:30 - 14:30	**Session C - Structure-Based Drug Design and Virtual Screening** Esther Kellenberger and Matthias Rarey, Presiding *Rotonde*
08:30 - 09:00	**C-1: In The Need of Bias Control: Evaluation of Chemical Data for Machine Learning Methods in Structure-Based Virtual Screening** Jochen Sieg, University Hamburg, Germany
09:00 - 09:30	**C-2: An Exhaustive Assessment of Computer-Based Drug Discovery Methods by High-Throughput Screening Data** Oliver Koch, TU Dortmund, Germany
09:30 - 10:00	**C-3: Lessons Learned in Benchmarking Virtual Screening for Polypharmacology** E.B. Lenselink, LACDR/Leiden University, Netherlands
10:00 - 10:30	**Coffee Break** *Atrium*
10:30 - 11:00	**C-4: Assisting Site-directed Mutagenesis in silico to Optimize Ligand-Binding** Hugo Gutierrez de Teran, Uppsala University, Sweden
11:00 - 11:30	**C-5: Structural Analysis of Chemokine Receptor-Ligand Interactions for Computational Modelling Integration in Drug Design** Marta Arimont, Vrije Universiteit Amsterdam, Netherlands
11:30 - 12:00	**C-6: Generation of Structure-based Pharmacophore Models in Protein Binding Sites Obtained from Molecular Dynamics Simulations: Application to Understanding Kd of Hsp90 Ligands** Thierry Langer, University of Vienna, Austria
12:00	**GROUP PHOTO**
12:00 - 13:00	**Lunch** *Atrium*
13:00 - 13:30	**C-7: How Significant are Unusual Intermolecular Interactions?** Bernd Kuhn, F. Hoffmann-La Roche, Switzerland
13:30 - 14:00	**C-8: Interaction Pattern Analysis – What are we Missing?** Alexandra Nass, FU Berlin, Germany
14:00 - 14:30	**C-9: Hydrogen Bonds as Determinants of Structural Stability** Maciej Majewski, University of Barcelona, Poland
14:30 - 15:00	**Coffee Break** *Atrium*
15:00 - 19:30	**Poster Session & Exhibition** *Atrium*
15:00 - 17:00	**Poster Presentations Blue** Lars Ridder, Presiding *Atrium*
18:30 - 19:30	**Reception** *Atrium*
19:30 - 21:30	**Conference Dinner** *Atrium*

Wednesday, May 30

08:30 - 10:30	**Structure-Based Drug Design and Virtual Screening II** Matthias Rarey, Presiding *Rotonde*
08:30 - 09:00	**C-10: Selectivity Determining Features in Proteins with Conserved Binding Sites - A Case Study Using N-myristoyltransferase as Model System** Ruth Brenk, University of Bergen, Norway
09:00 - 09:30	**C-11: Active Search for Computer-Aided Drug Design** Steven Andrew Oatley, University of Nottingham, United Kingdom
09:30 - 10:00	**C-12: Conformational sampling of macrocycles in both the solid- and solution-states** Paul Hawkins, OpenEye Scientific, United States
10:00 - 10:30	**C-13: Automated Fragment Evolution (FrEvolAted) Applied to Fragments Bound to NUDT21** Moira Michelle Rachman, University of Barcelona, Spain
10:30 - 11:00	**Coffee Break** *Atrium Lounge*
11:00 - 13:00	**Session D - Analysis of Large Chemical Datasets** Peter Ertl, Presiding *Rotonde*
11:00 - 11:30	**D-1: Hit Dexter 2.0: Machine Learning for Triaging Hits from Biochemical Assays** Johannes Kirchmair, University of Hamburg, Germany
11:30 - 12:00	**D-2: Recent Advances in Chemical and Biological Search Systems: Evolution vs. Revolution** Roger Sayle, NextMove Software, United Kingdom
12:00 - 12:30	**D-3: Advancing Automated Synthesis Via Reaction Data Mining and Reuse** Christos Nicolaou, Eli Lilly and Company, United States
12:30 - 13:00	**D-4: Revealing Important Molecular Fragments in Drug Discovery Using Time Trend Analyses** Barbara Zdrazil, University of Vienna, Austria
13:00	**Box Lunch**
13:00 - 23:00	**Excursion** Cruise the IJsselmeer on two traditional sailing boats and visit the Batavia Yard. A banquet dinner will be served on the boats on the way back.

Thursday, May 31

07:30 - 08:30	**Hotel Check-Out**
08:30 - 10:30	**Session E - Dealing with Biological Complexity** Andrea Volkamer, Presiding *Rotonde*
08:30 - 09:00	**E-1: Strategies for Assembling an Annotated Library for Phenotypic Screening** Henriette Willems, University of Cambridge, United Kingdom
09:00 - 09:30	**E-2: Targeting of the Disease Related Proteome by Small Molecules** Modest von Korff, Idorsia Pharmaceuticals Ltd., Switzerland
09:30 - 10:00	**E-3: Gearing Transcriptomics Towards High-Throughput Screening: Compound Shortlisting from Gene Expression Using in silico Information** Natalia Aniceto, University of Cambridge, United Kingdom
10:00 - 10:30	**E-4: Discrimination of G–protein Coupled Receptors and their Conformational States Using Intramolecular Interaction** Florian Koensgen, University of Strasbourg, France
10:30 - 11:00	**Coffee Break & Hotel Check-Out** *Atrium Lounge*
11:00 - 13:10	**Cheminformatics** ***Dedicated to Peter Willett*** Val Gillet, Presiding *Rotonde*
11:00 - 11:10	**Some remarks** Val Gillet
11:10 - 11:40	**F-1: Comparison and Analysis of Molecular Patterns on the Example of SMARTS** Robert Schmidt, Universität Hamburg, Germany
11:40 - 12:10	**F-2: Anisotropic Atom Reactivity Descriptors for the Prediction of Liver Metabolism, Ames Toxicity and Hydrogen Bonding** Andreas Hans Göller, Bayer AG, Germany
12:10 - 12:40	**F-3: Exploring 3D Molecular Shape Using Spectral Geometry** Matthew Seddon, University of Sheffield, United Kingdom
12:40 - 13:10	**F-4: Creating Atom-to-Atom Mapping in Chemical Reaction Using Machine Learning Methods** Timur Madzhidov, Kazan Federal University, Russia
13:10 - 13:15	**Closing Remarks**
13:15 - 14:00	**Lunch or Box Lunch**
13:30	**Shuttle Busses leave for Shiphol Airport**
14:30	**Shuttle Busses leave for Shiphol Airport**
14:00 - 16:00	POST-CONFERENCE WORKSHOPS
	Maximizing the Impact of Computational Modelling on Drug Design Workshop by Schrödinger *Boston 13*
	Chemistry Data Workflows – Leveraging Explorative Native KNIME Technology and Xemistry Custom Nodes Joint Workshop by Xemistry & KNIME *Boston 15*
16:30	**Shuttle Busses leave for Shiphol Airport**

Poster Session RED

Integration of Chemical Information

Structure-Activity and Structure-Property Prediction

Structure-Based Drug Design and Virtual Screening

Poster Session BLUE

Analysis of Large Chemical Data Sets

Dealing with Biological Complexity

Cheminformatics

Plenary Session Abstracts

Keynote Address CSA Trust Mike Lynch Award

From Teletype Structure Input to Biology and Chemistry Intelligent Knowledge Graphs: My 45 Years in Cheminformatics

Rudy Potenzone, Ph. D.

Ingentium Inc.

A short review will be presented of the changes and incredible advances that have occurred over the past 45 years in cheminformatics, and related scientific informatics. While our scientific knowledge has developed at an incredible pace, it has come alongside of the advances in computer hardware and software. Advances in the capabilities and accessibility of chemical and biological information has been amazing but dwarfed by the increasing volume. As we enter the Fourth Paradigm of scientific research and discovery, the availability of machine learning and cybernetics offers an opportunity to leverage the vast amounts of information. Finding relevant sources is challenging as they are spread across normal scientific channels as well as the press, social media, and in various forms including audio and video. At Ingentium, we have been studying how to organize and mine this information as well as how today's scientists want to consume it. We have creating disease focused knowledge bases and 'magazines' for browsing, a portal for searching and extended tools for reviewing our collections. In focusing on specific topics, a richer, more focused context can be mapped and made available to research scientists in the various forms including new content alerts, readable content summaries, related items and knowledge graphs. Examples of our magazines and knowledge graphs will demonstrate the value of our approach.

Session A:

INTEGRATION OF CHEMICAL INFORMATION

A-1: Synthetically Accessible Virtual Inventory (SAVI) – Reaction Generation and Handling at the One-Billion Compounds Scale

H. Patel [1], W. D. Ihlenfeldt [2], M. C. Nicklaus [1]

[1] Computer-Aided Drug Design Group, Chemical Biology Laboratory, Center for Cancer Research, National Cancer Institute, National Institutes of Health, NCI-Frederick, Frederick, MD 21702, United States, [2] Xemistry GmbH, D-61462 Königstein, Germany.

The Synthetically Accessible Virtual Inventory (SAVI) project is an international collaboration between partners in government laboratories, small companies, not-for-profits, and large corporations, to computationally generate a very large number of reliably and inexpensively synthesizable novel screening sample structures. SAVI handles reactions not by virtue of applying simple SMIRKS to a set of building blocks of unknown availability. It instead combines a set of transforms richly annotated with chemical context, coming from, or being newly developed in the mold of, the original LHASA project knowledgebase, with a set of highly annotated, reliably available, purchasable starting materials. These components are tied together for SAVI product generation with the chemoinformatics toolkit CACTVS with custom developments for this project. Each product is annotated with a number of computed properties seen as important in current drug design, including rules for identifying potentially reactive or promiscuous compounds. After having produced and made publicly available the first (beta) set of 283 million SAVI products annotated with proposed one-step syntheses, we will be reporting on the second full production run aimed at creating a database of one billion high-quality, easily synthesizable screening samples. We will present the current status, ongoing developments, as well as scientific and technical challenges of the project.

A-2: Fast molecular searching tools and their extension at GSK

P Pogany [1], T Kostrzewski [1], S Senger [1], S Pickett [1]

[1] GlaxoSmithKline, Stevenage, Hertfordshire SG1 2NY UK

Searching pharmaceutically relevant chemical datasets is an integral part of the lead discovery workflow and an important source of ideas for lead optimization. It is important that such tools are readily available at the desktop and that results can be obtained in an interactive fashion. Some of these datasets can contain hundreds of millions of drug-like molecules and pose a challenge to traditional database systems for similarity searching and other cheminformatics tasks.

We have implemented the ChemAxon tool MadFast[1,2] for this purpose at GSK. We have used MadFast with large datasets (ca. 180 million) vendor compound collections and connected to an in-house adapted SureChEMBL[3,4] database containing patent compounds and their mapping to the patent information. Reduced graph fingerprints[5] have been added to the existing datasets as an alternative to the regular chemical hashed fingerprints and extended connectivity fingerprints. MadFast search has been made available through the LiveDesign[6] platform and a separate implementation used for patent compound lookup in the SureChEMBL database. From both implementations it is possible to link out to the original datasources to provide additional information and context for any hit compounds. In this work we present our implementation and and illustrate the use in lead discovery and lead optimization stages of drug discovery programs.

1. https://chemaxon.com/products/madfast
2. Pickett S., ChemAxon UGM, Budapest 2016. https://chemaxon.com/presentation/fast-similarity-searching-making-the-virtual-real
3. Papadatos, G.; Davies, M.; Dedman, N.; Chambers, J.; Gaulton, A.; Siddle, J.; Koks, R.; Irvine, S. A.; Pettersson, J.; Goncharoff, N.; Hersey, A.; Overington, J. P. SureChEMBL: a large-scale, chemically annotated patent document database. *Nucleic acids res.* **2015**, 44.D1, D1220-D1228.
4. Senger, S.; Bartek, L.; Papadatos, G.; Gaulton, A. Managing expectations: assessment of chemistry databases generated by automated extraction of chemical structures from patents., *J. Cheminform.* **2015**, 7, 49.
5. 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.
6. https://www.schrodinger.com/livedesign

A-3: Analysis of the ToxCast & Tox21 compound set using regulator-derived GHS toxicity annotations and *in silico*-derived protein-target descriptors

C. H. G. Allen[1], L. H. Mervin[1], A. Bender[1]

[1] *Centre for Molecular Informatics, Department of Chemistry, University of Cambridge, U.K.*

Accurate *in silico* prediction of compound toxicity is of value to the chemical industry because traditional toxicity testing is slow and expensive, and there are societal and legal incentives to minimize *in vivo* experimentation. Meanwhile, the regulatory demand for toxicity information is higher than ever. Conventional *in silico* approaches seek relationships between chemical structure and adverse outcomes, and the integration of high-throughput *in vitro* screening data and protein target annotations has been shown to increase toxicity classification models' applicability and interpretability.[1] However, a challenge in developing such heterogeneous models is the collation of suitable datasets, which require the presence of a toxicological endpoint, chemical structure, protein targets, and *in vitro* readouts for each compound. Finding a suitably large number of compounds with such an overlap of data on which to train predictive models is not trivial.

The Globally Harmonized System of Classification and Labelling (GHS) is an international framework for standardising chemical hazard information. *Inter alia*, the GHS facilitates the collation of the outcomes of independent acute oral toxicity studies into internationally-recognised categories, for the purposes of producing globally-recognized hazard labels. Five acute oral toxicity categories are defined, each corresponding to a quantitative LD_{50} interval specified in mg/kg, with three categories labelled "toxic", one "hazardous" and one requiring no label. The European Chemicals Agency (ECHA), Japan's National Institute of Technology and Evaluation and New Zealand's Environmental Protection Authority provide public access to governmentally mandated/recommended acute oral toxicity classifications under the GHS. Further, ECHA publishes the industrial submissions it receives under the requirements of the EU's REACH legislation, which include declaring GHS classifications. The common classification standards provided by the GHS system enable the collation of acute oral toxicity data from all of these resources with confidence that they are mutually commensurate. This represents a valuable means of annotating arbitrary compound sets with toxicity labels.

In our study, we annotated 8,003 unique standardized chemical structures from the ToxCast and Tox21 datasets[3,4] with toxicity classifications derived from regulatory GHS information; toxicity classifications could be found for 2736 (34%) of the structures, illustrating the coverage of this technique. For these compounds, a set of 206 physiochemical and structural descriptors were calculated using MOE.[5] The dataset was further annotated with 1,651 *in silico*-derived protein-target descriptors using an in-house random forest (RF) protein-ligand prediction algorithm trained on over 13 million bioactivity datapoints.[6] We next analysed the predictability of regulator-derived toxicity annotations using clustering and linear discrimination analysis on the chemical and protein-target descriptors, the ToxCast/Tox21 qHTS assay data, and combinations of these spaces. We show the performance of RF classifiers (evaluated by the ROC and precision-recall curves) and the effect of the inclusion of the different combinations of heterogeneous descriptors on these models' interpretability and applicability domains.

1. Allen, C. H. G.; Koutsoukas, A.; Cortés-Ciriano, I.; Murrell, D. S.; Malliavin, T. E.; Glen, R. C.; Bender, A. Improving the prediction of organism-level toxicity through integration of chemical, protein target and cytotoxicity qHTS data. *Toxicol. Res.* **2016**, *5*, 883–894.
2. United Nations. *Globally Harmonized System of Classification and Labelling of Chemicals (GHS)*, 7th revised ed.; New York and Geneva, 2017; Chapter 3.1, pp 115–125. http://www.unece.org/trans/danger/publi/ghs/ghs_rev07/07files_e0.html (accessed Jan 20, 2018).
3. Kavlock, R. J.; Austin, C. P.; Tice, R. R. Toxicity testing in the 21st century: Implications for human health risk assessment. *Risk Anal.* **2009**, *29*, 485–487**.**
4. Dix, D. J.; Houck, K. A.; Martin, M. T.; Richard, A. M.; Setzer, R. W.; Kavlock, R. J. The ToxCast program for prioritizing toxicity testing of environmental chemicals. *Toxicol. Sci.* **2007**, *95*, 5–12**.**
5. *Molecular Operating Environment (MOE)*, 2013.08; Chemical Computing Group ULC, 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, **2018**.
6. Mervin, L. H.; Bulusu, K. C.; Kalash, L.; Afzal, A. M.; Svensson, F.; Firth, M. A.; Barret, I.; Engkvist, O.; Bender, A. Orthologue chemical space and its influence on target prediction. *Bioinformatics*. **2018**, *34*, 72–79.

Session B:

STRUCTURE-ACTIVITY AND STRUCTURE-PROPERTY PREDICTION

B-1: How do you build and validate 1500 models and what can you learn from them? An automated and reproducible system for building predictive models for bioassay data

G.A. Landrum[1], D. Goldmann[2], A. Martin[3]

[1] KNIME AG, Zurich, Switzerland, [2] KNIME GmbH, Berlin, Germany, [3] KNIME GmbH Konstanz, Germany

Here we describe an automated workflow for building and training predictive models for bioassay data and the application of that workflow to train and validate more than 1500 predictive models for the assay data present in ChEMBL 23. The workflow is implemented with the KNIME Model Factory [1] in the open-source KNIME Analytics Platform. Since we know that there is no single best machine-learning algorithm/chemical fingerprint combination for all datasets [2], our workflow tries a variety of different fingerprints and algorithms for each assay and selects the one that performs best. The breadth of methods we use, and the automation of the process sets this effort apart from other large-scale modeling exercises with ChEMBL [3].

We begin with an overview of the KNIME Model Factory itself, and then describe the individual steps used to build and validate the predictive models:

1) Selection and extraction of the datasets
2) Feature generation
3) Model building: parameter optimization, model building, model selection
4) Model validation
5) Model deployment

We've also started analyzing the models themselves and will close with a presentation of what we've learned so far about combinations of fingerprints/algorithms/parameters which seem to work well across this very large collection of different datasets.

The KNIME Model Factory is open source and can be freely downloaded from our website (URL provided during the presentation). The files for the final models and datasets are large, but we will also make those available upon request. Although we have worked with public data (ChEMBL), applying the workflow described here to other data sources (for example an internal data warehouse) would only require modification of the section that selects and extracts the data from the database.

1. Adä, I.; Winters, P.; Berthold, M.R. The KNIME Model Factory: Scaling Modeling Processes for the Enterprise. [Online] **2017**. https://files.knime.com/sites/default/files/inline-images/Model_Process_Management_20170404_1.pdf (accessed 14 Feb 2018).
2. Riniker, S.; Fechner, N.; Landrum, G. A. Heterogeneous classifier fusion for ligand-based virtual screening: or, how decision making by committee can be a good thing. *J. Chem. Inf. Model.* **2013**, 53 2829–2836.
3. http://chembl.blogspot.com/2014/04/ligand-based-target-predictions-in.html, http://chembl.blogspot.ch/2016/03/target-prediction-models-update.html

B-2: Machine Learning of Partial Charges From QM Calculations and the Application in Fixed- Charge Force Fields and Cheminformatics

Sereina Riniker [1]

[1] Laboratory of Physical Chemistry, ETH Zurich, Vladimir-Prelog-Weg 2, 8093 Zurich, Switzerland

Partial charges are a highly important component of fixed-charge force fields, which are used in classical molecular dynamics (MD) simulations. Partial charges are also widely used as descriptors in quantitative structure-activity relationship (QSAR) or quantitative structure-property relationship (QSPR) models. To obtain partial charges one has typically to choose between speed and accuracy. The vastness of the chemical space makes fast approaches using building blocks or connectivity information challenging. Therefore, a common approach used for force fields

is to extract partial charges from semi-empirical or *ab initio* calculations. However, the high computational cost of QM methods limits the use of this approach to low throughput applications. In order to obtain high- quality partial charges in a fast manner, we have developed a machine-learning (ML) based approach for predicting partial charges extracted from density functional theory (DFT) electron densities.[1] The training set was chosen with the goal to provide a broad coverage of the known chemical space of drug-like molecules. In addition to the speed of the approach, the partial charges predicted by ML are not dependent on the three-dimensional conformation in contrast to the ones obtained by fitting to the electrostatic potential (ESP). The quality and the compatibility of the ML-predicted partial charges with standard force fields is assessed by calculating thermodynamic properties of organic liquids. In addition, the chemically meaningful partial charges obtained by the presented ML-based approach are tested in high-throughput ligand-based virtual screening.

1. Bleiziffer, P.; Schaller, K.; Riniker, S. Machine Learning of Partial Charges Derived From High-Quality Quantum-Mechanical Calculations. *J. Chem. Inf. Model.* **2018**, submitted.

B-3: Artificial Intelligence for Predicting Molecular Electrostatic Potentials (ESPs): A Step Towards Developing ESP-guided Knowledge-based Scoring Functions

P. C. Rathi [1], R. Lewis [2], A. Bender [2], and M. L. Verdonk [1]

[1] Astex Pharmaceuticals, 436 Cambridge Science Park, Milton Road, Cambridge CB4 0QA, United Kingdom, [2] Department of Chemistry, Centre for Molecular Informatics, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom

Molecular electrostatic potential (ESP) surfaces are a useful tool for optimizing protein-ligand interactions in drug design.[1] It has been shown that binding efficiency of a ligand can be greatly improved by careful modifications that lead to a better complementarity between protein and ligand electrostatics.[2] We believe that knowledge-based scoring functions can be improved by leveraging the information about electrostatic potentials around atoms. However, time-intensive quantum mechanical calculations are required for generating molecular ESP surfaces. Therefore, to develop a fast ESP-guided knowledge-based scoring function, a quick and accurate model is required for the prediction of molecular ESPs.

We will present a model for predicting ESPs around atoms (specifically, in the direction of atomic features, e.g., lone pairs, hydrogens, p orbitals, etc.) developed using graph convolutional deep neural network techniques.[3] The network was trained on ~48,000 diverse molecules (mean heavy atom count ≈ 11). For each molecule, ESP surfaces were generated by running quantum mechanical (QM) calculations (B3LYP with 6-31G* basis set for optimization and 6-311G** basis set for energy calculation). The input layer of the model receives 209 atomic descriptors calculated using the scikit-chem library built on RDKit[4] plus atomic connectivity matrices. The output layer returns ESP values in the direction of the atomic features. The optimized model performs very well on predicting ESP values for a validation set of ~12,000 molecules (R^2 = 0.95, p << 0.001 for a correlation with ESP values derived using QM calculations). The mean absolute error in predicting ESP values is ~3 kcal/mol for the validation set suggesting that our model can provide good estimates of ESP values obtained using time-intensive QM calculations, but in a fraction of the computing time. This level of precision should also allow a successful application in guiding knowledge-based scoring functions, and we will exemplify how this may be achieved for the Protein Ligand Informatics force field (PLIff).[5]

1. Vinter, J. G. Extended electron distributions applied to the molecular mechanics of some intermolecular interactions. II. Organic complexes. *J. Comput-Aided. Mol. Des.* **1996,** *10*, 417-426.
2. Chessari, G.; Buck, I. M.; Day, J. E.; Day, P. J.; Iqbal, A.; Johnson, C. N.; Lewis, E. J.; Martins, V.; Miller, D.; Reader, M. Fragment-based drug discovery targeting inhibitor of apoptosis proteins: discovery of a non-alanine lead series with dual activity against cIAP1 and XIAP. *J. Med. Chem.* **2015,** *58*, 6574-6588.
3. Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular graph convolutions: moving beyond fingerprints. *J. Comput-Aided. Mol. Des.* **2016,** *30*, 595-608.
4. The scikit-chem library (https://github.com/richlewis42/scikit-chem), built on RDKit (http://www.rdkit.org).
5. Verdonk, M. L.; Ludlow, R. F.; Giangreco, I.; Rathi, P. C. Protein-Ligand-Informatics force field (PLIff): towards a fully knowledge driven "force field" for biomolecular interactions. *J. Med. Chem.* **2016,** *59*, 6891-6902.

B-4: Next-Generation MD-QSAR Models of Dynamic Kinase-Inhibitor Interactions Based on Machine Learning and Molecular Dynamics

Denis Fourches [1]

[1] *Department of Chemistry, Bioinformatics Research Center, North Carolina State University, Raleigh, USA*

Quantitative Structure-Activity Relationships (QSAR) typically rely on the two- and three-dimensional structures of molecules to assess their bioactivity. These models have proven to be accurate enough for screening large chemical libraries but have also shown poor performances when it comes to lead optimization and in-depth assistance to medicinal chemists. In this presentation, I will present the MD-QSAR modeling approach that uses machine learning and MD descriptors directly computed from the molecular dynamics trajectories of kinase-inhibitor complexes. I will discuss the rationale of the approach, its origin with early attempts of 4D-QSAR modeling, and two case studies involving a set of 85 ERK2 kinase inhibitors[1] and another large set of 925 Bcr-Abl tyrosine-kinase inhibitors (all being imatinib analogues). Our MD-QSAR modeling workflow includes *(i)* the structure-based docking of all compounds in the binding site of the kinase, *(ii)* the independent molecular dynamics (MD) simulations of each protein-ligand complex (Desmond-GPU, 15 ns, NTP, 300K, TIP3P, 1fs), *(iii)* the computation of MD fingerprints to characterize ligands' conformational flexibility and the dynamic kinase-inhibitor interactions over the trajectories, and *(iv)* both training and cross-validation of MD-QSAR models using machine learning techniques (random forests and artificial neural networks). Not only MD-QSAR models afforded similar or better prediction performances compared to classical 2D and 3D QSAR models, but the interpretation of MD descriptors was facilitated with the direct visualization of their associated dynamic kinase-inhibitor interactions. This next-generation modeling workflow combining machine learning, 3D docking, and molecular dynamics simulations could provide key knowledge for the design of more potent and selective small molecule inhibitors.

1. Ash, J.; Fourches, D. Characterizing the Chemical Space of ERK2 Kinase Inhibitors Using Descriptors Computed from Molecular Dynamics Trajectories. *J Chem Inf Model*. **2017**, 57, 1286−1299.

B-5: Automated selectivity inversion of kinase inhibitors

Simone Fulle

BioMed X Innovation Center, Heidelberg, Germany
Current affiliation: Novo Nordisk, Copenhagen, Denmark

Elimination of inadvertent binding is crucial for inhibitor design targeting conserved protein classes like kinases. In turn, compounds in clinical trials provide a rich source for initiating drug design efforts by exploiting such secondary binding events. Considering both aspects, we shifted the selectivity of a kinase inhibitor, originally developed against a cancer target, towards a pain target. In line with the design objectives, the top-ranked compound has a significant selectivity improvement against a selected off-target and is highly selective in a kinase panel. This was achieved in a single round of automated *in silico* optimization, highlighting the power of recent advances in computer-aided drug design technologies to automate design and selection processes.

The presentation will describe the employed multi-objective selection scheme that filters for selective and highly active compound based on orthogonal methods grounded in computational chemistry and machine learning. The benefit of the underlying technologies (e.g. ref. [1-3]), primarily developed for the design of selective inhibitors, will be exemplarily demonstrated and discussed using our novel compound series for a pain target.

1. Merget, B.; Turk, S.; Eid, S.; Rippmann, F.; Fulle, S. Profiling Prediction of Kinase Inhibitors: Toward the Virtual Assay. *J Med Chem*. **2017**, 60, 474-485.
2. Turk, S.; Merget, B.; Rippmann, F.; Fulle, S. Coupling Matched Molecular Pairs with Machine Learning for Virtual Compound Optimization. *J Chem Inf Model*. **2017**, 57, 3079-3085.
3. Eid, S.; Turk, S.; Volkamer, A.; Rippmann, F.; Fulle, S. KinMap: a web-based tool for interactive navigation through human kinome data. *BMC Bioinformatics*. **2017,** 18:16.

B-6: Multivariate Regression with Left-censored Data – Efficient Use of Incompletely Measured Bioactivity Data for Predictive Modelling

K. Baumann [1], M. Mathea [1,#], W. Klingspohn [1], A. ter Laak [2], N. Heinrich [2]

[1] *Institute of Medicinal and Pharmaceutical Chemistry, Technische Universität Braunschweig, Beethovenstraße 55, 38106 Braunschweig, Germany*

[2] *Bayer AG, Drug Discovery, Pharmaceuticals, 13342 Berlin, Germany*

[#] *Current address: BASF SE, 67056 Ludwigshafen am Rhein, Germany*

In industrial drug discovery research, bioactivity data are often incompletely measured so that for weakly active compounds no exact pIC_{50} or pK_i value is known. In these cases, it is only known that the pIC_{50} or pK_i value is smaller than a certain cut-off value. Data of this type are called left-censored. Such data frequently occur in econometrics and environmental chemistry (values below the determination or detection limit) and efficient regression algorithms to include those data into calibration models are well known[1]. However, in the latter two application areas only few predictors are typically processed. Regression algorithms for handling hundredths or thousands of predictors are not available. For right-censored survival data, high-dimensional regression algorithms have been published. Yet, the censoring mechanism in these cases is often very different from the one at work for the aforementioned left-censored data. Hence, not all available algorithms can efficiently be adapted to the left-censored case.

Here, we describe the adaption of the Buckley-James algorithm[2] to principal component regression (PCR) and partial least-squares regression (PLS), as well as to other penalized regression algorithms for handling high-dimensional structure descriptor data with left-censored bioactivity data. Two different implementations are conceivable: In the first case, the regression algorithm is left untouched and the Buckley-James imputation scheme is adapted to left-censored data using a reverse Kaplan-Meier estimator. In the second case, the aforementioned regression algorithms are decomposed into many univariate regression steps for matrix decomposition and each univariate regression step is replaced by the respective Buckley-James regression. This has the advantage that cross-validation schemes can be implemented more efficiently.

Numerical stability and predictive capability is equivalent for the different implementations. Although the adaption is rather straightforward, pitfalls are possible. Critical issues with respect to estimating intercepts and the predictive ability will be discussed. The specifically tailored regression algorithms will be compared to the naïve case where the censored data are handled as if they were uncensored and to the case where simply all censored data are removed from the data set. Not surprisingly, the tailored regression algorithms use the data more efficiently and thus perform better. The performance differences will be discussed with simulations and real data.

1. Helsel, D. R.. *Statistics for censored environmental data using Minitab® and R*. John Wiles & Sons, Hoboken, NJ, USA, 2012, 2nd Ed.
2. Buckley, J.; James, I. Linear regression with censored data. *Biometrika*. **1979**, 66, 429-436.

Session C:

STRUCTURE-BASED DRUG DESIGN AND VIRTUAL SCREENING

C-1: In The Need of Bias Control: Evaluation of Chemical Data for Machine Learning Methods in Structure-Based Virtual Screening

Jochen Sieg[1], Florian Flachsenberg[1], Matthias Rarey [1]

[1] *Center for Bioinformatics, Hamburg, Germany*

Currently, machine learning (ML) methods receive increasing attention. This includes the field of structure-based virtual screening, where these methods are used for predicting binding of small molecules to protein targets. Improved predictions for the scoring of protein-ligand complexes in comparison to established empirical scoring functions are reported for example with convolutional neural networks (CNNs)[1-3]. However, trained ML models are treated often as black boxes and are not straightforward interpretable[4]. The difficulty of interpretation makes it laborious to identify which features and patterns are responsible for activity prediction and makes these methods prone to unnoticed bias.

New methods are usually evaluated by retrospective validation on benchmark datasets[5]. Different ML methods have achieved impressive results on commonly used benchmark datasets. Exemplary, utilizing CNNs on the *Directory of Useful Decoys* (DUD)[5] and *Directory of Useful Decoys – Enhanced* (DUD-E)[6] values of the area under the receiver operating characteristic curve (AUC) of 0.81[2] and 0.86[3] have been reported, respectively, for discriminating active and inactive molecules. Thus, it seems that these datasets are no challenge for these ML methods. Nevertheless, the question of the true prospective predictive capability and the applicability domain remains.

Benchmark datasets are usually designed for a specific evaluation scenario. While DUD and DUD-E both have been developed for the evaluation of structure-based virtual screening (SBVS) methods, the *Maximum Unbiased Validation* (MUV)[7] dataset is a benchmark for ligand-based virtual screening. A benchmark dataset can be seen as a selected chemical subspace that is appropriate to constitute good test cases for a specific group of methods and descriptors. Although the prediction task might remain the same, a benchmark might be inapplicable once methods or descriptors change. A benchmark dataset designed for SBVS with empirical scoring functions is not necessarily be suited for ML. As a consequence, unphysical bias might become the cause for over-estimated performance.

We show exemplary on current literature that it is possible to learn bias unobserved and implicitly. Specifically, we show that the molecules property of being active against any target can be learned with only ligand features for example from an established benchmark datasets like DUD with an AUC of 0.83. Here, we present a new approach aiming at more realistic estimates on SBVS performance. Our approach utilizes domain knowledge to recognize good performance caused by unphysical data patterns when applying a specific composition of methods and descriptors to a given dataset as illustrated in Figure 1. Therefore, it is possible to identify descriptor and method combinations that cause unreasonable good performance on the given dataset, which helps to choose a suitable dataset for validation.

1. Our findings suggest that there is a need for bias control in the validation of machine learning methods. For this reason we propose best practice guidelines for designing validation experiments to identify and control bias. Furthermore, these steps can be used to create new benchmark datasets with reduced risk for implicit bias. Ragoza, M.; Hochuli, J.; Idrobo, E.; Sunseri, J.; Koes D. R. Protein–Ligand Scoring with Convolutional Neural Networks *Journal of Chemical Information and Modeling* **2017**, 57 (4), 942-957
2. Pereira, J. C.; Caffarena, E. R.; Santos, C. N. Boosting Docking-Based Virtual Screening with Deep Learning *J*ournal of Chemical Information and Modeling 2016**,** 56 (12), 2495-2506
3. Wallach, I.; Dzamba, M.; Heifets, A. AtomNet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery **2015**, *arXiv preprint* arXiv:1510.02855
4. Polishchuk, P. Interpretation of Quantitative Structure–Activity Relationship Models: Past, Present, and Future Journal of Chemical Information and Modeling 2017**,** 57 (11), 2618-2639
5. Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking *Journal of Medicinal Chemistry* **2006**, 49 (23), 6789–6801
6. Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking *Journal of Medicinal Chemistry* **2012**, 55 (14), 6582–6594
7. Rohrer, S. G.; Baumann, K. Maximum unbiased validation (MUV) data sets for virtual screening based on PubChem bioactivity data *Journal of Chemical Information and Modeling* **2009**, 49 (2), 169–184

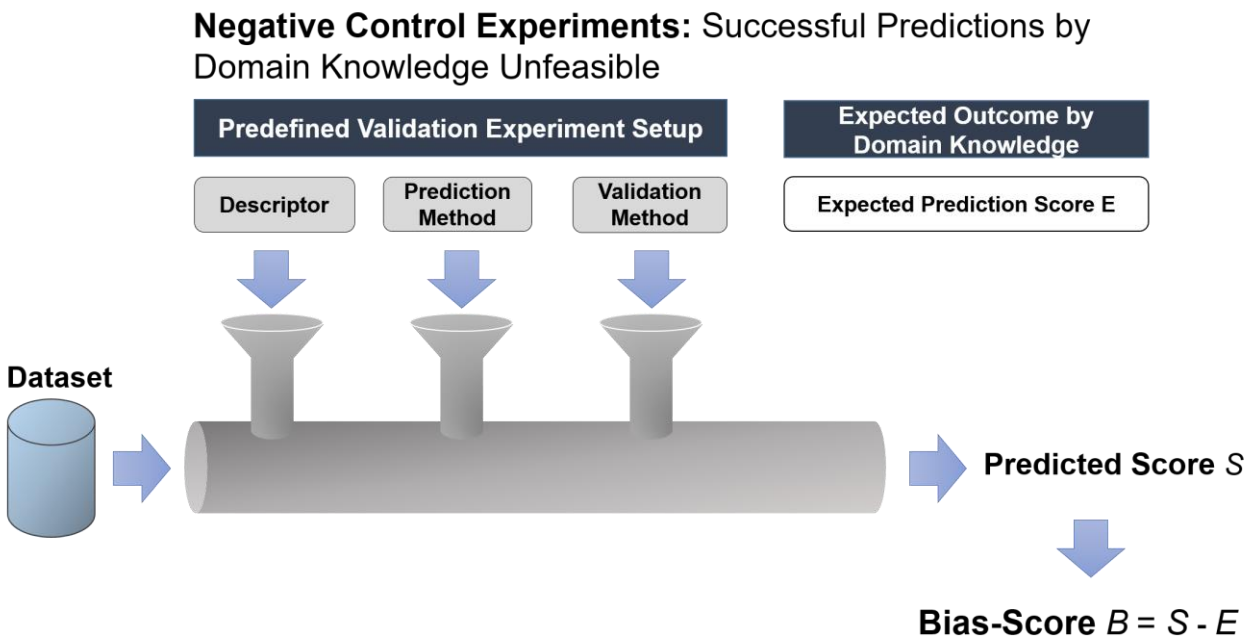


Figure 1: Workflow of our approach to identify bias. This workflow uses domain knowledge on predefined validation experiments to identify property distributions in a given dataset which are violating physical reality. The difference of the expected score and the predicted score are used to calculate a bias score.

C-2: An Exhaustive Assessment of Computer-Based Drug Discovery Methods by High-Throughput Screening Data

Christiane Ehrt[1], Dennis M. Krüger[1,2], Tom Mejuch[3], Sonja Sievers[3], Herbert Waldmann[3], Oliver Koch[1]

[1]Faculty of Chemistry and Chemical Biology, TU Dortmund, Dortmund, Germany, [2]Chemical Genomics Centre of the Max Planck Society, Dortmund, Germany , [3]Max Planck Institute of Molecular Physiology, Dortmund, Germany

In silico methods, especially virtual screening approaches, proved to be useful sources of inspiration for drug discovery as well as suitable tools for modern drug design [1], although there are potential pitfalls [2]. Most *in silico* screening approaches highly rely on experimental data which can be either three-dimensional structures of the proteins of interest or knowledge about ligands binding to the respective target. In the ideal case, the research is guided by the knowledge of protein-ligand complex structures.

Here, the outcome of a virtual screening study is presented that aimed to identify small molecule ligands for a protein whose structure was solved in the presence of a peptide ligand. The impact of different computer-based methods on the virtual screening performance was assessed without previous knowledge of small molecule binders and exclusively based on one X-ray structure. A combination of MD simulations, hot-spot analyses, pharmacophore searches and docking approaches was used to identify potential ligands and circumvent virtual screening pitfalls.

In contrast to the popular method of performance assessment using benchmarking data sets [3] the screened database of about 150,000 compounds was subsequently tested experimentally. This data enabled a detailed analysis of the performance of this exhaustive structure-guided virtual screening approach in an unbiased manner. In addition, the abundance of available experimental data provides the opportunity to oppose ligand-based and structure-based in silico screening approaches in a retrospective manner. Thus, the experimental data was used to analyse the maximum enrichment that could be obtained during pharmacophore screening and molecular docking.

I will present and discuss the outcome of this prospective screening and compare these results to the knowledge-based results to finally answer the question: How much knowledge is needed to save time and money during drug discovery?

1. Tanrikulu, Y., Krueger, B., Proschak, E.: The Holistic Integration of Virtual Screening in Drug Discovery. *Drug Discov. Today* **2013**, 18, 358-364.
2. Scior, T., Bender, A., Tresadern, G., Medina-Franco, J.L., Martínez-Mayorga, K., Langer, T., Cuanalo-Contreras, K., Agrafiotis, D.K. Recognizing Pitfalls in Virtual Screening: A Critical Review. *J. Chem. Inf. Model.* **2012**, 52, 867-881.
3. Lagarde, N., Zagury, J-F., Montes, M. Benchmarking Data Sets for the Evaluation of Virtual Ligand Screening Methods: Review and Perspectives. *J. Chem. Inf. Model.* **2015**, 55, 1297-1307.

C-3: Lessons learned in benchmarking Virtual Screening for polypharmacology.

E.B. Lenselink [1], L. Burggraaff[1], B.J. Bongers[1], X. Liu[1], M. Gorostiola-Gonz Ález[1], J. van Engelen[2], H. Hoos[2], J.K. Wegner[3], M. Steijaert[4], W. Jespers[1,5], Hugo Gutiérrez-de-Terán[5], H.W.T van Vlijmen[1,3], A.P. IJzerman[1], G.J.P van Westen[1].

1 Division of Drug Discovery and Safety, Leiden, The Netherlands, 2 Leiden Institute of Advanced Computer Science, Leiden, The Netherlands, 3 Janssen Pharmaceutica NV, Beerse, Belgium, 4 Open Analytics NV, Antwerp, Belgium 5 Department of Cell and Molecular Biology, Uppsala, Sweden

Polypharmacology is typically regarded as a drawback in drug discovery, as side effects might occur due to interactions with other targets than the main target. However, it has been estimated that on average a drug will interact with at least 6 targets, questioning the fact if true selectivity exists.[1] In November 2017 the Multi-Targeting DREAM challenge was launched with the aim to Virtually Screen the ZINC database for molecules that adhere to a profile of targets and anti-targets.[2]

Here the results will be presented of the rigorous benchmarking we performed prior to running the actual Virtual Screen for the DREAM challenge. Our philosophy was to select an optimal workflow per protein target, consisting of three successive stages: statistical modelling, ensemble docking, and metadynamics. For the statistical modelling we used models[3] created on public data (i.e. ExCAPE[4] and ChEMBL[5]). Models were benchmarked and compared, and predictions of the best models were used to filter compounds to proceed to the docking stage. Benchmarking of docking was performed based on active compounds, inactive compounds, and decoys.[6] We selected 5 high enriching X-rays for an ensemble, using the Z2 score[7] of both the docking scores and SPLIF.[8, 9] This ensemble yielded high, predictive BEDROC and ROC scores for most targets and anti-targets. Finally, for the primary targets the top 100 ranking compounds were also scored using binding pose metadynamics enriching the results even further.[10] In general this successive Virtual Screening workflow can be applied to any target with sufficient data.

1. J. Mestres; E. Gregori-Puigjane; et al. Data completeness—the Achilles heel of drug-target networks. *Nat. Biotechnol.* **2008**, 26, 983-984.
2. Schlessinger, A.; Abagyan, R.; et al., Multi-targeting Drug Community Challenge. *Cell Chem. Biol.* **2017**, 24, 1434-1435.
3. E.B. Lenselink; N. ten Dijke; et al., Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. *J. Cheminf.* **2017**, 9, 45.
4. Sun, J.; Jeliazkova, N.; et al., ExCAPE-DB: an integrated large scale dataset facilitating Big Data analysis in chemogenomics. *J. Cheminf.* **2017**, 9, 17.
5. A. Gaulton; L.J. Bellis; et al., ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, 40, D1100-7.
6. M.M. Mysinger; M. Carchia; et al., Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J. Med. Chem* **2012**, 55, 6582-6594.
7. Sastry, G. M.; Inakollu, V. S.;et al., Boosting virtual screening enrichments with data fusion: coalescing hits from two-dimensional fingerprints, shape, and docking. *J. Chem. Inf. Model.* **2013**, 53, 1531-1542.
8. E.B. Lenselink; W. Jespers; et al., Interacting with GPCRs: Using Interaction Fingerprints for Virtual Screening. *J. Chem. Inf. Model.* **2016**, 56, 2053-2060.
9. Da, C.; Kireev, D., Structural protein–ligand interaction fingerprints (SPLIF) for structure-based virtual screening: method and benchmark study. *J. Chem. Inf. Model.* **2014**, 54, 2555-2561.
10. A.J. Clark; P. Tiwary; et al., Prediction of protein–ligand binding poses via a combination of induced fit docking and metadynamics simulations. *J. Chem. Theory Comput.* **2016**, 12, 2990-2998.

C-4: Assisting site-directed mutagenesis in silico to optimize ligand-binding

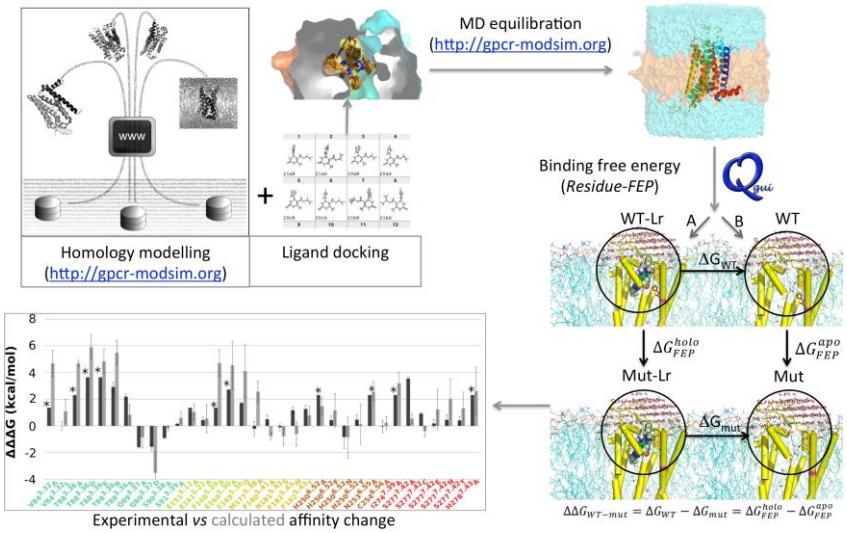
Hugo Gutiérrez de Terán, Willem Jespers, Silvana Vasile, Johan Åqvist.

Department of Cell and Molecular Biology, Uppsala University, Biomedical Center, Box 596, SE-751 24 Uppsala, Sweden. E-mail: hugo.gutierrez@icm.uu.se

Site-directed mutagenesis (SDM) is a powerful and widely used tool to understand ligand-binding at the structural and molecular level. The characterization of ligand binding affinities against a set of mutant proteins, interpreted by computational modeling, is a process that has been used in the hit-to-lead optimization of many drug targets, with the GPCR superfamily of membrane receptors being a paradigmatic example due to the traditional lack of structural information.

I will outline our recently developed computational scheme, based on free energy perturbation (FEP) simulations, to quantitatively and routinely assess the effects of point-mutations on ligand binding (Fig 1).[1] The procedure is based on an MD sampling of the protein-ligand binding site, using spherical boundary conditions centered on the binding site, which makes it computationally efficient. The methodology is now automated and will be soon released as part of the Q-gui graphical interface of our MD software Q, where it can be combined with classical FEP simulations on ligands series, providing a full picture of the energetics of ligand binding in the scope of mutagenesis data or ligand-SAR.

Recent applications include assisting on antagonist design on the A_{2A} and A_3 adenosine receptors, and deorphanization of receptor GPR139,[2] in collaborative projects with medicinal chemists and pharmacologists. I will focus here on explaining our the most recent results of this application, centered on understanding agonist binding to the type 2 (Y2) neuropeptide Y receptor, to assist on further ligand optimization in collaboration with scientists from Novo Nordisk.



1. Keranen, H.; Aqvist, J.; Gutierrez-de-Teran, H. *Chem Commun.* **2015**, 51, 3522
2. Nøhr AC, Jespers W, Shehata MA, et al. *Sci Rep.* **2017**, 7, 1128.
3. Xu, Vasile et al. *Mol Pharmacol.* **2018**, *in press*

C-5: Structural analysis of chemokine receptor-ligand interactions for computational modelling integration in drug design.

