# ICCS

International Conference on Chemical Structures

## 12th International Conference on Chemical Structures

June 12 – 16, 2022 | Noordwijkerhout, The Netherlands

## Program & Abstracts

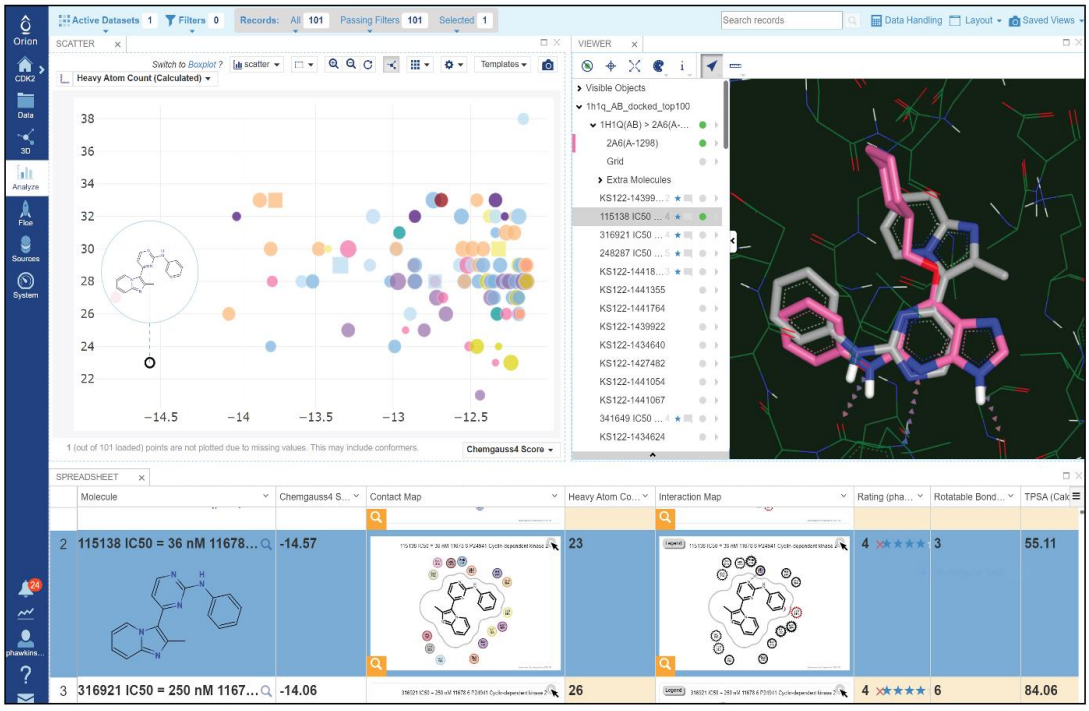
iccs-nl.org


## The Evolution of Molecular Design

Deliver trusted science with Orion®, the only cloud-native fully integrated software-as-a-service molecular modeling platform that offers unprecedented capabilities for the advancement of pharmaceuticals, biologics, agrochemicals, flavors, and fragrances.

Harness the power of Amazon cloud computing to run molecular dynamics simulations, free energy predictions, quantum mechanics calculations, sequence analysis, ultra-fast ligand-based or structure-based 2D and 3D virtual screening of more than 5 billion stereoenumerated commercially available molecules and 40 billion conformers, and more from just a web browser. With a cloud-native design and a proprietary scheduler, Orion was built to handle even the most demanding computing needs.

David LeBard, PhD, Head of Enhanced Sampling at OpenEye Scientific, will discuss at ICCS 2022 the therapeutic challenges with membrane permeation in pharmaceutical development, and the new kinetic model developed in the Orion platform for passive permeability.




*The Orion platform provides all the tools and data needed to allow calculation, analysis, collaboration, and decision-making in one environment.*

Visit our booth to learn more!