M. Arimont[1], S. Sun[1], M. Vass[1], A.J. Kooistra[1,2], R. Leurs[1,3], I.J.P. de Esch[1,3], C. de Graaf[1]

[1] Division of Medicinal Chemistry, Faculty of Sciences, Amsterdam Institute of Molecules, Medicines and Systems (AIMMS), Vrije Universiteit Amsterdam, Amsterdam, The Netherlands. [2] Centre for Molecular and Biomolecular Informatics (CMBI) Radboudumc, Nijmegen, The Netherlands. [3] Griffin Discoveries BV, Department of Medicinal Chemistry, Amsterdam, The Netherlands.

Construction and application of structural chemokine receptor models are essential for the elucidation of molecular determinants of chemokine receptor modulation and the structure-based discovery and design of chemokine receptor ligands[1,2]. We present a comparative analysis of ligand binding pockets in chemokine receptors and their implication on modeling receptor-ligand interactions[2]. This is specially challenging in chemokine receptors as they present multiple druggable binding sites, including the minor and major pocket in the orthosteric site (small molecules/peptides), the extracellular vestibule (chemokines, antibodies), and even an intracellular binding site (small molecules, nanobodies). We will show how this data can be integrated with pharmacological data using structural chemoinformatics worflows[3,4] and applied in drug discovery. We will also present how the integration of new structural information of chemokine receptors with extensive structure-activity relationship and site-directed mutagenesis data is necessary for the prediction of the structure of chemokine receptor-ligand complexes that have not been crystallized yet. Finally we will illustrate how molecular dynamics simulations and analyses combined with the use of structural interaction fingerprints are key tools for optimization of molecular models that can be used for protein-based virtual screening approaches. Structure-based ligand discovery and design studies based on chemokine receptor crystal structures and homology models illustrate not only the possibilities, but also the challenges to find novel ligands for chemokine receptors.

1. Scholten, D. J.; Canals, M.; Maussang, D.; Roumen, L.; Smit, M. J.; Wijtmans, M.; de Graaf, C.; Vischer, H. F.; Leurs, R., Pharmacological modulation of chemokine receptor function. *Br J Pharmacol* **2012,** *165* (6), 1617-1643.
2. Arimont, M.; Sun, S. L.; Leurs, R.; Smit, M.; de Esch, I. J. P.; de Graaf, C., Structural
3. Analysis of Chemokine Receptor-Ligand Interactions. *J Med Chem* **2017,** *60* (12), 4735-4779.
4. McGuire, R.; Verhoeven, S.; Vass, M.; Vriend, G.; de Esch, I. J.; Lusher, S. J.; Leurs, R.; Ridder, L.; Kooistra, A. J.; Ritschel, T.; de Graaf, C., 3D-e-Chem-VM: Structural Cheminformatics Research Infrastructure in a Freely Available Virtual Machine. *J Chem Inf Model* **2017,** *57* (2), 115-121.
5. Kooistra, A. J.; Vass, M.; McGuire, R.; Leurs, R.; de Esch, I. J.; Vriend, G.; Verhoeven, S.; de Graaf, C., 3D-e-Chem: Structural Cheminformatics Workflows for Computer-Aided Drug Discovery. *ChemMedChem* **2018**.

C-6: Generation of Structure-based Pharmacophore Models in Protein Binding Sites Obtained from Molecular Dynamics Simulations: Towards Understanding K_D of Hsp90 Ligands

T. Seidel[1], D. Schütz[2], M. Körbel[1], A. Garon[1], M. Wieder[1], G. Ibis[2], G. F. Ecker[1], T. Langer[1]

[1] University of Vienna, Vienna, Austria, [2] Inte:Ligand GmbH, Vienna, Austria

Structure-based pharmacophore models are usually derived from known three-dimensional structures of active ligands (i.e. small organic molecules) bound to a protein target of interest in their active conformation[1]. In many different application domains such models have been proven to be useful as selective *in silico* screening filters.[2]

Recently, we have extended the static pharmacophore approach by a dynamic one, deriving interaction models from molecular dynamics trajectory snapshots[3] and including also a novel consensus screening approach, which was shown to be superior to previous pharmacophore-based virtual screening methods.[4]

One of the main benefits of performing molecular dynamics simulations of protein-ligand complexes is the possibility to detect global changes in protein geometry, and thus enabling the observation of emerging pockets of

potential interest for the formation of additional ligand-protein interactions. To address this challenge, we have developed an algorithm that is able to detect transient protein pockets and to place pharmacophore features in such empty target binding sites without the guidance of a known bound-state ligand structure. The generated features are placed and oriented in the protein pocket in a way that an optimal interaction with complementary binding partners in the receptor environment is ensured. The thus derived dynamic *apo pharmacophore models* provide invaluable information that can be put to good use for the *de novo* design of new ligands as well as for the refinement of existing lead structures in the drug development process.

Details about the algorithm developed together with the results of its validation with a series of protein conformation snapshots obtained from molecular dynamics simulations of Hsp90 ligand complexes will be presented.

1. Wolber, G.; Langer T. LigandScout: 3D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J. Chem. Inf. Model*. **2005**, 45, 160-169.
2. Langer, T. Pharmacophores in Drug Research. *Mol. Inf.*, **29**, 470-475.
3. Wieder, M.; Perricone, U.; Boresch, S.; Seidel, T.; Langer, T. Evaluating the stability of pharmacophore features using molecular dynamics simulations. *Biochem. Biophys. Res. Comm.* **2016**, 470, 685-689.
4. Wieder, M.; Garon, A.; Perricone, U.; Boresch, S.; Seidel, T.; Almerico, A.M.; Langer, T. Common Hits Approach: Combining Pharmacophore Modeling and Molecular Dynamics Simulations. *J. Chem. Inf. Model*. **2017**, 57, 365-385.

C-7: How significant are unusual intermolecular interactions?

Bernd Kuhn, Oliver Korb

Roche Pharmaceutical Research and Early Development, Innovation Center Basel, F. Hoffmann-La Roche Ltd, 4070 Basel, Switzerland

In recent years a large number of novel interaction types have been postulated to have a stabilizing effect on protein-ligand complex formation. However, the significance for some these "unusual" interactions has yet to be validated with experimental and theoretical studies of model systems as well as statistical analyses of crystallographic databases. We have pursued the latter approach and extended the recently published line-of-sight analysis by Taylor[1] to protein-ligand complexes from the Protein Data Bank. With this method confounding secondary interactions are pruned out and statistically significant interaction propensities for different functional groups can be derived. In addition this approach provides insights into the geometric preferences of intermolecular contacts.

As a result of our studies we will present crystal structure based statistical analyses of different interaction types and highlight preferred protein environments of selected functional groups of relevance for medicinal chemistry. This will be complemented by illustrative examples from drug discovery projects.

1. Taylor, R. Which Intermolecular Interactions Have a Significant Influence on Crystal Packing? *CrystEngComm* **2014**, 16, 6852-6865.

C-8: Interaction Pattern Analysis – What are we Missing?

A. Naß [1], G. Wolber [1]

[1] *Molecular Design Lab, Pharmaceutical and Medicinal Chemistry, Institute of Pharmacy, Berlin, Germany*

The main principles found in almost all high-affinity protein-ligand complexes are high steric complementarity, high complementarity of surface properties and an energetically favourable ligand conformation.

However, static structures only show a small part of the whole picture: For example is the entropic contribution to the binding energy of a ligand usually not observable in static structures. Further have flexible protein parts been shown to prefer flexible ligand moieties over rigid ones which can hardly be investigated with one static structure. Only a few methods like the recently developed Dynophores[1] (dynamic pharmacophores based on molecular dynamics data) take into account the flexibility of both protein and ligand for interaction analysis.

Unfortunately, classic interaction analysis tools are mainly focusing on one of the three principles discovered in high affinity binding complexes neglecting the steric complementarity and favourable ligand conformation. Consequently, a tool was created and implemented in R that can quantify shape complementarity of a ligand to a protein over a molecular dynamics simulation of the bound ligand. The tool further supports the calculation of shape fit per ligand atom and therefore allows statements about favourable and unfavourable parts of the ligand in terms of shape fit. Ligand strain energy is monitored simultaneously to detect good shape fit on the expense of extremely unfavourable ligand conformations due to trapping of the ligand in the protein binding site.

The tool was validated on slightly selective PTP1B ligands to explain the observed activity differences in PTP1B and the closely related TC-PTP which are not explainable by interaction based methods.

Since shape complementarity could also be a valuable means to increase selectivity in cases where no ligand is available as starting point, a second tool was developed: It allows identification of selectivity relevant binding site areas especially for cases where interaction feature patterns are highly similar and facilitates exploitation of the discovered differences for virtual screening. Calculations are based on clustering of binding site shape point maps extracted from molecular dynamics frames with the help of the open source tool POVME2[2]. The binding site shape clustering tool was also validated on the test case of PTP1B/TC-PTP in order to identify active site inhibitors of PTP1B with increased selectivity.

Both tools developed in this study address the issue of selectivity in a flexible protein-ligand context with different scenarios of available input data and therefore provide novel opportunities to design ligand selectivity in challenging cases.

1. Bock, A.; Bermudez, M.; Krebs, F.; Matera, C.; Chirinda, B.; Sydow, D.; Dallanoce, C.; Holzgrabe, U.; De Amici, M.; Lohse, M. J.; Wolber, G.; Morh, K. Ligand Binding Ensembles Determine Graded Agonist Efficacies at a G Protein-Coupled Receptor. *J. Biol. Chem.* **2016,** 291(31), 16375-16389.
2. Durrant, J. D.; de Oliveira, C. A.; McCammon, J. A. Povme: an Algorithm for Measuring Binding-Pocket Volumes. *J. Mol. Graph. Model.* **2011**, 29(5), 773-776.

C-9: Hydrogen bonds as determinants of structural stability

M. Majewski [1], S. Ruiz-Carmona [1], X. Barril [2]

[1]Facultat de Farmàcia and Institut de Biomedicina, Universitat de Barcelona, Barcelona, Spain, [2]Catalan Institution for Research and Advanced Studies (ICREA), Spain

Structural stability is a fundamental property of protein-ligand complexes that so far has been ignored in drug design. It can be provided by hydrogen bonds (HBonds), thanks to their sharp distance and angular dependencies[1]. Certain HBonds present strong opposition to small structural distortions and can act as kinetic traps. The local environment hinders the transition from a direct HBond to a water-bridged interaction[2]. As an early unbinding event, rupture of the so-called water-shielded HBonds can influence the whole dissociation process. The concept has been recently implemented in the Dynamic Undocking (DUck)[3], a new method consists of series of steered molecular dynamics. During the simulation, the ligand is being pulled from the bound to the Quasi-Bound state, in which the ligand has just broken the most important HBond with the receptor. The value of work consumed in the process (W_{QB}) is an effective factor associated with structural stability.

Here we present a first large scale assessment of robustness of HBonds. We have calculated W_{QB} for every single HBond in a subset of 77 protein-ligand complexes from the Iridium data set[4] (total 341 HBonds). HBond-driven structural stability is very common in protein-ligand complexes. Strong HBonds can be found in 75% of complexes and tend to group in fragment-sized structural anchors. For the remaining structures, with weak HBonds, other stability-providing interactions have been identified. Furthermore, additional calculations have shown that we can modulate the strength of the HBond by modifying the ligand. Manipulating the microenvironment around a HBond has important implication for structural stability and is a useful drug design principle.

1. Bissantz, C.; Kuhn, B.; Stahl, M. A medicinal chemist's guide to molecular interactions. *J. Med. Chem.* **2010,** 53, 5061-5084.
2. Schmidtke, P.; Luque, F. J.; Murray, J. B.; Barril, X. (2011). Shielded hydrogen bonds as structural determinants of binding kinetics: application in drug design. *J. Am. Chem. Soc.*, **2011,** 133, 18903-18910.

3. Ruiz-Carmona, S.; Schmidtke, P.; Luque, F. J.; Baker, L., Matassova, N.; Davis, B.; Roughley, S.; Murray, J.; Hubbard, R.; Barril, X. Dynamic undocking and the quasi-bound state as tools for drug discovery. *Nature Chemistry*, **2017,** 9, 201.
4. Warren, G. L.; Do, T. D.; Kelley, B. P.; Nicholls, A.; Warren, S. D. Essential considerations for using protein-ligand structures in drug discovery. *Drug Discovery Today*, **2012,** 17, 1270-1281.

C-10: Selectivity determining features in proteins with conserved binding sites - a case study using N-myristoyltransferase as model system

R Brenk [1], FC Kersten [2]

[1] *University of Bergen, Department of Biomedicine, Bergen Norway, City, Country,* [2] *Johannes Gutenberg University, Institute of Pharmacy, Mainz, Germany*

One particular challenge in structure-based design is how to derive selective inhibitors for proteins with conserved binding sites. To investigate this topic on the molecular level, we studied a model system of two related enzymes, namely N-myristoyltransferase (NMT) from *L. major* and *H. sapiens* The binding sites of both enzymes are highly conserved (Figure 1). Nevertheless, unselective and selective inhibitors were developed.[1–3]

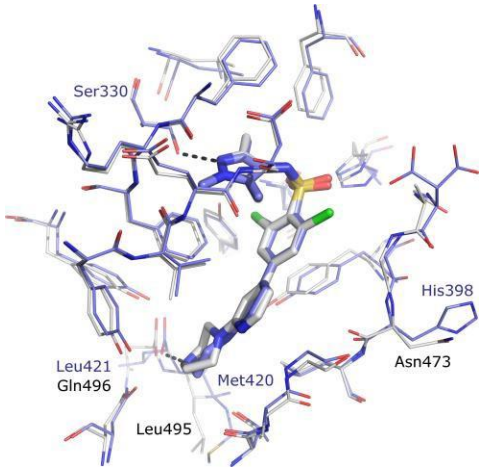


Figure 1 Superposition of *L. major* and *H. sapiens* NMT binding sites together with unselective ligand.

We used a combination of molecular dynamic simulations, isothermal titration calorimetry, enzyme inhibition assay, site-directed mutagenesis, and X-ray crystallography to analyse protein dynamics, water network formation and their changes upon ligand binding. Using this approach, for two compound series to two different selectivity-determining features were identified. For one series, a change in protein flexibility upon ligand binding seemed to responsible for selective inhibition. In the other compound series, selectivity was caused by the ability to displace a highly conserved water molecule. Based on these finding, a virtual screening for selective compounds was conducted resulting in three hit compounds with the desired selectivity profile.

1. Frearson, J. A.; Brand, S.; McElroy, S. P.; Cleghorn, L. A.; Smid, O.; Stojanovski, L.; Price, H. P.; Guther, M. L.; Torrie, L. S.; Robinson, D. A.; Hallyburton, I.; Mpamhanga, C. P.; Brannigan,
2. J. A.; Wilkinson, A. J.; Hodgkinson, M.; Hui, R.; Qiu, W.; Raimi, O. G.; van Aalten, D. M.; Brenk, R.; Gilbert, I. H.; Read, K. D.; Fairlamb, A. H.; Ferguson, M. A.; Smith, D. F.; Wyatt, P. G. *Nature* **2010**, *464* (7289), 728.
3. Brand, S.; Norcross, N. R.; Thompson, S.; Harrison, J. R.; Smith, V. C.; Robinson, D. A.; Torrie, L. S.; McElroy, S. P.; Hallyburton, I.; Norval, S.; Scullion, P.; Stojanovski, L.; Simeons,
4. F. R.; van Aalten, D.; Frearson, J. A.; Brenk, R.; Fairlamb, A. H.; Ferguson, M. A.; Wyatt, P. G.; Gilbert, I. H.; Read, K. D. *J. Med. Chem.* **2014**, *57* (23), 9855.
5. Brannigan, J. A.; Roberts, S. M.; Bell, A. S.; Hutton, J. A.; Hodgkinson, M. R.; Tate, E. W.; Leatherbarrow, R. J.; Smith, D. F.; Wilkinson, A. J. *IUCrJ* **2014**, *1* (4), 250.

C-11: Active Search for Computer-Aided Drug Design

S. Oatley [1], D. Oglic [2,3], S. Macdonald [4], T. McInally [1], R. Garnett [5], T. Gärtner [2], J. Hirst [1]

[1] School of Chemistry, University of Nottingham, Nottingham, UK, [2] School of Computer Science, University of Nottingham, Nottingham, UK, [3] Institut für Informatik III, Universität Bonn, Bonn, Germany, [4] GlaxoSmithKline, Stevenage, UK, [5] Department of Computer Science and Engineering, Washington University in St. Louis, St. Louis, USA

Chemical space is large, to the point of precluding its explicit enumeration. Thus, it represents a so-called intensionally defined design space. Search strategies for intensionally designed spaces are a current area of interest in machine learning. In the context of a drug design problem, we have investigated the application of a data-driven adaptive Markov chain approach, where the acceptance probability is given by a probabilistic surrogate of the target property, modelled with a maximum entropy conditional model.[1,2] We apply the approach to a lead discovery search for inhibitors of an α_v integrin, using a molecular docking score as the optimisation function. α_v integrins are currently an important target for the treatment of a number of fibrotic diseases e.g. idiopathic pulmonary fibrosis, an increasingly prevalent lung disease. These integrins are large, bidirectional transmembrane signalling proteins that share a common RGD binding motif. Our algorithm is (i) soundly based in machine learning; (ii) proposes structures from an implicitly defined space of potential designs; (iii) is guaranteed to converge; and (iv) achieves a large structural variety of proposed target structures, some of which provoke significant interest from a medicinal chemistry perspective.

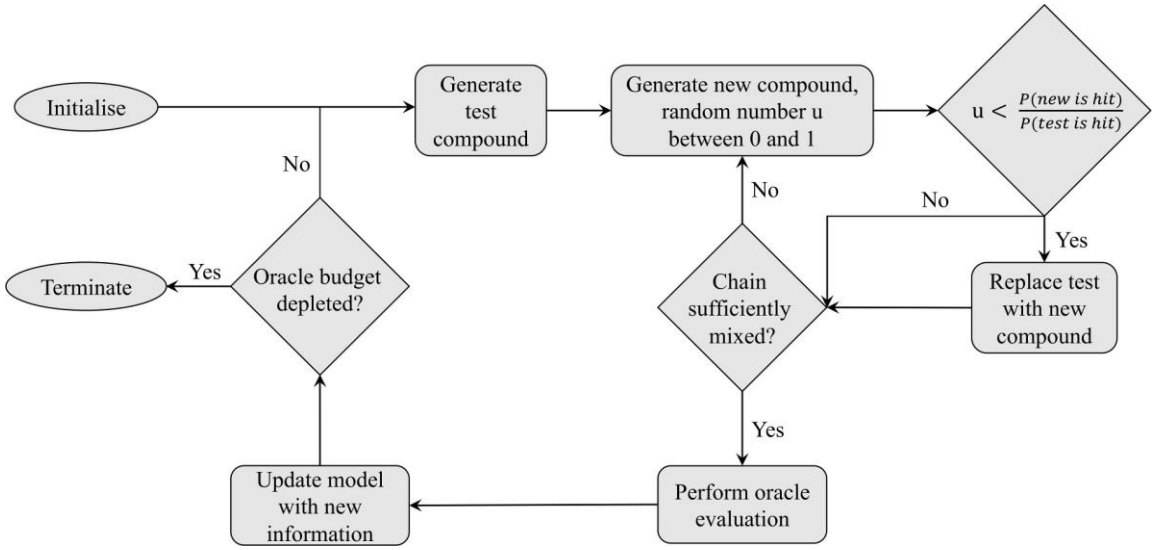


Figure 1. Overview of the machine learning alogirthm used.

The algorithm is summarised in Figure 1. The parent compound (Figure 2) is substituted, with a bias toward lower molecular weight, along with other restrictions from synthetic and medicinal chemistry considerations. The Markov chain Monte Carlo algorithm is designed to propose compounds that maximally increase the known information. This is achieved by accepting new compounds according to a Metropolis criterion based on an estimate of theprobability the current model predicts the compounds as hits. Once the Markov chain has been sufficiently mixed, the compound is evaluated *in silico* and the information returned used to update the model. This iterates until a number (budget) of evaluations has been reached. Using this active algorithm shows, with as few as 100 evaluations, an approximately two-fold improvement in predicting hits over standard Monte Carlo.

Molecules are represented as nodes (atoms) and vertices (bonds), i.e., as graphs, according to the Weisfeiler-Lehman graph kernel. Compounds are proposed and generated by the algorithm and passed to the *in silico* 'oracle', where 3D coordinates are generated and the protonation state is set to that at pH 7.4. Conformers are generated using OpenEye Omega. These are passed to the molecular docking program, OpenEye FRED[3], to be docked to an $\alpha_v\beta_6$ crystal structure, 4UM9, from the RCSB database. The search space was centred on the Thr221 residue in the centre of the binding site and extended past important features, the Mg^{2+} ions and the Asp218 residue. This gave a total search volume of 17,010Å^3. The search was performed using the *chemgauss4* scoring function with a final grid spacing of 0.5Å.

A docked compound can be seen in Figure 3, showing a 3-Cl substituent. Important interactions with the Asp218 residue and chelation with the central Mg^{2+} are present, with distances of 1.83 Å and 2.65 Å, 2.94Å, respectively. Across five simulations, with a search space of around 185,000 compounds and an oracle budget of 500, compounds with high activity from previous synthetic efforts[4] were discovered, some multiple times[2] in addition to previously mentioned novel compounds of significant interest to medicinal chemists.

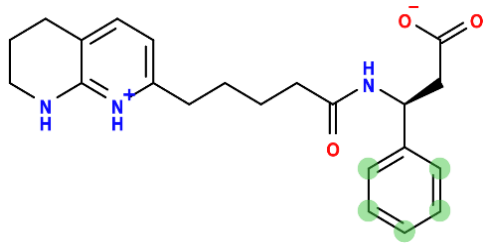


Figure 2. The parent compound considered in this study; green circles denote points where substituents could be attached.

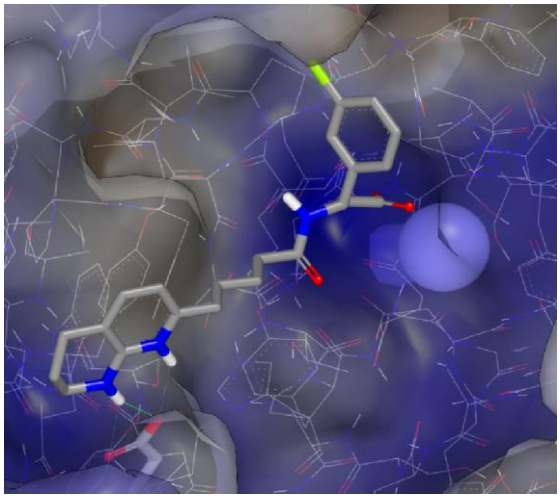


Figure 3. 3-Cl substituent docked in the binding site, Mg^{2+} shown as CPK and Asp218 as stick, displayed using OpenEye VIDA.

1. Oglic, D.; Garnett, R.; Thomas, G. Active Search in Intensionally Specified Structured Spaces. *Proc. 31th Conf. Artif. Intell. (AAAI 2017)* **2017**, 2443–2449.
2. Oglic, D.; Oatley, S. A.; Macdonald, S. J. F.; McInally, T.; Garnett, R. Active Search for Computer-Aided Drug Design. *Mol. Inform.* **2018**, *In Press*.
3. McGann, M. FRED Pose Prediction and Virtual Screening Accuracy. *J. Chem. Inf. Model.* **2011**, *51* (3), 578–596.
4. Adams, J.; Anderson, E. C.; Blackham, E. E.; Chiu, Y. W. R.; Clarke, T.; Eccles, N.; Gill, L. A.; Haye, J. J.; Haywood, H. T.; Hoenig, C. R.; Kausas, M.; Le, J.; Russell, H. L.; Smedley, C.; Tipping, W. J.; Tongue, T.; Wood, C. C.; Yeung, J.; Rowedder, J. E.; Fray, M. J.; McInally, T.; Macdonald, S. J. F. Structure Activity Relationships of α_v Integrin Antagonists for Pulmonary Fibrosis by Variation in Aryl Substituents. *ACS Med. Chem. Lett.* **2014**, *5* (11), 1207–1212.

C-12: Conformational sampling of macrocycles in both the solid- and solution-states

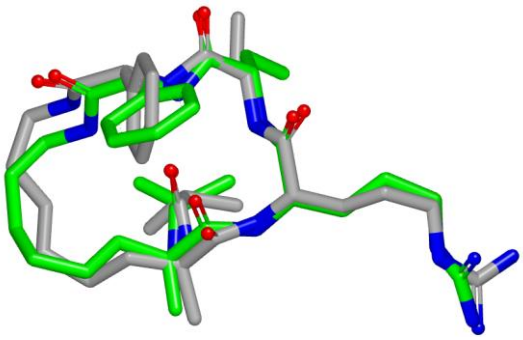
Paul C. D. Hawkins [1] & Stanislaw Wlodek [1]

[1] *OpenEye Scientific, Santa Fe, USA.*

Molecules containing large rings, or macrocycles, have become of greater and greater interest to the drug discovery community over the past decade. A key part of productively exploiting this class of molecules as therapeutics is understanding their conformational landscape, and there have been a number of different approaches to this problem presented recently.[1,2] Here we will present a new approach to macrocycle conformation sampling based on distance geometry, OMIGEN. In the most extensive comparison performed to date in this area we evaluate OMIGEN against a wide variety of other algorithms in reproducing conformations found in the solid-state, the most popular approach to validating conformer generators.

While conformations found in the solid-state are easy to validate against, and are relevant to a number of problems in macrocycle design, including pose prediction by docking and structure-guided lead optimization, generating conformations relevant to the solution state is also important. We will present preliminary data on the use of distance geometry to generate conformations consistent with experimental data from NMR experiments.

1. Sindikhara, D.; Spronk, S. A.; Day, T.; Borrelli, K.; Cheney, D. L.; Posy, S. L. Improving Accuracy, Diversity and Speed with Prime Macrocycle Conformational Sampling. *J. Chem. Inf. M odel.* **2017**, 57, 1881-1894.
2. Coutsias, E. A.; Lexa, K. W.; Wester, M. J.; Pollock, S. N.; Jacobsen, M. P. Exhaustive Conformational Sampling of Complex Fused Ring Macrocycles Using Inverse Kinematics. *J. Chem. Theory Comput.* **2016**, 12, 4674-4687.



C-13: Automated Fragment Evolution (FrEvolAted) Applied to Fragments Bound to NUDT21

Moira Rachman [1], Serena Piticchio [1], Xavier Barril [1,2]

1 Facultat de Farmàcia and Institut de Biomedicina, Universitat de Barcelona, Av. Joan XXIII 27-31, 08028 Barcelona, Spain, 2 Catalan Institution for Research and Advanced Studies (ICREA), Passeig Lluís Companys 23, 08010 Barcelona, Spain

In the last twenty years, FBDD has proven to be a successful method, evident from resulting drugs that have already been marketed and those currently undergoing clinical trials.[1] FBDD is an appealing approach due to its ability to explore a broader chemical space, however, once a fragment has been found to bind, growing the fragment into a lead-like compound is a challenge.[2,3]

For this reason, we have developed FrEvolAted (Automated Fragment Evolution), a computational procedure that automatically evolves fragments into lead-like molecules, whereby, the evolved ligands are extracted from commercially available or synthetically tractable ligand databases. The FrEvolAted workflow (*Figure 1*) includes, I) similarity searching of ligands containing a maximum of two heavy atoms more, II) tethered docking with rDock[4], whereby the main scaffold does not deviate from the initial fragment, III) dynamic undocking (DUck[5]) utilizing crucial receptor-ligand information and IV) MMGBSA-minimization for consensus scoring. In this work, we apply FrEvolAted to fragments bound to the NUDT21 protein provided by XChem and compare the results to a more traditional fragment growing approach in terms of hit rate and novelty.

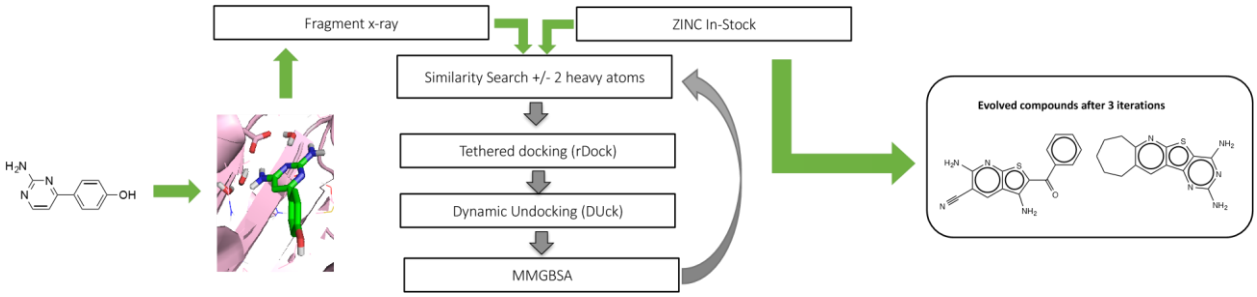


Figure 1. Schematic overview of the FrEvolAted (Automated Fragment Evolution) platform.

1. Erlanson, D. A.; Fesik, S. W.; Hubbard, R. E.; Jahnke, W.; Jhoti, H. *Nat. Rev. Drug Discov.* **2016**.
2. Schulz, M. N.; Hubbard, R. E. *Curr. Opin. Pharmacol.* **2009**, *9* (5), 615–621.
3. Hall, R. J.; Mortenson, P. N.; Murray, C. W. *Prog. Biophys. Mol. Biol.* **2014**, *116* (2–3), 82–91.
4. Ruiz-Carmona, S.; Alvarez-Garcia, D.; Foloppe, N.; Garmendia-Doval, A. B.; Juhos, S.; Schmidtke, P.; Barril, X.; Hubbard, R. E.; Morley, S. D. *PLoS Comput. Biol.* **2014**, *10* (4), 1–7.
5. Ruiz-Carmona, S.; Schmidtke, P.; Luque, F. J.; Baker, L.; Matassova, N.; Davis, B.; Roughley, S.; Murray, J.; Hubbard, R.; Barril, X. *Nat. Chem.* **2017**, *9* (3), 201–206.

6.

Session D:

ANALYSIS OF LARGE CHEMICAL DATASETS

D-1: Hit Dexter 2.0: Machine Learning for Triaging Hits from Biochemical Assays

J. Kirchmair [1], C. Stork [1], J. Wagner [1], N.-O. Friedrich,[1] C. de Bruyn Kops[1] and M. Šícho [1,2]

[1] Universität Hamburg, MIN Faculty, Department of Computer Science, Center for Bioinformatics, Hamburg, Germany, [2] CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Laboratory of Informatics and Chemistry, Faculty of Chemical Technology, University of Chemistry and Technology Prague, Prague, Czech Republic

High-throughput screening is a key technology in early drug design that enables the experimental testing of tens of thousands of compounds per day.[1] However, false-positive signals triggered by badly behaving compounds—frequent hitters, pan-assay interference compounds (PAINS), aggregators and others—continue to pose a major pitfall in early drug discovery and still lead to a substantial number of false hits reported as valid active compounds in the scientific literature.[2] Few computational approaches that allow the identification of badly behaving compounds exist, and those that do offer limited applicability and accuracy.

In this contribution we report the further development of Hit Dexter, a free web service that allows the identification of badly behaving compounds with high accuracy.[3] The initial release of Hit Dexter included two extremely randomized tree classifiers trained on a well-prepared dataset of 311k compounds that have been tested on at least 50 different proteins. Hit Dexter is able to discriminate non-promiscuous from promiscuous and highly-promiscuous compounds of large external test sets with MCC and AUC values of up to 0.67 and 0.96, respectively.

Since the initial release of Hit Dexter we have refined the data preparation and modeling procedures. We have also added several new components that allow, e.g., the identification of true promiscuous binders that may be of particular interest in the context of polypharmacology and drug repurposing. In this talk we will also provide evidence that the reach and accuracy of the established methods for the identification of badly behaving compounds are not sufficient and showcase their limitations in case studies.

The talk will conclude with the introduction of the Hit Dexter 2.0 web service, which, for the first time, will provide researchers a simple tool for testing the likelihood of their hit compounds of being (i) true promiscuous binders, (ii) badly behaving compounds or (iii) "dark chemical matter". Importantly, Hit Dexter 2.0 reports detailed information on the data underlying a prediction, which will enable researchers to make better-informed decisions on the further perusal of their hit compounds.

1. Macarron, R.; Banks, M. N.; Bojanic, D.; Burns, D. J.; Cirovic, D. A.; Garyantes, T.; Green, D. V. S.; Hertzberg, R. P.; Janzen, W. P.; Paslay, J. W.; Schopfer, U; Sittampalam, S. Impact of High-Throughput Screening in Biomedical Research. *Nat. Rev. Drug Discov.* **2011**, 10, 188–195.
2. Baell, J.; Walters, M. A. Chemistry: Chemical Con Artists Foil Drug Discovery. *Nature* **2014**, 513, 481–483.
3. Stork, C.; Wagner, J.; Friedrich, N.-O.; de Bruyn Kops, C.; Šícho, M.; Kirchmair, J. Hit Dexter: A Machine-Learning Model for the Prediction of Frequent Hitters. *ChemMedChem* **2017**, DOI 10.1002/cmdc.201700673.

D-2: Recent Advances in Chemical and Biological Search Systems: Evolution vs. Revolution

R. A. Sayle [1], J. W. Mayfield [1], N. M. O'Boyle [1]

[1] NextMove Software, Cambridge, United Kingdom

The fields of cheminformatics and bioinformatics are both embroiled in perpetual wars against the exponential growth of the scientific data on which they build. The number of small molecules available to chemists and of sequences (and genomes) analyzed by biologists appear to double at astounding rates. For example, the number of "make-on-demand" molecules available for purchase from Enamine doubles every 6 months (from 171M in April 2017, to 337M in October 2017, to 647M in April 2018). This rate of increase is significantly faster than the rates of technological hardware advances, such as those predicted by Moore's law, creating an ever increasing challenge to scientific researchers and the informatics and IT groups that support them.

The fundamental crux of the problem is that the performance of many applications typically scales proportionally to the size of the input data, a situation termed $O(n)$ in computer science. This means that if a database grows to ten times its original size, searching it takes ten times as long (or requires ten times as many computers). This presents two possible strategies; to either make the existing $O(n)$ searches as fast as possible, or to radically change the approaches used by switching to "sublinear" methods which have better scaling properties. In this presentation, we term these strategies "evolution" and "revolution" respectively, and describe examples of recent progress in both.

Chemical similarity searching, using the Tanimoto coefficient to compare binary ECFP fingerprints, is an example of $O(n)$ search. Despite much research effort, and attempts to apply advanced spatial indexing data structures, all modern chemical database search systems need to inspect the majority of a database when searching for non-trivial numbers of nearest neighbor compounds. Hence state-of-the-art search systems such as ChemAxon's madfast and Dalke Scientific's ChemFP attempt to solve the problem by pure brute-force speed. In this talk, we describe the use of optimizing Just-In-Time compilation and advanced sorting techniques to push modern multicore hardware, such as Intel/AMD CPUs and NVidia GPUs as fast as they (their memory) will go.

A more revolutionary strategy in 2D chemical similarity searching is the use of graph databases to greatly accelerate the calculation of the Graph Edit Distance (GED) [and Maximum Common Subgraph (MCS)] between query molecules and their nearest neighbors in a chemical database. At the cost of pre-calculating and storing a large number of subgraphs, chemical database searching then requires consideration of only a tiny fraction of the original database. This significantly sublinear behavior promises to solve the challenge of exponential database growth. We report progress on constructing a graph database index of chemical space with less than 99 bonds, which currently contains about 200 virtual subgraphs for each real input molecule, and which corresponds to over 80 billion subgraphs, but requires only 6 terabytes of disk space using advanced compression techniques. This space-time tradeoff would have been impractical a few years ago, but can fit on an single external USB disk today, and should even fit in memory (RAM) within the next few years.

A similar space-time tradeoff is also applicable to the problem of bioinformatics sequence searching. The traditional $O(n)$ sequential scan approaches of BLAST and FASTA can sometimes be replaced with a more efficient sublinear search based upon a data structure known as a suffix array. This data structure effectively encodes all of the subsequences in a database efficiently, in much the same way as the chemical subgraphs above. Although suffix arrays have been known for some time, recent advances in storage technology now make them practical for protein sequence applications, though such indexing of typical nucleic acid and genomic sequences (probably) remains impractical.

Exponential database growth will always be a technical challenge but evolutionary strategies offer to hold off the inevitable in the short term and revolutionary strategies promise a longer-term solution.

D-3: Advancing Automated Synthesis Via Reaction Data Mining and Reuse

C. A. Nicolaou, T. Masquelin, J. Wang

Lilly Research Laboratories, Eli Lilly & Company, Indianapolis, IN 46285, USA

Ongoing attempts to identify optimal synthetic routes for compounds of interest, prepare virtual libraries of synthesizable compounds[1] and automate drug design[2] require better tools for synthesis predictability as well as robust synthetic route planning and optimization approaches. Algorithmic advances combined with the availability of large collections of reaction data has enabled the development of several computational tools for chemical synthesis support ranging from simple organic reaction lookup to rule-based reaction planning and retrosynthetic analysis[3]. Of particular interest are retrosynthetic analysis (RA) tools which design synthetic routes by recursively identifying synthesizable chemical bonds in a target structure, removing a bond, converting the resulting fragments to the necessary reactants and checking for reactant availability. Typical RA methodologies may provide multiple theoretical synthetic routes for a target structure and, often, require human expert knowledge to define reaction mechanisms and synthesizable bonds to break. Expert chemists are also the recipients of such system results and are tasked with the assessment of the proposed routes and the selection of the one(s) with the highest feasibility potential. Recent advances in automated synthesis systems[4] presents an opportunity to fully automate compound design-to-synthesis by submitting select routes for robotic execution. In order to achieve this goal the most appropriate route for each target needs to be identified and custom reaction execution workflows need to be implemented.

In this presentation we describe our efforts to (i) mine corporate reaction data, stored in electronic laboratory notebooks (eLN) and automated synthetic systems databases, and compile a corporate synthetic knowledge repository; (ii) develop a data-driven RA engine aiming to provide feasible synthetic routes for input chemical structures; (iii) assess the proposed synthetic routes using a neural network predictive model to select samples for automated synthesis execution. We thoroughly discuss our reaction mining process, the implementation and design of our RA engine and the deep learning approach used for synthetic route feasibility assessment. We present results from the training of the RA engine using a patent reaction dataset and the application to a collection of approved drugs. The RA tool, originally developed to serve in-house needs, is provided to the cheminformatics community in an effort to facilitate research in synthetic route design and reaction informatics in general. A discussion on lessons learned, issues to be resolved, and future development directions including ongoing work to instantiate system-specific synthetic workflows for automated synthesis execution concludes the presentation.

1. Nicolaou, C. A.; Watson, I. A.; Hu, H.; Wang, J. The Proximal Lilly Collection: Mapping, exploring and exploiting feasible chemical space. *J. Chem. Inf. Model.* **2016**, 56 (7), 12531266.
2. Schneider, G. Automating drug discovery. *Nat. Rev. Drug Discov.* **2017**.
3. Ravitz, O. Data-driven computer aided synthesis design. *Drug Discov. Today: Technol.* **2013**. 10(3), 443-449.
4. Godfrey, A. Masquelin, T. Hemmerle, H. A remote-controlled adaptive medchem lab: an innovative approach to enable drug discovery in the 21st Century. *Drug Discov. Today* **2013**, 18, 795−802.

D-4: Revealing important Molecular Fragments in Drug Discovery Using Time Trend Analyses

B. Zdrazil [1], N. Brown [2], R. Guha [3]

[1] Department of Pharmaceutical Chemistry, University of Vienna, Vienna, Austria, [2] BenevolentAI, London, UK, [3] National Center for Advancing Translational Sciences (NCATS), National Institutes of Health (NIH), Rockville, Maryland, US

Recently, we analysed data from ChEMBL[1] to examine the evolution of scaffold-derived properties, such as the number of enumerated compounds, biological activity, and liabilities, over 17 years. Our analysis highlights that certain properties such as the number of enumerated compounds, but not liabilities, show statistically significant increasing trends for some scaffolds. We also attempted to explain why a scaffold receives more attention over time and highlighted that obvious aspects such as synthetic feasibility do not explicitly drive attention.[2] A parallel investigation on the origins of three-dimensionality in drug-like molecules, revealed a tendency towards molecular planarity.[3]

Next, we were interested how different parameters for drug-likeness (such as QED, Lipinski rules), molecular complexity (such as Fsp^3, Principal moment of inertia, Plane of best fit), and solubility have evolved over time. After fragmentation of the data compounds were grouped by fragment. Next, we examined the trends of these drug-discovery relevant properties per fragment. We were interested how well the investigated parameters correlate on average and per fragment and which parameters have experienced greatest changes over time. Also, differences in the time trends for the major target classes were investigated.

In summary, trend analyses inform on distinct tendencies in drug-discovery related properties. This analysis can suggest directions that the drug discovery community is heading, in terms of relevant fragment and property space. In addition, such analyses support the prioritization of fragments in small molecule development projects.

1. Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2017**, 45 (Database issue), D945–D954.
2. Zdrazil, B.; Guha, R. The Rise and Fall of a Scaffold: A Trend Analysis of Scaffolds in the Medicinal Chemistry Literature. *J. Med. Chem.* **2017** Dec 27.
3. Meyers, J.; Carter, M.; Mok, N. Y.; Brown, N. On the Origins of Three-Dimensionality in Drug-like Molecules. *Future Med. Chem.* **2016**, *8* (14), 1753–1767.

Session E:

DEALING WITH BIOLOGICAL COMPLEXITY

E-1: Strategies for assembling an annotated library for phenotypic screening

H. Willems [1], S. Andrews[1], S. Ashenden[2], A. Bender[2], A. Gandhi-Kohli [2], A. Merritt[3], C. Mpamhanga[3], J. Skidmore[1], P. Sterk [1]

[1] Alzheimer's Research UK Drug Discovery Institute and [2] Centre for Molecular Informatics, University of Cambridge, Cambridge, UK, [3]LifeArc, Stevenage, UK

Phenotypic drug discovery involves screening with a functional cellular assay or a cell-based disease model, where the specific molecular target is not known. This approach can be more representative of the human disease than a target-based assay, but when screening small molecules, subsequent optimization steps can be more difficult because there is no protein structure or pharmacophore hypothesis to guide design. One approach to identifying which target mediates the response observed for a phenotypic hit is to screen an annotated, or chemogenomic, library. This is a collection of small molecules that are known to interact relatively selectively with their primary target. Ideally, compounds in an annotated library would be potent, selective, cell permeable and soluble. Multiple, structurally diverse ligands with affinity for the same target can be included to help target deconvolution.

This report describes our efforts to create an annotated library for phenotypic screening by mining the ChEMBL23 database and the issues we encountered along the way. A key decision point in the process was the definition of selectivity: how many targets does a compound need to be tested at before selectivity is a meaningful concept? What is an acceptable number of targets to hit if a compound has been in lots of assays? Is selectivity over subtypes important for a phenotypic library? Another point of discussion was how to achieve a wide target coverage. GPCR and kinase targets have many ligands, and could be overrepresented. Other target types only have few known ligands, and target diversity may need to be traded off against drug-likeness of the annotated ligands. Also, several tool compounds from our internal collections were not in ChEMBL. This raises questions on what is missed by focussing on ChEMBL as a datasource, and what other datasources could be explored.

On the technical side, we used SQL to extract all potent (<300 nM in a binding assay) small molecules (MW <1000) from ChEMBL . A window score and ranking score were then calculated in R using an algorithm published by Bosc, Meyer and Bonnet[1] to assess selectivity. The scored output was then processed in KNIME to extract 3 sets of 'selective' compounds: those with 100-fold selectivity over the second best target; those that were tested at more than 20 targets and were less than 100-fold selective at no more than 10% of these; and those that had 10-fold selectivity over 1 target and > 100-fold over at least 1 other. The 30,000 selective compounds and all potent SGC and chemical probe tools compounds[2] were checked for commercial availability and pushed through a further KNIME workflow to select the most potent and/or selective structures for each target. The library's target coverage was assessed using iTol[3]. Coverage for GPCR and kinase targets is shown in Figure 1. A total of 448 compounds were selected for purchase, covering 297 targets.

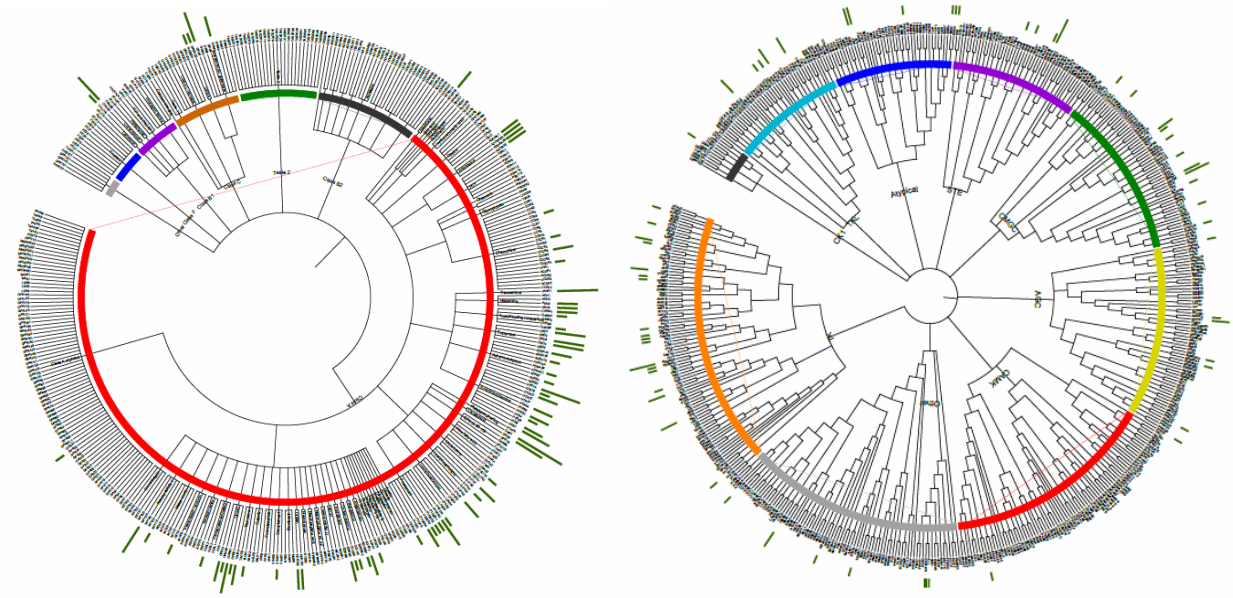


Figure 1. GPCR family (left) and kinase family (right) target coverage for the annotated library. The bars on the outside of the circle indicate the number of ligands that have been included in the library for that target

1. Bosc, N., Meyer, C., Bonnet, P. The use of novel selectivity metrics in kinase research. *BMC Bioinformatics* **2017**, 18, 17-29. https://doi.org/10.1186/s12859-016-1413-y
2. http://www.thesgc.org/chemical-probes; http://www.chemicalprobes.org/
3. Letunic, I. and Bork, P. Interactive Tree Of Life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res* **2016** doi: 10.1093/nar/gkw290

E-2: Targeting of the disease related proteome by small molecules

Modest von Korff, Thomas Sander

Idorsia Pharmaceuticals Ltd, Allschwil, Switzerland,

How many disease related proteins encoded by the human genome have been already targeted by small molecules? This is an open question of high interest in the pharmaceutical industry. Nowadays, databases like ChEMBL and SwissProt can be used to answer the question. However, an in-depth analysis of the targets available in the ChEMBL database showed that many proteins used in biological test assays were derived from species other than human. This triggered a second question. What is the similarity of the human proteins to their non-human counterparts that were used for biological testing? If the homology between the human protein and its non-human analog is high enough, it can be assumed that the ligand space of both proteins overlaps.

An exhaustive analysis of the actual ChEMBL 23 library was done to answer these two questions in the dimensions of chemical space and proteome space. All molecules in the ChEMBL 23 library were mapped to their target proteins. Scaffolds of all mapped compounds were analyzed. From the mapped compounds, the scaffolds were analyzed. For all target proteins, the most similar ones from the human genome were searched with the BLAST engine from SwissProt. It is a rule of thumb in molecular modeling that up to a BLAST similarity of 0.4 two proteins are similar enough to allow homology modeling. The disease related proteome was compiled by the Gene2Disease tool.[1] This in-house tool relates all approved genes by the Human Genome Organisation Gene Nomenclature Committee with all diseases defined by the MeSH tree from the NIH. Gene2Disease analyzed approximately 26 million records available in the PubMed database to detect protein-gene-disease relations.

Results of the analysis were captured in a large and sparse ChemProteome matrix, with rows containing molecular structures and columns containing disease related proteins connected to the respective diseases. If the biological activity of a small molecule was tested on a target protein, the corresponding field in the matrix was filled with the biological test value. Analysis of this matrix showed that no bioactivity value was found in the ChEMBL database for > 80% of disease related proteins. Including the chemical space which covered homologous proteins did not significantly increase the percentage of covered proteome.

In conclusion, there is a huge unexplored chemical space potentially targeting disease related proteome. Many protein targets are waiting to be exploited, while our ChemProteome matrix shows which potential non-exploited targets are linked to diseases.

1. Korff, M. v.; Fink, T.; Sander, T., A new relevance estimator for the compilation and visualization of disease patterns and potential drug targets. In *Pacific Symposium for Biocomputing*, Hawaii, 2017.

E-3: Gearing Transcriptomics Towards High-throughput Screening: Compound Shortlisting From Gene Expression using in silico information

Natalia Aniceto [1,2], Andreas Bender [1], Florian Nigsch [2]

[1] Centre for Molecular Informatics, University of Cambridge, Cambridge, United Kingdom, [2] Chemical Biology and Therapeutics Informatics, Novartis, Basel, Switzerland.

The ability to efficiently screen large libraries compounds in order to find candidates that will exhibit a specific gene expression profile of interest is potentially very useful, and doing so without requiring experimental data is very appealing. Here we propose a new approach that creates shortlists of compounds for a query gene expression signature. This approach uses a previously proposed variant of the k-NN procedure (called variable-NN) where a test signature results from the weighted average from the signatures of surrounding neighbours.[1] However we propose an added modification to this method, where predicted signatures are submitted to a quality (or confidence) criterion based on standard deviation of predictions.

Two transcriptomics datasets were used in this work, LINCS[2] and a Novartis internal transcriptomics panel (PANOMICS), and two different similarity measures were tried, calculated from structural and biological (predicted) signatures, in order to explore the feasibility of a recently available in silico biological signature.[3]

The signatures for the full dataset, obtained from the variable-NN procedure, were ranked by similarity to target (desired) signatures. This process showed to consistently enrich the top of the ranked compound lists with the compounds that match each of the target signatures. Obtained median ranking percentiles of true compound-signature matches, which were originally consistent with random sorting when confidence is not taken into account, were as high as 90% when confidence was used to correct similarity. This translates into obtaining shortlists formed of the top 10% of ranked compounds that are likely to contain the best compound candidate. This process allows filtering target signature queries according to reliability in such a way that correlates with compound enrichment, and different levels of trade-off between coverage and compound enrichment can be selected by the user.

The use of confidence criteria to filter signatures was pivotal in the ability to enrich the top of the ranked lists with the compounds correctly matching the target signatures, where, for example, we were able to locate 20% of the LINCS signatures for which a shortlist formed by the top 10% of the full compound list would likely contain the compound that more closely yields each of those query signatures. This approach allows locating candidates for a target gene signature purely from in silico information, thus enabling high-throughput virtual screening of compounds libraries to find compounds associated with a gene expression profile of interest.

1. Liu, R.; AbdulHameed, M. D. M.; Wallqvist, A. Molecular Structure-Based Large-Scale Prediction of Chemical-Induced Gene Expression Changes. *J. Chem. Inf. Model*. **2017**, 57, 2194–2202.
2. Subramanian, A.; *et al*. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell*. **2017**, 171, 1437–1452.e17.
3. Martin, E. J.; Polyakov, V. R.; Tian, L.; Perez, R. C. Profile-QSAR 2.0: Kinase Virtual Screening Accuracy Comparable to Four-Concentration IC50s for Realistically Novel Compounds. *J. Chem. Inf. Model*. **2017**, 57, 2077–2088.

E-4: Discrimination of G–protein coupled receptors and their conformational states using intramolecular interaction

F. Koensgen [1], F. Da Silva, E. Kellenberger

[1] Laboratoire d'Innovation Thérapeutique, UMR 7200 CNRS- University of STRASBOURG, [2] Medalis Drug Discovery Center, Illkirch, France

G–protein coupled receptors (GPCR) are membrane receptors able to transmit stimuli to cells. The molecular mechanism of signal transmission involves the receptor coupling to effector in response to ligand binding, and this depends on the receptor conformational state. To date, the 3D-structures of 50 different GPCRs have been characterized by X-ray crystallography and the dynamics of some of them have been extensively studied by molecular dynamics simulation, suggesting general mechanism of activation/inactivation.[1]

Here we propose a new method to compare different GPCR structures, independently of predefined structural or functional determinants. This method is based on the detection and comparison of intramolecular non-covalent interactions in the seven transmembrane domains (TM).

In more details, the analysis of a 3D–structure involves the extraction of TM coordinates followed by the representation of hydrogen bonds (labeled with *inter–helix* or *intra–helix*, and with *with sidechain* or *within backbone*), ionic bonds and aromatic bonds as either a graph or a fingerprint built from Ballesteros-Weinstein numbering.[2] Comparing two 3D–structures does not require that they are described in a common frame. Two graphs are aligned for the bestfit superimposition of the maximum common substructure. Similarity between two fingerprints is calculated using the Tanimoto coefficient.

We have applied the method to the classification of 215 GPCR structures available in the Protein DataBank. Networks built from the comparison of graphs showed that *with sidechain inter/intra–helix* hydrogen bonds are sufficient to differentiate GPCRs. All polar interactions except *within backbone intra–helix* hydrogen bonds well differentiate the activation states of a GPCR. Global analysis of interactions suggested specific signatures of GPCRs and their activation state.

We have also applied the method to the analysis of two molecular dynamics trajectories where a GPCR experiences a transition from the active to the inactive state[3,4] The all-against all comparison of frames delimited a few clusters. The characterization of clusters by consensus interaction fingerprints revealed which interactions are state–specific.

In conclusion, we developed a new method able to discriminate GPCR from a simplified 3D-representation (8-46 interaction points). The same approach also distinguishes conformational states, and has proved to successfully cluster and describe the conformational states generated by molecular dynamics simulation.

1. Manglik, A.; Kobilka, B. The Role of Protein Dynamics in GPCR Function: Insights from the B2AR and Rhodopsin. *Curr. Opin. Cell Biol.* **2014**, *27*, 136–143.
2. Desaphy, J.; Raimbaud, E.; Ducrot, P.; Rognan, D. Encoding Protein–Ligand Interaction Patterns in Fingerprints and Graphs. *J. Chem. Inf. Model.* **2013**, *53* (3), 623–637.
3. Miao, Y.; McCammon, J. A. Graded Activation and Free Energy Landscapes of a Muscarinic G-Protein–coupled Receptor. *Proc. Natl. Acad. Sci.* **2016**, *113* (43), 12162–12167.
4. Dror, R. O.; Arlow, D. H.; Maragakis, P.; Mildorf, T. J.; Pan, A. C.; Xu, H.; Borhani, D. W.; Shaw, D. E. Activation Mechanism of the B2-Adrenergic Receptor. *Proc. Natl. Acad. Sci.* **2011**, *108* (46), 18684–18689.

Session F:

CHEMINFORMATICS

F-1: Comparison and Analysis of Molecular Patterns on the Example of SMARTS

Robert Schmidt [1], Emanuel S. R. Ehmki [1], Matthias Rarey [1]

[1] *Universität Hamburg, ZBH – Center for Bioinformatics, Bundesstraße 43 20146 Hamburg, Germany*

Chemical patterns are widely used to filter structural properties in molecular design endeavors. These properties are defined in filter sets like PAINS[1] or company-specific filter lists[2]. Although frequently applied, filter sets and their use are currently under discussion[3]. For the analysis of filter sets, an algorithmic approach to compare chemical pattern with each other would be highly desirable, however, has not been published so far. Here we present a novel algorithm, its implementation and application for the calculation of pattern equality, inclusion, and similarity. SMARTScompare, the accompanying tool, allows for filter set analysis, pattern hierarchy verification and chemical pattern feasibility tests.

A chemical pattern P can be understood as a description of the infinite set of all molecules matched by it. Two patterns can be considered equal or isomorphic when their molecule sets are the same. Based on the subset relationship between the sets of matched molecules, patterns can be characterized as more or less specific. Similarity of two patterns is expressed via a probability model for atom matching. Since these relations are defined by the set of molecules matched, they are independent of the language the patterns are formulated in.

The comparison of molecular patterns is based on the following strategy: Within a chemical model, the spaces of all feasible atom or bond states are enumerated. Based on these enumerations, each node and edge of a chemical pattern can be described by a fingerprint representing all compatible atom or bond states. Equality, subset relations and similarity are easy to calculate on fingerprints. After the assignment of fingerprints, a maximum common subgraph (MCS) algorithm is applied resulting in the *maximum common subpattern (MCSP).* Based on the MCSP, subset relations and similarity scores are computed. The approach naturally allows for asymmetric similarity scores, providing an estimation of the coverage of one pattern in another.

The algorithm is implemented in a new software named SMARTScompare designed for similarity assessments and subset classifications of SMARTS[4] patterns. It supports most of its features, including recursion and the negation of properties. Although possible in theory, isotopes, radicals and chirality are not supported by the algorithm so far. Besides pattern comparison, SMARTScompare detects chemically infeasible structures in patterns, meaning all kinds of incorrect valence states or impossible property combinations. Beyond that, redundant formulations of SMARTS recursion are detected.

One main application of SMARTScompare is the analysis of structural filters for molecular screening applications. We analized similarity and pattern inclusion in between eight structural filter sets[5]. For example, quinone patterns showed high similarity scores and are manually labelled what makes them excellent test cases. Additionally, their heterosubstituted derivatives (e.g. quinonimines) cause complex patterns making them incredibly hard to read for humans. SMARTScompare found 14 patterns in seven filter sets including dyes with quinone-like substructures.

Besides general pattern comparisons, SMARTScompare can verify hierarchically structured pattern collections. Knowledge-based conformation generation relies on statistical analysis of available data, like for example the Torsion Library[6]. SMARTScompare can verify an assigned class membership and even reorder patterns within such classes. We were able to identify a few examples of formerly missed subset relations. The software helps to avoid common mistakes when writing SMARTS and removes redundancy in recursive SMARTS patterns as well as pattern collections. Furthermore SMARTScompare allows for similarity search and in-depth discussion of patterns for similar structural properties.

1. Bell, J. B.; Holloway, G. A. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays *Journal of Medicinal Chemistry* **2010**, 53, 2719-2740.
2. Bruns, R.; Watson, I., Rules for Identifying Potentially Reactive or Promiscuous Compounds *Journal of Medicinal Chemistry* **2012**, 66, 9763-9772
3. Capuzzi, S.; Muratov, E. N.; Tropsha, Alexander, Phantom PAINS: Problems with the Utility of Alerts for Pan- Assay interference CompoundS *Journal of Chemical Information and Modeling* **2017**, 57, 417-427.
4. Daylight Theory Manual. http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (accessed Feb. 6, 2018)
5. Gaulton, A.; Hersey, A.; Nowotka, M; et al. The ChEMBL database in 2017, *Nucleic Acids Reseach* **2017**, 45, D945-D054
6. Schärfer, C.; Schulz-Gasch, T. Ehrlich, H.C.; Guba W.; Rarey M.; Stahl M., Torsion angle preferences in druglike chemical space: A comprehensive guide, *Journal of Medicinal Chemistry* **2013**, 56, 2016-2028

7. Schomburg, K; Ehrlich, H. C.; Stierand, K; Rarey, M., From Structure Diagrams to Visual Chemical Patterns, *Journal of Chemical Information and Modeling* **2010**, 50, 1529-1535

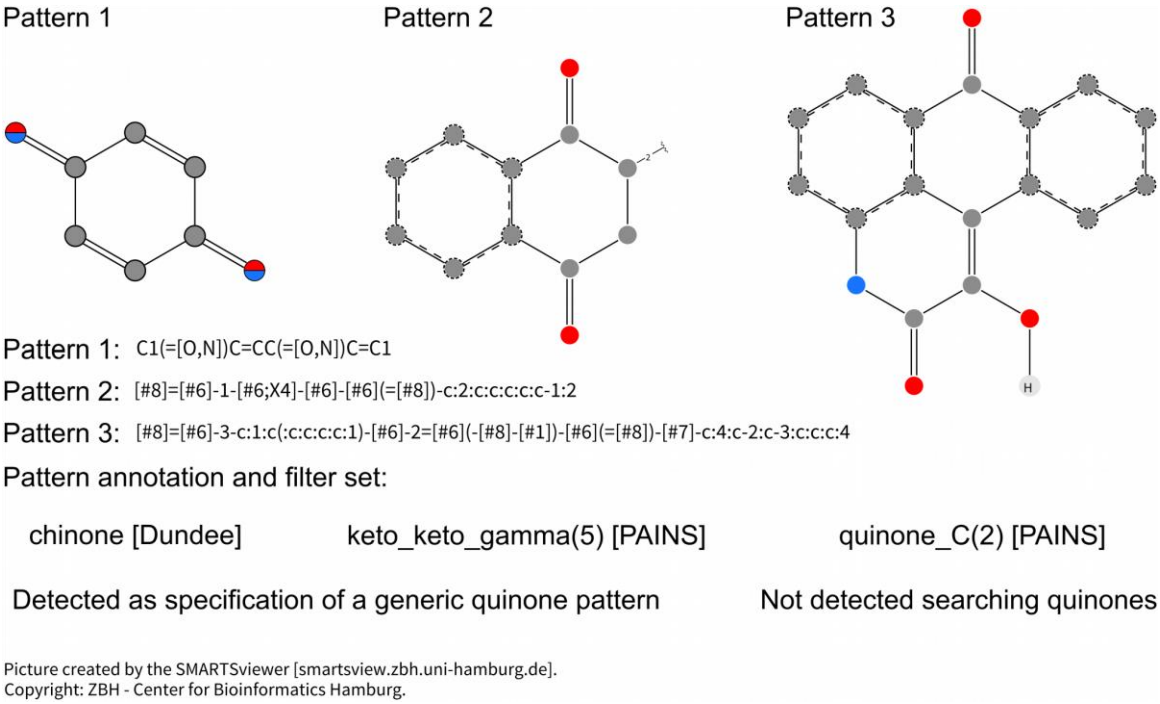


Figure 1: Results of the search for Quinones in the filter sets referenced by ChEMBL[5]. The pattern on the left is annotated as Quinone and was detected by SMARTScompare. The second pattern is not annotated as Quinone, but contains a Quinone substructure. The pattern on the right was not found by SMARTScompare. We consider its annotation as false or incomplete since it contains a quinomethane substructure. Pattern depictions are generated with SMARTSViewer[7].

F-2: Anisotropic Atom Reactivity Descriptors for the Prediction of Liver Metabolism, Ames Toxicity and Hydrogen Bonding

Andreas H. Göller [1], Arndt Finkelmann [2], Lara Kuhnke [3], Christoph Bauer [2]

[1] Bayer AG, Computational Chemistry, Wuppertal,Germany, [2] ETH Zürich, Dept. of Chemistry and Applied Biosciences, Switzerland, Zurich, Switzerland, [3] Bayer AG, Computational Chemistry, Berlin,Germany

Contrary to many other ADMET properties of small molecules which are well-described by molecular descriptors, the identification of atoms susceptible to metabolic reactions or the activation of primary aromatic amines (pAA) via N-hydroxylation requires atomic reactivity descriptors and atomic resolution machine learning.

We here report on the application of our recently developed sets of atomic descriptors that encode the anisotropic electron density distribution using conformation-independent quantum-mechanical atomic charge schemes (see Figure 1).[1]

First example is site-of-metabolism prediction. We have extended our cytochrome P450 model[2] to phase II metabolism by incorporation of about 25,000 carefully cleaned metabolic transformations from the Accelerys Metabolite database, resulting in cross-validated atom-position Matthews correlation coefficients of 0.61 and 0.76 for phase I and phase II, respectively, and MCC of 0.41 for a validation set of recent compounds collected from 2015 and 2016, not being part of the training set.

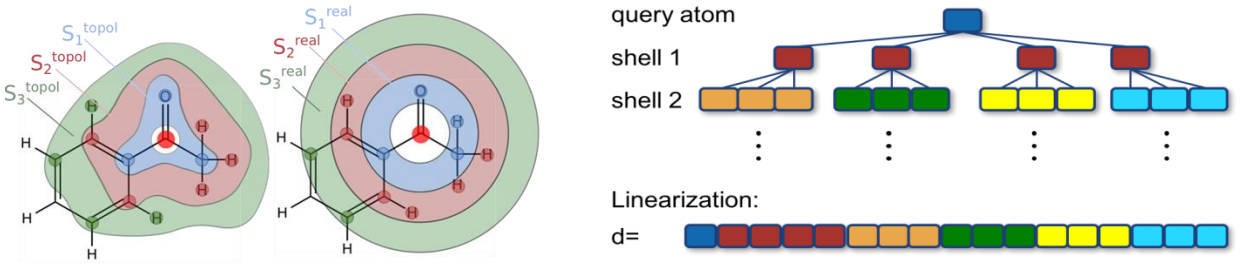


Figure 1: Anisotropic circular descriptors are created be either topological or 3D real-space binning of atomic properties like e.g. atomic charges (left) and mapping to a linear vector (right).

Second example is the prediction of the activation of PAA in the Ames assay with S9 mix. There, the first reactive step is N-hydroxylation, leading finally to the formation of a reactive nitrenium ion that can bind to the negatively charged DNA. Here, we successfully combined an ECFC-6 fingerprint counts, HOMO-LUMO gap and an atom reactivity model.

Third example is the prediction of maximal hydrogen bond acceptor strengths from chemical structure alone, again taking advantage of the anisotropy and conformer-independence of our novel descriptors, with the aim to extend the concept to the prediction of hydrogen bond donors.

1. Finkelmann, A.R.; Göller, A.H.; Schneider, G. Robust molecular representations for modelling and design derived from atomic partial charges. *Chem. Commun.* **2016**, *52*, 681-684.
2. Finkelmann, A.R.; Göller, A.H.; Schneider, G. Site of metabolism prediction based on ab initio derived atom representations. *ChemMedChem*, **2017**, 12, 606-612.

F-3: Exploring 3D molecular shape using spectral geometry

M Seddon [1], D Cosgrove[2], M Packer [3], V Gillet [1]

[1] University of Sheffield, Sheffield, UK, [2] CozChemIx Limited, Macclesfield, UK, [3] AstraZeneca, Cambridge, UK

Three-dimensional molecular shape is a key determinant of molecular interactions[1]. To date, widespread use of 3D similarity methods in drug development has been hampered by computational complexity surrounding structure alignment and molecular flexibility. Structure alignment is either carried out at runtime, thus incurring a computation cost, or avoided through the use of 3D shape descriptors, which results in a loss of information. Typically, 3D shape comparison treats the molecules as rigid bodies and flexibility is taken into account using conformation ensembles to sample conformational space, which significantly increases the computational cost of 3D similarity searching on large molecular databases.

Spectral geometry provides a framework for exploring concepts of flexible 3D shape[2]. In brief, spectral geometry treats the surface of a 3D shape as a curved 2D surface and encodes the geometric properties in the spectrum of the Laplace-Beltrami Operator over that surface. The field of spectral geometry originates in 3D computer vision where typical use cases are to identify the same figure in different poses. These methods are of particular interest for high throughput virtual screening because they produce rich descriptors of 3D shape that are alignment-invariant and also invariant to a specific class of flexibility, called *isometric deformation*. Furthermore, the conceptual framework has a large amount of promise for investigating the relationship between 3D molecular shape and conformational variation in a mathematically robust manner.

We have used the spectral geometry framework to explore the relationship between conformational variation and molecular shape. In particular, the extent to which the variation in conformational shape can be captured by the isometric deformation assumption was explored. Furthermore, the chemistry of these variations was investigated to identify the cases in which these methods are optimal. Our results suggest a method of determining when two conformations of a molecule are sufficiently different to be considered different shapes in spectral geometry. Finally, we implemented an alignment invariant shape descriptor for the purpose of high throughput virtual screening and compared its performance to open source implementations of a standard alignment based shape comparison method[3] and a shape descriptor[4]. We show that the spectral geometry descriptors outperform these methods using the Directory of Useful Decoys Enhanced[5] (DUD-E) data set.

1. Nicholls, A.; McGaughey, G. B.; Sheridan, R. P.; Good, A. C.; Warren, G.; Mathieu, M.; Muchmore, S. W.; Brown, S. P.; Grant, J. A.; Haigh, J. A.; Nevins, N.; Jain, A. N.; Kelley, B. *J. Med. Chem.* **2010**, *53* (10), 3862–3886.
2. Biasotti, S.; Cerri, A.; Bronstein, A.; Bronstein, M. *Comput. Graph. Forum* **2015**, n/a-n/a.
3. Grant, J. A.; Gallardo, M. A.; Pickup, B. T. *J. Comput. Chem.* **1996**, *17* (14), 1653–1666.
4. Ballester, P. J. *Future Med. Chem.* **2011**, *3* (1), 65–78.
5. Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. *J. Med. Chem.* **2012**, *55* (14), 6582–6594.

F-4: Creating atom-to-atom mapping in chemical reaction using machine learning methods

T. Madzhidov [1], A. Khayrullina [1], R. Nugmanov [1], I. Baskin [2], A. Varnek [3]

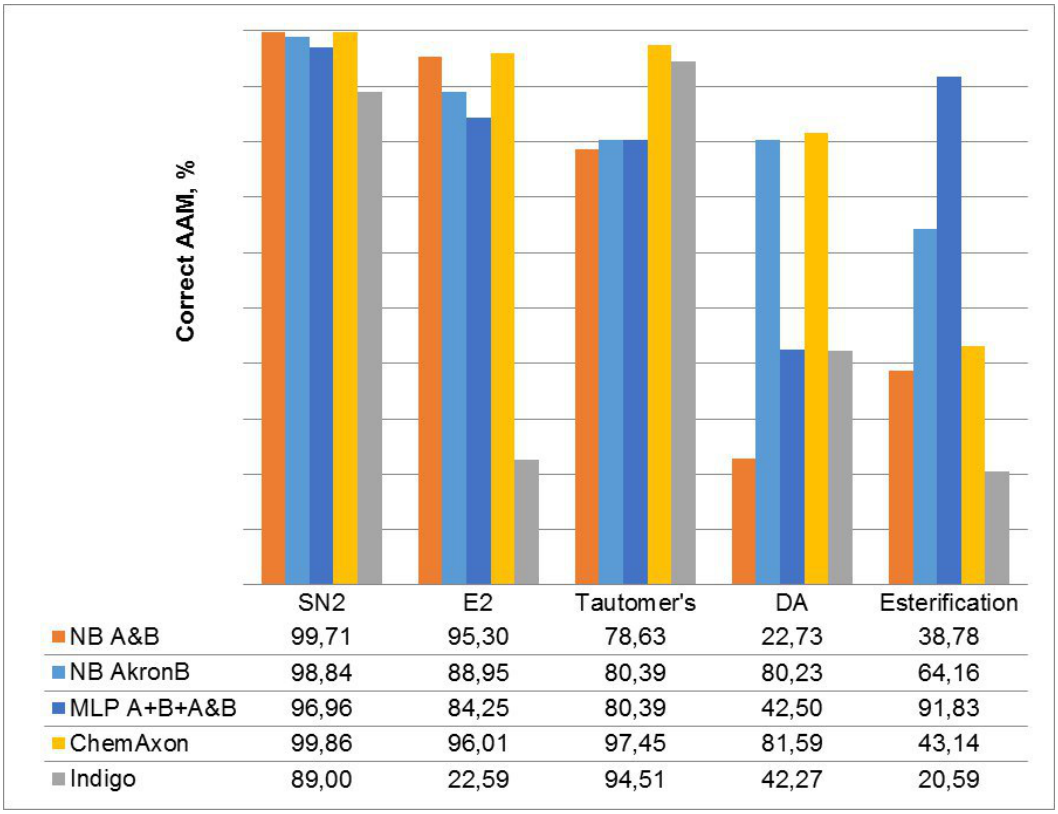
[1] A.M. Butlerov Insitute of Chemistry, Kazan Federal University, Kazan, Russia, [2] Faculty of Physics, M.V.Lomonosov Moscow State University, Moscow, Russia, [3] Laboratory of Chemoinformatics, University of Strasbourg, Strasbourg, France

The fundamental first step in the computer analysis of chemical reactions is determination of the correspondence between atoms of substrats and products, called the atom-atom mapping (AAM). AAM is used to find the changing part of substrate and product molecules, i.e. the reaction center[1]. Knowing reaction center it is possible to run advanced reaction search, like substructure and similarity search, establish reaction type, etc. The most well-known and consummate algorithms are implemented in EPAM Indigo, Accelrys Automapper, JChem Standardizer and ICMap programs. All of them are based on maximum common substructure (MCS) detection; however other approaches exist as well[1].

Known programs are based on complex heuristics that guarantee correctness of AAM on most cases. Our idea was to create an algorithm that could learn how to create AAM based on known reactions with correct AAM. In this work we propose a novel approach to find optimal AAM that is based on application of machine learning techniques. The task is formulated as classification: for every pair of reagent-product atoms one need to establish whether their mapping is correct. To train the classifier for each reaction, pairs of atoms were generated that correspond to the correct and incorrect AAM. A simple probabilistic "Naive" Bayesian (NB) and shallow neural network (Multi-layer Perceptron) classifiers were used. The attribute vector for every reagent-product atom pair contains information on environment of both atoms, represented by fragment descriptors of different topology: sequences, augmented atoms and their combinations. For a given atom pair from the test set the probability that this pair corresponds to correct AAM is returned. Using Munkres algorithm mapping of atoms from product to reagent that correspond to maximum likelihood was identified. Special approaches were added to correctly handle molecular symmetry.

The proposed approach was implemented and tested on 5 reaction types: substitution (SN2), elimination (E2), rearrangement (tautomeric transformation), cycloaddition (Diels-Alder) and esterification reactions. Cross-validation was used for validation of our approach and the ratio of correctly assigned AAM was used as a quality metric. Our approach was compared with other programs for AAM identification. Despite in this work we used the simplest machine learning methods it already showed quality at the level of commercial tools in the creation of AAM. The quality of produced AAM is almost the same as for ChemAxon Automapper on most datasets and in 4 of 5 reactions significantly outperforms Indigo Automapper. Failure in tautomeric reaction AAM creation by our approach is caused by the fact that dataset was small and diverse. For esterification reaction our approach outperforms both commercial tools.

Thus, we proposed and implemented the first atom-to-atom mapping determination tool that learns how to create AAM on the basis of known reactions.



	SN2	E2	Tautomer's	DA	Esterification
NB A&B	99,71	95,30	78,63	22,73	38,78
NB AkronB	98,84	88,95	80,39	80,23	64,16
MLP A+B+A&B	96,96	84,25	80,39	42,50	91,83
ChemAxon	99,86	96,01	97,45	81,59	43,14
Indigo	89,00	22,59	94,51	42,27	20,59

Fig.1 Percentage of correct AAM produced by our approach using Naïve Bayes classifier (NB A&B and NB AkronB, that differ in a way of fingerprint generation), shallow neural network (MLP A+B+A&B) vs commercial MCS-based tools ChemAxon JChem Automapper (ChemAxon) and EPAM Indigo Automapper.

1. Chen, W. L.; Chen, D. Z.; Taylor, K. T. Automatic Reaction Mapping and Reaction Center Detection. *Wiley Interdisciplinary Reviews: Computational Molecular Science.* **2013**, pp 560–593.

Poster Session Abstracts RED

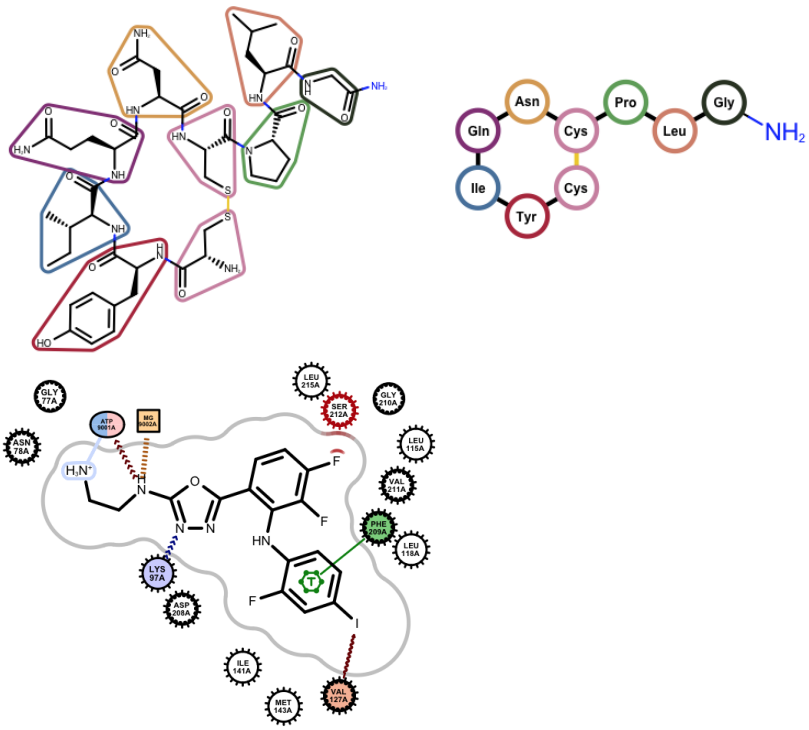
P-01: Accelerating problem solving and decision making in medicinal chemistry through visualisation

Paul C. D. Hawkins [1] & Krisztina Boda [1]

[1] *OpenEye Scientific, Santa Fe, USA.*

Modern ligand discovery and optimization projects, and chemists involved in those projects, rely heavily on complex three-dimensional data for success. Whether this data is obtained from experiment (structural data from crystallography), or computation (active site pose and interaction predictions, molecular simulations, quantum mechanics) it is valuable and frequently expensive to obtain. Efficient conversion of this 3D data into comprehensible information and then into actionable knowledge to drive the project is a problem that is exacerbated by a language barrier; the natural language of project chemists is 2D, while the native form of the highest value data that they use is 3D.

Here we present an approach to the effective and efficient visualization of 3D data in 2D in order to accelerate the process of decision making in ligand design. Examples of this approach will be provided for data often analysed in 2D (crystallography, pose prediction) and from methodologies that, while commonly used in medicinal chemistry, are not generally interpreted in 2D, including molecular dynamics and quantum mechanics.



P-03: Nanomaterial safety data integration with substance data model and federated search

Nina Jeliazkova[1], Nikolay Kochev[1,2] , Vesselina Paskaleva[2], Gergana Tancheva[2], Penny Nymark[3,4] , Margarita D. Apostolova[5], Andrea Haase [6]

[1]Ideaconsult Ltd, 4 A. Kanchev str., Sofia 1000, Bulgaria, [2]Department of Analytical Chemistry and Computer Chemistry, University of Plovdiv, 24 Tzar Assen Str., 4000 Plovdiv, Bulgaria, [3]Karolinska Institutet,Institute for Environmental Medicine, Nobelsväg 13 Stockholm, Sweden, [4]Misvik Biology, Toxicology Division, Karjakatu 35b, Turku, Finland, [5]Medical and Biological Research Lab., Institute of Molecular Biology – Bulgarian Academy of Sciences, Acad. G. Bonchev Str. Bl. 21, Sofia 1113, Bulgaria, [6]German Federal Institute for Risk Assessment (BfR), Department of Chemical and Product Safety, Max-Dohrn-Strasse 8-10, 10589 Berlin, Germany

The basis of most public chemical databases is the direct link between the chemical structure and properties. This paradigm has been used for several decades, providing a platform for virtual screening and modelling of the properties of small molecules. However this approach is too restrictive for many challenging cases, including nanomaterials and industrial chemicals, which may have complex compositions. The REACH definition of a substance encompasses all forms of substances and materials on the market, including nanomaterials. The nanomaterial safety assessment has become an important task following the growth in production of engineered nanomaterials (ENMs) and the increased interest in ENMs from various academic, industry and regulatory parties. Nanomaterials data management is also challenged by the lack of agreed representation of nanomaterials, e,g, the graph theoretic representation of well-defined chemical structures and linear notations such as SMILES and InChI are unsuitable for representing nanomaterials. We present experience with integrating large sets of nanosafety data generated from past NanoSafety Cluster projects with the help of a substance data model, implemented in the eNanoMapper database [1]. This data model is also successfully used to handle chemical substances and safety data from ECHA dossiers [2].

Data generated by multiple nanosafety projects is compiled, annotated and imported into separate eNanoMapper database instances. These databases offer a user friendly web interface and REST API [1] and serve as building blocks to provide federated search across all or subsets of the database instances, enabled by Apache Solr backend. The eNanoMapper ontology[3] is used for harmonisation of the terminology and as a synonym list for query expansion. While multiple structured import formats are supported (IUCLID, RDF, JSON), the nanosafety data from past and ongoing projects use custom spreadsheet templates, currently encompassing over 1000 Excel files. Import of Excel files is enabled by a configurable parser that maps the spreadsheet data via external configuration files. Multiple export formats are supported, including tab delimited files, RDF and ISA-JSON. Free text and faceted search applications, with public and restricted access for different subsets of data, are available at https://search.data.enanomapper.net. The NanoReg2 integrated database (Figure 2) is online at https://search.data.enanomapper.net/nanoreg2 and allows project partners to access data from past EU FP7 funded projects (NANoREG http://www.nanoreg.eu/ , MARINA, NanoGenotox, Nanotest) through a common view and faceted search. The database is actively used by project partners, helping to identify and, where possible, resolve a range of data quality and completeness issues.

Acknowledgment: This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No. 646221.

1. Jeliazkova, N.; Chomenidis, C.; Doganis, P.; Fadeel, B.; Grafström, R.; Hardy, B.; Hastings, J.; Hegi, M.; Jeliazkov, V.; Kochev, N.; Kohonen, P.; Munteanu, C. R.; Sarimveis, H.; Smeets, B.; Sopasakis, P.; Tsiliki, G.; Vorgrimmler, D.; Willighagen, E. *Beilstein J. Nanotechnol.* **2015**, *6*, 1609–1634.
2. Jeliazkova, N.; Koch, V.; Li, Q.; Jensch, U.; Reigl, J. S.; Kreiling, R.; Georgiev, I.; Hubesch, B. *Toxicol. Lett.* **2016**, *258*, S114–S115.
3. Hastings, J.; Jeliazkova, N.; Owen, G.; Tsiliki, G.; Munteanu, C. R.; Steinbeck, C.; Willighagen, E. *J. Biomed. Semantics* **2015**, *6* (1), 10.

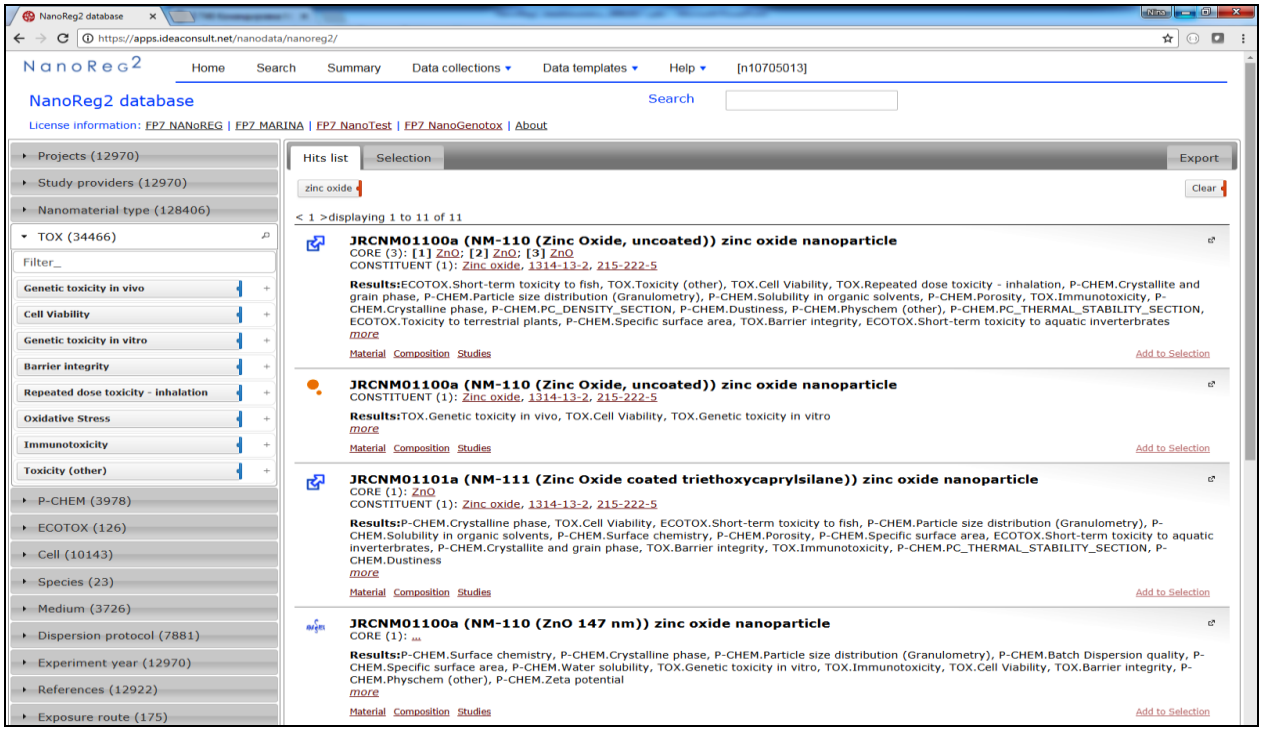


Figure 2. Screenshot of NanoReg2 database free faceted search application. The material, composition and studies links for each nanomaterial entry lead to the corresponding eNanoMapper database instance.

P-05: Can we agree on the structure represented by a SMILES string? A benchmark dataset

N O'Boyle [1], J Mayfield [1], R Sayle [1]

[1] *NextMove Software, Cambridge, UK*

Let us start with a question for the reader: How many hydrogens are on the nitrogen in the molecule described by the SMILES string "N(C)(C)(C)C"? When we asked this question recently on Twitter, only two out of 25 respondents correctly answered.

In February 1988, Dave Weininger published a description of the SMILES language in the Journal of Chemical Information and Computing Science,[1] and in the 30 years since then, SMILES has become one of the *de facto* standards for exchanging chemical information. It has a concise yet expressive form that remains reasonably human-readable, and it is convenient for a broad spectrum of use cases whether copying and pasting a single SMILES string into a webapp, or for storing millions of molecules in a database. As a result, SMILES readers and writers abound in every chemistry toolkit and application.

Here we investigate to what extent these SMILES readers agree on the structure represented by a SMILES string. Our goal is to highlight corner cases and foster a consensus on their handling, with the ultimate goal of improving information exchange between chemistry tools. For this reason, we focus exclusively on SMILES reading – until we can agree on the meaning of a particular SMILES string, there is little point in discussing SMILES writing.

Our benchmark dataset[2] consists of almost 50 thousand scaffolds derived from structures in ChEMBL, and converted to aromatic SMILES by a variety of tools. The benchmark test itself is simple: each tool to be tested must read each SMILES string and report the number of hydrogens on each atom.

We describe results for more than a dozen tools, from both the open source and commercial worlds, and highlight both the areas of contention and agreement. Furthermore, we show how testing against the benchmark has already led to improvements in several toolkits.

1. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31-36.
2. SMILES reading benchmark. https://github.com/nextmovesoftware/smilesreading (accessed Feb 15, 2018).

P-07: Computational Studies of Integrin Inhibitors

S. Alarfaji [1], T. McInally [1], S. Macdonald [2], J. Hirst [1]

[1] School of Chemistry, University of Nottingham, UK, [2] GlaxoSmithKline, Stevenage, UK

Compounds containing amides play a key role in the pharmaceutical industry and have been used widely in the treatment of diabetes and have been shown to limit tumour growth. Two- and three-dimensional quantitative structure-activity relationships (2D/3D-QSARs) were used to model the correlation between the physicochemical properties of some amides and their biological activity (pIC_{50}) to predict the activities of new molecules. An autocorrelation based method, topological maximum cross correlation[1] (TMACC), was employed to build our 2D-QSARs models. We have generated models across four sets of data, ranging in size from 25 to 47 molecules on a number of integrins subtypes. The results were cross-validated using a leave-one-out (LOO) approach. Using partial-least-squares regression PLS, a TMACC model with good predictive ability was generated based on training set of 25 compounds and showed satisfactory statistical results ($q^2 = 0.57, r^2 = 0.82$). Other TMACC models for 40 and 47 molecules were generated and showed similar but slightly lower predictivity: ($q^2 = 0.71, r^2 = 0.93$), and ($q^2 = 0.49, r^2 = 0.70$), respectively. 3D-QSARs models were also established for the same data sets using comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA). The CoMFA and CoMSIA models were not sensitive to changes in the orientation of the amide structure. The TMACC QSARs showed better predictive ability than the 3D-QSARs.

1. Melville, J.L.; Hirst, J.D., TMACC: Interpretable Correlation Descriptors for Quantitative Structure-Activity Relationships. *J. Chem. Inf. Mod*., 2007, **47**, 626–634.

P-09: Fast prediction of the specific conductivity of electrolytes from the molecular structure of the solvent

R. Bouteloup [1], D. Mathieu [1]

[1] CEA Le Ripault, Monts 37260, France

With the development of battery utilization in a lot of devices, the needs to improve their safety and performance are in expansion. For this purpose, new liquid electrolytes are investigated, which require new solvents and/or additives. In view of screening efficiently the chemical space for suitable candidate compounds, this work exposes a way to predict the specific conductivity of a nonaqueous electrolyte solution with a $LiPF_6$ salt.

The purpose of this method is to predict this property, quickly and simply, with only the 2D structure of each molecule of the solvent and their proportions as input parameters. To this aim, we have chosen the Casteel-Amis empirical equation [1] to represent the specific conductivity as a function of the salt concentration and the solvent composition.

The four parameters of this equation can be related to properties of the solvent, since in our case, the salt is always the same. The two properties that determine the specific conductivity are the ionic mobility and the ionic association. To connect them with solvent properties, we approximate that these two can be represented on the basis of the viscosity and the dielectric constant of the solvent, respectively. If we can predict them for each solvent, we can calculate the specific conductivity.

For the viscosity, we have developed an additive model, based only on the 2D structure of pure solvents. For the dielectric constant of pure compounds, we use the Fröhlich equation [2], that allows to calculate the dielectric constant from the molar volume, the refractive index and the orientational parameter $g\mu^2$ (with μ the dipole moment). To this aim, we have developed additive models for these three properties.

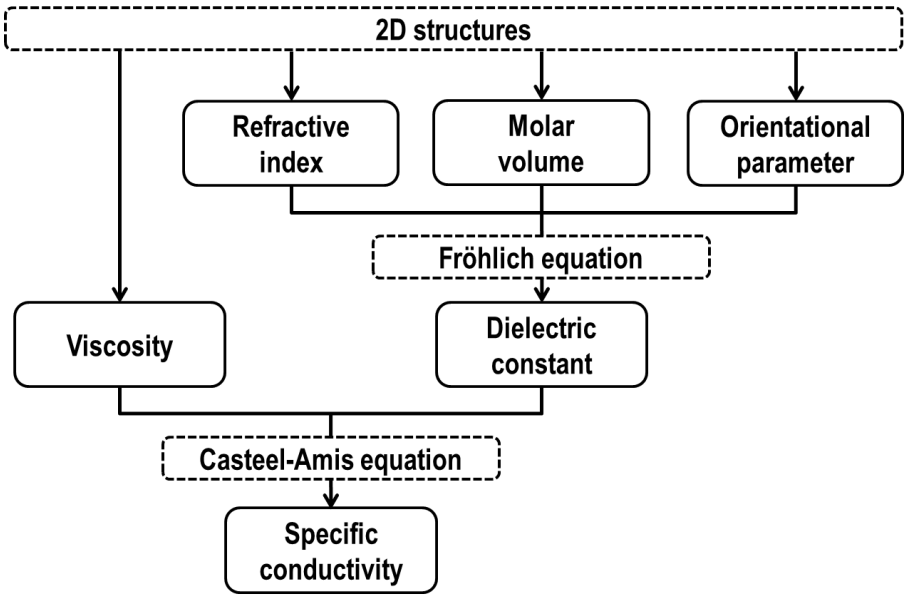


Figure: Method to predict the specific conductivity from 2D structures

1. Casteel, J. F.; Amis, E. S. Specific Conductance of Concentrated Solutions of Magnesium Salts in Water-Ethanol System. *J. Chem. Eng. Data* **1972**, 17, 55-59.
2. Fröhlich, H. *Theory of Dielectrics: Dielectric Constant and Dielectric Loss*, 2nd ed.; Oxford at the Clarendon Press, 1958.

P-11: Identification of novel sodium-dependent glucose co-transporter 1 inhibitors using proteochemometrics

Lindsey Burggraaff[1], Paul Oranje[2], Robin Gouka[2], Pieter van der Pijl[2], Marian Geldof[2], Guus Duchateau[2], Herman W.T. van Vlijmen[1,3], Adriaan P. IJzerman[1], and Gerard J.P. van Westen[1]

[1]Division of Drug Discovery & Safety, Leiden Academic Centre for Drug Research, Leiden University, Einsteinweg 55, 2333 CC, Leiden, The Netherlands, [2]Unilever Research & Development, Olivier van Noortlaan 120, 3133 AT, Vlaardingen, The Netherlands, [3]Janssen Research & Development, Turnhoutseweg 30, 2340 Beerse, Belgium.

Sodium-dependent glucose co-transporter 1 and 2 (SGLT1/SGLT2) are solute carriers responsible for glucose (re)absorption. SGLT2 is a target in the treatment of diabetes type 2 because of its high glucose transporting capacity [1,2]. Additionally, dual inhibitors blocking both SGLT1 and SGLT2, are currently in clinical development [3,4]. SGLT2 blockers exert their function at the renal tubules, whereas SGLT1 is mainly present at the apical side of the small intestine [5]. In contrast to SGLT2, selective SGLT1 inhibitors have not yet been marketed and are relatively little explored.