**OpenEye Scientific is an industry leader in computational molecular design through rapid, robust, and scalable software and consulting services.**

## Future Focused.

OpenEye SCIENTIFIC

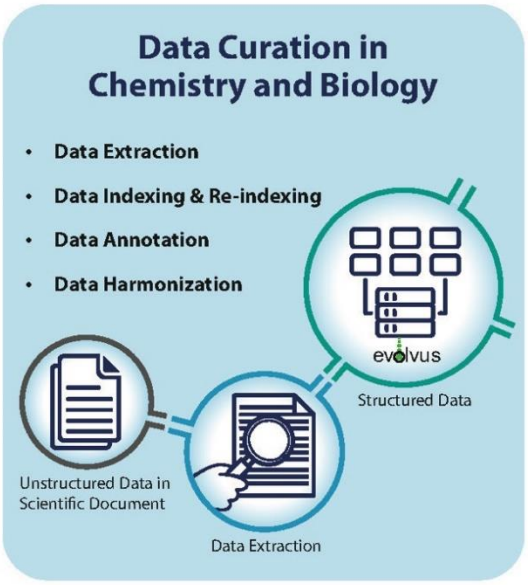
www.openeye.inc

evolvus

# Curating Chemical Structures Since 2001....

## WE CAN HELP YOU CHANGE THIS !!!

**There are a plethora of tools available for ML or AI, but very few reliable data sources.**




**Getting access to scalable, very large customised chemical databases.**

**Our Offerings**

**ADME Database:** Our small molecule database consisting of pharmacokinetics data, including microsomal stability.

**Liceptor Database:** Our small molecule bioactivity database curated from patents has 9 million compounds and the associated target based biological metadata for every compound.

**Protein Degraders Database:** Our manually curated database of bi-functional and hetero bi-functional compounds which lead to targeted protein degradation and their associated bioactivity metadata.

**ThEME Database:** Our metabolism or biotransformation dataset involving drug liver breakdown metabolic bio-pathways based in the specific therapy areas.

- Since 2001, Evolvus is the world's premier chemistry curation powerhouse. Bulk of our work is designed around database curation and managing unstructured data, and everything we do, ideally revolves around -- Generation, Management, Correlation and Structuring of Chemical Data.
- We perform Indexing and annotation of Chemical, Biological and Genomic data from patents and literature, on an industrial scale.
- Our Liceptor database is the world's largest curated small molecule bioactivity dataset. With more than 9 million chemical structures, it's the foremost patent derived chemical bioactivity data on the planet.


www.evolvus.com

Contact: aniket@evolvus.com

# Preface

Welcome to the 2022 International Conference on Chemical Structures (ICCS). This is the 12th ICCS to be organized and was delayed by one year due to the COVID19 pandemic. The conference builds on a long and successful history, which started with a NATO Advanced Study Workshop in 1973 [1]. 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 almost 100 abstract submissions which were all subjected to a review process carried out by our Scientific Advisory Board of 20 international reviewers from academia and industry. This allowed us to compile an outstanding scientific program of 34 plenary and 52 poster, welcoming participants from 20 countries from 3 continents. 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 which will bring us back from a visit to the Zuiderzee museum.

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. Greg Landrum in recognition of his work on the development of RDKit and his fostering of the community around it, a transformative software resource for cheminformatics and machine learning [2]. Dr. Landrum has agreed to give a keynote lecture on Sunday evening.

Keeping in line with tradition, 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 Willem Jespers. Papers can be submitted at any date up to the 1st of November 2022, and authors should mention in their cover letter that the manuscript is intended to be included in the 2022 ICCS 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 2022. Note that in the online program ORCID identifiers and Twitter accounts are provided where available, allowing you to learn more about past research by presenters and contact them. 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!

Gerard van Westen (ICCS Chair), Willem Jespers, Egon Willighagen, Frank Oellien, Markus Wagener and Francesca Grisoni

1. An overview of all ICCS meetings. https://tools.wmflabs.org/scholia/event-series/Q47501052
2. Scholarly output of Dr. Greg Landrum https://scholia.toolforge.org/author/Q42716526
3. 2022 ICCS page: https://scholia.toolforge.org/event/Q111749081

## Contents

# The Conference

## Organizing Committee

- Gerard JP van Westen, University of Leiden, The Netherlands
- Willem Jespers, University of Leiden, The Netherlands
- Egon Willighagen, Maastricht University, The Netherlands
- Markus Wagener, Grünenthal, Germany
- Francesca Grisoni, TU Eindhoven, Netherlands
- Frank Oellien, AbbVie, Germany