Here we aim at finding novel SGLT1 inhibitors to reduce intestinal dietary glucose absorption. We hypothesize that inhibition of intestinal SGLT1 requires a lower effective dose compared to inhibition of renal SGLT as the inhibitor is not subject to absorption, distribution and metabolism before reaching its target at an effective concentration. Solute carriers are complex and their hydrophobic nature in the cell membrane makes them difficult to crystalize. Hence, we applied machine learning to detect novel SGLT1 inhibitors as it does not require structural information [6]. We performed proteochemometrics by implementing 1D protein information into our models using Zscales [7,8].

We obtained a predictive model with a Matthews correlation coefficient of 0.52, sensitivity of 0.45, specificity of 0.97, positive predictive value of 0.78, and negative predictive value of 0.87. Subsequent to model training, we applied our model in virtual screening to select SGLT1 hit compounds. Of the 40 purchased compounds, 15 were experimentally validated *in vitro* leading to a hit rate of 38% with activities in the low micromolar range.

1. Clar, C.; Gill, J. A.; Court, R.; Waugh, N. *BMJ Open* **2012**, *2* (5), e001007.

2. Rosenstock, J.; Seman, L. J.; Jelaska, A.; Hantel, S.; Pinnetti, S.; Hach, T.; Woerle, H. J. *Diabetes, Obes. Metab.* **2013**, *15* (12), 1154–1160.
3. Sands, A. T.; Zambrowicz, B. P.; Rosenstock, J.; Lapuerta, P.; Bode, B. W.; Garg, S. K.; Buse, J. B.; Banks, P.; Heptulla, R.; Rendell, M.; Cefalu, W. T.; Strumph, P. *Diabetes Care* **2015**, *38* (7), 1181–1188.
4. Rendell, M. S. *Expert Opin. Pharmacother.* **2017**, 14656566.2017.1414801.
5. Gorboulev, V.; Schürmann, A.; Vallon, V.; Kipp, H.; Jaschke, A.; Klessen, D.; Friedrich, A.; Scherneck, S.; Rieg, T.; Cunard, R.; Veyhl-Wichmann, M.; Srinivasan, A.; Balen, D.; Breljak, D.; Rexhepaj, R.; Parker, H. E.; Gribble, F. M.; Reimann, F.; Lang, F.; Wiese, S.; Sabolic, I.; Sendtner, M.; Koepsell, H. *Diabetes* **2012**, *61* (1), 187–196.
6. Tresadern, G.; Trabanco, A. A.; Pérez-Benito, L.; Overington, J. P.; van Vlijmen, H. W. T.; van Westen, G. J. P. *J. Chem. Inf. Model.* **2017**, acs.jcim.7b00338.
7. van Westen, G. J. P.; Wegner, J. K.; IJzerman, A. P.; van Vlijmen, H. W. T.; Bender, A. *Med. Chem. Commun.* **2011**, *2* (1), 16–30.
8. Sandberg, M.; Eriksson, L.; Jonsson, J.; Sjöström, M.; Wold, S. *J. Med. Chem.* **1998**, *41* (14), 2481–2491.

P-13: Application of 3D-QSAR Methods in Drug Design & Discovery: Two Case Studies

Giulia Chemi [1], Simone Brogi [1], Margherita Brindisi [1], Stefania Butini [1], Sandra Gemma [1], Giuseppe Campiani [1]

[1] *European Research Centre for Drug Discovery and Development (NatSynDrugs) and Department of Biotechnology, Chemistry, and Pharmacy, DdE 2018-2022, University of Siena, Siena, Italy.*

The purpose of 3D-QSAR (three-dimensional structure activity relationships) technique is to derive the statistically significant relationships between molecular structures and biological activities by chemometric methods leading to the development of predictive mathematical models. During the time, numerous QSAR approaches have been developed for different purposes including the identification or design of new chemical entities for a selected target, and the prediction of specific properties or undesirable effects of novel molecules.[1] Here we describe two successful applications of 3D-QSAR method: i) the discovery of novel compounds by virtual screening with anti-prion profile and ii) the assessment of undesirable toxic effect of new chemical entities such as *h*ERG K^+ channels liability.

In the first work, a 3D-QSAR model was developed to screen a library of compounds to find novel chemicals able to prevent prion protein misfolding. Prion (PrP) is a protein that, after an incorrect folding, causes not curable neurodegenerative disorders: the transmissible spongiform encephalopathies (TSE). There are two forms of that protein: the cellular one (PrP^C) formed by three α-helices and two short β-sheets; and the pathological variant, the misfolded one (PrP^{Sc}) β-sheet motif rich protein that tends to accumulate in the brain of infected patients. The 3D-QSAR model was used as a first filter in a virtual screening protocol to select a limited number of potential molecules able to prevent the misfolding of the PrP^C. Then the model was combined with a molecular docking procedure and the prediction of ADME properties, to choose only that molecules able to bind the protein and to cross the blood-brain barrier. *In vitro* tests led us to select 9 hit compounds that effectively reduced the level of PrP^{Sc} and showed a non-toxic profile. Among them, one hit showed an interesting activity in preventing the pathological transition of PrP^C to PrP^{Sc} (IC_{50} = 1.6 μM). This compound can also bind and stain PrP^{Sc} aggregates in infected ScN2a cells. The combination of this interesting anti-prion cellular profile with a fluorescence imaging behavior, and the good brain permeability, suggests that this compound could be considered as a novel prototypic tool useful for the development of diagnostic and therapeutic probes for TSE.[2]

While in the second protocol, 3D-QSAR has been used to identify an *in silico* method for predicting, at an early stage of the drug discovery trajectory, the capability of compounds to interfere with *h*ERG K^+ channels, a well-known antitarget.[3] Blockade of *h*ERG K^+ channel has become a severe limitation for the introduction of new drugs in the market. In the past, several drugs have been withdrawn due to their relevant affinity for this channel and the consequent possible cardiotoxicity. In our work, a 3D-QSAR model was developed, employing a common pharmacophore as an alignment rule built on a subset of 22 highly active compounds (threshold K_i: 50 nM) active against *h*ERG K^+ channel. The sequential model developed with a set of 421 compounds with different span of activity (randomly divided in training and test set) proved to be predictive with respect to an external test set of 309 compounds ($r^2_{ext\_ts}$ = 0.86). The model was further validated by applying a decoys set, evaluating the Güner and Henry score (GH) and the Enrichment Factor (EF), and by the ROC curve analysis. The outcome demonstrated the high predictive power of the inclusive 3D-QSAR model, confirming the validity of this approach to obtain an *in house* tool useful for the design of new molecules with reduced *h*ERG-related cardiotoxicity.

1. Cherkasov, A.; et al. QSAR Modeling: Where Have You Been? Where Are You Going To?. *J. Med. Chem.*, **2014**, 57 (12).
2. Zaccagnini, L.; et al. Identification of Novel Fluorescent Probes Preventing PrP^{Sc} Replication in Prion Diseases. *Eur. J. Med. Chem.* **2017**, 127, 859-873.
3. Chemi, G.; et al. Computational tool for fast in silico evaluation of *h*ERG K^+ channel affinity. *Front. Chem.* **2017**, 5:7.

P-15: Applications of in silico approaches to decipher the structure and functions of ADAMTS13: En route to novel therapeutics of TTP

Bogac Ercig [1,2,3], Johana Hrdinova [1,2,3] Kanin Wichapong [1], Chris Reutelingsperger [1,3], Karen Vanhoorelbeke [4], Jan Voorberg [2], Gerry A.F. Nicolaes [1,3]

[1] Department of Biochemistry, Cardiovascular Research Institute Maastricht (CARIM), Maastricht University, Maastricht, The Netherlands, [2] Department of Plasma Proteins, Sanquin-AMC Landsteiner Laboratory, Amsterdam, The Netherlands, [3] PharmaTarget, Maastricht, The Netherlands, [4] iLaboratory for Thrombosis Research, IRF Life Sciences, KU Leuven Campus Kulak Kortrijk, Kortrijk, Belgium

A cryptic epitope in the ADAMTS13 spacer domain is targeted in most of the immune TTP patients. Based on these findings an auto-antibody resistant, gain-of-function variant (GoF) of ADAMTS13 was designed containing the following amino acid substitutions in spacer domain: R568K / F592Y / R660K / Y661F / Y665F[1]. Structural studies revealed that GoF-ADAMTS13 exists in an open (active), while wt-ADAMTS13 is found in closed (inactive) conformation, with the C-terminal CUB domains bound to the spacer domain. Our aim is to employ in silico approaches to predict the network interactions between the spacer domain and its binding partners (i.e: CUB domains and autoantibodies) which will provide structural and functional data that will be translated into therapeutic use.

The experimental structure of C-terminal CUB domains of ADAMTS13 are not available in Protein Data Bank. The YASARA Structure tool was used for homology modeling of the C-terminal CUB1-2 domains[2]. The anti-ADAMTS13 autoantibodies were modelled by Bioluminate module in Schrodinger suite. HADDOCK protein-protein docking were employed to study the interactions of both CUB domains and autoantibodies against spacer domain of ADAMTS13. GoF ADAMTS13 mutations were in silico introduced to final poses, next both WT- and GoF- ADAMTS13 were subjected to binding free energy calculation with AMBER16 over a 100ns molecular dynamics simulation. Subsequently, these poses were investigated to reveal which residues are contributing to conformational changes of ADAMTS13.

A pose with relatively higher binding affinity against WT-ADAMTS13 and lower binding affinity against GoF-ADAMTS13 at the same time was found to be informative to predict which residues are important for binding of CUB domains and autoantibodies. These residue predictions are subjected to in vitro mutation studies in order to test changes on conformation, proteolytic activity and resistance against auto-antibodies.

We have used the available structural bioinformatics tools to predict the nature of conformational changes which switches the human ADAMTS13 protein between active and inactive states. The derived knowledge from the current study will further be used for the design of novel therapeutics of immune TTP.

1. Jian, C.; Xiao, J.; Gong, L.; Skipwith, C. G.; Jin, S. Y.; Kwaan, H. C.; Zheng, X. L. Gain-ofFunction ADAMTS13 Variants That Are Resistant to Autoantibodies against ADAMTS13 in Patients with Acquired Thrombotic Thrombocytopenic Purpura. *Blood* **2012**, *119* (16), 3836– 3843.
2. Ercig, B.; Wichapong, K.; Reutelingsperger, C. P. M.; Vanhoorelbeke, K.; Voorberg, J.; Nicolaes, G. A. F. Insights into 3D Structure of ADAMTS13 : A Stepping Stone towards Novel Therapeutic Treatment of Thrombotic Thrombocytopenic Purpura. *Thromb. Haemost.* **2018**, 28–41.

P-17: Confidence estimation of ADME properties using conformal prediction

C. Founti [1], V. Gillet [1], G. Vessey [2]

[1] Information School, The University of Sheffield, Sheffield, UK, [2] Lhasa Limited, Leeds, UK

The impact of predictive models to guide the drug discovery cycle is broadly accepted, particularly for the optimisation of ADME properties. The application of QSAR models for property prediction reduces the need for iterative in vivo and in vitro testing and consequently saves a significant amount of resources. Building a QSAR model that will predict at a suitable level of accuracy for the intended application, however, can be a complex task and depends on the composition of the dataset. Furthermore, an inaccurate prediction could lead to a lost opportunity for a potent compound and further delays in development.

Conformal prediction provides a machine-learning framework that integrates confidence estimation in QSAR modelling[1]. For QSAR regression models, the main objective of the framework is to produce compound-specific prediction intervals that represent the reliability of the prediction at a user-defined level of confidence. Prediction intervals are obtained by training a machine learning algorithm on the proper training set and generating a ranked list of nonconformity scores from a calibration set. Compound-specific prediction intervals are obtained by normalising the scores with estimates obtained from an error model. The nonconformity score corresponding to a user-defined confidence level on the list is then used to calculate the prediction intervals for all future predictions.

In this study, the validity and efficiency of prediction intervals produced by random forest and support vector machine conformal predictors is evaluated for ADME datasets. The normalisation of standard prediction intervals using different error models is investigated[2, 3] and evaluated against the already established k-nearest neighbor algorithm[4].

Abbreviations:
QSAR: Quantitative Structure Activity Relationship
ADMET: Absorption, Distribution, Metabolism, Excretion

*The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement nº 612347.

1. Eklund, M.; Norinder, U.; Boyer, S.; Carlsson, L. Application of Conformal Prediction in QSAR. In: *Artificial Intelligence Applications and Innovations. AIAI 2012. IFIP Advances in Information and Communication Technology*; Iliadis L., Maglogiannis I., Papadopoulos H., Karatzas K., Sioutas S., Ed.; Springer: Berlin, Heidelberg, 2012; Vol 382.
2. Sheridan, RP. Three useful dimensions for domain applicability in QSAR models using random forest. *J. Chem. Inf. Model.* **2012**, 52: 814-823.
3. Toplak, M.; Močnik, R.; Polajnar M.; et al. Assessment of Machine Learning Reliability Methods for Quantifying the Applicability Domain of QSAR Regression Models. *J. Chem. Inf. Model.* **2014**, 54: 431-441.
4. Papadopoulos, H.; Vovk, V.; Gammerman, A. Regression Conformal Prediction with Nearest Neighbours. *J. Artif. Intell. Res*. **2011**, 40: 815-840.

P-19: Selectivity profiles in Activity Atlas

M. Mackey [1], P. Tosco [1]

[1] Cresset, Litlington, UK

At the last ICCS we presented methods for computing activity cliff metrics based on 3D similarity and showed how these can be utilized to identify pairs of molecules where small changes in steric or electrostatic potential have a disproportionate effect on activity. More recently we have developed the Activity Atlas method, which summarizes information obtained from multiple 3D activity cliff pairs. This provides an analysis of the molecular shape and electrostatic features which correlate with a change in activity across multiple pairs, highlighting the regions critical to binding (Fig. 1)

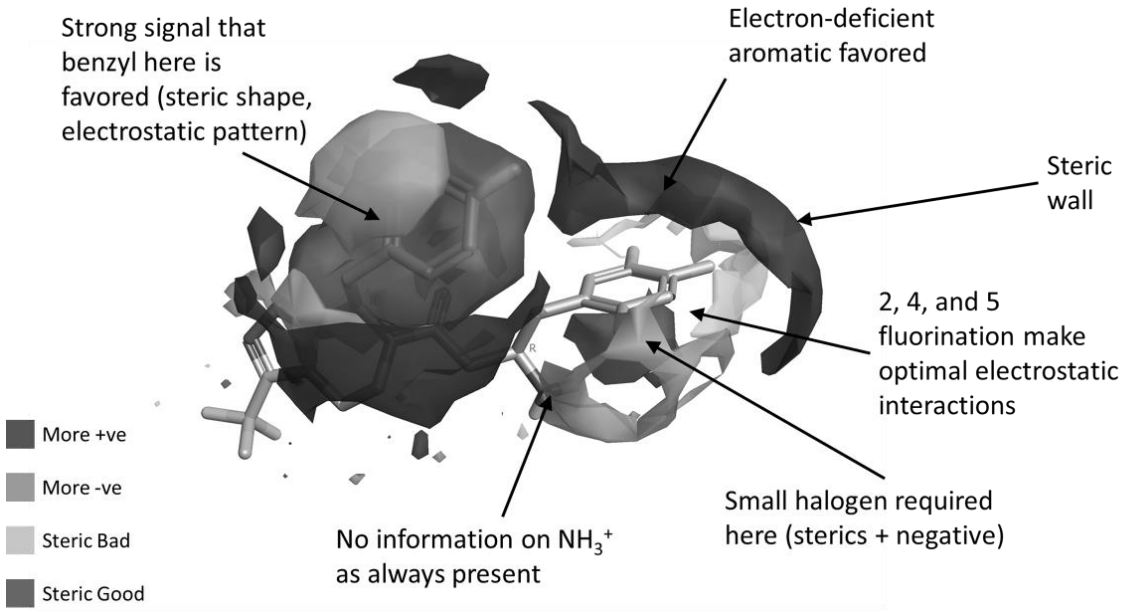


Fig. 1: The Activity Atlas method provides interpretable maps of the SAR around a series.

In this talk we show how to extend the Activity Atlas method to investigate selectivity. For closely related proteins, a reasonable assumption is that the ligand alignment will be conserved. In this case, the Activity Atlas method can be used to investigate selectivity. Two methods are presented. In the first, the analysis proceeds separately on the different activity values. The resulting models can be compared to give insights into the different activity requirements for the subtypes, and hence what changes might increase activity at the desired subtype and/or reduce it at the undesired one (Fig. 2).

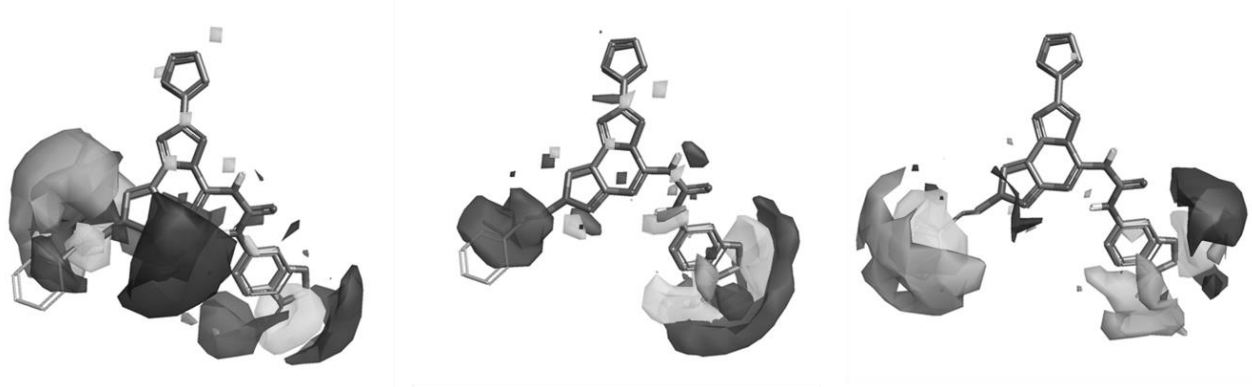


Fig. 2: The Activity Atlas maps for adenosine A1, A2a and A3.

Alternatively, the analysis can be performed on activity differences giving a direct readout of the changes leading to selectivity (Fig. 3). This method is significantly more powerful at teasing out the direct influences on selectivity, at the expense of losing information on whether the highlighted differences increase or decrease the affinity across all subtypes.

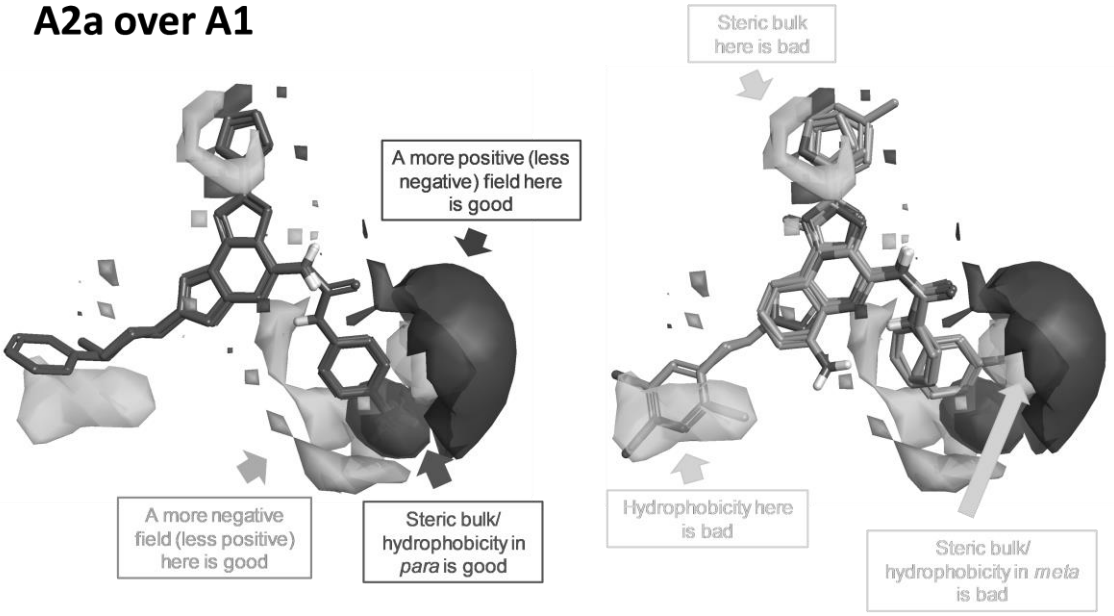


Fig. 3: Activity cliff summary maps for adenosine A2a over A1 selectivity. The maps are superposed to the most selective (left, shown in darker gray) and the least selective (right, shown in lighter gray) compounds in the training set.

P-21: *KnowTox*: Risk Assessment by Automated Read-Across and Machine Learning

Andrea Morger [1], Janosch Achenbach [2], Miriam Mathea [2], Antje Wolf [2], Robert Landsiedel [2], Klaus-Jürgen Schleifer [2], Andrea Volkamer [1]

[1] In Silico Toxicology Group, Institute of Physiology, Charité, Berlin, Germany, [2] BASF SE, Ludwigshafen, Germany

With new chemicals being synthesized every year, assessment of their toxicological potential, *i.e.*, their harmful effects on humans, and the environment is a prerequisite for production and marketing. Most of the toxicological testing required by regulations is still requesting animal studies. In this context, *in silico* methods have great potential to reduce time, cost, and ultimately animal testing as they make use of the ever-growing amount of available toxicity data.[1]

In our *KnowTox* project, we develop a toxicity prediction tool that makes use of available knowledge from external and in-house data to provide rational support for Read-Across, including modern machine learning (ML) techniques to come closer to the vision of transforming toxicology into a predictive science. Our major data source is the freely available *ToxCast*[2] dataset, consisting of ~8300 compounds, such as pesticides, pharmaceuticals, and industrial chemicals, tested on up to 1000 different endpoints, *e.g.* effects on cell cycle, cytotoxicity, or steroid receptor interactions. We will present a workflow – together with a case study – to search the entire *ToxCast* dataset for substances, which are most similar in terms of features and structure to any query compound. Information about these substances' commonalities such as toxicological key events and adverse outcomes, and common structural features can automatically be generated. Furthermore, previously identified substructures associated with toxic effects[3,4] are highlighted to warn the user and either guide the design of less toxic compounds or target subsequent *in vitro* and *in vivo* testing. For ML application, we adapted an open source standardization workflow to remove duplicates, salts, and mixtures, yielding a reduced set of ~7500 clean compounds. Random Forest models for toxicity predictions on this dataset are currently trained and evaluated - by exhaustively sampling different combinations of fingerprints, ML parameters and data balancing strategies per endpoint - showing promising prediction accuracies.

Identification of sufficiently similar chemicals will support rationales for Read-Across[5], and accurate toxic mechanism or endpoint predictions will guide further toxicity testing or the deselection of most likely harmful compounds in an early stage of the often lengthy research and development process. Our combined prediction tool

can, together with the experience of toxicologists, help to improve efficiency and reduce the need for animal testing for toxicological assessments in development projects and regulatory product registration.

1. Mayr, A. *et al.* DeepTox: Toxicity Prediction using Deep Learning. *Front. Environ. Sci.* **2016**, 3, 80.
2. Richard, A. M. *et al.* ToxCast chemical landscape: paving the road to 21st century toxicology. *Chem. Res. Toxicol.* **2016**, 29(8), 1225-1251.
3. Sushko, I. *et al.* ToxAlerts: a web server of structural alerts for toxic chemicals and compounds with potential adverse reactions. *J. Chem. Inf. Model.* **2012**, 52(8), 2310-2316.
4. Baell, J. B.; Holloway, G. A. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.* **2010**, 53(7), 2719-2740.
5. Teubner W.; Landsiedel R. Read-across for hazard assessment: The ugly duckling is growing up. *Altern. Lab. Anim.* **2015**, 43, 67-71.

P-23: Machine learning to predict the recruitment profile of intracellular binding partners of G Protein Coupled Receptors

Trung Ngoc Nguyen, Marcel Bermudez, Gerhard Wolber

Freie Universität Berlin, Molecular design lab, Institute of Pharmacy, Berlin, Germany

G protein-coupled receptors (GPCRs) are generally characterized by seven α-helical transmembrane domains (TM1–7).[1] They transfer diverse extracellular stimuli into intracellular downstream signals. The perception of GPCRs as mere on-off switches for a given pathway has changed drastically over the last decade.[2] Today we know that GPCRs are highly dynamic signaling machines with multiple signaling pathways like G protein activation and β-Arrestin recruitement.[3] Due to significant progress in structural elucidation of GPCRs, ligand design using structure-based modeling became more feasible. However, the identification of small organic molecules binding to the extracellular part is only loosely coupled to the desired pharmacological effect. A prominent example is TRV-130 (Oliceridine)[4], which is a biased orthosteric ligand for μ-opioid receptor and has a preferred activation of G_i-protein instead of ß-arrestin resulting in a more effective pain medication with less side effects. Currently, such ligands can only be discovered by serendipity, because no comprehensive models for predicting intracellular effects upon extracellular ligand binding exists so far.

The presented study aims at identifying and utilizing structural properties of class A GPCRs to predict specific signaling induced by different ligands. Starting with the extraction of three-dimensional structural descriptors from 195 different crystal structures of 40 GPCRs, supervised machine learning[5] algorithms are used to predict the recruitment profile of intracellular binding partners. Additional descriptors are generated through molecular dynamics simulations[6] with different ligands with known intracellular recruitment profiles. We present a comparison of static and dynamic structural descriptors in the context of different machine learning algorithms with respect to their ability to predict intracellular binding partner selection and the resulting signaling bias.

1. Szczepek, M.; Beyrière, F.; Hofmann, K. P.; Elgeti, M.; Kazmin, R.; Rose, A.; Bartl, F. J.; Stetten, D. von; Heck, M.; Sommer, M. E. *et al.* Crystal structure of a common GPCR-binding interface for G protein and arrestin. *Nature communications* **2014**, *5*, 4801.
2. Pavlos, N. J.; Friedman, P. A. GPCR Signaling and Trafficking: The Long and Short of It. *Trends in endocrinology and metabolism: TEM* **2017**, *28*, 213–226.
3. Hilger, D.; Masureel, M.; Kobilka, B. K. Structure and dynamics of GPCR signaling complexes. *Nature structural & molecular biology* **2018**, *25*, 4–12.
4. Chen, X.-T.; Pitis, P.; Liu, G.; Yuan, C.; Gotchev, D.; Cowan, C. L.; Rominger, D. H.; Koblish, M.; Dewire, S. M.; Crombie, A. L. *et al.* Structure-activity relationships and discovery of a G protein biased μ opioid receptor ligand, (3-methoxythiophen-2-yl)methyl({2-(9R)-9-(pyridin-2-yl)-6-oxaspiro-4.5decan-9-ylethyl})amine (TRV130), for the treatment of acute severe pain. *Journal of medicinal chemistry* **2013**, *56*, 8019–8031.
5. Jordan, M. I.; Mitchell, T. M. Machine learning: Trends, perspectives, and prospects. *Science (New York, N.Y.)* **2015**, *349*, 255–260.
6. Bermudez, M.; Mortier, J.; Rakers, C.; Sydow, D.; Wolber, G. More than a look into a crystal ball: Protein structure elucidation guided by molecular dynamics simulations. *Drug discovery today* **2016**, *21*, 1799–1805.

P-25: Estimation of electrophilicity for warheads of covalent protease inhibitors

S. Pach, C. Tauber, J. Rademann, G. Wolber

Pharmaceutical and Medicinal Chemistry, Institute of Pharmacy, Freie Universität Berlin, Königin-Luise Straße 2+4, Berlin 14195, Germany

Enteroviral infections are associated with increased risk of neurological and cardiac complications. The enteroviral cysteine protease $C3^{pro}$ is involved in processing of viral polyprotein during the replication into functional components of viral particles. Therefore, it represents a promising target to fight enteroviral infections efficiently. Our goal is to target viral proteases by covalent inhibition.

Covalent inhibitors have improved efficiency of enzyme inhibition compared to non-covalent ones [1]. Furthermore, they possibly allow targeting of "undruggable" binding sites and lower the risk of resistance development [2]. Efforts to find a suitable warhead are typically based on "trial and error" chemical synthesis and in-vitro testing. Hence, there is a need for a rational *in-silico* method describing mechanisms of covalent binding to enzymes.

We investigate the activity of eleven Michael acceptors on C3-Protease. Considering only electronic effects as described by classical chemical theories (e.g. the HSAB-Concept) only fails to explain activity trends. For this reason, we developed a geometric descriptor based on the reaction mechanism between electrophile warhead and nucleophile protease-active site. The change of essential angles between different steps of reaction shows that reactivity of electrophiles and thereby reaction kinetics depends on steric parameters (Fig. 1). Our novel approach expands the classical view on covalent ligand-enzyme-interaction (Fig. 2) and allows prediction of electrophile quality, including electronic and steric effects during the binding reaction.

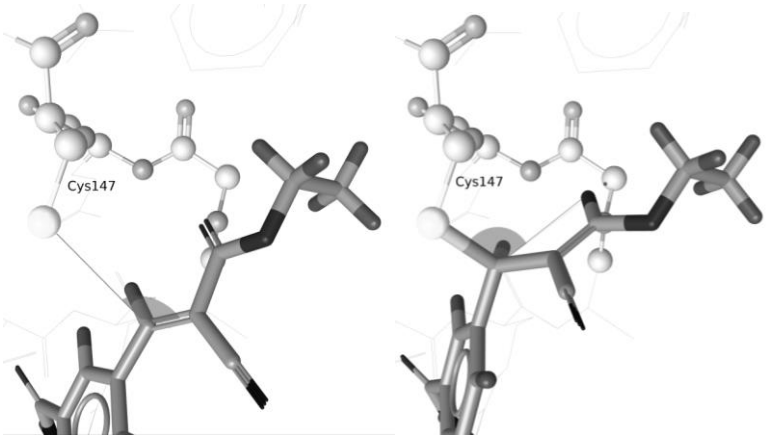


Fig. 1: Measurement of double-bond-plane-angle for non-covalent complex and intermediate

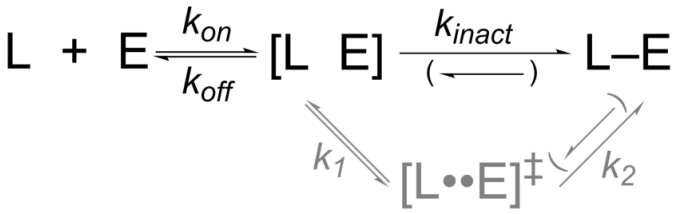


Fig. 2: Classic description of reaction mechanism (black) and extended path (grey) between ligand (L) and enzyme (E) [mod. After 1]

1. Singh, J.; Petter, R. C.; Baillie, T. A.; Whitty, A. The resurgence of covalent drugs. *Nat. Rev. Drug Discovery* [Online] **2011**, 10, 307−317.
2. Bauer, R. A. Covalent inhibitors in drug discovery: from accidental discoveries to avoided liabilities and designed therapies *Drug Discovery Today* [Online] **2015**, 20 (9), 1061-1073.

P-27: A web-based informatics platform for PhysChem/ADME/Tox property predictions

A. Sazonovas [1,2], K. Lanevskij [1,2], R. Didziapetris [1,2]

[1] VšĮ „Aukštieji algoritmai“, A.Mickevičiaus 29, LT-08117 Vilnius, Lithuania, [2] ACD/Labs, Inc., 8 King Street East, Toronto, Ontario, M5C 1B5, Canada

Percepta for ACD/Portal is a new platform that builds upon the well established components of ACD/Labs Percepta desktop software – reliable predictive algorithms for a multitude of physicochemical, ADME, and safety-related properties, powerful data mining, visualization, compound profiling and risk assessment capabilities, as well as ACD/Structure Design Engine for generating libraries of virtual analogs compatible with the desired characteristics. Percepta for ACD/Portal combines these features with flexible network-based deployment, raising software interactivity to a new level and offering some exciting features. This work brings particular focus to the components of the web version of Percepta that leverage the power of high performance computing in a server environment. The server-side architecture of ACD/Portal relies on multiple calculation units (kernels) that enable parallel processing of very large amounts of data in real time. These capabilities paved the road for new developments in several key areas. The addition of a quick exploration of the predicted property values for a multitude of structural analogs of a compound enables on-the-fly liability checking, i.e. identifying the areas of the molecule potentially responsible for unfavorable ADME/Tox characteristics. Adaptation of the ACD/Structure Design Engine to the employed architecture gave rise to a new generation of this tool that enables extensive enumeration of substituent property space in accordance with specific user-defined constraints at up to four independently varying substituent positions at the same time. Along with a built-in database of more than 10^4 building blocks, this leads to exploration of up to 10^{16} virtual analogs, which is actually feasible in real time inside Percepta for ACD/Portal. Such broadened scope of the chemical space investigated greatly enhances the potential of encountering new compounds with the most favorable property profiles.

P-29: Development of a novel structure descriptor combining molecular shape and surface properties

A. Schultz [1], K. Baumann [1]

[1] Institute of Medicinal and Pharmaceutical Chemistry, Technische Universität Braunschweig, Beethovenstr. 55, 38106 Braunschweig, Germany

Molecular shape and the spatial localization of potential binding partners play a major role in a molecule's interaction with its complementarily shaped target. Methods using the 3D geometry to represent a molecule can be divided into alignment-based and alignment-independent approaches. The alignment-based methods have the drawback of high computational costs and potential bias due to the process of finding the optimal alignment prior to computing their respective descriptors, while the alignment-independent approaches are often more difficult to interpret visually. A challenge faced by both, alignment-based and alignment-independent methods, is the conformational flexibility that needs to be taken into account when dealing with 3D representations of molecules. The standard approach is to create a relatively large ensemble of conformers per molecule and to use this ensemble for the intended study. The downside of the ensemble method is the computational cost required to generate appropriate conformers as well as the exponentially rising amount of calculations for encoding and numerically processing the ensembles for machine learning.

For the representation and comparison of the 3D shapes of macroscopic objects, Gal et al. developed the local diameter function which is not only invariant to scaling, translation and rotation but is also insensitive to transformations which are based on skeletal articulated movement[1].

In the following, we describe the adaptation of the local diameter function to molecular shapes and the development of a descriptor based on this adapted function. Starting with a triangulated mesh of the molecular surface the molecule's diameter in the neighborhood of each vertex is calculated using a cone of rays traveling through the molecule yielding a shape representation. Furthermore, properties of RDKit's Chemical Features[2] and local hydrophilicity/lipophilicity are mapped onto the molecular surface. Properties on the local surface touched by a cone's origin and base are recorded together with the local diameter value and merged into one descriptor. The novel

descriptor is applied to several benchmark datasets and compared to established approaches. The most critical hyperparameters (cone angle, number of rays) of the novel descriptor are systematically varied and their influence on predictive power is demonstrated.

1. Gal, R.; Shamir, A.; Cohen-Or, D. Pose-Oblivious Shape Signature. *IEEE Trans. Vis. Comput. Graph*. **2007**, 13, 261-271.
2. The RDKit: Open-source cheminformatics; http://www.rdkit.org (accessed Jan 30, 2018).

P-31: Classification of corneal permeability of drug-like compounds using data mining and machine learning

João Meireles [1], Carlos Simões[1], Rui Brito [1,2]

[1] BSIM Therapeutics, Coimbra, Portugal, [2] Chemistry Department, University of Coimbra, Coimbra, Portugal

The eye cornea works as a protecting barrier against the penetration of xenobiotics, including drugs. The determination of corneal permeability by experimental means is a time-consuming and expensive process, requiring fresh biological tissue and significant amounts of pure compounds. As such, the ability to predict the corneal permeability of drug candidates from their molecular properties *in silico* would be a valuable instrument in the development of new drugs for ophthalmologic indications, as illustrated by previous efforts by a few researchers [1,2].

The main goal of this work is to find mathematical functions that map selected chemical features onto corneal permeability through a careful use of machine learning and data mining techniques.

Through interaction with an ophthalmologist and mining of scientific literature and patent data, we assembled a set of 70 compounds: 35 of which are known to be capable of penetrating the cornea at therapeutic concentrations (cornea_positive or "+" class) and 35 are unable to do so (cornea_negative or "–" class). The molecules in the set were built, clean and processed into lowenergy three-dimensional structures using open source tools, including Avogrado, OpenBabel and MOPAC. For each compound, 3082 molecular descriptors were then computed using well-known packages such as CDK, MOLD2, OpenBabel and PaDEL. With the exception of the compound name and its ocular permeability class, all data is numeric.

The characteristics of the resulting data set pose several challenges to mathematical analysis: 1) there are many more features (variables) than samples; 2) the variables are highly correlated, and 3) many variables show non-symmetry, significantly different scales and ranges, a great number of zero entries, modes, etc. Thus, our first concern was the cleaning-up and transformation of the data using robust statistical tools. Since our main goal is to understand how corneal permeability is related to certain chemical features, yet not with others, in this study we use feature selection and global optimization methods to uncover the most relevant attributes. We explore many different types of variable selection to derive multiple predictive models: 1) filtering information based on the distribution properties of the chemical features, using less restrictive statistical assumptions like nonparametric tests, 2) filtering the attributes based on ensemble strategies, 3) averaging the quality of the variables according to their conditional dependencies, 4) building robust linear latent variables, and 5) exploring multidimensional scaling, stochastic proximity embedding, and other nonlinear dimensionality reduction methods.

Given the small size of the data set, model validation is carried out via leave-one-out cross-validation (LOO-CV) and bootstrap with several repetitions – yielding highly accurate predictions. Moreover, we compare the performance of our machine learning-based classifiers using robust non-parametric tests.

Considering previous attempts to predict cornea permeability, our strategy brings the advantage of not making unrealistic assumptions about the molecular data and combines the strengths of multiple approaches to interpreting the proposed predictive models.

1. Kidron, H.; Vellonen, K. S.; Del Amo, E. M.; Tissari, A.; Urtti, A. Prediction of the Corneal Permeability of Drug-like Compounds. *Pharm. Res.* **2010**, *27* (7), 1398–1407.
2. Ghorbanzad'E, M.; Fatemi, M. H.; Karimpour, M.; Andersson, P. L. Quantitative and Qualitative Prediction of Corneal Permeability for Drug-like Compounds. *Talanta* **2011**, *85* (5), 2686–2694.

P-33: Coarse-grained approaches for prediction of solubility and membrane permeability of large drugs: The Why and the How

Johannes (Hans) G.E.M. Fraaije [1]

[1] *Culgi BV and Leiden University, Leiden, The Netherlands.*

Drug discovery is currently addressing molecular compounds that are larger than could be anticipated at the times of the seminal Lipinsky contribution[1]. One may think of protein-protein interaction inhibitors, kinase inhibitors, and all biologicals. Age-old statistical methods to predict development qualifiers such as solubility[2] and membrane permeability (reviewed in [3-4]), date from the times that pre-date the new chemistries. They are, without exception, calibrated on data for small molecules, and for that reason must be considered of less relevance now. Unfortunately, many a commercial or free software bases its algorithms on the older methods, for lack of anything better. The necessary experimental data for the new molecules is simply not available (in the public domain) to the extent necessary for a data-driven approach. For this reason, more modern data-driven approaches such as artificial intelligence, to supplant the existing QSPR models, are very difficult if not impossible to develop.

Even if data would be public, one realizes quickly that the necessary database would have to be extensive to be of use. The drift to larger and larger molecules is driven by the quest for selectivity, to address targets that before were considered undruggable. Selectivity translates into large molecules, with many handles to find an exquisite and unique binding pattern. The same many-handles property translates into great many more potential combinations of side-groups (and scaffolds) than any calibration subset could cover. At the same time, a literature survey points to reliable solubility data for large drug-like molecules measured in the hundreds, and, similar, public membrane permeability datasets have perhaps an even smaller number[3-4] of trustworthy datapoints. An illustrative example is from the burgeoning field of protein-protein inhibitors. Such inhibitors (modified peptides) suffer greatly from lack of membrane permeability[5]. A data-driven approach would not only need to take into account the extremely large chemical space but also the flexibility of the molecules, and potential folding in the membrane; this seems almost impossible. Therefore, however unfortunate, purely datadriven approaches cannot work, not now, and not in the foreseeable future.

When data is scarce, the only alternative we have is a physics-based approach[3-4], such as we discuss here. The physics-based modeling we propose rests on the coarse-grained paradigm[6-7], that lumps groups of atoms into small fragments. Efficient simulation algorithms then use the fragments for the calculation of thermodynamic interactions. Our coarse-grained algorithms permit the complete calculation of solubility and permeability for a large drug-like molecule (up to MW 10000) in one go, in a mere few minutes on an ordinary desktop computer. The algorithm includes automated ways for calculation of charge distribution, the cutting of molecules into pieces and, the calculation of thermodynamics and diffusion through thermodynamic integration. A dynamics simulation is a basis for the thermodynamics and diffusion calculation, and automatically includes flexibility, differential folding, regrouping, etc., and makes no a priory assumption regarding positioning of fragments. Apart from the original calibration on datasets outside the pharma-domain, very few additional parameters are needed for finetuning to the applications at hand, which makes the method excellently suited for the scarce-data problems in solubility and membrane permeability.

Application of the method includes calculation of diffusion coefficients of more than 11k molecules[7], with a rather astonishing agreement with earlier empirical laws. The figure shows the correlation between simulated and experimental results. We obtained the results in a few minutes per molecule on a single PC-core. While the new method is physics-based, the computational speed is on par with typical QSPR calculation times and makes the method suitable for virtual screening studies. The inserted Wilke-Chang empirical relation is from the 1950's. Our study is the first time to recover such correlation by either theory or simulation.

From a theoretical perspective, there is no difference in our method between that of solubility and membrane permeability prediction. Both algorithms rely on the same coarse-graining protocol, with the same set of parameters. We will present the method of calculation and illustrative examples from the prediction of kinase-inhibitor and peptide solubility, and membrane permeability.

Culgi is sponsored by a consortium of industries from personal and home care industries, oil industries and chemical industries.

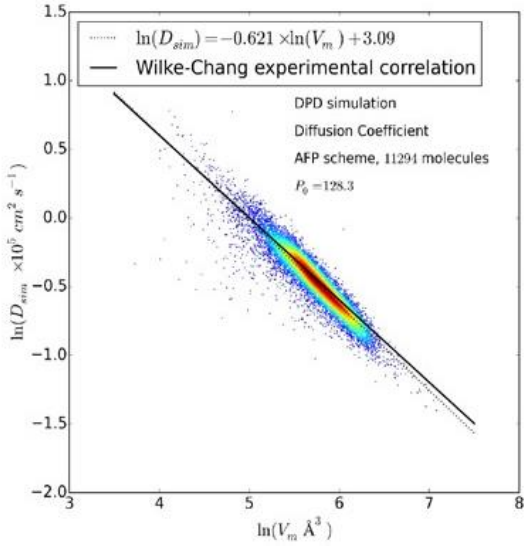


Figure 1 Coarse-grained prediction of diffusion coefficients. Each point takes a few minutes calculation time on one core.

1. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J., Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews* **1997,** *23* (1-3), 3-25.
2. Jorgensen, W. L.; Duffy, E. M., Prediction of drug solubility from structure. *Advanced Drug Delivery Reviews* **2002,** *54* (3), 355-366.
3. Leung, S. S. F.; Mijalkovic, J.; Borrelli, K.; Jacobson, M. P., Testing Physical Models of Passive Membrane Permeation. *J Chem Inf Model* **2012,** *52* (6), 1621-1636.
4. Leung, S. S. F.; Sindhikara, D.; Jacobson, M. P., Simple Predictive Models of Passive Membrane Permeability Incorporating Size-Dependent Membrane-Water Partition. *J Chem Inf Model* **2016,** *56* (5), 924-929.
5. Matsson, P.; Kihlberg, J., How Big Is Too Big for Cell Permeability? *J. Med. Chem.* **2017,** *60* (5), 1662-1664.
6. Fraaije, J. G. E. M.; Van Male, J.; Becherer, P.; Serral Gracià, R., Coarse-Grained Models for Automated Fragmentation and Parametrization of Molecular Databases. *J. Chem. Inf. Model.* **2016,** *56* (12), 2361-2377.
7. Fraaije, J. G. E. M.; van Male, J.; Becherer, P.; Serral Gracià, R., Calculation of Diffusion Coefficients through Coarse-Grained Simulations Using the Automated-Fragmentation-Parametrization Method and the Recovery of Wilke–Chang Statistical Correlation. *Journal of Chemical Theory and Computation* **2017**.

P-35: Molecular Dynamics Fingerprints (MDFP): Combining MD and Machine Learning to Predict Physicochemical Properties

Shuzhe Wang[1], Sereina Riniker [1]

[1] *Laboratory of Physical Chemistry, ETH Zurich, Vladimir-Prelog-Weg 2, 8093 Zurich, Switzerland*

Molecular dynamics fingerprints (MDFP) are a novel approach that combines MD with cheminformatics modeling. In brief, short simulations are performed on small molecules followed by extracting statistical moments of calculated terms (e.g. potential energy components, radius of gyration, solvent-accessible surface) to form fingerprint vectors. Such MDFP are constructed for a set of molecules in different solvents and physical states. Then, using supervised machine learning (ML), models can be trained with MDFP as inputs to predict various physicochemical properties, e.g. solvation free energy, partition coefficient, vapor pressure, melting point and solubility, which are important quantities for both pharmaceutical and environmental research. MDFP are information-rich descriptors that are highly versatile and can easily be adapted to the property to be predicted (i.e. which physical states are simulated and which terms are calculated).

Recently,[1] we have shown that solvation free energies in different solvents can be predicted accurately using the MDFP-based approach. The models performed similarly to other more rigorous *in silico* methods such as free-energy perturbation (FEP) and COSMO-RS, with the added benefits of being easier to implement and computationally less expensive. From the predicted solvation free energies, partition coefficients in numerous mixtures can be obtained. When applied to the molecules in the SAMPL5 blind challenge[2] retrospectively, the

MDFP-based approach performed better than the submitted approaches.[1] We are further testing our approach by participating in the current SAMPL6 blind challenge.[3] In addition, we present a MDFP-based approach to predict vapor pressure.

1. Riniker, S. Molecular Dynamics Fingerprints (MDFP): Machine Learning from MD Data To Predict Free-Energy Differences. *J. Chem. Inf. Model*. **2017**, 57, 726-741.
2. Bannan, C. C.; Burley, K. H.; Chiu, M.; Shirts, M. R.; Gilson, M. K.; Mobley, D. L. Blind Prediction of Cyclohexane–Water Distribution Coefficients from the SAMPL5 Challenge. *J. Comput. Aided Mol. Des*. **2016**, 30, 927-944.
3. Mobley, D. L. SAMPL6 Challenge Homepage, https://github.com/MobleyLab/SAMPL6 (accessed Jan 26, 2018).

P-37: Towards Small Molecule Inhibition of HSP90 Dimerization

D. Bickel[1], E. Ciglia[1], S. Bhatia[2], F. Hansen[3], T. Kurz[1], J. Hauer[2], H. Gohlke[1]

[1] Institute for Pharmaceutical and Medicinal Chemistry, Heinrich Heine University Düsseldorf, Düsseldorf, Germany, [2] Department of Pediatric Oncology, Hematology and Clinical Immunology, Medicinal Faculty, Heinrich Heine University Düsseldorf, Düsseldorf, Germany, [3]Pharmaceutical/Medicinal Chemistry, Institute of Pharmacy, Leipzig University, Leipzig, Germany

Protein-protein interactions are known to be involved in a wide variety of physiological and pathophysiological processes. Thus, the interest in targeting protein-protein interfaces in drug discovery is increasing steadily.