## Scientific Advisory Board

- Andreas Bender, University of Cambridge, UK
- Anirudh Ranganathan, Stockholm University, Sweden
- Antony Williams, EPA, US
- Barabara Zdrazil, EMBL-EBI, UK
- Christoph Steinbeck, University of Bonn, Germany
- Christos Nicolau, Recursion, US
- Elif Ozkirimli, Roche, Switzerland
- Ester Kellenberger, University of Strasbourg, France
- Herman van Vlijmen, Janssen Pharmaceutica NV, BE
- John Overington, Exscientia, UK
- Matthias Rarey, University of Hamburg, DE
- Ola Engkvist, AstraZeneca, Sweden
- Pat Walters, Relay Therapeutics, US
- Peter Ertl, Novartis, CH
- Rajarshi Guha, National Institutes of Health, US
- Sereina Riniker, ETH Zurich, Switzerland
- Teodoro Laino, IBM Zurich, Switzerland
- Val Gillet, University of Sheffield, UK
- Zoe Cournia, Biomedical Research Foundation, Greece

## 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
https://www.eyesopen.com/

## Platinum Sponsors

Chemical Abstract Service
https://www.cas.org/

Collaborative Drug Discovery
https://www.collaborativedrug.com/

## Gold Sponsors

Chemaxon
https://chemaxon.com/

Chemical Computing Group
https://www.chemcomp.com/

KNIME
https://www.knime.com/

NextMove Software
https://www.nextmovesoftware.com/

Galapagos
https://www.glpg.com/

## Silver Sponsors

**CCDC**
**https://www.ccdc.cam.ac.uk/**

**Discngine**
https://www.discngine.com/

## Conference Bag Sponsor


**KNIME**
**https://www.knime.com/**

## Notepad Sponsor


**KNIME**
**https://www.knime.com/**

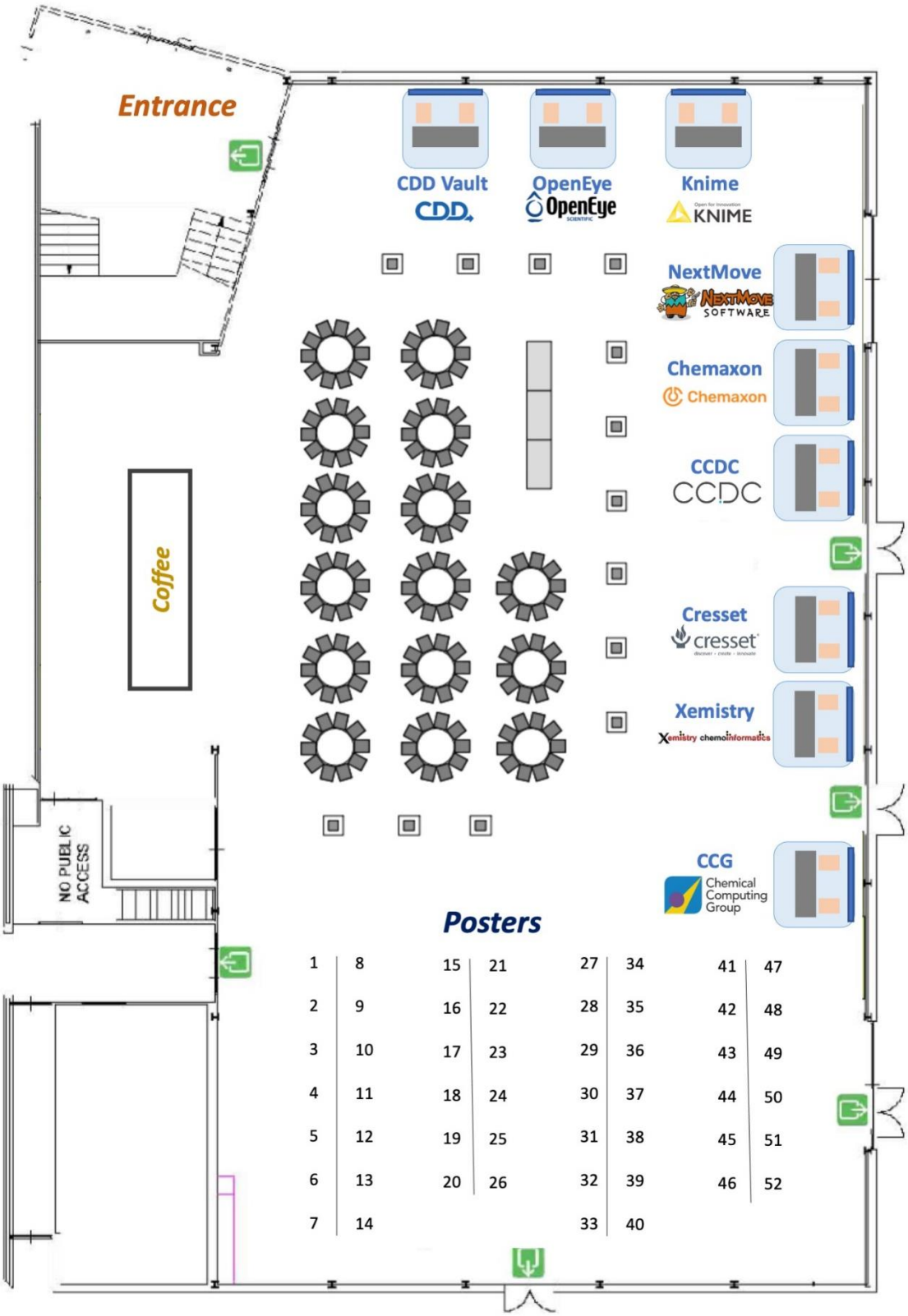
## Poster Awards Sponsor


**Evolvus**
**http://www.evolvus.com/**

## Other Sponsors

We would like to thank CCL.NET and Jan Labanowski for adding the conference to the CCL Conferences webpage.

# Exhibition




## Exhibition Hours

- Monday, June 13th, 2022, 15:30 – 19:30
- Tuesday, June 14th, 2022, 15:30 – 19:30

# Exhibitors


KNIME
https://www.knime.com/


NextMove Software
https://www.nextmovesoftware.com/


Chemical Computing Group
https://www.chemcomp.com/