The heat shock protein of 90 kDa (HSP90) was shown to be involved in malignant transformation and tumor progression in several cancer cell lines. As such, HSP90 represents an attractive target for cancer therapy[1]. Targeting the dimerization interface in the C-terminal domain should provide a novel way to interfere with HSP90 function. Here we present an approach for the *de novo* design of dimerization inhibitors for HSP90.

The approach is based on a structural analysis of the dimerization interface of HSP90 and prediction of binding hotspots, using a structural decomposition of the effective energy of binding, computed by MM-GB/SA calculations[2]. Afterwards, these hotspots were used to guide the development of medium-sized peptide[3] and peptidomimetic inhibitors[4]. These were tested in enzymatic and cellular assays, demonstrating their ability to inhibit HSP90 dimerization as well as anti-proliferative activity against various tumor cell lines.

In order to find a set of drug-like small molecule inhibitors, we used the most active peptidic compounds as lead structures for a pharmacophore- and shape-based virtual screening. Refining the queries to emphasize interactions with previously identified hotspots, we were able to obtain smaller molecules with improved physicochemical properties. In a preliminary screening, anti-proliferative activity in the lower µM range could be confirmed for some of these compounds. This highlights the possibility to identify drug-like protein-protein interaction inhibitors from structural analysis of the dimerization interface.

1. Whitesell, L.; Lindquist, S. L., HSP90 and the chaperoning of cancer. *Nat Rev Cancer* **2005,** *5* (10), 761-72.
2. Ciglia, E.; Vergin, J.; Reimann, S.; Smits, S. H.; Schmitt, L.; Groth, G.; Gohlke, H., Resolving hot spots in the C-terminal dimerization domain that determine the stability of the molecular chaperone Hsp90. *PLoS One* **2014,** *9* (4), e96031.
3. Bopp, B.; Ciglia, E.; Ould-Chaib, A.; Groth, G.; Gohlke, H.; Jose, J., Design and biological testing of peptidic dimerization inhibitors of human Hsp90 that target the C-terminal domain. *Biochim Biophys Acta* **2016,** *1860* (6), 1043-55.
4. Diedrich, D.; Moita, A. J. R.; Rüther, A.; Frieg, B.; Reiss, G. J.; Hoeppner, A.; Kurz, T.; Gohlke, H.; Lüdeke, S.; Kassack, M. U., α-Aminoxy Oligopeptides: Synthesis, Secondary Structure, and Cytotoxicity of a New Class of Anticancer Foldamers. *Chemistry-A European Journal* **2016,** *22* (49), 17600-17611.

P-39: Reverse Virtual Screening Procedure for Identifying the Target of an Antiplasmodial Hit Compound

Simone Brogi [1], Giulia Chemi [1], Stefania Butini [1], Margherita Brindisi [1], Giuseppe Campiani [1], Sandra Gemma [1], Soon Goo Lee [2], Joseph Jez [2]

[1]European Research Centre for Drug Discovery and Development (NatSynDrugs) and Department of Biotechnology, Chemistry and Pharmacy DdE 2018-2022, University of Siena, 53100 Siena, Italy, [2]Department of Biology, Washington University, St. Louis, USA

Phenotypic screening has become a crucial approach to discovery novel compounds, especially in the field of anti-cancer and anti-infective agents. Despite recent advances, phenotypic-driven target identification remains a challenging task that could be approached through several methodologies including biochemical methods, genetic interactions, computational approaches, or most likely by combining them. In the framework of antimalarials drug discovery in my research group, we focused our attention on the MMV Malaria Box compound MMV019918 (**1**, 1-[5-(4-bromo-2-chlorophenyl) furan-2-yl]-*N*-[(piperidin-4-yl)methyl]methanamine) for its dual activity against asexual stages and gametocytes. After a structure-activity relationship study based on phenotypic screening and cytotoxicity evaluation, we selected derivative **2** (1-(5-(2-phenyl-4-chlorophenyl)furan-2-yl)-*N*(piperidin-4-ylmethyl)methanamine) as a promising antimalarial agent. To further optimize derivative **2**, our next challenge was to understand its target and its mechanism of action. To this end, a computational effort for identifying the drug target of **2** is described herein. Hit compound **1** and its optimized analogue **2** were used in a reverse docking procedure, also known as reverse virtual screening or target fishing. Firstly, we retrieved from the PDB database all the proteins relevant for *Plasmodium* biology (108 crystal structures). Next, compounds **1** and **2** were used to "fish" potential targets, employing a high-throughput docking procedure[1] using Glide and Prime software[2] against the mentioned proteins after appropriate preparation. The *in silico* results indicated a small number of potential targets for **1** and **2** due to unfavorable docking scores and ligand binding energies found among the examined docking complexes. For one protein (phosphoethanolamine methyltransferase, *Pf*PMT) we noted for both compounds favorable *in silico* scores. Interestingly *Pf*PMT, an enzyme necessary for the phospholipid biosynthesis in all stages of the parasite life cycle, is a validated drug target for which few inhibitors, mainly related to the natural substrate (AdoMet), have been described so far.[3] Gratifyingly, enzymatic assays established that both **1** and **2** inhibited *Pf*PMT. Although **1** showed a weak inhibitory profile, compound **2** displayed a significant inhibition of *Pf*PMT (70% of residual activity of *Pf*PMT at 100 µM), so it was used as hit compound for developing a novel series of antiplasmodial and transmission-blocking agents. According to the composition of the binding site, a structure-based approach based on molecular docking and molecular dynamics[1] was performed, exploring different decorations of the hit molecule **2**, also considering the synthetic accessibility. With the introduction of a methoxy group on the phenyl ring we obtained one of the best performing molecules (**3**, (1-(5-(5-chloro-3'-methoxy-[1,1'-biphenyl]-2yl)furan-2-yl)-*N*-(piperidin-4-ylmethyl)methanamine)) in terms of computational scores, confirmed by *in vitro* tests against *Pf*PMT (**3**: 38.7% of residual activity of *Pf*PMT at 100 µM). Briefly, employing diverse computational methods we identified *Pf*PMT as drug target for the antiplasmodial hit compound **2**, found through a phenotypic screening, and we optimized its activity obtaining more potent inhibitors typified by **3**. The acquired knowledge about *Pf*PMT will allow the rational design of potent *Pf*PMT inhibitors for developing antimalarials with an innovative mechanism of action.

1. Brindisi, M.; *et al*. Structure-based discovery of the first non-covalent inhibitors of *Leishmania major* tryparedoxin peroxidase by high throughput docking. *Sci. Rep.* **2015**, 5, 9705
2. Glide, version 6.6; Prime, version 3.9; Desmond, version 4.1, Schrödinger, LLC, Release 2015
3. Lee, S. G.; *et al*. Structure and reaction mechanism of phosphoethanolamine methyltransferase from the malaria parasite Plasmodium falciparum: an antiparasitic drug target. *J. Biol. Chem.* **2012**, 287(2), 1426-134

P-41: Conformational Sampling and Binding Affinity Prediction of Macrocycles

Daniel Cappel [1]

[1] *Schrödinger GmbH, Q7 23, 68259 Mannheim, Germany*

When optimizing ligand binding to a target protein during the drug design process a macrocyclic structure of the ligand can provide advantages. Macrocyclisation is an effective way to restrict a compound's conformational space compared to acyclic inhibitors with the potential to improve potency, selectivity and metabolic stability.

In the context of computationally-driven drug design this diverse class of chemical structures provides some challenges when it comes to conformational flexibility. Here we will discuss a method for exploring macrocyclic conformational space and the results of a benchmarking study[1] for this algorithm. A dataset of 208 structures was curated from the Cambridge Structural Database, the Protein Data Bank and the Biologically Interesting Molecule Reference Dictionary. A conformational search algorithm using the program Prime reproduces the crystal structure conformations in a highly accurate way and is fast compared to other published approaches. The sampling algorithm is also used in the context of a membrane permeability prediction protocol for macrocyles.

Furthermore, results for binding affinity prediction using the FEP+ framework for macrocycles are presented.[2] We have applied the method to 7 pharmaceutically interesting data sets taken from recent drug discovery projects including 33 macrocyclic ligands covering a diverse chemical space. The predicted binding free energies are in excellent agreement with experimental data, with an overall root mean square error (RMSE) of the predictions below 1 kcal/mol.

1. Sindhikara, D.; Spronk, S. A.; Day, T.; Borelli, K.; Cheney, D. L.; Posy, S. L. Improving Accuracy, Diversity, and Speed with Prime Macrocycle Conformational Sampling. *J. Chem. Inf. Model.* **2017**, *57*, 1881-1894.
2. Yu, H. S.; Deng, Y.; Wu, Y.; Sindhikara, D.; Rask, A. R.; Kimura, T; Abel, R.; Wang, L. Accurate and Reliable Prediction of the Binding Affinities of Macrocycles to Their Protein Targets. *J. Chem. Theory Comput.* **2017**, *13*, 6290-6300.

P-43: Using FEP (Free Energy Perturbation) Calculations to estimate relative binding affinities and selectivity for GPCR targets

Francesca Deflorian, Benjamin G Tehan, Jonathan S Mason, and Miles Congreve

Heptares Therapeutics Ltd

G protein-coupled receptors (GPCRs) are the largest family of membrane proteins. GPCRs are involved in a wide variety of cellular functions, serving as key players in cellular signalling. Endogenous ligand binding from outside the cell leads to conformational changes of the receptor and consequential pairing with signalling partners in the intracellular environment, such as G proteins and β-arrestins, initiating signal transduction and other cellular responses. GPCRs mediate an abundant variety of physiological responses throughout the body representing pivotal candidate drug targets for the pharmaceutical industry.

The accurate estimation of protein-ligand binding free energy can be crucial at the lead generation and optimization stages of a drug discovery program. Free energy calculations take into account protein flexibility and system solvation, both very important aspects in the ligand binding process for GPCRs. In this poster we present the results from FEP studies using FEP+ in collaboration with Schrodinger on GPCR targets offering different challenges to the methodology.

The Orexin receptors have binding sites located in the TM domain and water-bridging interactions are crucial for ligand binding. We were also interested in selectivity between the subtypes OX1 and OX2 receptors with high sequence similarity in the ligand binding region. Using X-ray structures solved using the Heptares StaR® technology of the OX1 receptor in complex with several ligands, we conducted both retrospective and prospective FEP+ studies with the aim to design selective OX1 antagonists.

The second target was the calcitonin gene-related peptide (CGRP) receptor, a receptor with a very shallow and solvent exposed binding site for non-peptide ligands. In this particular study, the ligands were small molecule antagonists and in-house crystallographic structures of the CGRP binding site were used to support the program.

P-45: Can I Have Seconds?

T. Brinkjost[1,2], C. Ehrt[1,2] , P. Mutzel[1], O. Koch[2]

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

The automated assignment of secondary structure elements has a long-term history since their first discovery. A reliable and in particular consistent assignment is of utmost importance for a multitude of applications in structure-based drug design and function elucidation such as protein structure alignment[1], polypharmacology and conserved motifs[2], secondary structure prediction[3], or the concept of ligand-sensing cores[4]. A variety of different approaches have been developed that rely on either hydrogen bond criteria, geometrical characteristics or a combination of both to allow for a reliable assignment.[5]

An alternative approach, SHAFT[6], was dedicated toward the assignment of helices based on a structural classification of different turn types. Herein we will present SCOT (Secondary structure Classification On Turns) which optimizes and extends this basic idea to assign various helix types and sheets via a combination of hydrogen bond and geometric characteristics.

Our novel method SCOT utilizes a hierarchical assignment of protein structural elements. Starting from the initial level of turn types, we identify right handed alpha-, 3_{10}-, pi-, and gamma-helices as well their left-handed counterparts for the first two types. Sheets are identified in seed/non-turn regions which interact with each other. Furthermore, both, helices and sheets, are annotated with additional kink information. As for helices, kinks are also annotated with an individual classification reflecting the classes of the incident helix segments. This information will additionally help to compare proteins on the secondary structure level, as in most cases the conformation of helices and sheets highly deviates from its ideal one.

Summarizing, with a classification for turns, helices and sheets, SCOT fulfills all basic needs for a reliable protein structure classification which will hopefully be of high interest in structure-based design and protein engineering.

1. Ma, J.; Wang, S. Algorithms, applications, and challenges of protein structure alignment. *Adv. Protein Chem. Struct. Biol.* **2014**, *94*, 121-175
2. Koch, O. The Use of Secondary Structure Element Information in Drug Design: Polypharmacology and Conserved Motifs in Protein-Ligand Binding and Protein-Protein Interfaces. *Future Med. Chem*. **2011**, *3(6)*, 699-708.
3. *Yang, Y.; Gao, J.; Wang, J.; Heffernan, R.; Hanson, J.; Paliwal, K.; Zhou, Y. Sixty-five years of the long march in protein secondary structure prediction: the final stretch? Briefings Bioinf.* ***2016****, bbw129.*
4. Koch, M. A.; Waldmann, H. Protein structure similarity clustering and natural product structure as guiding principles in drug discovery. *Drug Discov. Today* **2005**, *10(7)*, 471-483.
5. Tyagi, M.; Bornot, A.; Offmann, B.; de Brevern, A. G. Analysis of loop boundaries using different local structure assignment methods *Protein Sci.* **2009**, *18(9)*, 1869-1881.
6. Koch, O.; Cole, J. An automated method for consistent helix assignment using turn information. *Proteins* **2011**, *79(5)*, 1416-1426.

P-47: Virtual Screening of CCR5 Inhibitors as Potential Anti- Colorectal Cancer Agents

M. El-Zohairy [1], Y. Mandour[1], H.Adwan[2,] D. Zolotos[1]

[1] Department of Pharmaceutical Chemistry, Faculty of Pharmacy and Biotechnology, The German University in Cairo, New Cairo, Egypt, [2] Department of Pharmacology and Toxicology, Faculty of Pharmacy and Biotechnology, The German University in Cairo, New Cairo, Egypt

CCR5 is G-protein-coupled receptor “GPCR” with seven transmembrane loops. This chemokine receptor interacts with the corresponding ligand leads to subsequent downstream signaling. CCR5 is commonly expressed by T lymphocytes, which acts as a co-receptor in the most commonly transmitted Human immunodeficiency virus “HIV”, for their entry to host cell. (1)

Recently, it was observed that CCR5 receptor is highly expressed on tumor cells in liver metastatic colorectal cancer. Inhibition of this receptor in the patients treated with Maraviroc a CCR5 inhibitor. Showed a decrease in growth signals for tumor cells and resulting in slowing down of tumor development. Through interfering with CCL5-CCR5 axis. (2) (3) (4) (5)

Our aim is to use virtual screening to detect new potential CCR5 inhibitors that will be more active as anti-colorectal cancer agents.

A Pharmacophore model for CCR5 inhibitors was generated by clustering of CCR5 binding database composed of 2827 Compounds based on their scaffolds and their finger print similarity. Then the most active representative from each scaffold was selected. The 39 selected representatives were aligned on the bioactive conformer of Maraviroc, which is obtained from the co-crystalized structure of Maraviroc bound to CCR5 receptor, PDB code "4MBS". After alignment this selected representatives generates a Pharmacophore model using MOE. The generated model was in consensus with the reported pharmacophoric features and point mutations of the receptor. The model was validated using a test set composed of 1255 compounds. 1160 compounds are actives and 95 compounds are in actives and decoys. This model was further used in virtual screening for potential CCR5 inhibitors.

1. Oppermann, M. (2004). Chemokine receptor CCR5: insights into structure, function, and regulation. Cellular Signalling, 16(11), 1201-1210. doi:10.1016/j.cellsig.2004.04.007
2. Bronte, V., & Bria, E. (2016). Interfering with CCL5/CCR5 at the Tumor-Stroma Interface. Cancer Cell, 29(4), 437-439. doi:10.1016/j.ccell.2016.03.019
3. Deming, D. A. (2016). Advances in immunotherapeutic strategies for colorectal cancer commentary on: tumoral immune cell exploitation in colorectal cancer metastases can be targeted effectively by anti-CCR5 therapy in cancer patients by Halama et al. Journal for ImmunoTherapy of Cancer, 4(1). doi:10.1186/s40425-016-0197-y
4. Kuritzkes, D., Kar, S., & Kirkpatrick, P. (2008). Maraviroc. Nature Reviews Drug Discovery, 7(15), 15-16. doi:10.1038/nrd2490
5. Pervaiz, A., Ansari, S., Berger, M. R., & Adwan, H. (2015). CCR5 blockage by maraviroc induces cytotoxic and apoptotic effects in colorectal cancer cells. Medical Oncology, 32(5). doi:10.1007/s12032-015-0607-x

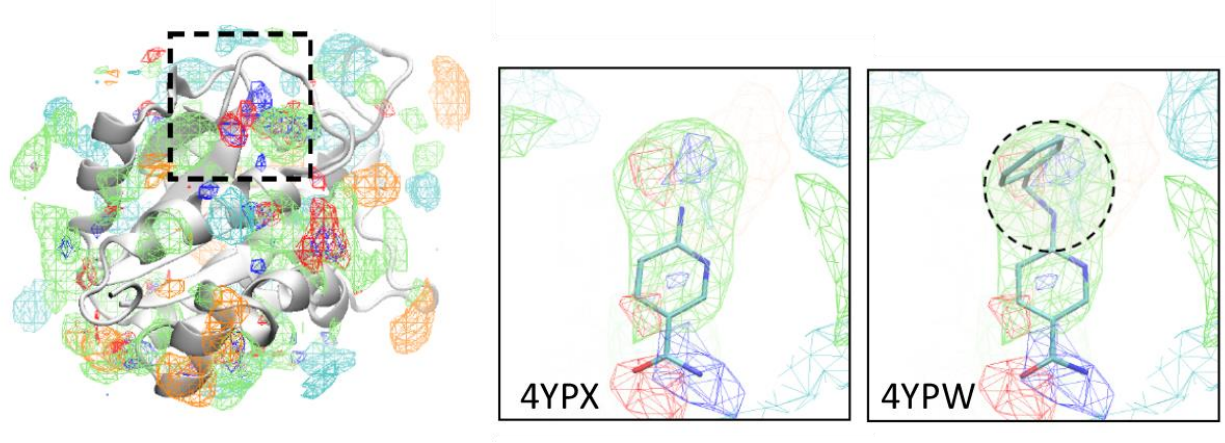
P-49: SILCS reproduces experimental binding trends for 31 TrmD ligands

SK Lakkaraju [1], O Guvench [1], S Jo[1], AD MacKerell, Jr. [1,2]

[1] SilcsBio, LLC, Baltimore, MD, USA, [2] University of Maryland, School of Pharmacy, Baltimore, MD, USA

Site-Identification by Ligand Competitive Saturation (SILCS)[1-4] computational functional group mapping provides insights into the binding preferences of a target protein that can be used both qualitatively and quantitatively to drive ligand design. SILCS is a robust structure-based approach that gives information-rich Grid Free Energy (GFE) FragMaps that account for critical aspects such as protein flexibility, desolvation penalties, as well as protein-functional group interactions.

Here we describe the use of the SILCS approach on t-RNA methyltransferase (TrmD) and 31 ligands belonging to two series made publicly available through the Community Structure-Activity Resource (CSAR) and the D3R Database. SILCS-MC sampling of ligands in the field of the FragMaps yields Ligand Grid Free Energy (LGFE) scores. SILCS scoring correctly predicts favorable vs. unfavorable modifications relative to a reference ligand (27/30 predictions correct). Additionally, SILCS FragMaps recapitulate functional group patterns of both series of ligands. This information can be used to drive design and optimization visually.



1. Guvench, O. and MacKerell, A.D. Jr. Computational Fragment-Based Binding Site Identification by Ligand Competitive Saturation. *PLoS Comput. Biol.* **2009**, 5, e1000435.
2. Raman, E.P., Yu, W., Guvench, O. and MacKerell, A.D., Jr., Reproducing Crystal Binding Modes of Ligand Functional Groups using Site-Identification by Ligand Competitive Saturation (SILCS) Simulations. *J. Chem. Inf. Model.* **2011**, 51, 877-890.
3. Lakkaraju, S.K., Raman, E.P., Yu, W., and MacKerell, A.D., Jr. Sampling of Organic Solutes in Aqueous and Heterogeneous Environments using Oscillating μ_{ex} Grand Canonical-like Monte Carlo-Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2014**, 10, 2281-2290.
4. Raman, E.P., Yu, W., Lakkaraju, S.K., and MacKerell, A.D., Jr. Inclusion of multiple fragment types in the Site Identification by Ligand Competitive Saturation (SILCS) approach. *J. Chem. Inf. Model.* **2013**, 53, 3384–3398.

P-51: Fuzzy ligands for allosteric target detection and lead identification

S. M. A. Hermans [1], C. Pfleger [1], D. Schmidt [1], M. Boehm [2], A. M. Mathiowetz [2], C. L. McClendon [2], K. Omoto [2], H. Gohlke [1]*

*[1] Department of Mathematics and Natural Sciences, Institute for Pharmaceutical and Medicinal Chemistry, Heinrich Heine University Düsseldorf, Düsseldorf, Germany, [2] Medicine Design, Pfizer Inc., 1 Portland Street, Cambridge, Massachusetts 02139, United States, *Email: gohlke@uni-duesseldorf.de*

Targeting allosteric regulation in biomolecules is a promising strategy in drug discovery, due to advantages over conventional orthosteric ligands.[1] However, the identification of novel allosteric pockets is complicated by the variety of allosteric mechanisms, differing by the extent of conformational change upon ligand binding. Particularly, dynamic allostery, which can occur in the absence of conformational change,[2] is difficult to detect from static crystal structures alone. Here, we developed an approach for generating fuzzy ligands as surrogates for "true" ligands, with which allosteric responses can be calculated by rigidity analysis.[3]

The performance of the fuzzy ligand approach was applied to 85 protein-ligand complexes.[4] For the *apo* states, generated by removing the original ligand from the protein, pockets were identified using PocketAnalyzer[PCA].[5] DrugScore pair-potentials[6] were calculated for each of the binding pockets, and used as input to guide the design of fuzzy ligands. The allosteric transmission caused by (fuzzy) ligand binding was determined by an ensemble-based perturbation approach that analyses biomolecular rigidity.[3] The fuzzy ligands were validated I) in terms of their influence on biomolecular rigidity compared to the "true" ligand and II) to what extent pharmacophore models based on fuzzy ligands allow for a successful identification of binders and non-binders in a retrospective virtual screening. Altered per-residue stability characteristics from rigidity analysis of our fuzzy ligands are in agreement with those from "true" ligands. The virtual screening results based on fuzzy ligands perform equally well or outperform the true ligands' results.

Analyzing unexplored pockets with our fuzzy ligand approach predicts whether binding a ligand to this pocket triggers an allosteric response to affect biomolecular function. If an allosteric response is present, the fuzzy ligand can be used for virtual screening to directly identify lead compounds for the identified target. The fuzzy ligand approach can thus be a promising step towards identifying novel allosteric drug targets and drugs.

1. Nussinov, R.; Tsai, C.J. Allostery in disease and in drug discovery. *Cell*, **2013**, *153*, 293-305
2. Cooper, A.; Dryden, D.T., Allostery without conformational change. A plausible model. *Eur. Biophys. J.*, **1984**, *11*, 103-109
3. Pfleger, C.; Minges, A.; Boehm, M.; McClendon, C. L.; Torella, R.; Gohlke, H. Ensemble- and Rigidity Theory-Based Perturbation Approach To Analyze Dynamic Allostery *J. Chem. Theory. Comput.*, **2017**, *13*, 6343-6357
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
5. Craig, I.R. ; Pfleger, C.; Gohlke, H.; Essex, J.W.; Spiegel, K. Pocket-space maps to identify novel binding-site conformations in proteins. *J. Chem. Inf. Model.*, **2011**, *51*, 2666-2679
6. Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.*, **2000**, *295*, 337-356

P-53: A fast and efficient rescoring method based on binding information of fragment and drug-like ligands

Célien Jacquemard[1], Malgorzata N. Drwal[1], Carlos Perez[3], Jérémy Desaphy[2], Esther Kellenberger[1]

[1]Laboratoire d'innovation thérapeutique, UMR7200 CNRS Université de Strasbourg, 67400 Illkirch, France, [2]Lilly Research Laboratories, Eli Lilly and Company, Indianapolis, IN 46285, USA, [3]Lilly Research Laboratories, Eli Lilly and Company, 28108 Alcobendas, Madrid, Spain

Protein structure-based computing approach to hit finding in Fragment-based drug design (FBDD) is not yet a reliable alternative to experiments, mostly because of our incomplete understanding of molecular interactions. Recently, we analyzed the binding modes of fragments and drug-like ligands bound to four diverse targets in the Protein Data Bank (PDB), and found that the two classes of compounds binding to the same cavity tend to have comparable interaction patterns.[1] Here we ask whether the binding mode information of fragments can improve the performance of molecular docking of drug-like ligands and *vice versa*. Our study compares two rescoring methods: the max GRIM method, which encodes reference ligand-protein interactions in individual graphs[3] and the new LID method based on a consensus 3D-density map built from the reference interactions.

Material and methods. Docking was performed using PLANTS.[2] Poses were scored using ChemPLP and rescored by similarity to interaction patterns found in the reference PDB complexes. The reference dataset includes 2702 crystallographic structures and describes 66 proteins, 727 drug-like ligands and 964 fragments. Each protein is represented by at least three 3D-structures of complexes with at least one drug-like ligand and one fragment.

Results (1) Pose selection. For all the compounds, we performed all possible cross-docking experiments and observed that drug-like ligand binding information always improved fragment docking, but the opposite was only true for difficult targets. Combining the binding information of drug-like ligands and fragments was the most robust rescoring option. *(2) Virtual screening.* We evaluated GRIM and LID performances in compound ranking using the DUD-e benchmark available for six of the 66 proteins in the reference set. The two rescoring methods better discriminates active compounds from decoys than the native scoring function.

Conclusion. GRIM and LID methods equally well improved docking predictions. LID is 100 times faster than GRIM thereby allowing large-scale calculations, such as the rescoring of multiples poses obtained for a ligand docked into multiple structures of the protein. Building of LID consensus 3D-density map however implies that all 3D-structures of the reference complexes are well 3D-aligned.

1. Drwal, M.; Jacquemard, C.; Perez, C.; Desaphy, J.; Kellenberger, E. Do Fragments And Crystallization Additives Bind Similarly To Drug-Like Ligands?. *Journal of Chemical Information and Modeling* **2017**, 57, 1197-1209.
2. Korb, O.; Stützle, T.; Exner, T. Empirical Scoring Functions For Advanced Protein−Ligand Docking With PLANTS. *Journal of Chemical Information and Modeling* **2009**, 49, 84-96.
3. Desaphy, J.; Raimbaud, E.; Ducrot, P.; Rognan, D. Encoding Protein–Ligand Interaction Patterns In Fingerprints And Graphs. *Journal of Chemical Information and Modeling* **2013**, 53, 623-637.
4. Mysinger, M.; Carchia, M.; Irwin, J.; Shoichet, B. Directory Of Useful Decoys, Enhanced (DUD-E): Better Ligands And Decoys For Better Benchmarking. *Journal of Medicinal Chemistry* **2012**, 55, 6582-6594.

P-55: Mapping Binding Site Thermodynamics by 3D RISM Theory for Drug Design

Julia Jasper, Yannic Alber, Florian Mrugalla, Stefan Kast, Oliver Koch

Faculty of Chemistry and Chemical Biology, TU Dortmund, Dortmund, Germany

The early stages of the drug discovery process require reasonably accurate and fast methods for optimising the binding affinity of protein-ligand complexes, taking into account direct and solvent-mediated interactions. Inspired by Goodford's GRID method[1] we here present a novel physics-based approach that incorporates (de-)solvation contributions to the binding thermodynamics of probe particles mimicking functional ligand groups in a protein binding site. To this end, we calculate the potential of mean force (PMF) and the distribution functions of different probes (uncharged C, charged N and O) inside the *apo* protein by 3D RISM (reference interaction site model) theory[2,3].

The method allows for an intuitive and easy visualization of probe density maps inside the binding site (Fig. 1) and can be exploited for various tasks in the drug development process. Applications range from pharmacophore and docking-based virtual screening up to defining design directions for medicinal chemists. In a first proof of concept study, the PMF results were embedded into the GOLD[4] docking process on a subset of the PDBbind dataset[5]. An uncharged C probe is used to calculate hydrophobic fitting points that are used for ligand placement throughout the docking process. These 3D RISM based points display a more detailed representation of hydrophobicity yielding improved docking success (Fig. 2).

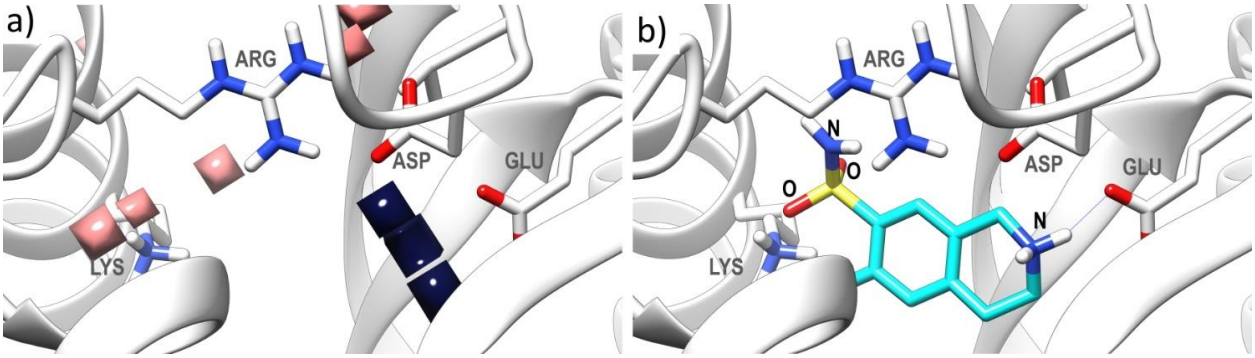


Figure 3: a) Density maps for charged N (dark volumes) and O (light volumes) probes calculated for the *apo* protein structure of 1hnn@pdb; b) binding site of 1hnn@pdb with the co-crystallized ligand.

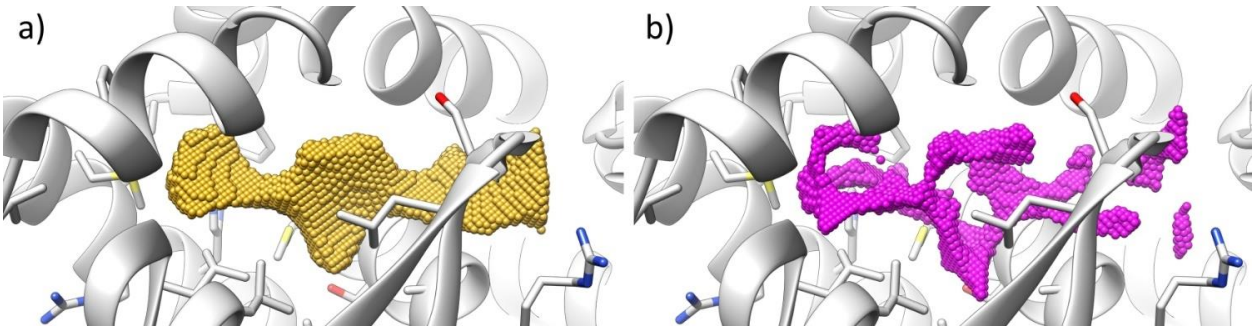


Figure 4: Fitting points for 1nav@pdb as calculated by a) GOLD (based on the van der Waals interaction energy between a bare C atom and the protein) and by b) our RISM based approach (uncharged C probe with a PMF threshold of -7.5 kJ/mol).

1. Goodford, P. J. A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *J. Med. Chem.* **1985**, 28, 849–857.
2. Mrugalla, F.; Kast, S. M. Designing Molecular Complexes Using Free-Energy Derivatives from Liquid-State Integral Equation Theory. *J. Phys. Condens. Matter* **2016**, 28, 344004.
3. Güssregen, S.; Matter, H.; Hessler, G.; Lionta, E.; Heil, J.; Kast, S. M. Thermodynamic Characterization of Hydration Sites from Integral Equation-Derived Free Energy Densities: Application to Protein Binding Sites and Ligand Series. *J. Chem. Inf. Model.* **2017**, 57, 1652-1666.
4. 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.

5. Li, Y.; Liu, Z.; Li, J.; Han, L.; Liu, J.; Zhao, Z.; Wang, R. Comparative Assessment of Scoring Functions on an Updated Benchmark: 1. Compilation of the Test Set. *J. Chem. Inf. Model.* **2014**, 54, 1700–1716.

P-57: Structure based design of potent and selective ligands for the adenosine receptor family

W. Jespers [1, 2], G. van Westen [2], R. Cooke[3], J. Mason[3], A. IJzerman[2], L. Heitman[2], J. Azuaje[4], J. Aqvist[1], E. Sotelo[4], H. Gutierrez-de-Teran[1]

[1] Uppsala University, Uppsala, Sweden, [2] Leiden University, Leiden, the Netherlands, [3] Heptares Ltd., Hertfordshire, UK, [4] Universidade de Santiago de Compostela, Santiago de Compostela, Spain

The four adenosine receptors (ARs), A_1, A_{2A}, A_{2B}, and A_3, constitute a subfamily of G protein-coupled receptors (GPCRs) with exceptional foundations for structure-based ligand design.[1] Recent advances in membrane protein engineering and crystallography have sparked a surge of experimental GPCR structures. Among these structures, ARs have emerged as one of the most thoroughly characterized families with A_1AR and A_{2A}AR inactive structures, and active structures of the latter.[2]

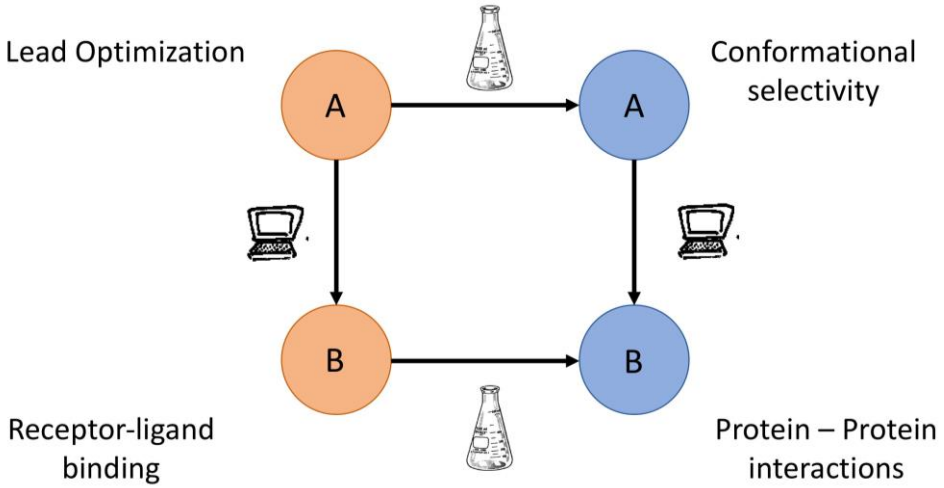


Figure 1: Thermodynamic cycle illustrating the calculation of the relative shift in binding free energy (vertical legs), which theoretically matches the data from experiments (horizontal legs) between two systems A and B. These can be two ligands binding to the same receptor or the affinity of one ligand for two constructs of the receptor (i.e. WT and single-point mutant), active and inactive receptor states or protein-protein interactions such as GPCR-G protein.

I will present results of our AR ligand-design program, where we combine advanced structure-based computational methods with efficient synthetic approaches (see Figure 1).[3] Particular emphasis is put on the development and application of free energy perturbation (FEP) protocols to modulate binding affinity, receptor selectivity and pharmacological profiles for our ligand series. With these protocols, we have recently provided a detailed understanding of the effects of point mutations on ligand binding on the A_{2A}AR[3] and A_1AR[4], the conformational preference of partial agonists of the A_{2A}AR[4], and assisted on the design of pyridines as a novel chemical structure for A_3AR antagonists from our previous series of pyridines.[5] Currently, we are investigating the role of mutations in the G protein binding site in receptor activation and their role in breast cancer. Additionally, we are designing the first non-ribose agonists for the A_{2B} receptor. Finally, we apply our protocol to understand the binding mode of a series of chromones for the A_{2A}AR, a method we recently applied to predict the binding mode of orphan GPCR GPR139 agonists.[6]

1. Gutiérrez-de-Terán, H. et al. *Curr. Top. Med. Chem.* **2017**, *17* (1), 40–58.
2. Jespers, W. et al. *Trends in Pharmacological Sciences*. Elsevier Ltd 2018, pp 75–89.
3. Vasile, S. et al. Humana Press, New York, NY, 2018; pp 23–44.
4. Jespers, W. et al. *Molecules* **2017**, *22* (11).
5. Azuaje, J.; Jespers, W. et al. *J. Med. Chem.* **2017**, *60* (17), 7502–7511.
6. Nøhr, A. C.; Jespers, W.; Shehata, M. A. et al. *Sci. Rep.* **2017**, *7* (1).

P-59: Transferable Neural Networks Architecture for Low Data Drug Discovery

Mun-Hwan Lee [1], Eung-Hee Kim, Ph.D. [2], Yong-Ju Lee[1], Hong-Gee Kim, Ph.D. [1]

[1] *Biomedical Knowledge Engineering Lab., Seoul National University, Seoul, Republic of Korea,* [2]*Dept. of Global Software Engineering., Sunmoon University, Asan-si, Republic of Korea*

Introduction. Machine learning (ML)'s advances in virtual screening (VS) has made significant contributions to drug discovery accompanied by novel approaches based on deep neural networks (DNN). However, such techniques require a considerable volume of both training and test data, in order to achieve comparable results. While certain models have managed to learn from small datasets[1], they optimize only within the defined range of data which is nontransferable. Thus, the current limitation in ML models necessitates the need of a universal architecture for generalization regardless of dataset size. In this study, we propose a transferable DNN to overcome such limitations in generalizing ability by incorporating binary response (positive or negative) rather than non-fixed output dimensions depending on a dataset. We demonstrate that the proposed architecture is capable to learn transferable information between varied datasets.

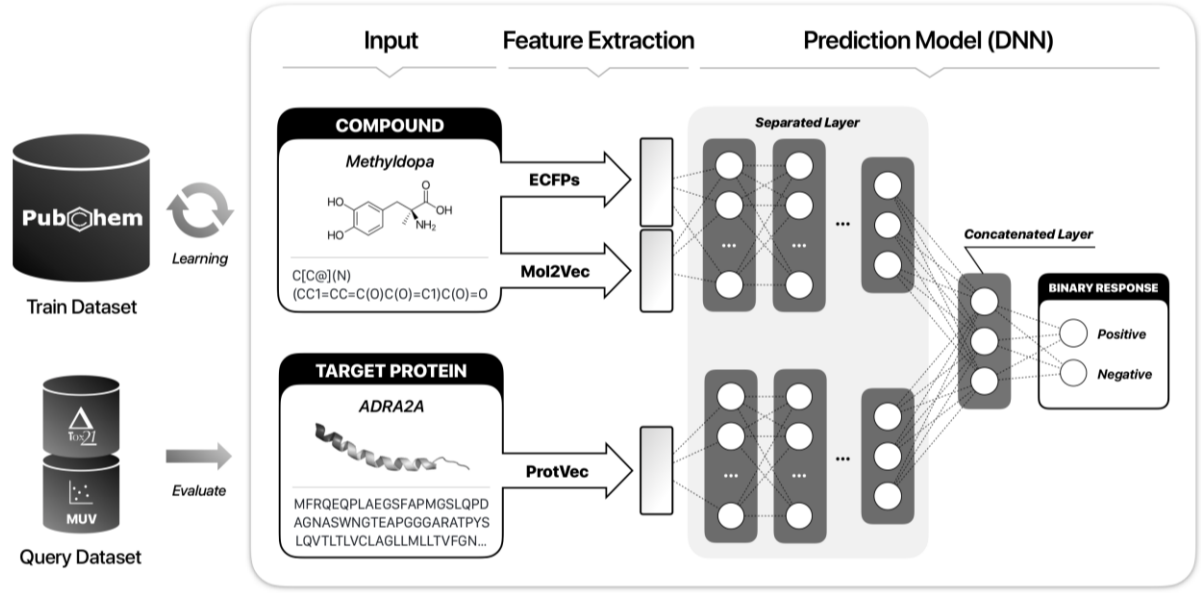


Figure 1. Overview of transferable DNN

Method. The proposed model consists of feature extraction and prediction model by implementing proteochemometric (PCM) approaches, which utilize the additional use of protein information as shown in Figure 1. First, Extended-Connectivity Fingerprints (ECFP) and Mol2Vec were exploited for compound feature extraction and ProtVec for target protein. The dataset comprised both positive and negative examples for each feature pairs, which can be mathematically defined as

$$\mathcal{D} = \left\{ \left[\text{Feature}_{\text{compound}_i}, \text{Feature}_{\text{protein}_i} \right], y_i \right\}_{i=1}^{|\mathcal{D}|}, y_i \in \{0,1\}.$$

Second, separated layers pass the paired data into concatenated layer for classification. By adjusting the dimensions of hidden nodes through separated layers, the model can prevent a disproportions between feature dimensions (e.g. between 2048-dimension for compounds and 256-dimesion for targets), and avoid bias.

Experiment and Result. The proposed models were trained on PubChem BioAssay (PCBA), and evaluated on Maximum Unbiased Validation (MUV) and Tox21 as a query dataset. To justify the generalization ability of the model, we removed overlapping data from evaluation dataset. In order to avoid learning bias in skewed distribution, the same number of positive and negative pairs were extracted in each compound. This resulted in 361,632 positive examples and 361,632 negative examples. Our proposed networks consist of 5 separated layers and 1 concatenated layer. For training the networks we used stochastic gradient descent with Adaptive Moment Estimation (Adam). We used 50% of dropout on the hidden layers to prevent overfitting of the networks. The proposed model shows better performance over the random-forest baseline as shown in Table 1, Table 2, and Table 3.

Table 1. ROC-AUC Scores of Models on Median Held-out Task for Each Model on Datasets [a]

Models	MUV	Tox21
Proposed Model	0.730 ± 0.079	0.772 ± 0.067
RF (100 trees)	0.661 ± 0.081	0.709 ± 0.100

[a] Numbers reported are medians and standard deviations.

Table 2. ROC-AUC Scores of Models on Each Tasks on MUV

Models	MUV-466	MUV-548	MUV-600	MUV-644	MUV-652	MUV-689	MUV-692	MUV-712	MUV-713	MUV-733	MUV-737	MUV-810	MUV-832	MUV-846	MUV-852	MUV-858	MUV-859
Proposed Model	**0.755**	**0.684**	**0.777**	**0.649**	**0.840**	**0.833**	**0.782**	**0.869**	**0.716**	**0.823**	**0.730**	**0.730**	**0.689**	**0.617**	**0.681**	**0.615**	**0.657**
RF (100 trees)	0.735	0.667	0.576	0.548	0.776	0.768	0.635	0.780	0.630	0.699	0.723	0.702	0.661	0.563	0.622	0.610	0.528

Table 3. ROC-AUC Scores of Models on Each Tasks on Tox21

Models	NR-AR	NR-AR-LBD	NR-AhR	NR-Aromatase	NR-ER	NR-ER-LBD	NR-PPAR-gamma	SR-ARE	SR-ATAD5	SR-HSE	SR-MMP	SR-p53
Proposed Model	**0.636**	**0.762**	**0.741**	**0.782**	**0.628**	**0.714**	**0.721**	**0.783**	**0.800**	**0.807**	**0.826**	**0.829**
RF (100 trees)	0.461	0.528	0.666	0.763	0.583	0.641	0.701	0.749	0.732	0.718	0.746	0.772

Conclusion. We propose a transferable DNN to improve a learning model by transferring information from abundant datasets to small datasets. This allows generalization ability for ML models with various feature extraction methods in a scalable way. We also demonstrated that the model learns transferable ability for various size of datasets especially in small one. In the future work, we are planning to augment transferability to multitask deep learning for robust performance by adopting conventional transfer learning approaches.

Acknowledgements: This work was partly supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIP) (No.2017-0-00398, Development of drug discovery software based on big data) and the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT and future Planning (No. NRF-2017R1A4A1014584, Epigenetic Regulation of Bone & Muscle Regeneration Lab)

1. Altae-Tran, H.; Ramsundar, B.; Pappu, A. S.; Pande, V. Low Data Drug Discovery with One-Shot Learning *ACS Cent. Sci.* **2017**, 3, 283– 293
2. Jaeger, S.; Fulle, S.; Turk, S. Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. *J. Chem. Inf. Model.*, **2018**, 58 (1), pp 27–35.
3. Asgari, E.; Mofrad, M. R. K. Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. *PLoS One* **2015**, 10, e0141287

P-61: Tetris of HDAC Inhibitor Design

J. Melesina[1], T. Heimburg[1], T. Bayer[1], E. Ghazy[1], M. Marek [2], P. Zeyen[1], K. Schmidtkunz[3], D. Robaa[1], R. Pierce[4], C. Romier[2], M. Jung[3], W. Sippl [1]

[1] Institute of Pharmacy, Martin Luther University Halle-Wittenberg, Halle, Germany, [2] Département de Biologie Structurale Intégrative, IGBMC, Université de Strasbourg, CNRS, INSERM, Illkirch Cedex, France, [3]Institute of Pharmaceutical Sciences, University of Freiburg, Freiburg, Germany, [4]University of Lille, CNRS, INSERM, Institute Pasteur de Lille, U1019 - UMR 8204 - CIIL - Centre d'Infection et d'Immunité de Lille, Lille, France

What is common between the puzzle game Tetris and design of histone deacetylase (HDAC) inhibitors? After several years of research dedicated to anti-parasitic and anti-cancer drug development it became clear that in both cases it is all about the shape. Just like Tetris tiles, HDAC inhibitors have specific shapes, which determine their selectivity (Fig. 1). Just like Tetris tiles, HDAC inhibitors fit to the cavity if their shape is complementary to it. And if not, then the unoccupied cavities and unemployed opportunities for drug design are left.

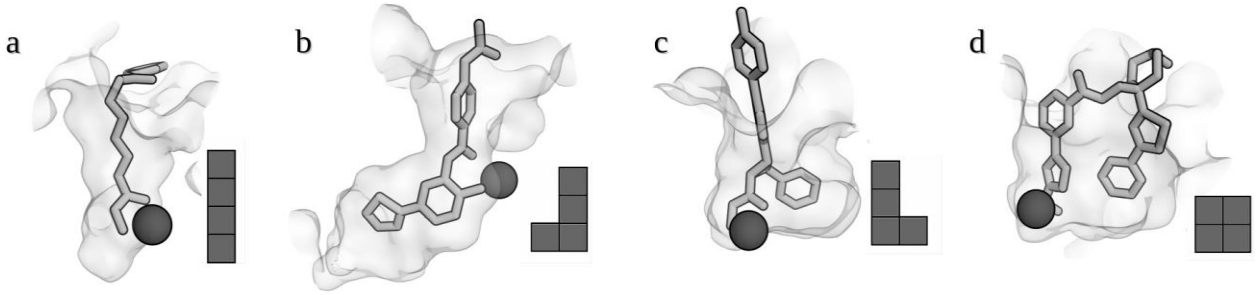


Figure 1. Shapes and binding modes of HDAC inhibitors: a) I-shaped inhibitor **SAHA**, PDB ID 4LXZ[1]; b) J-shaped inhibitor **20Y**, PDB ID 4LY1[1]; c) L-shaped inhibitor **9F4**, PDB ID 4CBT[2]; d) Oshaped inhibitor **NU7**, PDB ID 3ZNS[3].

Our quest for novel HDAC inhibitors started with virtual screening campaign on a newly validated anti-parasitic target *Schistosoma mansoni* HDAC8 (SmHDAC8). Around 15 million compounds have been screened *in silico* to find first SmHDAC8 inhibitors (Fig. 2a)[4]. One of the virtual screening hits - a fragment-sized molecule **J1038** (Fig. 2b) - has been chosen for further optimization. An open-ring Γ-shaped analog **UV4** was designed (Fig. 2c), which targeted HDAC8specific side pocket and unique SmHDAC8 amino acid residue H292. Its crystal structure with the target protein confirmed the predicted binding mode (Fig. 2d)[5].

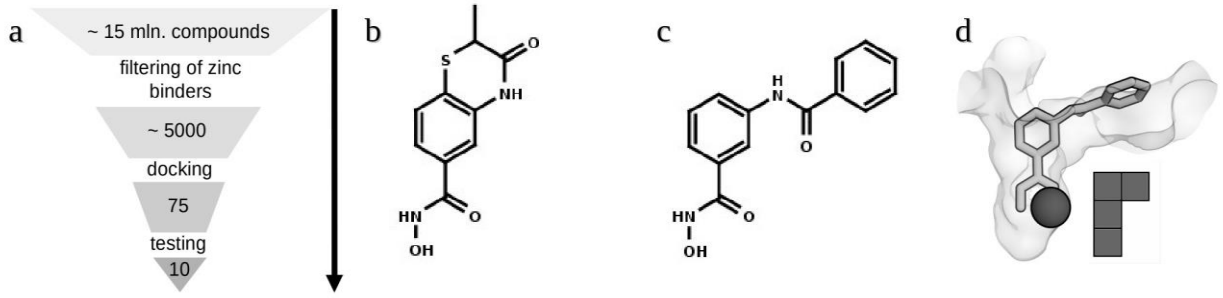


Figure 2. Design of novel Γ-shaped HDAC inhibitors: a) virtual screening workflow used to find first SmHDAC8 inhibitors[4] ; b) chemical structure of virtual screening hit **J1038**[4] ; c) chemical structure of optimized Γ-shaped inhibitor **UV4**[5]; d) binding mode of Γ-shaped inhibitor **UV4**, PDB ID 5FUE[5].

A library of further Γ-shaped ligands has been designed and docked to different HDACs. The most promising candidates were synthesized and tested. Their potency and selectivity has been optimized[5].

Homology models of various parasitic HDACs have been prepared and analyzed to see if they could also accept Γ-shaped inhibitors[6]. Furthermore, a series of Γ-shaped inhibitors were developed as antineuroblastoma agents[7]. All in all, our computer-aided molecular design approach yielded Γ-shaped HDAC inhibitors, which proved to be promising drug candidates.

1. Lauffer, B. E.; Mintzer, R.; Fong, R.; Mukund, S.; Tam, C.; Zilberleyb, I.; Flicke, B.; Ritscher, A.; Fedorowicz, G.; Vallero, R.; Ortwine, D. F.; Gunzner, J.; Modrusan, Z.; Neumann, L.; Koth, C. M.; Lupardus, P. J.; Kaminker, J. S.; Heise, C. E.; Steiner, P. Histone Deacetylase (HDAC) Inhibitor Kinetic Rate Constants Correlate with Cellular Histone Acetylation but not Transcription and Cell Viability. J. Biol. Chem. **2013**, 288, 26926-26943.
2. Bürli, R. W.; Luckhurst, C. A.; Aziz, O.; Matthews, K. L.; Yates, D.; Lyons, K. A.; Beconi, M.; McAllister, G.; Breccia, P.; Stott, A. J.; Penrose, S. D.; Wall, M.; Lamers, M.; Leonard, P.; Müller, I.; Richardson, C. M.; Jarvis, R.; Stones, L.; Hughes, S.; Wishart, G.; Haughan, A. F.; O'Connell, C.; Mead, T.; McNeil, H.; Vann, J.; Mangette, J.; Maillard, M.; Beaumont, V.; Munoz-Sanjuan, I.; Dominguez, C. Design, Synthesis, and Biological Evaluation of Potent and Selective Class IIa Histone Deacetylase (HDAC) Inhibitors as a Potential Therapy for Huntington's Disease. J. Med. Chem. **2013**, 56, 9934-9954.
3. Lobera, M.; Madauss, K. P.; Pohlhaus, D. T.; Wright, Q. G.; Trocha, M.; Schmidt, D. R.; Baloglu, E.; Trump, R. P.; Head, M. S.; Hofmann, G. A.; Murray-Thompson, M.; Schwartz, B.; Chakravorty, S.; Wu, Z.; Mander, P. K.; Kruidenier, L.; Reid, R. A.; Burkhart, W.; Turunen, B. J.; Rong, J. X.; Wagner, C.; Moyer, M. B.; Wells, C.; Hong, X.; Moore, J. T.; Williams, J. D.; Soler, D.; Ghosh, S.; Nolan, M. A. Selective Class IIa Histone Deacetylase Inhibition via a Nonchelating Zinc-Binding Group. Nat. Chem. Biol. **2013**, 9, 319-325.
4. Kannan, S.; Melesina, J.; Hauser, A. T.; Chakrabarti, A.; Heimburg, T.; Schmidtkunz, K.; Walter, A.; Marek, M.; Pierce, R. J.; Romier, C.; Jung, M.; Sippl, W. Discovery of Inhibitors of Schistosoma mansoni HDAC8 by Combining Homology Modeling, Virtual Screening, and in vitro Validation. J. Chem. Inf. Model. **2014**, 54, 3005-3019.
5. Heimburg, T.; Chakrabarti, A.; Lancelot, J.; Marek, M.; Melesina, J.; Hauser, A. T.; Shaik, T. B.; Duclaud, S.; Robaa, D.; Erdmann, F.; Schmidt, M.; Romier, C.; Pierce, R. J.; Jung, M.; Sippl, W. Structure-Based Design and Synthesis of Novel Inhibitors Targeting HDAC8 from Schistosoma mansoni for the Treatment of Schistosomiasis. J. Med. Chem. **2016**, 59, 24232435.
6. Melesina, J.; Robaa, D.; Pierce, R. J.; Romier, C.; Sippl, W. Homology Modeling of Parasite Histone Deacetylases to Guide the Structure-Based Design of Selective Inhibitors. J. Mol. Graph. Model. **2015**, 62, 342-361.
7. Heimburg, T.; Kolbinger, F. R.; Zeyen, P.; Ghazy, E.; Herp, D.; Schmidtkunz, K.; Melesina, J.; Shaik, T. B.; Erdmann, F.; Schmidt, M.; Romier, C.; Robaa, D.; Witt, O.; Oehme, I.; Jung, M.; Sippl, W. Structure-Based Design and Biological Characterization of Selective Histone Deacetylase 8 (HDAC8) Inhibitors with Anti-Neuroblastoma Activity. J. Med. Chem. **2017**, 60, 10188-10204.

P-63: Applications of Binding Free Energy Calculations and QSAR Modeling to Design Novel Inhibitors for Human Myt1 Kinase

A. Najjar, C. Platzer, M. Schmidt, W. Sippl

Institute of Pharmacy, Martin Luther University of Halle-Wittenberg, Wolfgang-Langenbeck Str. 4, 06120, Halle (Saale), Germany

Membrane-associated inhibitory kinase Myt1 belongs to Wee1-kinase family and regulates the cell cycle at G2/M transition [1]. Myt1 is responsible for inhibitory Cdk1 phosphorylation [1]. As result, the cell cycle is restricted until DNA damage is repaired [2]. A new strategy for cancer treatment is to keep the cell going in the cell cycle with unrepaired DNA damage in premature mitosis. The abrogation of the G2 checkpoint results mitotic catastrophe and immediately causes apoptotic or non-apoptotic cell death [3].

In the current project we used a combination of in silico and in vitro screening to identify novel Myt1 inhibitors. The in-silico screening was done using the available Myt1 crystal structure (PDB 3P1A) and several docking methods [4,5]. As databases for screening we used in-house libraries of already tested inhibitors as well as focused kinase inhibitor libraries (e.g. Selleckchem and GSK kinase inhibitor dataset I and II). The docking solutions were analyzed and re-scored using binding free energy calculations. The tested inhibitors were used to derive a quantitative structure-activity relationships (QSAR) including different descriptors and scoring methods. The QSAR models were validated using external test sets and showed good predictivity. Several scaffolds were identified as starting point for the development of novel Myt1 inhibitors. To optimize the identified hits, we used the fragment-based approach. The most promising docking solutions were used to identify putative binding groups for the individual binding pockets of Myt1. The first set of inhibitors was synthesized and submitted to the biological evaluation. Novel active Myt1 kinase inhibitors have been identified.

1. Mueller, P. R.; Coleman, T. R.; Kumagai, A.; Dunphy, W. G. Myti: A Membrane-Associated Inhibitory Kinas That Phosphorylates Cdc2. *Science.* **1995**, *270, 86-90.*
2. Booher, R. N.; Holman, P. S.; Fattaey, A. Human Myt1 is a cell cycle-regulated kinase that inhibits Cdc2 but not Cdk2 activity. *J. Biol. Chem.* **1997**, *272*, 22300–22306.
3. De Witt Hamer, P. C.; Mir, S. E.; Noske, D.; Van Noorden, C. J. F.; Würdinger, T. WEE1 kinase targeting combined with DNA-damaging cancer therapy catalyzes mitotic catastrophe. *Clin. Cancer Res.* **2011**, *17*, 4200–4207.
4. Rohe, A.; Göllner, C.; Wichapong, K.; Erdmann, F.; Al-Mazaideh, G. M. A.; Sippl, W.; Schmidt, M. Evaluation of potential Myt1 kinase inhibitors by TR-FRET based binding assay. *Eur. J. Med. Chem.* **2013**, *61*, 41–8.
5. Schmidt, M.; Rohe, A.; Platzer, C.; Najjar, A.; Erdmann, F.; Sippl, W.; Regulation of G2/M transition by inhibition of WEE1 and PKMYT1 Kinases. *Molecules.* **2017**, *22 (12)*, 2045.

P-65: Estimation of solvation free energies by continuum methods: How to tackle halogenated species?

R. Nunes [1,2,3], P. J. Costa [1,2]

[1] Centro de Química e Bioquímica, Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal, [2] BioISI - Biosystems & Integrative Sciences Institute, Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal, [3] Centro de Química Estrutural, Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal

The incorporation of halogens into drug candidates has occupied an important role in drug discovery and development processes. While traditionally this design strategy mainly aimed at improving drug-like properties (e.g. biomembrane permeability or pharmacokinetic stability), the pharmaceutical potential of halogenated compounds has been increasingly explored for their ability to modulate protein–ligand binding affinity by establishing halogen bonds (XB)[1]. These are highly directional non-covalent interactions explained by the existence of a positive region on the electrostatic potential (ESP) of heavier halogens (X), called σ–hole, which is available to interact with electron-rich species (i.e. Lewis bases). The development of computational methods that accurately model the charge anisotropy of halogenated compounds is therefore of great importance, in view of their use in computer-aided drug design and virtual screening routines. Particularly challenging is the case of molecular mechanics (MM)-based methods since these rely on point charges, therefore typically failing to represent XBs. The simplest approach to describe the ESP anisotropy in halogenated species involves the addition of an off-centre positive extra-point (EP) of charge mimicking the σ−hole.[2] We have successfully applied this type of methodology to the study of protein–ligand complexes by means of molecular dynamics (MD) simulations.[3] Regarding the prediction of absolute protein–ligand binding free energies, the use of molecular mechanics energies combined with Poisson-Boltzmann surface area (MM-PBSA) continuum solvation is a popular methodology. While EP addition has been shown to improve the molecular mechanical description of halogen-containing systems, its effect on the accuracy of binding free energy as estimated by MM-PBSA is yet to be assessed. This method relies on the estimation of the solvation free energy of the ligand, amongst other terms, for which an empirical assignment of halogen parameters, such as the PB radius, is required. Hence, we conducted a comprehensive study on the effect of varying the X···EP distance, together with the halogen PB radii, on the performance of MM-PBSA-based solvation free energy calculations for a library of halogenated ligands. The results, highlighting the dramatic impact of varying the two parameters on the computed error, when compared with experimental data, will be disclosed. Implications for computer-aided drug design will also be addressed.

Acknowledgments: This work was supported by Fundação para a Ciência e a Tecnologia (FCT), Portugal, through fellowship SFRH/BD/116614/2016 and projects IF/00069/2014/CP1216/CT0006, UID/MULTI/04046/2013 and UID/MULTI/00612/2013.

1. Lu, Y.; Shi, T.; Wang, Y.; Yang, H.; Yan, X.; Luo, X.; Jiang, H.; Zhu, W. Halogen Bonding–A Novel Interaction for Rational Drug Design? *J. Med. Chem.* **2009**, 52, 2854-2862.
2. Wolters, L. P.; Schyman, P.; Pavan, M. J.; Jorgensen, W. L.; Bickelhaupt, F. M.; Kozuch, S. The many faces of halogen bonding: a review of theoretical models and methods. *WIREs Comput. Mol. Sci.* **2014**, 4, 523-540.
3. a) Nunes, R.; Vila-Viçosa, D.; Machuqueiro, M.; Costa, P. T4 Lysozyme/Halobenzene: A Test System for Modeling Biomolecular Halogen Bonds. In Proceedings of the MOL2NET, International Conference on

Multidisciplinary Sciences, 15 January–15 December 2017; Sciforum Electronic Conference Series, Vol. 3, 2017, doi:10.3390/mol2net-03-05075; b) Nunes, R.; Costa, P. J.; et al., under submission.

P-67: A multi-target approach to neurodegenerative diseases

Sebastian Oddsson[1], Thomas Balle[2],Elín Soffía Ólafsdóttir[1]

[1]Faculty of Pharmaceutical Sciences, University of Iceland, Iceland, [2]Faculty of Pharmacy, The University of Sydney

In order to address the lack of new strategies for drugs that can potentially treat neurodegenerative diseases such as Alzheimer's disease (AD), which affect the world's population increasingly due to the demographic changes, we decided to target more than one molecular player at once. Inhibition of the enzyme Acetylcholinesterase (AChE) is currently the treatment of choice for most AD patients and its mode of action is commonly explained with the cholinergic hypothesis, whereby it is assumed that function of cholinergic synapses is impaired. Secondly, the drug memantine, which is the only other approved drug for AD acting by a different mechanism than the aforementioned, blocks N-methyl-D-aspartate (NMDA) receptor channels and was therefore also targeted. Thirdly, nicotinic acetylcholine receptors (nAChRs) are also implicated in the disease mechanism selected because of this. A Virtual Screening has been performed on a database of 5 million compounds against these three targets independently of one another. Based on constraints and a variety of properties potential hits have been selected and are currently undergoing *in vitro* testing. Preliminary results are presented.

P-69: A Computational Platform For Fragment Evolution

S. Piticchio,[1] M. Martinez,[1] S. Scaffidi,[1] M. Rachman,[1] X. Barril.[1,2]

[1] Physical Chemistry Department, Faculty of Pharmacy, Barcelona University, Av. De Joan XXIII, 27-31, 08028 Barcelona, Spain, [2] Catalan Institution for Research and Advanced Studies (ICREA), Passeig Lluís Companys 23, 08010 Barcelona, Spain

Fragment-based drug design has gained ground as a hit identification strategy and is increasingly being used by researchers in industry and academia. With relatively small collections, fragment screening explores a large portion of chemical space and achieves higher hit rates than traditional drug-like collections. Fragments can form optimal interactions with particular subpockets and attain better ligand efficiencies than the bigger HTS hits[1].

However, due to their size, the binding potency is usually weak, leading to the challenge of evolving it to a more potent drug-like compound. The pool of synthetically accessible and drug-like compounds that the medicinal chemists have to explore is vast. Here we present an automatic protocol to facilitate and direct this process.

Given an initial fragment, which binding mode is known (e.g. by X-ray crystallography) the protocol searches in a database for molecules that are chemically related and slightly bigger in size. These are then tethered docked to the target protein to identify those that are complementary. Dynamic Undocking[2] is then applied to filter out false positives and the top candidates are selected. The process is repeated until drug-sized molecules are attained.

We applied the protocol prospectively to the bromodomain BRD4(1). Starting from a published fragment, we identify active molecules that are different from existing BRD4 inhibitors, even those that were evolved from the same fragment[3]. Active molecules are being tested with complementary biophysical methods and characterized by X-ray crystallography.

1. Scott, D. E.; Coyne A. G.; Hudson S. A.; Abell C. "Fragment-Based Approaches in Drug Discovery and Chemical Biology" *Biochemistry*, **2012**, 51 (25), 4990–5003
2. Ruiz-Carmona, S.;Schmidtke P.; Luque F. J.; Baker L.; Matassova N.; Davis B.; Roughley S.; Murray J.; Hubbard R. & Barril X. "Dynamic undocking and the quasi-bound state as tools for drug discovery" *Nature Chemistry*, **2017**, 9, 201–206
3. Gehling V. S.; Hewitt M. C.; Vaswani R. G.; Leblanc Y.; Côté A.; Nasveschuk C. G.; Taylor A. M.; Harmange J.; Audia J. E.; Pardo E.; Joshi S.; Sandy P.; Mertz J. A.; Sims R. J., III; Bergeron L.; Bryant B.

M.; Bellon S.; Poy F.; Jayaram H.; Sankaranarayanan R.; Yellapantula S.; Bangalore Srinivasamurthy N.; Birudukota S. and AlbrechtB. K. "Discovery, Design, and Optimization of Isoxazole Azepine BET Inhibitors" ACS Medicinal Chemistry Letters, **2013,** 4 (9), 835-840

P-71: NAOMInext - Reaction-Driven Probing of Protein Binding Sites

Kai Sommer[1], Florian Flachsenberg[1], Matthias Rarey[1]

[1] ZBH-Center for Bioinformatics University of Hamburg, Hamburg, Germany

After identification of initial active compounds for a target of interest, medicinal chemists usually explore the surrounding chemical space for interesting lead compounds. To support this process in a structure-based design scenario, we developed NAOMInext a program combining organic synthesis rules, structural sampling (growing), and primary target- and user-constraints in an easy-to-use graphical user interface (GUI) to design the next generation of lead compounds.

Our powerful and efficient SMARTS processing library allows the integration of user-defined reactions encoded in Daylight's Reaction SMARTS format[2,3] on demand, which enables individual reaction steps for a desired target. Even complex SMARTS expressions – using recursion to clearly define the surrounding or special properties of a reacting atom – are supported. As a beginning, we integrated 58 published robust organic synthesis reactions[1] encoded in the Daylight's Reaction SMARTS notation[3] into NAOMInext (see Figure 1 for an example).

The success of performing synthetic reactions for the generation of *de novo* molecules has already been shown in several other studies.[4,5] In our work we apprehend the process described in SCUBIDO[6], which showed good results of synthetic tractability in *de novo* drug design studies. Herein, moving the described workflow into 3D space using a straightforward condensed algorithm for fragment growing, further increases the significance of the produced results and minimizes false positives. Based on our NAOMI framework[7] we are able to generate structurally flawless synthesis results, considering both, stereo- and regioisomers, covering all relevant reaction results simultaneously. A structural sampling of the grown reaction product is performed under consideration of the primary target and anchor constraints (see Figure 1). Furthermore, user defined constraints – to guide the sampling procedure – can be defined. Generating synthetic feasible *de novo* molecules, starting with a single building block (anchor) at one of the key interaction sites of the target protein, enables medicinal chemists to explore the chemical space of a screened fragment.

The combinatorial explosion of the structural sampling is solved using a heuristic and knowledgebased approach. To cope with large and highly flexible molecules we implemented a combination of a Breadth- and Depth-First-Search algorithm, which only proceeds with the best *n* partial solutions. We start with the statistically most relevant torsion angles and at each atom all possible solutions are examined. If no solution is possible, the algorithm dynamically extends the relevant torsion angles[8] by using tolerance values. This reduces the search space significantly but ensures a good solution in an acceptable time frame. The evaluation of our sampling approach was performed on a dataset of 297 co-crystallized ligands with their putative precursor.[9] The results were then compared with docking results to demonstrate the benefit of the spatial information of the initial anchor fragment.

Our approach provides the user with an easy and at the same time powerful instrument to rapidly generate new ideas in early stage fragment-based drug discovery projects. Our tool implicitly combines several constraints at a time without the need but the possibility for the user to take action. First, available synthetic reactions are filtered, retaining only those that match the anchor fragment. Second, only building blocks that are compatible with the anchor fragment in terms of further synthesis are used. And third, only results which fit the target binding site and perform favorable protein-ligand interactions are retained. Thus, fragment chemical space and structural space are pruned at an early stage, which tremendously reduces the number of false positives and speeds up further investigation of the results.

Pictet-Spengler: [cH1:1]1:[c:2](-[CH2:7]-[CH2:8]-[NH2:9]):[c:3]:[c:4]:[c:5]:[c:6]:1.[#6:11]-[CH1;R0:10]=[OD1]>>[c:1]12:[c:2](-[CH2:7]-[CH2:8]-[NH1:9]-[C:10]-2(-[#6:11])):[c:3]:[c:4]:[c:5]:[c:6]:1

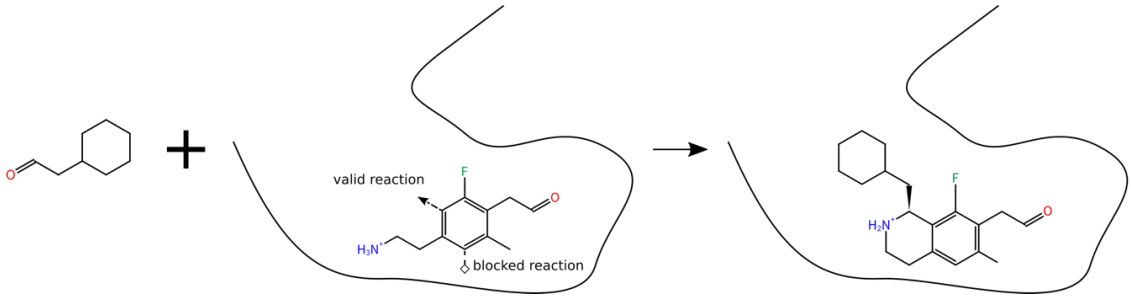


Figure 1: Exemplary Pictet-Spengler reaction within a schematically drawn binding site. The possible reaction center in the phenyl ring next to the methyl group is blocked due to spatial restrictions of the binding site. The reaction center next to the fluorine is not blocked. Only one possible stereoisomer of the product is shown here.

1. Hartenfeller, M.; Eberle, M.; Meier, P.; Nieto-Oberhuber, C.; Altmann, K. H.; Schneider, G.; Jacoby, E.; Renner, S. A Collection of Robust Organic Synthesis Reactions for in Silico Molecule Design. *J. Chem. Inf. Model.* **2011**, *51* (12), 3093–3098.
2. Daylight Chemical Information Systems, Inc., Laguna Niguel, CA, USA.
3. Daylight SMARTS Documentation, http://www.daylight.com/ dayhtml/doc/theory/theory.smarts.html (accessed February 1th, 2018)
4. Pottel, J.; Moitessier, N. Customizable Generation of Synthetically Accessible, Local Chemical Subspaces. *J. Chem. Inf. Model*. **2017**, 57 (3), 454–467.
5. Chevillard, F.; Rimmer, H.; Betti, C.; Pardon, E.; Ballet, S.; van Hilten, N.; Steyaert, J.; Diederich, W. E.; Kolb, P. Binding-Site Compatible Fragment Growing Applied to the Design of β 2 -Adrenergic Receptor Ligands. *J. Med. Chem*. **2018**, 61(3), 1118-1129
6. Chevillard, F.; Kolb, P. SCUBIDOO: A Large yet Screenable and Easily Searchable Database of Computationally Created Chemical Compounds Optimized toward High Likelihood of Synthetic Tractability. *J. Chem. Inf. Model.* **2015**, *55* (9), 1824–1835.
7. Urbaczek, S.; Kolodzik, A.; Rarey, M. The Valence State Combination Model: A Generic Framework for Handling Tautomers and Protonation States. *J. Chem. Inf. Model.* **2014**, *54* (3), 756–766.
8. Schärfer, C.; Schulz-Gasch, T.; Ehrlich, H. C.; Guba, W.; Rarey, M.; Stahl, M. Torsion Angle Preferences in Druglike Chemical Space: A Comprehensive Guide. *J. Med. Chem.* **2013**, *56* (5), 2016–2028.
9. Malhotra, S.; Karanicolas, J. When Does Chemical Elaboration Induce a Ligand To Change Its Binding Mode? *J. Med. Chem.* **2017**, *60* (1), 128–145.

P-73: Effects of MD-MM/GBSA Parameters on the Rank-Ordering of Ligands in Drug Design

N. Stiefl [1], P. Pacak [1], S. Riniker [2], R. Wolf [1]

[1] Global Discovery Chemistry, Novartis Institutes for Biomedical Research, Novartis Pharma AG, Basel, Switzerland, [2]Department of Chemistry and Applied Biosciences, Laboratory of Physical Chemistry, Computational Chemistry, ETH Zurich, Switzerland

In the course of early stage medicinal chemistry projects, often large collections of compounds need to be evaluated (e.g., in rescoring of virtual screening docking results). For medium sized result sets, MM/(GB)(PB)SA approaches are applicable due to their simplicity and computational speed.

However, when working with large data sets, the choice of parameters like simulation protocols, force fields, protonation and tautomeric states, various implementations of Poisson-Boltzmann (PB) or Generalized-Born (GB) can be a major challenge. When ranking is more relevant than the closest fit to experimental free energies of binding, some of these choices become less important. Still, a systematic evaluation can be computationally very expensive especially when MM/(GB)(PB)SA is combined with molecular dynamics (MD) simulations.

In this work, we highlight the impact of specific parameters on the protocols to be followed to take into account the variability of MD-based results, also considering multiple short MD trajectories versus single longer ones. In addition, we check the possible influence of keeping a small number of selected water molecules as part of the MM/GBSA scheme. Comparisons to single-point, energy-refined starting structures also reveal that for "correct" initial poses as in most examples reported here, a simple minimization followed by an MM/GBSA evaluation can be sufficient (or even superior) to a lengthy MD treatment.

P-75: Can I make this into a macrocycle? Effective methods for fragment growing, joining and cyclisation.

P. Tosco[1], M. Mackey[1]

[1] *Cresset, Litlington, UK*

An increasingly common medicinal chemistry technique is conformational restriction through macrocyclization, in order to attain higher affinity and selectivity for the target coupled to improved oral bioavailability.[1] Although the concept is simple, the execution is difficult: the proposed linker must have enough flexibility that it can join the proposed cyclisation sites without introducing too much steric strain, but not so much that the entropic benefits of cyclisation are lost. It must be synthetically feasible, must fit into the available space in the active site, and ideally should make favourable (or at least not unfavourable) interactions with the protein.

Macrocyclisation can be seen as a special case of fragment linking, which in turn is a constrained form of fragment growing. Fragment linking strategies have been recently reported as a highly successful route to lead optimization.[2,3] In both linking and macrocyclisation the design problem is to find a moiety which enforces the required geometry between the two link sites. If available, a template ligand with known affinity can be used to help choosing the best linker and achieve the desired binding properties in the final product.

We present the application of Cresset's Spark bioisostere search methodology to this problem. Spark has the ability to search for bioisosteric replacements in a molecule while scoring the results against a separate reference molecule. In a fragment growth experiment, the goal is to decorate a starter fragment known to bind the target with functional groups that allow to further enhance its affinity. Very often, one or more drug-like molecules with good affinity for the target which may act as templates are already known in the literature, but they may have been already patented or have other drawbacks (e.g., unfavourable pharmacokinetic properties). The fragment growing workflow allows to grow the starter fragment in a user-defined direction with moieties that allows to retain as much as possible the interaction and shape properties of the template but using a different chemistry to escape IP and PK issues (Fig 1).

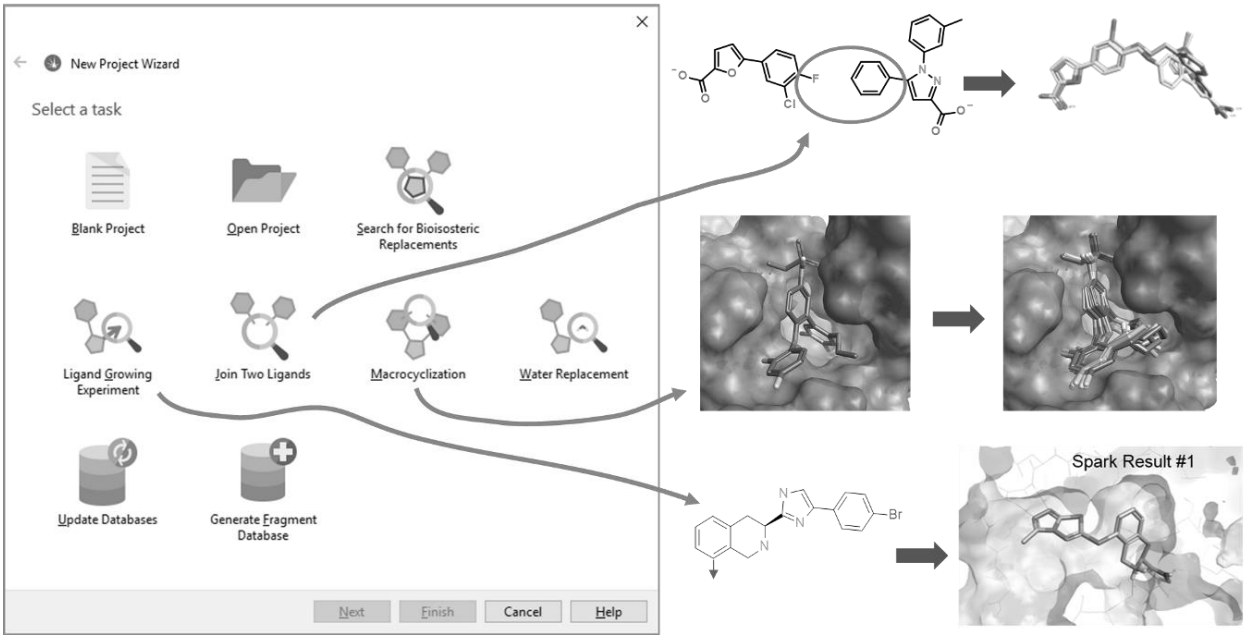


Fig 1. Fragment growing, joining and macrocyclization experiments through the Spark wizards.

In the context of a fragment linking experiment, the Spark method can look for a "bioisosteric" replacement joining the gap between the two fragments, with the bioisosteric similarity computed against a reference molecule known to interact favourably with the protein in the region of the proposed linker. The results can be guided with additional constraints, such as the shape of the active site cavity, the known pharmacophores for activity, and regions of electrostatic potential that are known to be crucial for binding.

In this talk we apply this technique to several data sets, including a set of recently-reported pyridine-based BRD4 inhibitors[4]. Known highly-active macrocyclic inhibitors were reliably obtained, especially if template molecules binding in the cyclisation region could be used to guide the experiment. However, even without this the application of both excluded-volume and pharmacophoric constraints from the protein structure provided excellent results (Fig 2). Results were obtained that matched known cyclisation strategies even when these involve a conformation change to the conserved part of the molecule. Analysis of the torsions of the newly-created bonds against torsional statistics from the CSD[5] provided confirmation that the macrocycle linker sizes suggested by the algorithms minimised conformation strain.

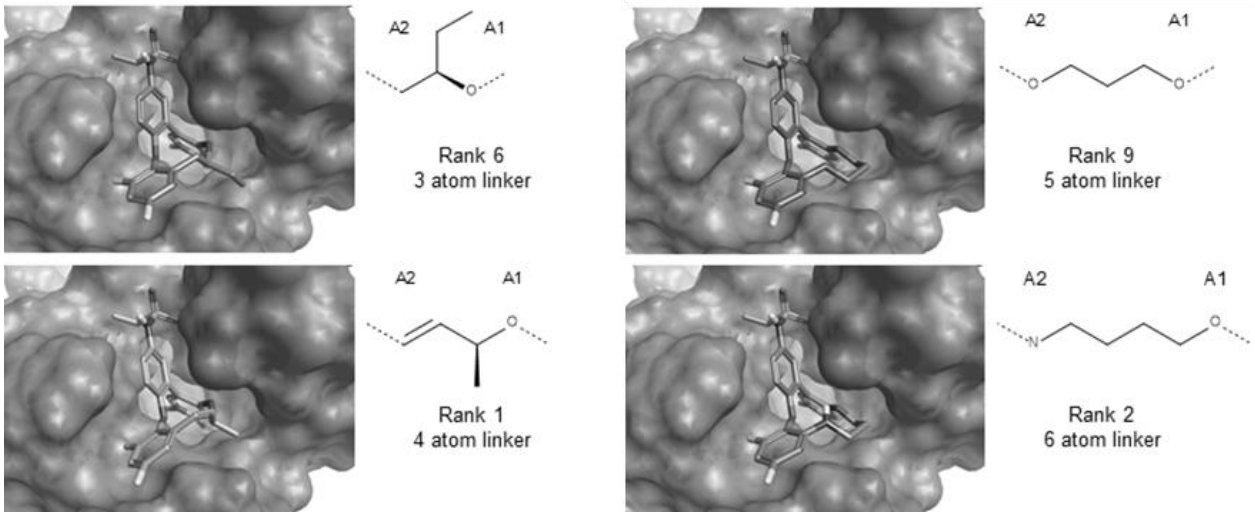


Fig 2. Spark macrocyclization results. Among the top 10 results Spark designed compounds with linker sizes between 3 to 6 atoms. The top ranking result for each linker size is shown

1. Marsault, E.; Peterson, M. L. Macrocycles Are Great Cycles: Applications, Opportunities, and Challenges of Synthetic Macrocycles in Drug Discovery. *J. Med. Chem.* **2011**, 54, 1961-2004.
2. Murray, C. W.; Rees, D. C. The rise of fragment-based drug discovery. *Nature Chemistry* **2009**, 1, 187-192.
3. Mondal, M.; Radeva, N.; Fanlo-Virgós, H.; Otto, S.; Klebe, G.; Hirsch, A. K. H. Fragment Linking and Optimization of Inhibitors of the Aspartic Protease Endothiapepsin: Fragment-Based Drug Design Facilitated by Dynamic Combinatorial Chemistry. *Angew. Chem. Int. Ed.* **2016**, 55, 9422-9426.
4. Wang, L.; McDaniel, K. F.; Kati, W. M. Fragment-Based, Structure-Enabled Discovery of Novel Pyridones and Pyridone Macrocycles as Potent Bromodomain and Extra-Terminal Domain (BET) Family Bromodomain Inhibitors. *J. Med. Chem.* **2017**, 60, 3828-3850.
5. Guba, W.; Meyder, A.; Rarey, M.; Hert, J. Torsion Library Reloaded: A New Version of Expert-Derived SMARTS Rules for Assessing Conformations of Small Molecules. *J. Chem. Inf. Model.* **2016**, 56, 1-5.

P-77: Truly Target-Focused Pharmacophore Modeling: A Novel Tool for Mapping Intermolecular Surfaces

Jérémie Mortier [1], Pratik Dhakal [1], Andrea Volkamer [1]

[1] *In silico Toxicology Group, Institute of Physiology, Charité Universitätsmedizin Berlin, Germany*

Various computational tools and molecular modeling platforms are known to support medicinal chemists in understanding bioactivities, predicting binding events and rationally designing drug molecules. Among them, the pharmacophore approach is an accurate and minimal tridimensional abstraction of chemical structures and intermolecular interactions.

Pharmacophore models are usually derived from a group of molecules in absence of structural information on their biological targets (ligand-based approach) or from a ligand-target complex (structure-based approach). However, only a limited amount of solutions exists to model comprehensive pharmacophores using the information of a particular target structure without knowledge of any binding ligand.[1] In the presented work, T2F-Pharm, a fully automated and customable tool for **T**ruly **T**arget-**F**ocused **Pharm**acophore modeling will be introduced. Using a grid-based approach,[2] this method samples the protein cavity, filters the grid points by energy level and clusters them into low energy hot spots. Subsequently, key features in the pocket required for optimal interaction in a 3D-pharmacophore model are derived. Using a variety of protein classes, the ability of this method to identify essential features was compared to structure-based pharmacophores derived from ligand-target interactions. Currently, we are extending our method to generate merged pharmacophores from molecular dynamics snapshots to capture protein flexibility.

The novel method represents a valuable instrument for drug discovery to investigate protein surfaces in absence of known binding partners, e.g. in cases of rather unexplored binding sites, protein allosteric pockets or protein-protein interactions.

1. Sanders MPA, et al. From the protein's perspective: the benefits and challenges of protein structure-based pharmacophore modeling. *Med.Chem.Comm.*, **2012**, 3:28-38
2. Trott O, Olson AJ. 2010. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J.Comput.Chem*, **2010**, 31:455-61

Poster Session Abstracts BLUE

P-02: Characterization of the Chemical Space of Known and of Readily Purchasable Natural Products

Y. Chen [1], M. Garcia de Lomana [1], N.-O. Friedrich [1], J. Kirchmair [1]

[1] *Universität Hamburg, MIN Faculty, Department of Informatics, Center for Bioinformatics, Hamburg, Germany*

Natural products are structurally diverse and exhibit a wide range of bioactivities, making them an important resource for drug discovery.[1-3] We have recently reviewed 25 virtual and 31 physical natural product libraries that are useful for applications in cheminformatics.[4] They cover a total of 250,000 natural products, at least 10% of which are readily purchasable.

In this follow-up study, we present a detailed analysis of the physicochemical property space of natural products that clearly goes beyond the reach of earlier reports. We implemented a new algorithm called "SugarBuster" that identifies and removes generally undesirable sugars and sugarlike moieties from natural products. This gives a more realistic view of the physicochemical properties of aglycons that may serve as templates for drug design. We also compare, for the first time, the physicochemical properties and scaffold diversity of purchasable natural products to those of all known natural products. This analysis provides valuable insights into the relevance of purchasable natural products for drug discovery and points out areas in the chemical space that are only covered by natural products that require on-demand sourcing, extraction or synthesis. Furthermore, a rule-based approach for the automated recognition of the structural classes of natural products (e.g. alkaloids or flavonoids) was implemented, which allowed us to quantify their abundance among various data sources.

1. Newman, D. J.; Cragg, G. M. Natural Products as Sources of New Drugs from 1981 to 2014. *J. Nat. Prod.* **2016**, 79, 629–661.
2. Harvey, A. L.; Edrada-Ebel, R.; Quinn, R. J. The Re-Emergence of Natural Products for Drug Discovery in the Genomics Era. *Nat. Rev. Drug Discov.* **2015**, 14, 111–129.
3. Rodrigues, T.; Reker, D.; Schneider, P.; Schneider, G. Counting on Natural Products for Drug Design. *Nat. Chem.* **2016**, 8, 531–541.
4. Chen, Y.; de Bruyn Kops, C.; Kirchmair, J. Data Resources for the Computer-Guided Discovery of Bioactive Natural Products. *J. Chem. Inf. Model.* **2017**, 57, 2099–2111.

P-04: Effects of missing data on multitask prediction performance

A de la Vega de León[1], V J Gillet[1]

[1] *University of Sheffield, Regent Court, 211 Portobello, S1 4DP Sheffield, United Kingdom*

Deep learning has become increasingly popular in chemoinformatics. Deep neural networks have been successfully applied to predict activity in large chemical data sets[1], as well as other chemical endpoints of interest[2,3]. One of the advantages of these techniques is their multitask nature; they are able to predict several outputs with a single model. This makes them interesting to support drug discovery projects, where molecules need to be optimized against a battery of different properties and activities. They can also be used to model multitarget data sets, such as those from kinase profiling exercises.

These types of data sets can be assembled using publicly available data sources. However, this leads to data sets that are sparse; where not all compounds have been tested against all targets. It is expected that when these data sets are used to train predictive models, their performance would be worse than if the data sets were complete. However, there has been little research into how much performance is lost when training data is removed. The aim of this work is to gain an understanding of how complete a data matrix should be in order to obtain models with acceptable levels of performance.

We have used two complete data sets to measure the effect of missing data in the performance of multitask methods. One data set is PKIS[4], a kinase profiling data set donated by GSK to ChEMBL, and the other is a PubChem based data set using a subset of assays from a previous publication[5]. Two different multitask methods were compared: deep neural networks and Macau[6], a technique based on probabilistic matrix factorization. A large set of models was trained for each data set and technique, where increasing amounts of training data were removed. Macau and deep neural networks showed very similar performance progression as increasing amounts of training data were removed.

In both cases, the decrease in performance was at first slow and it did not increase until almost three quarters of the training data were removed. Our results suggest the multitask nature of these techniques is the origin of their beneficial performance progression.

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n°612347.

1. Ma, J.; Sheridan, R.P.; Liaw, A.; Dahl, G.E.; Svetnik, V., Deep Neural Nets as a Method for Quantitative Structure–Activity Relationships. *J. Chem. Inf. Model*. **2015**, 55, 263-274.
2. Lusci, A.; Pollastri,G.; Baldi, P. Deep Architectures and Deep Learning in Chemoinformatics: The Prediction of Aqueous Solubility for Drug-Like Molecules. *J. Chem. Inf. Model.* **2013**, 53, 1563-1575.
3. Xu, Y.; Dai, Z.; Chen, F.; Gao, S.; Pai, J.; Lai, L. Deep Learning for Drug-Induced Liver Injury. *J. Chem. Inf. Model.* **2015**, 55, 2085-2093
4. https://www.ebi.ac.uk/chembldb/extra/PKIS/ (accessed 18th January 2018)
5. Helal, K. Y.; Maciejewski, M.; Gregori-Puigjané, E.; Glick, M.; Wassermann, A. M. Public Domain HTS Fingerprints: Design and Evaluation of Compound Bioactivity Profiles from PubChem's Bioassay Repository. *J. Chem. Inf. Model.* **2016**, 56, 390-398.
6. Simm, J.; Arany, A.; Zakeri, P.; Haber, T.; Wegner, J. K.; Chupakhin, V.; Ceulemans, H.; Moreau, Y. Macau: Scalable Bayesian Multi-Relational Factorization with Side Information Using MCMC. *arXiv* **2015**, 1509.04610.

P-06: Compound enumeration using Reaction Workflows

J Hussain, G Bravi & M Hartshorn

Dotmatics, Bishop's Stortford, UK

The ability to enumerate virtual chemical structures is vital in the design and synthesis of chemical arrays. It is now straight-forward to enumerate many compounds for a given chemical reaction. However, the synthesis of compounds typically involves several steps. In addition, the drive to improve the efficiency of multistep synthesis, convergent synthesis is increasingly common. These more complex workflows can present a challenge for compound enumeration systems.

In this poster, a new chemical enumeration application called Reaction Workflows (RW) will be presented. The application uses a graph to represent a reaction workflow with nodes to represent the reagents and reactions. The products of a reaction can also be passed to another reaction. These reaction workflows are akin to the reaction schemes we are familiar with in compound synthesis. This means it is straight-forward and intuitive for a chemist to drag and drop reagents and reactions nodes to represent a complex convergent multi-step synthesis within RW. The application also contains nodes to perform other functions such chemical property calculation, structure normalization and substructure filtering. The graphical nature of the application and the functionality available means it becomes possible for chemists to build complex workflows needed for their compound filtering and enumeration needs.

P-08: chem2vec : vector embedding of atoms and molecules

N. Jeliazkova [1], V. Chupakhin [2], Hugo Ceulemans[2], Jose-Felipe Golib-Dzib [3], V. Jeliazkov [1]

[1] Ideaconsult Ltd., 4 Angel Kanchev Str. Sofia, Bulgaria, [2] Computational Biology, Discovery Sciences, Janssen Pharmaceutica NV, Beerse, Belgium, [3] Computational Biology, Discovery Sciences, Janssen Cilag SA, Toledo, Spain

We present chem2vec, a method to generate a novel type of real-valued chemical structure descriptors with user specified dimension, based on the word2vec algorithm. The word2vec method, a neural network with one hidden layer[1], takes on input sequences of words (sentences) and uses the information of the environment a given word is found in to derive a vector representation of the word in a user specified dimension *D*. Similarly, chem2vec takes on

input linear sequences of objects, which represent parts of the molecule, namely paths of atoms, or more precisely, paths of atom types (as implemented in The CDK library[2]).

Chem2vec consists of two steps 1) generating a dictionary of vectorized atom types, preferably from a large dataset of chemical structures 2) generating molecular descriptors using the dictionary. The dictionary is transferable, i.e. it could be generated once from a large set of chemical structures and reused subsequently for generating descriptors of arbitrary sets of chemical structures, not necessary the same as the ones used for generating the dictionary. Dictionaries of vectorized atom types have been generated using several datasets (e.g. ChEMBL, ExCAPEDB and industry data) and approaches for comparison are presented. The molecular descriptors are compositional, built by combining the atom type vectors. The result is a vector representation of the molecule, encompassing low dimensional space. We show that while associating a single vector dimension to a molecular moiety is not possible, the vectors can be decoded into the familiar count of atom types and tuples of atom types.

The new chem2vec descriptors can be used for similarity assessment, and as input for unsupervised (clustering) and supervised (regression, classification) machine learning methods. Experiments with several datasets are performed on large scale public datasets (e.g. chemogenomics dataset ExCAPEDB[3]) and industry data. The predictive performance of supervised models using the low dimensional real valued chem2vec descriptor space is comparable or exceeds the performance of models using traditional high dimensional sparse fingerprint-based descriptors.

Acknowledgment: This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No. 671555.

1. Mikolov, T.; Corrado, G.; Chen, K.; Dean, J. In Proceedings of the International Conference on Learning Representations (ICLR 2013); 2013; pp 1–12.
2. Willighagen, E. L.; Mayfield, J. W.; Alvarsson, J.; Berg, A.; Carlsson, L.; Jeliazkova, N.; Kuhn, S.; Pluskal, T.; Rojas-Chert?, M.; Spjuth, O.; Torrance, G.; Evelo, C. T.; Guha, R.; Steinbeck, C. J. Cheminform. **2017**, 9 (1), 33.
3. Sun, J.; Jeliazkova, N.; Chupakin, V.; Golib-Dzib, J.-F.; Engkvist, O.; Carlsson, L.; Wegner, J.; Ceulemans, H.; Georgiev, I.; Jeliazkov, V.; Kochev, N.; Ashby, T. J.; Chen, H. J. Cheminform. **2017**, 9 (1), 17.

P-10: Building and searching large chemistry spaces

U. Lessel [1], C. Lemmen [2]

[1] Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach an der Riss, Germany, [2] BiosolveIT, St. Augustin, Germany

Virtual screening in large chemistry spaces was first popular with the rise of High Throughput Screening and Combinatorial Chemistry. It then lost its attraction for various reasons and gains now another wave of interest through technologies like DNA encoded libraries as well as the recognition that traditional compound libraries are a quite limited resource, while the know how is there to break current size limits.