Cresset
http://www.cresset-group.com/


OpenEye
https://www.eyesopen.com/


Xemistry
http://www.xemistry.com/


Chemaxon
http://www.chemcomp.com/


CDD Vault
https://www.collaborativedrug.com/


CCDC
https://www.ccdc.cam.ac.uk/

# Workshops Sunday, June 12th

## Chemical Computing Group Workshop: De novo design of novel compounds to meet multiple property constraints

**Sunday June 12th 2022, 15:00-17:00, NH Conference Hotel Noordwijkerhout**

The workshop describes SBDD workflows in drug discovery projects and encompasses a range of topics from pharmacophore query generation to protein-ligand interaction fingerprints. More specifically, the workshop will cover the application of pharmacophores in the context of protein-ligand docking, scaffold replacement and R-group screening. A method for querying a 3D project database will also be presented along with the generation and analysis of protein-ligand interaction fingerprints (PLIF).

## KNIME Workshop

**Sunday June 12th 2022, 15:00-17:00, NH Conference Hotel Noordwijkerhout**

KNIME Analytics Platform offers an interactive environment to process your data. With the diverse array of open source and commercial life sciences extensions you can build cheminformatics applications.

In this hands-on workshop we will learn how to process chemical data based on a common cheminformatics problem: library enumeration. We will start by reading in and cleaning up a database of building blocks from a catalog. Then we will define a two-component reaction and filter the building blocks to match it. Next we will filter the products from any reference compounds or substructures. To visualize and interactively explore the products, we will teach you how to build a component which you can share and reuse in future workflows. Last, we will save the selected products to an Excel table.

On top of learning something new in KNIME, by the end of the workshop you will come out with a workflow you can adapt, reuse and share with you colleagues.