Feature Trees Fragment Spaces came up about 10 years ago[1, 2, 3, 4] as a software to search chemistry spaces. While the searching with this technology is quite simple and effective, building a Fragment Space used to be a significant effort. Last year a new version of the software tool CoLibri[5] came out, which turned building Fragment Spaces into a reasonably straight forward process.

With the new CoLibri as a tool to easily build chemistry spaces the question arises, how to compare different spaces. To the best of our knowledge so far no technique fulfilling this task has been published. Here we compare two different chemistry spaces, namely the Knowledge Space[6], a literature-based resource, and the BICLAIM space from Boehringer Ingelheim, by means of different application scenarios. Additionally we look at the ZINC-database[7] as a traditional compound resource.

We assessed similarity and diversity within hit sets, number of Murcko cores, as well as chemical feasibility. This way we detected interesting differences, partially caused by the diverging design principles behind these resources.

In this presentation we present the study and its results. We discuss the value of the different parameters analysed for characterizing chemistry spaces and for their comparison.

1. Rarey, M.; Dixon, J.S. Feature trees: a new molecular similarity measure based on tree matching. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 471-90.
2. FTrees Version 3.3, BioSolveIT. BioSolveIT GmbH, http://www.biosolveit.de/.

3. Boehm, M.; Wu, T.-Y.; Claussen, H.; Lemmen, C. Similarity Searching and Scaffold Hopping in Synthetically Accessible Combinatorial Chemistry Spaces *J. Med. Chem.*, **2008**, *51*, 2468–2480.
4. Lessel U.; Wellenzohn, B.; Lilienthal, M.; Claussen, H. Searching Fragment Spaces with Feature Trees *J. Chem. Inf. Model.*, **2009**, *50*, 1-21.
5. CoLibri Version 3.1, BioSolveIT. BioSolveIT GmbH, http://www.biosolveit.de/.
6. Knowledge Space Version 2.4, BioSolveIT. BioSolveIT GmbH, http://www.biosolveit.de/.
7. Sterling, T.; Irwin, J.J. ZINC 15 – Ligand Discovery for Everyone *J. Chem. Inf. Model.*, **2015**, *55*, 2324-2337.

P-12: Learning from Extant Medicinal Chemistry to Accelerate Hit Identification and Optimisation in Drug Discovery

N Y Mok [1], J Meyers [2], M Carter [1], T Kaserer [1], N Brown [2]

[1] Cancer Research UK Cancer Therapeutics Unit, Division of Cancer Therapeutics, The Institute of Cancer Research, 15 Cotswold Road, London SM2 5NG, United Kingdom. [2] BenevolentAI, 40 Churchway, London NW1 1LW, United Kingdom.

Data mining of publicly available chemical structure databases can enable us to understand the historic exploration for medicinal chemistry relevant structures. This presentation will discuss recent analyses of extant medicinal chemistry and learning from these studies that can be applied to inform and accelerate hit identification and optimisation in drug discovery.

Hit identification strategies in drug discovery often rely on the screening of small-molecule compound libraries to discover hit matter as starting points for development into lead molecular series. A major source of compounds constituting such screening libraries originates from commercial compound vendor collections of small molecules and target-focussed screening sets. Using ChEMBL[1] and eMolecules as exemplar repositories of extant medicinal chemistry and commercially available compounds respectively, over 9 million medicinal chemistry compounds were analysed to understand the coverage of biologically relevant medicinal chemistry space using commercial compound screening libraries. Applying various complementary molecular comparison methods, extant medicinal chemistry space with enrichment in bioactive molecules is identified, and the corresponding molecular and physicochemical properties are characterised. Results from this analysis can inform on the design of effective screening collections that would provide us with greater confidence in identifying high-quality medicinal chemistry starting points in drug discovery projects.

In addition, mining of the ChEMBL database can also provide valuable insights in the exploration and exploitation of chemical space during compound optimisation, as highlighted in recent publications analysing the molecular shape diversity and molecular scaffolds of medicinal chemistry relevant space.[2,3] The appropriate stage at which molecular shape diversity should be introduced in molecular design and the systematic exploration of molecular scaffolds during medicinal chemistry optimisation will be presented. Based upon these results, emerging strategies that can modulate relevant drug-like properties of molecular scaffolds and substituent spaces to accelerate hit optimisation will be discussed.

1. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A large-scale bioactivity database for drug discovery. Nucleic Acids Res. 2012, 40, D1100−D1107.
2. Meyers, J.; Carter, M.; Mok, N. Y.; Brown, N. On the origins of three dimensionality in drug-like molecules. Future Med. Chem. 2016, 8, 1753-1767.
3. Mok, N. Y.; Brown, N. Applications of systematic molecular scaffold enumeration to enrich structure-activity relationship information. J. Chem. Inf. Model. 2017, 57, 27-35.

P-14: HTS workup at AZ – state of the art

J.W.M. Nissink[1]

[1] *Oncology, IMED Biotech Unit, Astrazeneca, Cambridge, United Kingdom*

Large scale high-throughput screening (HTS) is a key lead-finding approach that underpins project starts both in industry and academia. HTS, alongside alternative lead-finding techniques like fragment screening, subset screening, and virtual screening can provide projects with a set of target binders, but it is by no means a trivial exercise.

Here we will discuss the state-of-the-art of HTS at AstraZeneca, with both successful examples and examples of problems that have been encountered. We will touch on design of the screening collection; development of and stress-testing hit-finding cascades with validation sets; analysis of emerging actives, and identification of false hits.

P-16: A Comprehensive Evaluation of ACD/LogD on a Pharmaceutical Compound Set

A. Sazonovas [1,2], K. Lanevskij [1,2], R. Didziapetris [1,2]

[1] *VšĮ „Aukštieji algoritmai“, A.Mickevičiaus 29, LT-08117 Vilnius, Lithuania,* [2] *ACD/Labs, Inc., 8 King Street East, Toronto, Ontario, M5C 1B5, Canada*

Lipophilicity, which is often expressed in terms of 1-octanol/water partitioning coefficient log*P*, or the corresponding pH-dependent distribution coefficient log*D*, is one of the key physicochemical characteristics of any new drug candidates, as it has a major influence on a variety of the compounds' properties constituting their ADME, pharmacokinetic, and drug safety profiles. Widely available *in silico* tools for predicting these properties are mostly based on experimental data for simple organic chemicals and marketed drugs. Consequently, as drug discovery projects are moving to increasingly novel regions of chemical space, utility of existing methods becomes more and more questionable. In several previously published evaluation studies[1,2], the mean log*P* prediction error for *in house* compound libraries of pharmaceutical companies was shown to exceed 1 log unit by almost all methods. Prediction of log*D* is even more challenging, as it requires accurate knowledge of both log*P* of neutral form and distribution of ionic forms of the compound in the relevant pH range. With these considerations in mind, the following objectives were set for the current study:

(1) Collecting a data set of experimental log*D* values from recent publications dealing with novel congeneric compound series from drug discovery projects;
(2) Evaluating the performance of ACD/LogD predictor[3] for the newly collected molecules using different combinations of available log*P* and pK_a calculation algorithms;
(3) Investigating the potential for improving prediction accuracy for unknown compound classes by application of automated model training.

The compiled data set consisted of ~1200 log*D* values measured at physiological pH conditions. According to the initial validation results, the highest accuracy of predictions based on the models employing only built-in compound libraries can be achieved using a combination of ACD/LogP Consensus and ACD/pKa Classic algorithms, yelding RMSE slightly under 1 log unit. However, utilizing the automatic training feature of ACD/LogP GALAS algorithm by the means of stepwise addition of collected data to the model self-training library allowed decreasing the RMSE of predictions for the reserved validation set to as low as 0.6 log units. Moreover, a significant improvement (RMSE ≈ 0.8) was already evident after adding the first portion of training data constituting less than 20% of the entire data set. These results demonstrate that performing experimental measurements for a relatively small number of molecules belonging to a novel chemical series is often sufficient to adapt ACD/LogP and ACD/LogD predictors to provide reliable property estimates for the entire class of compounds.

1. Mannhold R., Poda G. I., Ostermann C., Tetko I. V. Calculation of molecular lipophilicity: State-of-the-art and comparison of log P methods on more than 96,000 compounds. *J. Pharm. Sci.* **2009**, 98, 861-893.
2. Tetko I. V., Poda G. I., Ostermann C., Mannhold R. Large-scale evaluation of log P predictors: local corrections may compensate insufficient accuracy and need of experimentally testing every other compound. B. *Chem. Biodivers.* **2009**, 6, 1837-1844.
3. ACD/LogD (part of Percepta® platform), v. 2017, ACD/Labs, Inc. (http://www.acdlabs.com/products/percepta/)

P-18: Halogens in protein-ligand binding mechanism: a structural perspective

N.K. Shinada, A.G. de Brevern, M. Oberlin, D. Alvarez Garcia, P. Schmidtke

Discngine S.A.S, 79 Avenue Ledru-Rollin, 75012 Paris, France

During the last decade, halogen atoms have become increasingly important in rational drug design. Fluorine is often used to enhance physico-chemical properties. Chlorine, bromine and iodine can influence interaction strength via directed halogen bonds[1]. Quantum mechanics studies and small molecule X ray data shed light onto the precise geometry necessary to favorable halogen interactions[2]. Several computational tools and models start to integrate data from such calculations and studies[3].

However, adding halogen to a ligand can have often overlooked effects. For example an aromatic quadrupole moment can shift from negative to positive, referred to as a πhole[4]. Although most of studies focus on halogen bonding through the σ-hole[5], recent studies suggest hydrogen bonding involving halogens may also have a significant impact with on intermolecular interactions[6].

Using the Discngine's *3decision*® structural knowledge base derived from the RCSB PDB, a refined analysis of halogen interactions reassesses the occurrence of multiple types of interactions made by halogen atoms. We furthermore underline various biases observed in previously used datasets to analyze halogen bonding. We further exemplify preferential interactions of halogenated fragments in small molecules and their implications for rational drug design.

Results shown here aim to complete the current understanding of halogen to biomolecule interaction preferences and can be used by medicinal chemists and molecular modelers to rationally place halogens on small molecules.

1. Hernandes, Marcelo Z., et al. "Halogen atoms in the modern medicinal chemistry: hints for the drug design." *Current drug targets* 11.3 (2010): 303-314.
2. Wilcken, Rainer, et al. "Addressing methionine in molecular design through directed sulfur–halogen bonds." *Journal of chemical theory and computation* 7.7 (2011): 23072315.
3. Ford, Melissa Coates, and P. Shing Ho. "Computational tools to model halogen bonds in medicinal chemistry." *Journal of medicinal chemistry* 59.5 (2015): 1655-1670.
4. Wang, Hui, Weizhou Wang, and Wei Jun Jin. "σ-hole bond vs π-hole bond: a comparison based on halogen bond." *Chemical reviews* 116.9 (2016): 5072-5104.
5. Wilcken, Rainer, et al. "Principles and applications of halogen bonding in medicinal chemistry and chemical biology." *Journal of medicinal chemistry* 56.4 (2013): 13631388.
6. Lin, Fang-Yu, and Alexander D. MacKerell Jr. "Do Halogen–Hydrogen Bond Donor Interactions Dominate the Favorable Contribution of Halogens to Ligand–Protein Binding?." *The Journal of Physical Chemistry B* 121.28 (2017): 6813-6821.

P-20: Interoperable and scalable data analysis in metabolomics

C Steinbeck[1], P Emami Khoonsari[2], K Kultima[2], O Spjuth[2] on behalf of the PhenoMeNal consortium

[1]Friedrich-Schiller-University, Jena, Germany, [2]University of Uppsala, Uppsala, Sweden

Metabolomics aims to characterise the biochemical stage of an organism or biological sample though simultaneous, (semi-) quantitative measurement of as many metabolites as possible. To this end, it uniquely employs cheminformatics and analytical chemistry methods to address questions in biology. Depending on the analytical methods and sample sizes used, metabolomics generates "big data", which can be time consuming to analyse and often exceeds the data processing capabilities of an individual laboratory.

As part of PhenoMeNal project (http://phenomenal-h2020.eu), funded by the European Commission, we have addressed this problem by developing a robust and performant data analysis workflow that integrates all necessary components whilst still being able to scale over multiple compute nodes.

The aim was to support flexible metabolomics data analysis. The system is designed as a virtual research environment which can be launched on-demand on cloud resources and desktop computers.

PhenoMenal is based on a microservice architecture, where software tools are encapsulated as Docker containers that can be connected into scientific workflows and executed in parallel using the Kubernetes container orchestrator. IT-expertise requirements on the user side are kept to a minimum, and established workflows can be re-used effortlessly by any novice user. We validated our method on two mass spectrometry studies, one nuclear magnetic resonance spectroscopy study and one fluxomics study, showing that the method scales dynamically with increasing availability of computational resources.

It is also noteworthy, that microservices are a generic methodology that can serve any scientific discipline and opens up for new types of large-scale integrative data analysis. Apart from offering a turnkey solution for metabolomics, PhenoMeNal therefore also presents an architecture to integrate individual tools into scalable workflows in public and private clouds.

1. Khoonsari, P. E., Moreno, P., Bergmann, S., Burman, J., Capuccini, M., Carone, M., Cascante, M., de Atauri, P., Foguet, C., González-Beltrán, A., Hankemeier, T., Haug, K., He, S., Herman, S., Johnson, D., Kale, N., Larsson, A., Neumann, S., Peters, K., Pireddu, L., Rocca-Serra, P., Roger, P., Rueedi, R., Ruttkies, C., Sadawi, N., Salek, R. M., Sansone, S.-A., Schober, D., Selivanov, V., Thévenot, E. A., van Vliet, M., Zanetti, G., Steinbeck, C., Kultima, K., and Spjuth, O. (2017) Interoperable and scalable metabolomics data analysis with microservices. bioRxiv 213603.

P-22: Supporting the assessment of the purging potential mutagenic impurities via analysis of patent literature

S Webb [1], M Burns [1], E Rosser [1]

[1] Lhasa Limited, Leeds, United Kingdom

Compounds introduced during synthesis including starting materials, intermediates and by-products may be carried through the synthesis to become impurities in the final product. If these are predicted or known to be mutagenic then they are subject to regulation under the ICH M7 guidelines[1]. These guidelines allow for a chemistry-based argument that the impurity will not survive the synthetic route and evidence of its absence may then not be necessary.

Text-mined reactions from the United States Patent Office patent applications and grants provided by NextMove software[2] have been used to support the development of a prototype tool which provides suggestions for potential reactivity-based purging of these impurities.

Reactions are automatically categorised via mapping and generation of a reaction core representing the atom and bond changes occurring within a single-step reaction. These reaction cores can be used generate clusters of reactions sharing common mechanisms without the need for a named reaction.

Reaction networks can be generated as tree structures, organising reaction cores into greater specification. The networks then provide an easy mechanism for identifying feasible reaction mechanisms and identifying the most relevant examples for a given set of reactants. Visualisation of reaction conditions such as yield, solvent, temperature, time, presence of acid/base etc. allow for the assessment of the suitability of a suggested purging mechanism.

1. Guideline, ICH Harmonised Tripartite. "Assessment and Control of DNA Reactive (Mutagenic) Impurities in Pharmaceuticals to Limit Potential Carcinogenic Risk, M7; 2017." (2017).
2. NextMove Software. https://www.nextmovesoftware.com/ (08/02/2017)

P-24: Metabolite Structure Prediction Benefits from Cytochrome P450 Regioselectivity Prediction

C. de Bruyn Kops [1], C. Stork [1], N. Jeliazkova [2], N. Kochev [2,3], J. Kirchmair[1]

[1] Universität Hamburg, Faculty of Mathematics, Informatics and Natural Sciences, Department of Computer Science, Center for Bioinformatics, Hamburg, Germany, [2] Ideaconsult Ltd, Sofia, Bulgaria, [3] University of Plovdiv, Department of Analytical Chemistry and Computer Chemistry, Plovdiv, Bulgaria

Knowledge of the metabolic fate of xenobiotics in humans is invaluable for the development of safe and effective drugs and other chemical substances, because biotransformation of small organic molecules can produce metabolites with biological and physicochemical properties that differ substantially from those of the parent compound.[1] Prediction of the atom positions in a molecule where metabolic reactions are initiated (i.e. sites of metabolism) is a popular aspect of metabolism prediction and can be used as a stepping stone for the prediction of the chemical structures of metabolites.

We have developed a strategy for metabolite structure prediction that is based on FAME 2,[2] our recently developed and highly effective machine learning method for human cytochrome P450 (CYP) regioselectivity prediction. Through the application of known CYP-mediated reactions to the sites of metabolism predicted by FAME 2, we are able to correctly predict the vast majority of known metabolites while keeping false-positive prediction rates low. Compared to CYP-mediated reactions applied to all atom positions in parent compounds, applying the site of metabolism predictions as a preceding filter results in an approximately ten-fold reduction in the number of false positive metabolite predictions on average.

1. Kirchmair J.; Göller A. H.; Lang D.; Kunze J.; Testa B.; Wilson I. D.; Glen R. C.; Schneider G. Predicting drug metabolism: Experiment and/or computation? *Nature Rev. Drug Discov.* **2015**, 14, 387-404.
2. Šícho, M.; de Bruyn Kops, C.; Stork, C.; Svozil, D.; Kirchmair, J. FAME 2: Simple and Effective Machine Learning Model of Cytochrome P450 Regioselectivity. *J. Chem. Inf. Model.* **2017**, 57, 1832-1846.

P-26: Small Molecule Binding Site Prediction – Know Your Needs

C. Ehrt[1,2], T. Brinkjost[1,2], O. Koch[1]

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

The automated prediction and visualization of potential protein binding sites is of interest for function annotation, the elucidation of "druggable" binding sites, and target identification.[1] During the past decades, a plethora of tools have been developed to cope with various challenges of binding site identification and to consider the binding site's nature (protein-protein, protein-DNA, etc.).[2]

Together with those tools, a high and still increasing number of small molecule cavity detection methods is nowadays available. Therefore, the question arises whether this can be attributed to major limitations of published methods or whether they evolve due to user-specific requisites. Although various methods were evaluated for publication purposes and were also quite recently reviewed[3], the choice for the most convenient methods is still a challenging task. A correlation of the results with various pocket properties seems indispensable to choose a suitable tool. Various comparisons of subsets of tools show contradictory results which can be attributed to the varying types of datasets and quality criteria which often highly depend on the output of the methods under investigation.

We applied two representative datasets of high-quality structures to gain a better understanding of the current limitations of binding site prediction. They cover a large spectrum of proteins and include binding sites, which are prone to conformational changes upon ligand binding, as well as comparative protein models. Subsequently, we investigated almost fifty available standalone tools with respect to run time and particularly performance by means of different quality criteria.

For the prediction of ligand binding sites of one single protein structure, the automated prediction should be supplemented by further analyses to obtain reliable results.[4] Nevertheless, one need which cannot be fulfilled by an elaborate pocket characterization workflow is the automated detection of cavities for huge databases of experimental

or theoretical protein structures. To this aim, we strived to identify the most flexible methods whose results do not highly depend on geometric or physicochemical pocket characteristics, or the applied parameter set.

Ultimately, we tested and evaluated various standalone small molecule binding site prediction methods to find answers to the aforementioned questions. This will be outlined by an exhaustive analysis and discussion of our final outcomes which will also include the analysis of ranking methodologies. The obtained results point towards a quite obvious trend which is crucial for all further developments of novel methodologies.

1. Nisius, B.; Sha, F.; Gohlke, H. Structure-based computational analysis of protein binding sites for function and druggability prediction. *J. Biotechnol.* **2012**, 159 (3), 123–134.
2. Watson, J. D.; Laskowski, R. A.; Thornton, J. M. Predicting protein function from sequence and structural data. *Curr. Opin. Struct. Biol.* **2005**, 15 (3), 275–284.
3. Krone, M.; Kozlíková, B.; Lindow, N.; Baaden, M.; Baum, D.; Parulek, J.; Hege, H.-C.; Viola, I. Visual Analysis of Biomolecular Cavities: State of the Art. *Comput. Graph. Forum* **2016**, 35 (3), 527–551.
4. Broomhead, N. K.; Soliman, M. E. Can We Rely on Computational Predictions To Correctly Identify Ligand Binding Sites on Novel Protein Drug Targets? Assessment of Binding Site Prediction Methods and a Protocol for Validation of Predicted Binding Sites. *Cell Biochem. Biophys.* **2017**, 75 (1), 15–23.

P-28: Molecular nature of the increased activity of the Uridine 5'-diphospho-glucuronosyltransferase nine-fold mutant 1A5*8 (UGT1A5*8)

D. Machalz[1], F. Yang[2], M. Bureik[2], G. Wolber[1]

[1] Pharmaceutical and Medicinal Chemistry, Institute of Pharmacy, Freie Universität Berlin, Königin-Luise-Str. 2+4, 14195 Berlin, Germany, [2] School of Pharmaceutical Science and Technology, Health Sciences Platform, Tianjin University, Tianjin 30072, China

Uridine 5'-diphospho-glucuronosyltransferase (UGT) 1A5 is a member of the UGT family catalyzing glucoronidation, a crucial mechanism of phase II metabolism, at low activity levels in different hepatic and gastrointestinal tissues[1]. Drug-induced increase and high intervariability of UGT1A5 expression[1] make it a phase II metabolism enzyme of interest. Besides the wildtype UGT1A5*1, which occurs with a frequency of > 80 % in the human population, a recent study[2] reports a nine-fold mutant UGT1A5*8 (frequency 11.5 %) and a six-fold mutant UGT1A5*9 (frequency 5 %). UGT1A5*9 contains six out of the nine variations present in UGT1A5*8. We expressed the wildtype UGT1A5*1 and the two mutants and incubated them with two Proluciferin UGT-substrates due to the scarcity of reported substrates.

While the six-fold mutant shows wildtype activity levels, we found that the nine-fold mutant of UGT1A5*8 exhibits significantly increased biotransformation activity. In order to investigate the molecular nature of the increased catalytic activity in UGT1A5*8 we conducted structural homology modeling for wildtype and nine-fold mutant. We identified the Gly259Arg mutation, only present in UGT1A5*8, as the likely cause for increased activity, since it introduces additional hydrogen bonding to Asp400 and Asn401 in the helix Q of the structural model of UGT1A5*8. We carried out molecular dynamics (MD) simulations of UGT1A5*1 and UGT1A5*8 in three replicas using Desmond[3] and two homology models served as input structures. We payed close attention to the hydrogen bond network proximal to the cofactor Uridine 5'-diphosphoglucuronic acid (UDPGA) and the Gly259Arg mutation. Simulation analysis showed that the postulated Arg259 hydrogen bonding is stable and thus rigidizes the helix Q observable in reduced root mean square fluctuation (RMSF) values. As a consequence, Asp397 and Gln398, situated in the helix Q, show hydrogen bonding to the cofactor UDPGA with higher occupancy in UGT1A5*8 than in the wildtype. In this study, we identify Arg259 mutation as the indirect cause for tighter UDPGA cofactor binding. This cofactor stabilization explains the increased the activity of UGT1A5*8 compared to the wildtype UGT1A5*1. These results provide new insights into the structure-function relationship of UGT1A5 and lead to the identification of two new substrates of this new potential target for xenobiotic metabolism.

1. Finel, M.; Li, X.; Gardner-Stephen, D.; Bratton, S.; Mackenzie, P. I.; Radominska-Pandya, A., Human UDP-glucuronosyltransferase 1A5: identification, expression, and activity. *J Pharmacol Exp Ther* **2005**, 315 (3), 1143-9.
2. Lek, M.; Karczewski, K. J.; Minikel, E. V.; Samocha, K. E.; Banks, E.; Fennell, T.; O'Donnell-Luria, A. H.; Ware, J. S.; Hill, A. J.; Cummings, B. B.; et al., Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **2016**, 536 (7616), 285-91.

3. Bowers, K.; Chow, E.; Xu, H.; Dror, R.; Eastwood, M.; Gregersen, B.; Klepeis, J.; Kolossvary, I.; Moraes, M.; Sacerdoti, F.; Salmon, J.; Shan, Y.; Shaw, D. In *ACM/IEEE SC 2006 Conference (SC'06)*; 2006; pp 43–43.

P-30: Searching within HELM

Eva Bültel [1,2], Anne Mund [1], Markus Weisser [1]

[1] quattro research GmbH, Planegg, Germany [2] University of Hamburg. Hamburg, Germany

The Hierarchical Editing Language for Macromolecules (HELM) is an open biomolecular representation standard created by Pfizer[1] and further developed by the Pistoia Alliance[2]. It provides a means to represent multiple types of complex macromolecules (e.g. nucleotides, proteins, antibodies and antibody-drug conjugates) including those that contain non-natural elements such as chemically modified amino acids or even abiotic components such as gold particles.

With many current biological entities being significantly more complex than those found in nature, HELM has proven its worth in handling the arising challenges. There are now toolkits for the easy depiction, conversion and storage of biomolecules using the HELM notation. Yet, we still lack the ability to search within HELM strings, which becomes increasingly important as databases grow in size. This searching capability has also been outlined as important by the Pistoia Alliance itself, with a proof of concept done by a student from Cambridge University.

quattro research has developed a new algorithm that allows users to not only query databases for exact matches or substructures, but also establishes a similarity measure between arbitrarily complex biomolecules. Using this similarity measure, biologists can infer behaviour of new substances based on the behaviour of known structures. The algorithm is also capable of taking natural analogues into account when comparing HELM notations. Our work closes an important gap researchers working with HELM have been facing. The extensive search and comparison capabilities demonstrated here will have a significant impact on the acceptance of HELM in the pharmaceutical community.

1. Zhang, T.; Li, H.; Xi, H.; Stanton, R. V.; Rotstein, S. H. HELM: a hierarchical notation language for complex biomolecule structure representation. *Journal of Chemical Information and Modeling* **2012** *52* (10), 2796-2806.
2. Milton, J.; Zhang, T.; Bellamy, C.; Swayze, E. E.; Hart, C. E.; Weisser, M.; Hecht, S.; Rotstein, S. HELM Software for Biopolymers. *Journal of Chemical Information and Modeling* **2017** *57* (6), 1233-1239.

P-32: HELM-driven Integration of Peptides into Structure-Based Drug Design and Cheminformatics

Conor C. G. Scully, Robert T. Smith, Benjamin G. Tehan

[1] Heptares Therapeutics, Welwyn Garden City, United Kingdom

Peptide-based therapeutics are undergoing a resurgence in popularity, currently making up more than 10% of marketed drugs and numbering over 140 in clinical trials. Research teams are incorporating peptides into ever increasing numbers of discovery programs. This boom in biologics has given us the impetus to develop tools aiding the inclusion of peptides into the operational environment of traditional small molecule drug discovery paradigms.

The HELM (Hierarchical Editing Language for Macromolecules) standard is rapidly gaining wide acceptance in industry and academia as an enabling tool for the sequence-based description of complex biological molecules, including peptides.

The implementation of HELM-based tools will be described for a variety of computational tasks including:

- alignment of peptide sequences containing complex networks of unnatural residues and chemical modifications;
- automated generation of templated 3D coordinates for complex peptides from HELM strings;

- generation of normalized 2D representations of complex peptide structures such as branched and cyclic peptides.

Finally, it will be shown how these tools can aid in extracting and parsing information for biologics from databases, in addition to facilitating the organization of this data in a manner that enables the application of differing learning techniques, in this way ensuring that we continue to extract knowledge from the ever-increasing amounts of data now available to everyday researchers.

P-34: Machine Learning Models of Hydrogen Bond Basicity Based on Anisotropy Atomic Reactivity Descriptors

Christoph Bauer [1], Andreas H. Göller [2], Gisbert Schneider [1]

[1] ETH Zurich, Department of Chemistry and Applied Biosciences, Zurich, Switzerland; [2]Bayer AG, Computational Chemistry, Wuppertal, Germany

The quantification of the hydrogen bond strength to a target relative to solvent facilitates the determination of substituent effect and hetero-atom replacement in a compound. The pK_{BHX} database[1] provides experimental values for approximately 1200 compounds, including polyfunctional molecules, *i.e.* molecules with at least two different H-bond acceptors (HBA). The pK_{BHX} data is of atomic resolution for the HBA atoms.

We report on the establishment of general, *i.e.* atom-type independent, machine learning models for the pK_{BHX} data. Our recently developed sets of atomic descriptors that encode the anisotropic electron density distribution using conformation-independent quantum-mechanical atomic charge schemes (Figure 1)[2] served as the feature space for machine learning.

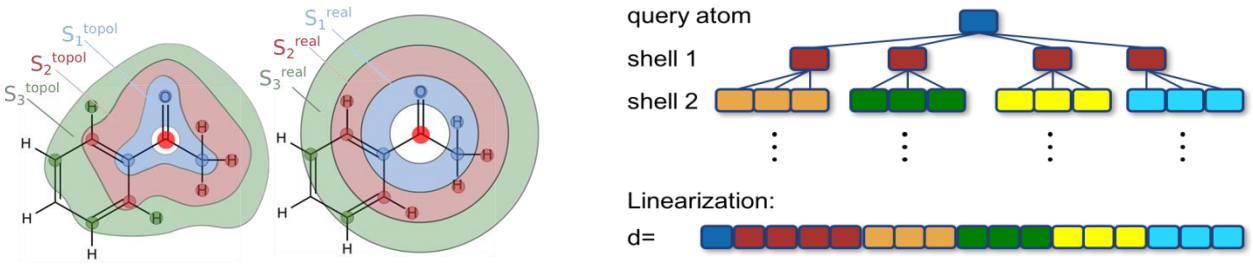


Figure 5: Anisotropic circular descriptors are created be either topological or 3D real-space binning of atomic properties like e.g. atomic charges (left) and mapping to a linear vector (right).

We present the model performances for several types of atomic descriptor vectors[3] computed for the HBA atoms using internal cross-validation. As the first result, we report a Gaussian Process Regression model performance on the pK_{BHX} monofunctional molecules subset, using a charge radial distribution function descriptor vector. The RMSE score for this preliminary test, obtained by 10-fold internal cross-validation, is equivalent to 0.60 ± 0.27 kcal mol^{-1}.

We further report on the quantum-mechanical calculation of donor-acceptor interaction energies with the aim to extend the dataset to larger, more complex and functionalized molecules.

1. Laurence, C.; Brameld, K. A.; Graton, J.; Le Questel, J.-Y.; Renault, E. The pK_{BHX} database: Toward a better understanding of hydrogen-bond basicity for medicinal chemists, *J. Med. Chem.* **2009**, *52*, 4073-4086.
2. Finkelmann, A. R.; Göller, A. H.; Schneider, G. Robust molecular representations for modelling and design derived from atomic partial charges. *Chem. Commun.* **2016**, *52*, 681-684.
3. Finkelmann, A.R.; Göller, A.H.; Schneider, G. Site of metabolism prediction based on ab initio derived atom representations. *ChemMedChem*, **2017**, *12*, 606-612.

P-36: International Chemical Identifier for Reactions (RInChI)
The key to effectively managing reaction databases

Gerd Blanke[1], Jonathan Goodman[2], Günter Grethe[3], Hans Kraut[4]

[1] StructurePendium Technologies GmbH, Essen, Germany, [2] University of Cambridge, Department of Chemistry, Cambridge, UK, [3] Dr. Guenter Grethe, Poway, CA 92064, US, [4] InfoChem Gesellschaft für chemische Information mbH, Munich, Germany

The Reaction-InChI (RInChI) extends the idea of the InChI, which provides a unique descriptor of molecular structures, towards reactions. Prototype versions of the RInChI supported by the IUPAC and the University of Cambridge have been available since 2011. The first official release (RInChI-V1.00), funded by the InChI Trust îs now available for download (http://www.inchi-trust.org/downloads/). This release defines the format and generates hashed representations (RInChIKeys) suitable for database and web operations. This talk will introduce the various RInChI representations and reports results of our work with reaction databases indexed by RInChI to demonstrate how RInChI may facilitate the manipulation and analysis of reaction data and will provide information how the RInChI will be further developed.

P-38: Characterizing Somatic Cancer Mutations in GPCRs

B.J. Bongers[1], X. Wang[1], X. Liu[1], H.W.T van Vlijmen[1,2], K. Ye[3], L.H. Heitman[1], A.P. IJzerman[1], G.J.P van Westen[1].

[1] Division of Drug Discovery and Safety, Leiden, The Netherlands,[2] Janssen Pharmaceutica NV, Beerse, Belgium, [3] Xi'an Jiaotong University, Xi'an, China

G Protein-Coupled Receptors (GPCRs) have recently gained interest as the second most mutated class of proteins in the context of cancer[1]. One of the key problems in cancer research remains the distinction between passenger and efficacious driver mutations. In the case of GPCRs, two decades of extensive research and the availability of crystal structures have led to insights into protein function and protein-ligand recognition which we aim to exploit.

Combining data from ChEMBL[2] and the BROAD institute[3] we constructed a dataset containing 9,144 patient samples data of 38 different cancer types. Subsequently, entropy analysis was performed to observe mutation prevalence in GPCRs, specifically missense mutations, and compare these to a control set based on the 1000 genomes dataset, which contained 2,504 samples [4,5].

First and foremost our analysis was able to retrieve and prioritize previously identified relevant GPCRs in a cancer context such as the Frizzled receptors and metabotropic glutamate receptors[6]. However, we were also able to gain new insights. More functional mutations were found in the TCGA data: Intracellular loops and transmembrane domains 3 and 6 (TMs) are most intensively mutated across GPCRs. Conserved residues in well-known motifs (such as the 'DRY' motif) are enriched for mutations. More specifically, residues flanking the highly-conserved and essential residues have a higher chance of mutation.

Secondly, we observe mutations to have a relatively low prevalence in the Class A GPCR ligand recognition site, opening the door for target modulation with small molecules. Small peptide receptors from both Class A and B GPCRs, such as neuropeptide receptors and angiotensin receptors show a large mutation rate compared to receptors recognizing small(er) molecules.

Thirdly, physicochemical changes resulting from these somatic mutations are on average neutral in the 1000 genomes set, whereas a difference is observed in the BROAD set.

Finally, several GPCRs are selected for follow-up experimental research to determine both the effect of mutations on GPCR function and the effect of different function on cell growth.

1. O'Hayre, M. *et al.* The emerging mutational landscape of G proteins and G-protein-coupled receptors in cancer. *Nat. Rev. Cancer* **13,** 412–424 (2013).
2. Bento, a. P. *et al.* The ChEMBL bioactivity database: An update. *Nucleic Acids Res.* **42,** D1083-90 (2014).
3. Broad Institute of MIT and Harvard. Firehose 2015_11_01 run. (2015). doi:10.7908/C1571BB1
4. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

5. Ye, K., Lameijer, E.-W. M., Beukers, M. W. & Ijzerman, A. P. A Two-Entropies Analysis to Identify Functional Positions in the Transmembrane Region of Class A G Protein-Coupled Receptors. doi:10.1002/prot.20899
6. Bar-Shavit, R. *et al.* G Protein-Coupled Receptors in Cancer. *Int. J. Mol. Sci.* **17,** 1320 (2016).

P-40: A Novel Approach to Assign Absolute Configuration Using Vibrational Circular Dichroism

Lennard Böselt [1], Dominik Sidler[1], Tobias Kittelmann [2], Jürgen Stohner [2], Sereina Riniker [1]

[1] Laboratory of Physical Chemistry, ETH Zurich, Vladimir-Prelog-Weg 2, 8093 Zurich, Switzerland,
[2] Institute of Chemistry and Biotechnology, Zurich University of Applied Sciences, Einsiedlerstrasse 31, 8820 Wädenswil, Switzerland

Vibrational circular dichroism (VCD) spectroscopy is a technique sensitive to the chirality of a molecule.[1] As VCD requires only a solution of the compound in question for measurement, it is an attractive alternative to single crystal X-ray diffraction. The interpretation of measured VCD spectra and thus the assignment of the absolute configuration rely on *ab initio* quantum-mechanical (QM) calculations.[1,2] For conformationally rigid molecules, the gas-phase QM calculations are straightforward and the calculated spectra agree well with the measured ones. However, for flexible molecules it becomes challenging to estimate the correct conformational ensemble. In addition, the large number of conformers, which may need to be considered, increases the computational cost.

In this work, we developed and evaluated a VCD sequence alignment (VSA) algorithm to match calculated and measured VCD spectra and assign the absolute configuration. The VSA algorithm has been parametrized on a set of rigid molecules, and tested on a set of flexible drug molecules taken from Ref. [3]. Using the VSA algorithm we obtained a success rate of 100% for predicting the correct absolute configuration. Furthermore, the number of relevant conformers for which a VCD spectrum has to be calculated can be reduced, lowering the computational cost of the approach substantially. A simple conformational search was found sufficient to obtain the relevant conformers. We anticipate that our approach will help research groups to determine the absolute configuration of chiral molecules in a robust and efficient manner.

1. Stephens, P. J.; Lowe, M. A. Vibrational Circular Dichroism. *Ann. Rev. Phys. Chem.* **1985**, 36, 213-241.
2. Magyarfalvi, G.; Tarczay, G.; Vass, E. Vibrational Circular Dichroism. *WILEY Comp. Mol. Sci*. **2011**, 1, 403-425.
3. Sherer, E. C.; Lee, C. H.; Shpungin, J.; Cuff, J. F.; Da, C.; Ball, R.; Bach, R.; Crespo, A.; Gong, X.; Welch, C. J. Systematic Approach to Conformational Sampling for Assigning Absolute Configuration Using Vibrational Circular Dichroism. *J. Med. Chem.* **2014**, 57, 477494.

P-42: A Novel Search Engine and Application for Very Large Chemistry Database Mining

R Brown, J Hussain, G Bravi & M Hartshorn

Dotmatics, Bishop's Stortford, UK

Chemical structure searching is an important technique within small molecule drug discovery. It is used to find available analogues to expand the SAR around a biological active as well as to identify appropriate reagents for compound synthesis. Modern search systems typically provide a few search types including exact structure, substructure and similarity searching and almost all provide these search services through an Oracle cartridge.

There are two challenges inherent in providing these types of workflows to chemistry teams

1. Building, maintaining and distributing very large (10M+) databases of compounds such as screening collections (e.g eMolecules) or public data (e.g. SureChEMBL), with up to date data.

2. Providing a range of search types to allow important tasks such as lead hopping or SAR expansion in a user interface appropriate for end user chemists rather than power users such as cheminformaticians and modellers.

In this talk we will show how we have addressed both problems: the first with a novel search engine (Minpoint) and the second with an application (Chemselector) that exploits that technology

Following early innovation in cheminformatics research, cartridges became the norm for database searching in the 1990s and since then the pace of innovation in this field has slowed dramatically. However, cartridges have definite limitations especially when needing to build and distribute very large databases on a regular basis

Minpoint is a novel search technology designed for very fast search performance – a substructure search again 10+ million structures can be run on a standard laptop. The search performance means almost all searches can be performed interactively as a structure is being drawn. The high performance can be utilized to help deal with tautomers when searching. Additionally, searching for matched molecular pairs of a query molecule in a large database becomes possible. The talk will also show different search modes that make it easier for chemists to formulate complex substructure queries and to select appropriate search techniques for a range of tasks including SAR expansion or lead hopping.

Minpoint does not rely on an Oracle cartridge for searching, instead holding structures and indexes as files on an application server. The scalability of this approach to very large databases is important since many tasks such as compound or reagent searching, or searching public datasets of patents, require the datasets that would be prohibitively expensive to build, maintain and distribute as Oracle databases. In the talk we will discuss the implementation of the Minpoint technology and its advantages for these type of applications

The powerful and often complex techniques employed when performing a chemical structure search mean designing an appropriate interface is paramount if the system is to be used by non-computational experts such as bench chemists. ChemSelector pulls together several powerful search methods in an interface appropriate for most bench chemists to use and the talk will discuss aspects of interface design and user experience that make it appropriate for that user community.

P-44: Designing of a "drug-like" natural compound library for secondary metabolites collected from the African flora.

Conrad V. Simoben[1], Fidele Ntie-Kang[1,2], Wolfgang Sippl[1]

[1]Institut für Pharmazie, Martin-Luther University of Halle-Wittenberg, Halle, Germany,[2]Department of Chemistry, University of Buea, Buea, Cameroon.

With the continuous search for new drugs to combat diseases, a topic of interest to medicinal chemist researchers is the search for new active compounds containing different core structures. Medicinal plants represent a potential source for the search of these new scaffolds. The criteria for choosing a particular natural product for studies were either based on the pre-existing traditional use of the plant in therapy (ethnobotanical knowledge) or the search for structurally related molecules using known pharmacologically active agents.[1] The African continent is very rich in biodiversity and some of the medicinal plants growing on the continent have been used by its local populations in traditional preparations for the treatment of several ailments. One of our research aims is to make available (freely online) the current knowledge on ethnobotanical uses of the medicinal plants as well as the three dimensional (3D) structures, physico-chemical properties and measured activities of the compounds isolated from medicinal plants collected from the African continent, with the view of assisting in the drug discovery process (http://african-compounds.org/). Previous surveys of the African flora, show that this part of the world could be a huge repository of bioactive natural products with diverse scaffolds and activities.[2-4]As a continuation of our ongoing database projects, we herein present a collection of ~2000 compounds isolated from East Africa (EA). Information about the said compounds was assembled from natural product journals and local African journals, as well as from M.Sc. and Ph.D. theses in African university libraries. These compounds were isolated mainly from about 300 medicinal plant species belonging to 60 families, harvested from EA and commonly used in the treatment of a variety of ailments. A majority of compounds reported were alkaloids, flavonoids, quinones, steroids and terpenoids. Computed physicochemical properties which are often relevant to predict pharmacokinetic and pharmacodynamic activities for compounds in this collection have been included.

1. Harvey, A. L.;Edrada-Ebel. R.; Quinn, R. J. The re-emergence of natural products for drug discovery in the genomics era. *Nat. Rev. Drug Dis.***2015**, 14(2), 111–129.

2. Ntie-Kang,F.;Telukunta, K. K.;Döring, K.;Simoben,C. V.;Moumbock,A. F. A.;Malange,Y. I.;Njume, L. E.;Yong,J. N.;Sippl,W.;Günther,S.; NANPDB: A Resource for Natural Products from Northern African Sources. *J. Nat. Prod*. **2017**, 80(7). 2067-2076.
3. Onguéné, P. A.;Simoben,C. V.;Fotso,G. W.; Andrae-Marobela, K.;Khalid, S. A.;Ngadgui, B. T.;Mbaze, L. M.;Ntie-Kang,F.*In silico* toxicity profiling of natural product compound libraries from African flora with anti-malarial and anti-HIV properties. *Comput. Biol. Chem*. **2017**, https://doi.org/10.1016/j.compbiolchem.2017.12.002
4. Ntie-Kang,F.;Nwodo, J. N.; Ibezim, A.; Simoben,C. V.;Karaman, B.; Ngwa, V. N.; Sippl, W.; Adikwu, M. U.; Mbaze, L. M. Molecular Modeling of Potential Anticancer Agents from African Medicinal Plants. *J. Chem. Inf. Model*. **2014**, 54(9), 2433-2450.

P-46: mmpdb: A Matched Molecular Pair Platform for Large Multi-Property Datasets

A. Dalke [1], C. Kramer [2], J. Hert[2]

[1] Dalke Scientific, Trollhättan, Sweden, [2] Roche Innovation Center, Basel, Switzerland

Matched Molecular Pair (MMP) analysis enables the automated and systematic compilation of medicinal chemistry rules from compound/property datasets. MMPDB is a new MMP platform to create, compile, store, retrieve and use MMP rules. MMPDB is suitable for the large datasets typically found in pharmaceutical and agrochemical companies, and provides new algorithms for fragment canonicalization and stereochemistry handling. The platform is implemented in Python using the RDKit toolkit[1] and is available as open source from https://github.com/rdkit/mmpdb.

The core molecular match algorithm is derived from the fragment-and-index approach of Hussain and Rea.[2] Structures are fragmented into a constant part and a variable part. The canonicalized constant is used as an index to find matching pairs of variable parts. Our new algorithm generates canonical SMILES for both the constant and variable parts, resulting in a canonical transformation description and improved analysis performance. It also handles stereochemistry through a process called up-enumeration to identify pairs between structures with partially specified stereochemistry.

MMP rules are highly dependent on the local environment around transformations.[3] A transformation which substitutes a hydrogen atom in a carboxylic acid with a methyl group, for example, results in different molecular property changes than the same substitution in an aliphatic chain. For each fragmentation record we include information about the circular environment around the attachment points, for each radius up to 5 bonds away, stored as a SHA256 hash. This provides an easy and effective means to stratify the transformation data without directly revealing its chemical structure, which may make it easier for organizations to share MMP data with others.

MMPDB is a command-line tool which can fragment and index a set of compounds to identify matched molecular pairs and store them, along with physical property or activity data, in a SQLite relational database. It implements two analysis features for ADMET and physico-chemical MMPA. The “transform” analysis helps identify the MMP transformations which may result in a structure with improved properties. The "predict" analysis estimates the property change between two molecules, typically an existing compound and a virtual one.

MMPDB is built with large corporate datasets in mind. The fragmentation step can reuse information from a previous run, if the structures haven't changed, and the fragmentation and parts of the analysis methods are parallelized. Using a benchmark dataset with the 20,267 compounds from ChEMBL with CYP3A4 or hERG data, a transform analysis of sofosbuvir takes 51 seconds to generate 1620 novel compounds. Most of that time is spent in startup overhead and random-access database seeks on a rotating drive; a web-service version backed by a RAM drive takes only 1.7 seconds. A predict analysis between sofosbuvir and p-Fluoro-phenyl-sofosbuvir with hERG as the query target property completes in 17 seconds, or 1.4 seconds through a web service.

1. Landrum, G. RDKit: Open-source cheminformatics **2006** http://rdkit.org/
2. Hussain, J.; Rea, C. Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets. *J. Chem. Inf. Model*. **2010**, 50 (3), 339–348.
3. Papadatos, G.; Alkarouri, M.; Gillet, V. J.; Willett, P.; Kadirkamanathan, V.; Luscombe, C. N.; Bravi, G.; Richmond, N. J.; Pickett, S. D.; Hussain, J.; Pritchard, J. M.; Cooper, A. W. J.; Macdonald, S. J. F. Lead Optimization Using Matched Molecular Pairs: Inclusion of Contextual Information for Enhanced Prediction of HERG Inhibition, Solubility, and Lipophilicity. *J. Chem. Inf. Model.* **2010**, 50 (10), 1872–1886

P-48: 3D-e-Chem: Structural Cheminformatics Workflows for Computer-Aided Drug Discovery

AJ Kooistra[1,2], M Vass[1], R McGuire[2,3], I de Esch[1], G Vriend[3], L Ridder[4], S Verhoeven[4], C de Graaf[1]

[1]Division of Medicinal Chemistry, Vrije Universiteit Amsterdam, The Netherlands; [2]Centre for Molecular and Biomolecular Informatics (CMBI) Radboudumc, Nijmegen, The Netherlands; [3]BioAxis Research, Oss, The Netherlands; [4] Netherlands eScience Center, Amsterdam, The Netherlands

eScience technologies are needed to process the information available in many heterogeneous types of protein-ligand interaction data and to capture these data into models that enable the design of efficacious and safe medicines. The 3D-e-Chem consortium has developed scientific KNIME tools and workflows[1-4] that enable the integration of chemical, pharmacological, and structural information, including: i) structure-based bioactivity data mapping, ii) structure-based identification of scaffold replacement strategies for ligand design, iii) ligand-based target prediction, iv) protein sequence-based binding site identification and ligand repurposing, v) structure-based pharmacophore comparison for ligand repurposing across protein families, and vi) *in silico* metabolic profiling. The modular setup of the workflows and the use of well-established standards allows the re-use of these protocols and facilitates the design of customized computer-aided drug discovery workflows.

1. McGuire R, Verhoeven S, Vass M, Vriend G, de Esch IJ, Lusher SJ, Leurs R, Ridder L, Kooistra AJ, Ritschel T, de Graaf C. 3D-e-Chem-VM: Structural Cheminformatics Research Infrastructure in a Freely Available Virtual Machine. *J Chem Inf Model* **2017**, 57, 115-121.
2. Kooistra AJ, Vass M, McGuire R, Leurs R, de Esch IJ, Vriend G, Verhoeven S, de Graaf C. 3D-e-Chem: Structural Cheminformatics Workflows for Computer-Aided Drug Discovery. *ChemMedChem* 2018. doi: 10.1002/cmdc.201700754
3. https://3d-e-chem.github.io
4. https://www.knime.com/3d-e-chem-nodes-for-knime

P-50: Analysis and inference within the molecular space: A visual approach using NAMS and multidimensional scaling

Samina Kausar[1,2], Andre O. Falcao[1,2*]

[1] LASIGE, Department of Informatics, Faculty of Sciences, University of Lisboa, Lisboa, Portugal,
[2] BioISI - Biosystems & Integrative Sciences Institute, Faculty of Sciences, University of Lisboa, Lisboa, Portugal

Molecular similarity quantification is a central task in cheminformatics and has numerous applications in drug discovery methods. The core concept of molecular similarity is based on Similar Property Principle, which states that similar compounds should have similar properties[1]. Irrespective of the specific analysis, molecular similarity values largely depend upon molecular structural representation and similarity coefficient. However, similarity quantification must be consistent for reliable application of molecular similarity in all situations.

Representing molecular similarity into a new reference system has been used previously[2, 3]. The basic idea is to capture the measured molecular distances according to any method and try to represent molecules in a reduced reference space for analysis and visualization. Many dimensionality reduction methods are extant, and some of the more popular are Principal Coordinates Analysis (PCooA), Kruskal Multidimensional Scaling (MDS) or Sammon mapping[4]. These constructs can be used to build a classification model for QSAR problems where molecules can be separated in two or more classes. The procedure to create such a model then can be described in the following steps. First, a full similarity matrix of a molecular dataset is computed. Secondly, similarities are transformed into distances and projected into a 2-Dimensional (2D) space using one of the above mapping functions. Finally, the probabilities of this reduced space are computed using a 2D kernel density estimation function to produce a probability map of a projected molecule for all classes. This 2D map can visually produce information as to where the more promising regions of the molecular space are located and can as well serve as a classification model. By projecting new molecules using the same transformation constructed, it is possible to attribute to any molecule the probability of it belonging to either class.

The designed methodology was validated using four human proteins (Table 1), retrieved from ChEMBL. To compute similarities, the Non-Contiguous Atom Molecular Similarity (NAMS)[5] was used as it has shown to be more precise than fingerprints-based approaches. The selected datasets were curated and divided into two classes using a cut-off of activity value (Ki) to separate highly active molecules (Ki<=10.0) as positives and less active and non-active molecules (Ki>10.0) as negatives. All datasets were randomly split into training and test sets. The 2D maps were generated only with the training sets and the class probability maps calculated for the 3 mapping functions (PCooA, Kruskal and Sammon). The 2D coordinates of the test molecules were computed into the new reference space using a data transformation matrix and the class probabilities were calculated. Each model's performance was assessed using Area Under Curve (AUC) and the Matthews Coefficient Correlation (MCC). AUC testing results range from 0.78 to 0.98 (Table 2), suggesting this methodology can validly capture the complexities of the molecular activity space. All three mapping functions provided generally good results with a slight more positive outcome for PCooA. Charts were produced for all four problems to make evident the visual nature of these models, aiding in the identification of the most promising molecular active regions (Figure 1).

Table 1: Dataset description

TARGET PROTEIN	UNIPROT ID.	TRAINING SET		TEST SET	
		Positives	Negatives	Positives	Negatives
Sigma non-opioid intracellular receptor 1 (Sigma1R)	Q99720	46	135	10	35
Histamine H1 receptor (HRH1)	P35367	184	783	46	195
Potassium voltage-gated channel subfamily H member 2 (HERG)	Q12809	39	1142	12	283
D(1B) dopamine receptor (DRD5)	P21918	41	231	5	62

Table 2: Results on validation set ((*) – best model)

TARGET PROTEIN	PCOOA		MDS		SAMMON	
	AUC	MCC	AUC	MCC	AUC	MCC
Sigma non-opioid intracellular receptor 1 (Sigma1R)	0.86(*)	0.63	0.80	0.60	0.79	0.55
Histamine H1 receptor (HRH1)	0.80	0.43	0.83(*)	0.42	0.80	0.35
Potassium voltage-gated channel subfamily H member 2 (HERG)	0.80	0.18	0.78	0.23	0.81(*)	0.29
D(1B) dopamine receptor (DRD5)	0.98(*)	0.72	0.85	0.33	0.81	0.42

1. Johnson, M.A.; Maggiora, G.M. Concepts and Applications of Molecular Similarity. Wiley, New York, **1990**
2. Teixeira, A.L.; Falcao, A.O. Structural similarity based kriging for quantitative structure activity and property relationship modeling. *J. Chem Inf Mod*, **2014**, 54, 1833–1849
3. Mahendra, A. Ricardo, V.; Daniel, P.; Josep, A.; Jean-Louis, R. Chemical Space: Big Data Challenge for Molecular Diversity. *Chimia*, **2017**, 71, 661-666
4. Venables, W. N.; Ripley, B. D. Modern Applied Statistics with S. Fourth Edition. Springer, New York, **2002**
5. Teixeira, A.L.; Falcao, A.O. Noncontiguous atom matching structural similarity function. *J. Chem Inf Mod*, **2013**, 53, 2511–2524.

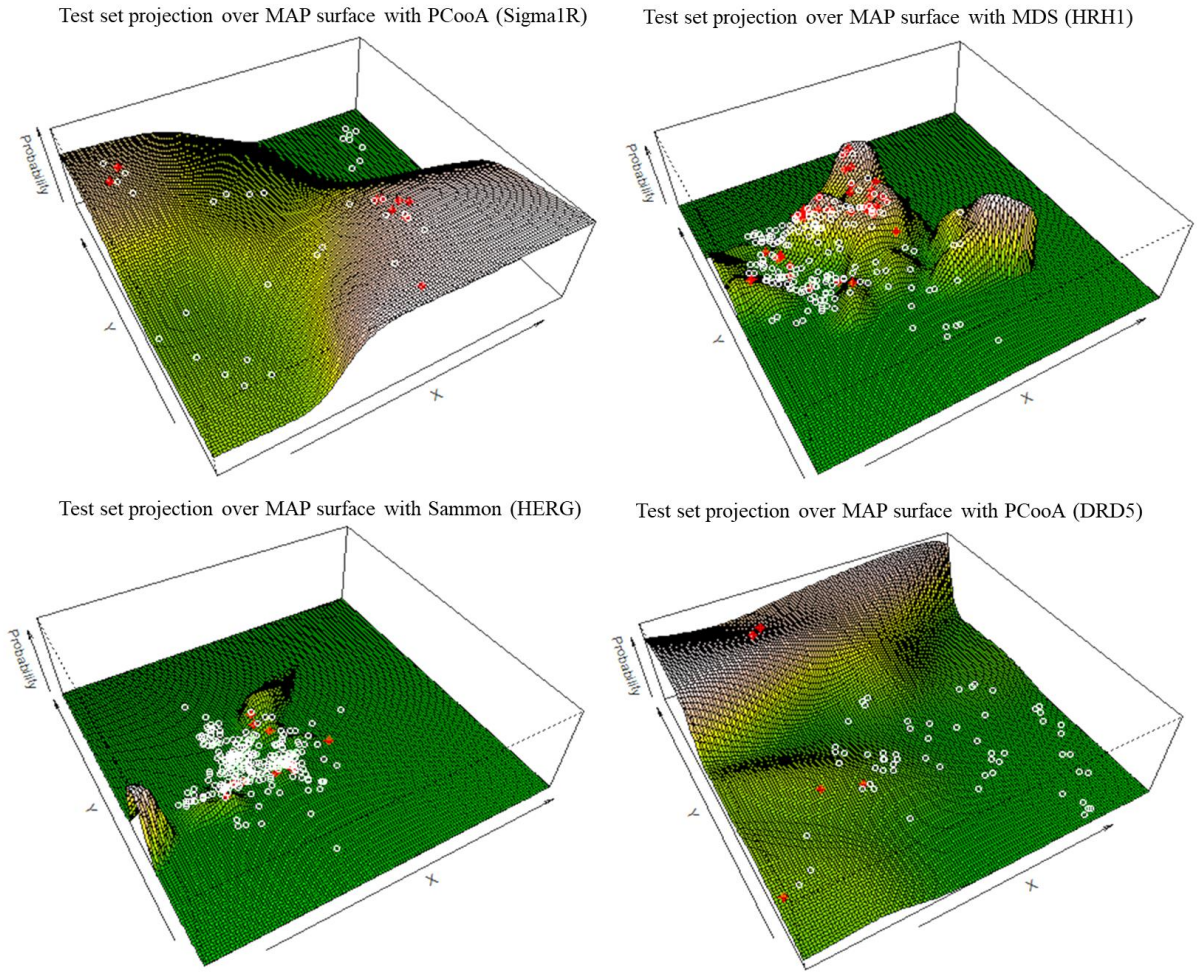


Figure 1: Test set projection over MAP surface of selected models with highest performance (red – circles are positives, white are negatives).

P-52: Reaction Classification by Reaction Vectors

G. Ghiandoni [1], B. Chen [2], M. J. Bodkin [3], V. J. Gillet. [1]

[1] Information School, University of Sheffield, Regent Court, 211 Portobello, Sheffield, S1 4DP, United Kingdom, [2] Chemistry Department, University of Sheffield, Dainton Building, Brook Hill, Sheffield, S3 7HF, United Kingdom, [3] Evotec (U.K.) Ltd, 114 Innovation Drive, Milton Park, Abingdon, OX14 4RZ, United Kingdom

De novo molecular design is the branch of chemoinformatics concerned with the rational design of tailored structures from scratch, combining desired pharmacodynamic and pharmacokinetic properties, in order to boost the identification of new chemical entities (NCEs).[1]

The size of potential drug-like chemical space has been estimated at 10^{60} molecules, according to Lipinski's Rule-of-Five (RO5), which determines roughly if a certain molecule may possess suitable characteristics to be an orally active drug.[2] However, the recent introduction of *de novo* design tools for the design of synthetically accessible compounds has resulted in a significant reduction of this number, earning its way among the fragment-based design methods.[3] In particular, reaction vectors, which are descriptors that incorporate the changes that occur in chemical reactions, have been implemented in a structure generation tool to produce new molecules by transforming a set of selected reactants with the use of a database of reaction examples.[4,5] Here, we propose a machine learning model for reaction classification, entirely based on the concept of reaction vector and specifically trained towards the

prediction of 336 medicinal chemistry reaction classes, that would offer fine-grained selection of products from sets of specific reaction classes. We then investigate how the reaction classification tool can be used to enhance *de novo* design through a more effective and direct exploitation of specific classes of reactions. For instance, reaction classification could be applied in order to group reactions into categories, such as bond formations or functional group conversions, thus enabling the use of specific reaction classes; or facilitating medicinal chemists in the direct selection of their favourite reactions from ELNs (electronic laboratory notebooks).

1. Hartenfeller, M. & Schneider, G., Enabling future drug discovery by de novo design. *Wiley Interdisciplinary Reviews: Computational Molecular Science*. **2011**, 1, 742-759.
2. Bohacek, R.S., McMartin, C. & Guida, W.C., The art and practice of structure-based drug design: A molecular modeling perspective. *Medicinal Research Reviews*. **1996**, 16, 3-50.
3. Hartenfeller, M., Reaction-Driven De Novo Design: a Keystone for Automated Design of Target Family-Oriented Libraries. In *De novo Molecular Design*; Schneider, G., Ed.; WileyVCH Verlag GmbH & Co. KGaA, Weinheim, Germany, **2013**; 245-266.
4. Patel, H., Bodkin, M. J., Chen, B. & Gillet, V. J., Knowledge-Based Approach to De Novo Design Using Reaction Vectors. *Journal of Chemical Information and Modeling*. **2009**, 49, 1163-1184.
5. Hristozov, D., Bodkin, M., Chen, B., Patel, H., & Gillet, V. J., Validation of Reaction Vectors for de Novo Design. *Library Design, Search Methods, and Applications of Fragment-Based Drug Design*, **2011**, 29-43.

P-54: Tautomeric Equilibria: Modeling and Visualization.

Marta Glavatskikh [1,2], Timur Madzhidov [2], Igor Baskin [3], Dragos Horvath [1], Ramil Nugmanov [2], Timur Gimadiev [2], Gilles Marcou [1] Alexandre Varnek [1*]

[1] University of Strasbourg, France, [2] Federal University of Kazan, Russia, [3] Lomonosov Moscow State University, Moscow, Russia

The existing tools for the prediction of ratio of tautomers are predominantly based on the calculation of pKa values of related tautomer. This may significantly affect the accuracy, especially, if the errors of the pKa predictions are comparable with the difference of tautomers' pKa values. Moreover, this calculation is usually restricted by aqueous solution and hence not applicable for other media. Here the prediction of tautomeric equilibria is performed directly for the equilibrium constant (logK) for the reactions proceeding in aqueous and organic solutions or their mixtures.

The models were built on a data set of 697 reactions of 10 tautomeric classes, for which logK values were measured in different solvents and at different temperatures[1]. Support Vector Machine[2] (SVM) and Generative Topographic Mapping[3] (GTM) were used as machine-learning methods. The structure of tautomers has been encoded by ISIDA fragments[4] whereas conditions were accounted for physico-chemical parameters of solvent and inverse temperature. Both SVM and GTM models perform well in cross-validation (RMSE *(5CV)*=0.63-0.67, R2 *(5CV)*=0.82-0.84). Validation of these models on two external test sets, either included the transformations under new reaction conditions (test 1) or new structures (test 2), lead to reasonable statistical parameters (RMSE=0.59 and 1.96, R2=0.62 and 0.65). Large RMSE value for test 2 is explained by the fact that more than half of the compounds were out of the model's applicability domain. The consensus SVM model is publicly available on our web-server: https://cimm.kpfu.ru/development/predictor.

As it is illustrated on Figure 1, a GTM map well separates both different tautomeric classes and the same equilibria proceeding in different solvents.

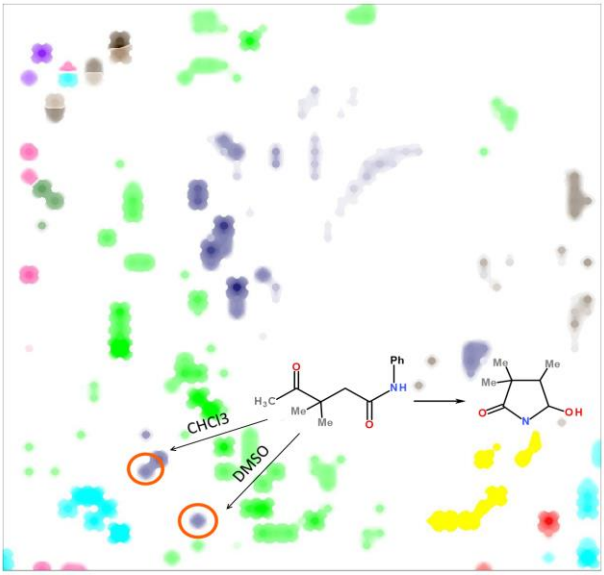


Figure 1. GTM map built for 697 tautomeric equilibria. The color code corresponds to 10 tautomeric classes. Selected data points correspond to the same equilibrium studied in CHCl3 (logK= -0.49) and DMSO (logK= 0.62).

1. Palm, V. A., Tables of Rate and Equilibrium Constants of Heterolytic Organic Reactions. VINITI: Moscow, 1978.
2. Chang, C.-C.; Lin, C.-J., LIBSVM: A library for support vector machines. ACM Trans. Intell. Syst. Technol. **2011,** 2 (3), 1-27.
3. Gaspar, H. A.; Marcou, G.; Horvath, D.; Arault, A.; Lozano, S.; Vayer, P.; Varnek, A., Generative Topographic Mapping-Based Classification Models and Their Applicability Domain: Application to the Biopharmaceutics Drug Disposition Classification System (BDDCS). J. Chem Inf. Model. **2013,** 53 (12), 3318-3325.
4. Varnek, A.; Fourches, D.; Horvath, D.; Klimchuk, O.; Gaudin, C.; Vayer, P.; Solov'ev, V.; Hoonakker, F.; Tetko, I. V.; Marcou, G., ISIDA - Platform for virtual screening based on fragment and pharmacophoric descriptors. Curr. Comput.-Aided Drug Des. **2008,** 4 (3), 191-198.

P-56: Artificial Intelligence in Medicinal Chemistry – Current Status at AstraZeneca

T. Kogej[1], H. Chen[1], C. Tyrchan[2], O. Engkvist[1], C. Green[3]

[1]Hit Discovery, Discovery Sciences, Innovative Medicines and Early Development Biotech Unit, AstraZeneca R&D Gothenburg, Sweden; [2]Respiratory, Inflammation and Autoimmunity, Innovative Medicines and Early Development Biotech Unit, AstraZeneca R&D, Gothenburg, Sweden; [3]Compound Management, Discovery Sciences, Innovative Medicines and Early Development Biotech Unit, AstraZeneca R&D, Cambridge, UK

Artificial intelligence (AI) is gaining importance in our modern society. Among AI algorithms, recurrent neural networks (RNNs) have emerged as powerful generative models for various applications such as natural language processing, images, video, music and speech recognition. Inspired by this development we started to use RNN recently for molecular de novo design[1,2]. Our current AI platform consists of not only the molecular de novo design component but also a set of novel machine learning models to score the de novo generated molecules according to potency, selectivity and ADME properties.

1. M. H. S. Segler, T. Kogej, C. Tyrchan, M. P. Waller, *ACS Cent. Sci*., **2018**, *4*, 120–131.
2. M. Olivecrona, T. Blaschke, O. Engkvist, H. Chen, *J. Cheminf*. **2017**, *9*, 48-62.

P-58: Compact descriptor sets for automatic annotation of natural products in large databases by pairwise variable screening

M. Kretzschmar [1], K. Baumann [1]

[1] *Institute of Medicinal and Pharmaceutical Chemistry, Technische Universität Braunschweig, Beethovenstraße 55, 38106 Braunschweig, Germany*

Natural products (NP), especially the subgroup of secondary metabolites, are of much interest in pharmaceutical research due to their highly optimized binding mechanisms with their macromolecular targets.

As NPs and NP-like molecules are much more diverse and cover a different chemical space than pure synthetics, different approaches for these molecular classes are needed for applications like target prediction. Unfortunately, large compound collections like ChEMBL lack an NP-annotation. Previous work[1,2] addressed this issue with a Naïve-Bayesian model based on a molecular fingerprint generating an (exhaustive) set of substructures. This study aims to identify NPs using a small set of easy-to-interpret chemical descriptors via machine-learning techniques.

NPs (approx. 220k) and synthetics (approx. 320k) were collected from several databases, carefully curated and split into training and test sets for validation. Employed descriptor sets to encode the compounds include manually selected features tailored towards NPs, molecular shape descriptors (WHIM), molecular fragments and other well-established descriptor sets like Dragon descriptors and MACCS keys. A pairwise variable screening was performed to identify those descriptors which show the largest differences in class means. Model performance was evaluated for each single descriptor set with varying variable numbers as well as for interesting combinations thereof. Finally, selected models were used to provide a predicted NP-annotation for the whole ChEMBL database.

It can be shown that even a very small set of descriptors in combination with a Random Forest Classifier is well capable of distinguishing between NPs and synthetics. The easy-to-interpret approach may help to explore the chemical space covered by NPs so that more specific cheminformatic applications could be developed for them.

1. Ertl, P.; Roggo, S.; Schuffenhauer, A. Natural Product-likeness score and Its Application for Prioritization of Compound Libraries. *J. Chem. Inf. Model.* **2008**, 48 (1), 68-74.
2. Jayaseelan, K. V.; Moreno, P.; Truszkowski, A.; Ertl, P.; Steinbeck, C. Natural product-likeness score revisited: an open-source, open-data implementation. *BMC Bioinf.* **2012**, 13, 106.

P-60: *De novo* drug-candidate molecule generation with generative adversarial networks

X. Liu[1], K. Ye[2], H. W. T. van Vlijmen[1, 3], A. P. IJzerman[1], G. J. P. van Westen[1].

[1] *Drug Discovery and Safety, Leiden Academic Center for Drug Research, Einsteinweg 55, Leiden, The Netherlands,* [2] *Omics and Omics informatics, Xi'an Jiaotong University, 28 Xianning W Rd, Xi'an, China,* [3] *Janssen Pharmaceutica NV, Beerse, Belgium.*

Over the last five years deep learning (DL) has progressed tremendously in both image recognition and natural language processing[1]. Based on these results DL has also been applied in cheminformatics to predict bioactivity[2]. A specific type of deep neural nets are generative adversarial networks (GANs)[3]. GANs were previously used for image generation and based on this SeqGAN[4] was constructed for sequence generation.

Here, we use SeqGAN to generate novel small molecules based on the SMILES format. This model contains two separate networks: a discriminator and a generator. Both were implemented with LSTM recurrent neural networks and trained simultaneously. In the reinforcement learning framework, the discriminator function is the reward function to measure whether the generated molecule has desired properties. The generator on the other hand is the policy function to determine which character to choose to stepwise construct a SMILES format string.

The dataset we used here contains two parts, the first one was the whole refined ChEMBL set as we used previously[2,5] to train the generator to learn the grammar of the SMILES format. The second one contained ligands that can bind the adenosine A_{2A} receptor. This data was used as the training set to train the discriminator.

Our proof of concept generated compounds that not only have a valid sequence following the SMILES format, but also possess predicted binding affinity for the A_{2A} receptor. Hence, we can enlarge chemical space for candidate drugs to search for the optimal molecular structure for further study. Follow up work includes selectivity optimization towards a single or multiple targets.

1. LeCun, Y., Y. Bengio, and G. Hinton, *Deep learning.* Nature, 2015. **521**(7553): p. 436-44.
2. Lenselink, E.B., et al., *Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set.* J Cheminform, 2017. **9**(1): p. 45.
3. Goodfellow, I.J., et al. *Generative Adversarial Networks*. ArXiv e-prints, 2014. **1406**.
4. Yu, L., et al. *SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient*. ArXiv e-prints, 2016. **1609**.
5. Gaulton, A., et al., *The ChEMBL database in 2017.* Nucleic Acids Res, 2017. **45**(D1): p. D945-D954.

P-62: The Need for Comprehensive Reaction Handling in SAVI and Beyond

M. Nicklaus [1], W. Ihlenfeldt [2], G. Blanke [3], P. Judson [4], V. Delannée [1]

[1] NCI, NIH, Frederick, USA, [2] Xemistry GmbH, Königstein, Germany, [3] StructurePendium Technologies GmbH, Essen, Germany, [4] Consultant, Harrogate, UK

With ever-increasing amounts of chemical data, fast and lossless chemical information processing is more important than ever. While small molecule representations have been an area of active research and significant advances – both public/open-source and commercial – during the past two decades, chemical reaction data have not received the same degree of attention. Reactions, while encompassing all the complexities of the chemical structures of the starting materials and products (plus possibly catalysts, solvents etc.), are yet more-complicated data structures.

Laboratory records registered in ELNs, reactions collected in large databases such as Reaxys and CASREACT, synthesis data submitted to FDA for drug ingredients etc. typically handle this wealth of information in a way that is targeted at the local needs of the software or organization, thus are neither comprehensive for all possible needs nor designed for data exchange.

We discuss the shortcomings of current reaction representations in the context of our Synthetically Accessible Virtual Inventory (SAVI) project (aimed at the creation in silico of 1 billion easily synthesizable molecules) as well as for general needs. The widely used RXN and RD file formats are not truly suited for comprehensive and semantic exchange of reaction data and cannot be used for searches and reaction comparison out of the box. The format of an RD file depends on the database it is exported from and varies accordingly. We briefly discuss inline formats such as RInChI and reaction SMILES in this context. To push for the development of a standardized, comprehensive, semantic reaction handling/representing format, we suggest criteria to be considered such as inclusion of reaction conditions and metadata, participating atoms/bonds information, representation of failed reactions, comprehensive and fine-grained handling of tautomerism. We discuss possible avenues to address these challenges.

P-64: Flavours in Aromaticity

M. Ott [1], D. Ponting [1], R. van Deursen [1]

[1] Lhasa Limited, Leeds, UK

In organic chemistry, the term aromaticity is used to describe a cyclic unsaturated structure that exhibits more stability and a different reactivity profile than a similar non-aromatic one. For the ring to be aromatic, its atoms must all be sp^2-hybridised to allow full delocalisation and the system should contain $4n + 2$ π-electrons ($n = 0, 1, 2, 3$; Hückel's rule). However, counting these electrons is less straightforward than it might seem when you start looking beyond the familiar group of benzene, pyrimidine, thiazole etc. For example, is 2-pyridinone aromatic, counting the carbonyl group for 0 electrons? While many compounds exhibit aromatic character to some extent, not all are as perfectly aromatic as benzene.

In order to adopt a flexible and useable approach to aromaticity, we chose to introduce the concept of a "degree of aromaticity", based on the HOMED approach.[1] The basic idea is that delocalisation causes double bonds to be "smeared" out, leading to shorter single bonds and longer double bonds, and this effect can be measured or calculated. Briefly, the calculations were performed as follows. A set of reference compounds for single, double and aromatic bond lengths for each of 29 potential pairs of atom types (*e.g.* for C-C the references were ethane, ethene and benzene respectively) were generated and optimised using DFT at the B3LYP/6-311G** level [2-6] of theory in NWChem,[7] and the bond lengths extracted. The structures from the PubChemQC project [8] (pubchemqc.riken.jp) were then downloaded, initially as a diverse test set of 36,000 and subsequently the entire database of around 4 million structures, and HOMED indices calculated by comparing the bond lengths in every ring system (using the Smallest Set of Smallest Rings, SSSR) with the reference compounds, then applying a filter based on Hückel's rule. The results indicated good performance in separating the "strongly" and "weakly" aromatic compounds. The method is both rapid and scaleable, however it does require either QM-optimised or crystal structures, since it is dependent on measuring the actual bond lengths in the ring.

Our ultimate aim is to be able to distinguish between "strongly" aromatic (*e.g.* pyridine), "weakly" aromatic (*e.g.* uracil), and non-aromatic (*e.g.* isocyanuric acid) structures. This categorisation allows us to describe chemical knowledge more accurately whilst not overstating the extent to which we can assess a degree of aromaticity.

1. Raczynska, E. D.; Hallman, M.; Kolczynska, K.; Stepniewski, T. M. On the harmonic oscillator model of electron delocalization (HOMED) index and its application to heteroatomic π-electron systems. *Symmetry* **2010**, 2, 1485-1509.
2. Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **1993**, 98, 648-5652.
3. Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **1988**, 37, 785-789.
4. Vosko, S. H.; Wilk, L; Nusair, M. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: A critical analysis. *Can. J. Phys.* **1980**, 58, 1200-1211.
5. Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields. *J. Phys. Chem.* **1994**, 98, 11623-11627.
6. Krishan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. Self-consistent molecular orbital methods. XX. A basis set for correlated wave functions. *J. Chem. Phys.* **1980**, 72, 650-654.
7. Valiev, M.; Bylaska, E. J.; Govind, N.; Kowalski, K.; Straatsma, T. P.; van Dam, H .J. J.; Wang, D.; Nieplocha, J.; Apra, E.; Windus, T. L.; de Jong, W. A. NWChem: A comprehensive and scalable open-source solution for large scale molecular simulations. *Comput. Phys. Commun.* **2010**, 181, 1477-1489.
8. Nakata, M.; Shimazaki, T. PubChemQC Project: A large-scale first-principles electronic structure database for data-driven chemistry. *J. Chem. Inf. Model.* **2017**, 57, 1300-1308.

P-66: Smooth Molecular Surfaces with Joined Marching Cubes

Thomas L. Sander

Idorsia Pharmaceutical Ltd.

The swift generation and visualisation of molecular surfaces is a crucial element of many cheminformatics and modelling applications.

One of the fastest algorithms to triangulate surface meshes from voxel data is the 'Marching Cubes' algorithm. In its original version some of the generated triangles are very tiny or skinny. Since these lead to rendering artefacts, there were some attempts to modify this algorithm such to avoid tiny and skinny triangles. These, however, increased complexity and caused significant performance losses.

We present a simple modification of the original algorithm that results in smoother surfaces without small and skinny triangles. This is achieved by buffering and merging triangle nodes of any previously processed voxel layer. This modification has little overhead, reduces the number of triangles by about 20 percent and results in much smoother surface renderings. To prove that the algorithm is robust and generalisable it was not only applied to molecules, but also to noisy MRI data. The source code is available as part of the DataWarrior open-source project. It contains classes to create a voxel density grid from 3-dimensional atom coordinates and classes to triangulate iso-value layers from voxel intensity data of any source.

P-68: Chemistry Identifier Mapping to Pathway Databases using Ontologies: Expanding metabolomics analysis in WikiPathways with ChEBI

D. Slenter[1], C. Evelo[1,2], E. Willighagen[1]

[1] Department if Bioinformatics - BiGCaT, Maastricht, Netherlands, [2] Maastricht Centre for Systems Biology - MaCSBio, Maastricht, Netherlands

Health research uses large scale (omics) methods to study the state of an individual, organs, and increasingly tissues and single cells. These methods can measure gene expression, epigenetic modification, and protein abundances. Metabolomics complement the aforementioned methods by studying the abundances of small molecular compounds in e.g. bodily fluids, tissue samples and breath.

Changes in metabolism are relevant for many diseases, such as metabolic diseases, hereditary diseases, various forms of cancers, and the symbiotic interaction of the gut microbiome and the (human) body. Pathway and network approaches are extensively used to integrate various data types and other information sources, in order to understand measurements and results in their biological context. Unfortunately, not all measured metabolites can be linked to metabolite identities present in biological pathway models, which make it more complicated to use metabolomics data in pathway and network analysis.

In order to overcome this intrinsic mismatch between metabolomics experiments and knowledge bases, we use the ontological information from ChEBI[1]. With this, we create additional mappings to metabolites in the pathway database WikiPathways[2]. With this approach, we can connect compounds classes (e.g. fatty acid, lipids), tautomers and/or charge states (e.g. ionisation into acid or base) to individual molecules in a data set. By applying this method on various publicly available datasets in the MetaboLights[3] repository, we want to estimate the increased mapping that chemical ontologies can provide.

1. Hastings J.; Owen G.; Dekker A.; Ennis M.; Kale N.; Muthukrishnan V.; Turner S.; Swainston N.; Mendes P.; Steinbeck C. ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucl. Acids Res. **2016**; 44: D1214–D1219.
2. Slenter D.; Kutmon M.; Hanspers K.; Riutta A.; Windsor J.; Nunes N.; Mélius J.; Cirillo E.; Coort S.; Digles D.; Ehrhart F.; Giesbertz P.; Kalafati M.; Martens M.; Miller R.; Nishida K.; Rieswijk L.; Waagmeester A.; Eijssen L.; Evelo C.; Pico A.; Willighagen E. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research, Nucl. Acids Res. **2018**; 46, D661–D667.
3. Haug K.; Salek R.; Conesa P.; Hastings J.; Matos P.; Rijnbeek M.; Mahendrakar T.; Williams M.; Neumann S.; Rocca-Serra P.; Maguire E.; González-Beltrán A.; Sansone S.; Griffin J.; Steinbeck C. MetaboLights-- an open-access general-purpose repository for metabolomics studies and associated meta-data. Nucl. Acids Res. **2013** 41, D781–D786.

P-70: Finding answers from chemical space extremely fast

A. Tarcsay[1], G. Imre [1], A. Volford [1]

[1] ChemAxon, Budapest, Hungary

The complex nature of chemical graphs offers an immense source of variability for drug designers to tackle optimization challenges along the project pathway towards candidates. The difficulty lies within the exploration of the chemical space either by chemical intuition of medicinal chemists or by using enabling technologies, like cheminformatics tools.

Real and virtual chemical spaces encompass broad scale of compound numbers and a vast potential to be exploited. An especially valuable sub-group is where measured data exists and stored, most commonly in relational databases. In our study both types, a very large compound collection and a medium size with extensive assay data were evaluated. As a read-out we used the cost associated with finding an answer for chemical questions, the search time.

In the first use-case, the aim was to suggest novel analogues of known drugs using the largest publicly available enumerated compound collection, the GDB-13 counting 977M unique entries. This collection was screened with ultra-fast similarity search technique, using a subset of marketed drugs, where ~4 sec elapsed search time was measured constantly on a commercially available server (EC2, r3.8xlarge) using standard 1k fingerprint. Top 100

most similar compounds were cross filtered with the database of exemplified structures from patents (SureChEMBL DB) to fetch novel moieties with higher tendency to be in freedom to operate space (Fig. 1.).

In the second part search performance on the entire data from ChEMBL DB was measured with three search types (duplicate, similarity and substructure) and joined queries. These joined queries represent complex questions asked from data warehouses in pharmaceutical industry, where performance is a key indicator due to massive load. The aim is to provide realistic speed statistics measured with chemical cartridge extending Oracle and the new generation engine running on PostgreSQL. Significant speed up was measured using the new search engine, especially on combined queries, where 100x speed up was achieved and median search time was in a range on ~100 milliseconds falling below the recognition time limit.

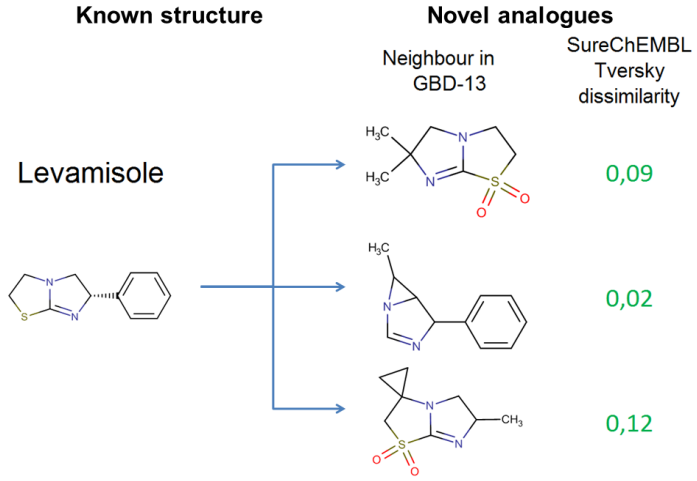


Figure 1.Example drug and its novel analogues identified from GDB-13. Tversky dissimilarity >0 rules out substructure match in SureChEMBL.

P-72: Structural Analysis of Protein Homomers – the Quest for Perfect Symmetry

Inbal Tuvi-Arad

Department of Natural Sciences, The Open University of Israel, Raanana, Israel

Symmetry has several advantages in the synthesis and function of protein homomers. It reduces synthetic errors, gives rise to faster oligomerization processes, increases the effectiveness of allosteric regulation, maximizes interaction between subunits and thus decreases the total energy, and in general contributes to the protein's stability. Yet, thermodynamic considerations and experimental conditions prevent proteins from achieving perfectly symmetric geometry. Here we present improved algorithms for estimating the level of symmetry of proteins by means of continuous symmetry measures. These are based on the Hungarian algorithm that solves the assignment problem in polynomial time. The amino acids sequence and the division into peptides is used to significantly reduce the size of the equivalent atoms groups and thus increase the speed and accuracy of the code. Analysis of the distortion levels of several sets of protein homomers extracted from the Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB), with various degrees of rotational symmetry will be presented. The new methodology launches the foundations for accurate, efficient and reliable large scale symmetry analysis of protein structure and oligomerization mechanisms.

P-74: Wikidata and Scholia as a hub linking chemical knowledge

E. Willighagen [1], D. Slenter [1], D. Mietchen [2], C. Evelo [1,3], F.Å. Nielsen [4]

[1] Department if Bioinformatics - BiGCaT, Maastricht University, Netherlands, [2] Data Science Institute, University of Virginia, Charlottesville, Virginia, USA, [3] Maastricht Centre for Systems Biology - MaCSBio, Maastricht University, Netherlands, [4] Cognitive Systems, DTU Compute, Technical University of Denmark, Denmark

Making chemical databases more FAIR (findable, accessible, interoperable, and reusable) benefits computational chemistry and cheminformatics. We here discuss Wikidata, a young sister project of Wikipedia but with one big difference: it is a machine readable database, making it far more useful for interoperability of molecular databases in systems biology[1]. Thanks to the Wikidata:WikiProject Chemistry community, there is a growing amount of information about chemical compounds: Wikidata currently has over 150 thousand chemical compounds, of which more than 95% is associated with InChIKeys and has more than 70 thousand CAS registry numbers. Ongoing work by this WikiProject includes capturing chemical classes and chemical compounds in the various Wikipedia's as machine readable data. Other projects include covering human drugs[2], MeSH Chemicals and Drugs, and volatile organic compounds. This work is supported the many tools around Wikidata, such as Mix'n'Match which is used to include ChEBI.

We here introduce our contributions to the WikiProject Chemistry to support FAIR-ification of open chemical knowledge. For example, we proposed new Wikidata properties to annotate compounds with external database identifiers for the EPA CompTox Dashboard[3], the SPLASH[4], and MetaboLights. Furthermore, we used a combination of Bioclipse and QuickStatements to add missing chemical compounds for biological pathways from WikiPathways[5]. Finally, we introduce an extension of Scholia [6], visualizing data about compounds and compound classes, including external identifiers, physicochemical properties, and an overview of the literature from which the knowledge is derived.

1. Mietchen D, et al. Enabling Open Science: Wikidata for Research (Wiki4R). Research Ideas and Outcomes. **2015** Dec 22;1:e7573.
2. Putman TE, et al. WikiGenomes: an open web application for community consumption and curation of gene annotation data in Wikidata. Database. **2017** Jan;2017(1).
3. Williams, AJ, et al. The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. J. Cheminform. **2017** Nov 18;9:61.
4. Wohlgemuth G, et al. SPLASH, a hashed identifier for mass spectra. Nature Biotechnology. **2016** Nov 8;34(11):1099–101.
5. Slenter DN, et al. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucleic Acids Research. **2018** Jan 4;46(D1):D661–D667.
6. Nielsen, FÅ, et al. Scholia, Scientometrics and Wikidata. The Semantic Web: ESWC 2017 Satellite Events, **2017**.

P-76: PSMILES – A particle-based Molecular Structure Representation for Mesoscopic Simulation

Karina van den Broek [1,2], Mirco Daniel [2], Matthias Epple [1], Jonas Schaub [2], Hubert Kuhn [3], Achim Zielesny [2]

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

Adequate molecular structure representations are at heart of cheminformatics developments: The various approaches like line notations, connection tables, XYZ tables or Z-matrices, fragment codes or fingerprints address the broad spectrum of different use cases which characterize current research efforts. The majority of existing structure representations are atom-based descriptions that comprise characteristic properties and topological or spatial aspects concerning a molecule's atomic composition[1].

In contrast, this contribution focusses on a particle-based molecular structure representation where a single particle may comprise several atoms, i.e. may represent a "molecular fragment": This kind of representation is an essential part of a molecular fragment cheminformatics roadmap[2] for particle-based mesoscopic simulation techniques like Dissipative Particle Dynamics (DPD) which aims at describing supramolecular phenomena at the nanometer (length) and microsecond (time) scale for large interacting physical ensembles (representing millions of atoms). DPD particles in particular may be identified with distinct small molecules of molar mass in the order of 100 Da where larger molecules are composed of multiple adequate "molecular fragment" particles that are bonded by harmonic springs to mimic covalent connectivities and spatial 3D conformations.

The proposed particle-based molecular structure representation is chosen to be an intuitive line notation which is similar to the well-established SMILES representation for atom-based molecular connectivity[3-5] and denoted Particle SMILES or PSMILES. An open Java library for PSMILES structure handling and mesoscopic simulation support in combination with an open Java Graphical User Interface viewer application for visual topological inspection of PSMILES molecule definitions are outlined.

1. Gasteiger J, Engel T. Chemoinformatics: A Textbook. Weinheim: WILEY-VCH; 2003.
2. Truszkowski A, Daniel M, Kuhn H, Neumann S, Steinbeck C, Zielesny A, Epple M. J Cheminf. 2014;6:45.
3. Weininger D. J Chem Inf Comput Sci. 1988;28:31−36.
4. Weininger D, Weininger A, Weininger JL. J Chem Inf Comput Sci. 1989;29(2):97−101.
5. Weininger D. J Chem Inf Comput Sci. 1990;30(3):237−243.

P-78: A new, improved model to predict kinase inhibition

Cornel Catana, Pieter Stouten

Galapagos NV, Mechelen, Belgium

Kinases constitute an important family of targets for Galapagos, as exemplified by filgotinib, which is currently in phase III clinical trials for RA and IBD. As part of its kinase HTS campaigns, Galapagos routinely and successfully screens a set of around 80,000 kinase-focused compounds, belonging to the categories shown in Fig. 1.

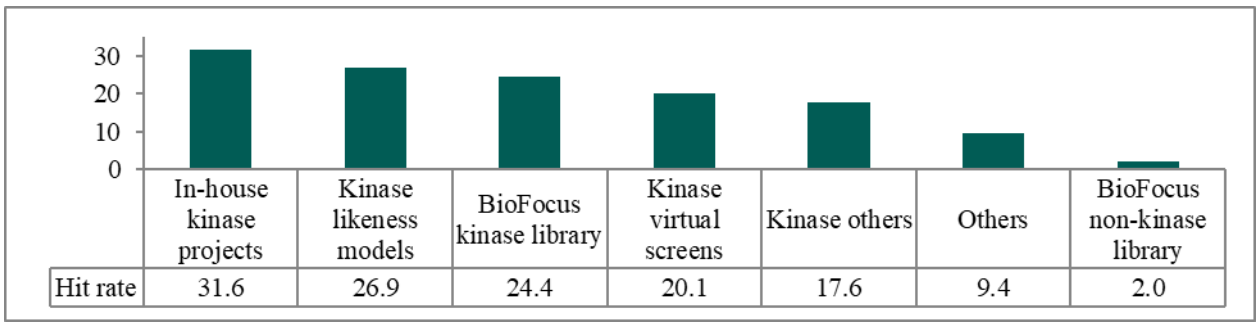


Fig. 1: Sources of hits (PIN > 50%) for a representative kinase HTS

As Fig. 1 shows, the set of compounds selected based on our home-grown kinase inhibition propensity ("kinase likeness") models, exhibits a high hit rate. In order to further enhance these models (dating back to 2008 and 2011), we have recently developed new models since:

larger and new data sets, new descriptors, and improved statistical techniques have become available; and

while we previously exclusively selected models on the basis of their performance on IC_{50} data, our current goal is to develop a model that performs well both on HTS PIN (%inhibition) and IC_{50} data.

The training set contained ~88 k kinase-active compounds taken from ref 1 and the Galapagos collection, and ~84 k kinase-inactive compounds taken from refs 1, 2 and 3. A random forest (RF 2018 all) classification model was developed using Pipeline Pilot[4]. The statistics on the training set are very good (kappa = 0.94; accuracy = 97%). In order to have an unbiased assessment of model performance against in-house data, a model was also developed without the ~22 k Galapagos compounds (RF 20018 NoG).

The models were tested on in-house HTS (PIN) data against 20 kinases. A compound was considered kinase-active if it was at least 2x a hit (PIN>75%) irrespective of the number of assays. It was considered kinase-inactive if it was

assayed at least 5x and never was a hit. Other compounds were ignored. This stringency was used to account for the variability in single dose experiments. This test set comprised a total of 45,569 unique compounds, of which 3,339 were active. Using the newly developed models, the following statistical results were obtained for the PIN test set (Table 1) and for two recently published IC_{50} data sets (Table 2).

Table 1: Model statistics for in-house test set (HTS PIN values)

Model	BEDROC $\alpha = 5$	Number of hits retrieved and enrichment factor		
		Top 1% (450)	Top 2% (900)	Top 5% (2,250)
Bayes 2008	0.265	184 / 6	259 / 4	444 / 3
RF 2011	0.310	216 / 7	324 / 5	562 / 3
RF 2018 NoG	0.354	162 / 5	297 / 5	612 / 4
RF 2018 all	0.358	191 / 6	428 / 6	920 / 6

Table 2: Model statistics for two literature test sets (IC50 values)

Model	Christmann (2,101 compounds) [5]		Martin (3,814 compounds) [6]	
	kappa	ROC score	kappa	ROC score
RF 2011	0.230	0.687	0.187	0.570
RF 2018 NoG	0.475	0.707	0.434	0.739
RF 2018 all	0.450	0.694	0.413	0.723

In conclusion, while our previous models (2008 and 2011) have been very useful in identifying kinase inhibitors (see Fig. 1), we have now developed two new and improved models. The ”RF 2018 all” model is already being used in the process of selecting and acquiring kinase-focused compounds to expand our kinase-focused collection.

1. Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos R.; Overington, J. P. The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* **2014**, 42, 1083-1090.
2. Bora, A.; Avram, S.; Ciucanu, I.; Raica, M.; Avram, S. Predictive Models for Fast and Effective Profiling of Kinase Inhibitors. *J. Chem. Inf. Model.* **2016**, 56, 895-905.
3. Rohner, S. G.; Baumann, K. Maximum unbiased validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioactivity Data. *J. Chem. Inf. Model.* **2009**, 49, 169-184.
4. Biovia Pipeline Pilot, 17.2.0, San Diego: Dassault Systèmes, **2017**.
5. Christmann-Franck, S.; Van Westen, G. J. P.; Papadatos, G.; Beltran Escudie, F.; Roberts, A.; Overington, J. P.; Domine, D. Unprecedently Large-Scale Kinase Inhibitor Set Enabling the Accurate Prediction of Compound-Kinase Activities: A Way toward Selective Promiscuity by Design? *J. Chem. Inf. Model.* **2016**, 56, 1654-1675.
6. Martin, E. J.; Polyakov, V. R.; Tian, L.; Perez, R. C. Profile-QSAR 2.0: Kinase Virtual Screening Accuracy Comparable to Four-Concentration IC50 for Realistically Novel Compounds. *J. Chem. Inf. Model.* **2017**, 57, 2077−2088.

Supporting Societies

- Division of Chemical Information (CINF)
  American Chemical Society (ACS)
- Royal Netherlands Chemical Society (KNCV)
- Computers in Chemistry Division (CIC)
  German Chemical Society (GDCh)
- The Chemical Structure Association Trust
  (CSA Trust)
- Chemical Information and Computer Applications Group (CICAG)
  Royal Society of Chemistry (RSC)
- Division of Chemical Information and Computer Science
  Chemical Society of Japan (CSJ)
- Swiss Chemical Society (SCS)
- European Association of Chemical and Molecular Sciences (EuCheMS)