# Workshops Thursday, June 16th

## Xemistry Workshop: Reaction Processing with the CACTVS Toolkit

**Thursday June 16th 2022, 14:00-16:00, NH Conference Hotel Noordwijkerhout**

CACTVS is an universal scripting environment for chemical information processing with a large collection of unique capabilities. In this workshop, we will explore the CHMTRN engine of the toolkit for processing reactions in forward and retrosynthetic direction. CHMTRN was originally developed by Corey et al. as the knowledge base language of the LHASA retrosynthesis planner software. We have re-implemented and modernized a byte-code compiler for this language in clean-room fashion and integrated it as a subsystem of the toolkit. You can now leverage the knowledge base of thousands of reaction schemes originally coded for LHASA and put into public domain by the Lhasa company – and also access recent transform additions for medicinal chemistry methods added in the last decades. We will examine why this engine does far more than traditional library enumerators and similar tools – in various environments, including stand-alone Python scripts, as a Python module loaded into other applications, Jupyter notebooks and KNIME nodes.

# Social media, photos, and good practices

**Rules of engagement**
The ICCS 2022 is meant to be an open platform where the latest cheminformatics is being discussed. In all cases, respect every person's opinion and be kind. When meeting new people, be like Inigo Montoya. Discrimination on age, gender, and ethnicitiy is strictly forbidden in The Netherlands.

**Photos**
Participants are allowed to take photos during the meeting UNLESS the presenter clearly indicates this is not allowed. Photos can be shared but if other people are identifiable, you are obliged to ask their permission before sharing. Generally, we encourage you to inform people of your intention and respect their positions before sharing a photo of people, posters, presentations.

**Conference photos**
The *Stichting Chemissche Congressen VI* reserves the right to use any photograph taken by the conference photographer at the 12th International Conference on Chemical Structures, without the expressed written permission of those included within the photograph. COH may use the photograph/video in publications or other media material produced, used, or contracted by *Stichting*, including but not limited to brochures, invitations, books, newspapers, magazines, television, websites, etc.

**Twitter**
The official Twitter hashtag for this meeting is #2022ICCS. Online coverage of presentations is encouraged UNLESS the presenter clearly indicates this is not allowed. The same rules of engagement apply online as they apply in person.

**Discord**
The 2022 ICCS has a Discord channel for participants and the full conference SAB, even if the cannot join in person this year. You can join via this link: https://discord.gg/3hh57ahj The same rules of engagement apply online as they apply in person.

# Excursion: Visit to the Zuiderzee Museum and Sailing Cruise on the IJsselmeer (Lake IJsel)

## Schedule

| | |
|---|---|
| 13:00 | Busses depart from the conference center, Noordwijkerhout |
| 14:00 | Arrive at the Zuiderzee Museum, Enkhuizen |
| 17:00 | Board the sailing boats *Willem Barentsz* and *Abel Tasman*, drinks and bites<br>Sail to Volendam |
| 19:30 | Dinner on board |
| 22:00 | Disembark at Volendam, return to Noordwijkerhout by bus |
| 23:00 | Arrive at the conference center |

## Zuiderzee Museum

The social event starts with a visit at the Zuiderzee Museum. Wikipedia writes:

"*The Zuiderzee Museum, located on Wierdijk in the historic center of Enkhuizen, is a Dutch museum devoted to preserving the cultural heritage and maritime history from the old Zuiderzee region. With the closing of the Afsluitdijk (Barrier Dam) on May 28, 1932, the Zuiderzee was split in two parts: the waters below the Afsluitdijk are now called the IJsselmeer, while the waters north of it are now considered to be part of the Waddenzee.*

*The indoor museum was opened on July 1, 1950. It consists of a string of (original and replicated) 17th century buildings of which some were used by the VOC and contains both temporary exhibitions as well permanent installations. Most notably is the 'Schepenhal' (ship's hall), which allows visitors a close-up view of some of the more common historical types of boats from the Zuiderzee's rich fishing industry as well as some recreational sailing ships. Among these beautiful boats is the Sperwer (sparrowhawk), a 'boeier' once owned by the English adventurer Merlin Minshall, who sailed this boat from England over the Danube to the Black Sea in the 1930s for his honeymoon, and a second time for the English secret service. Also a historic 'Midzwaardjacht' (Centreboard) is on display.*", CC-BY-SA 3.0 Unported, https://en.wikipedia.org/wiki/Zuiderzee_Museum

Image: By Rijksdienst voor het Cultureel Erfgoed, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=24107956

# Scientific Program

# Plenary Session

## Sunday, June 13

| | |
|---|---|
| 12:00 - 18:00 | **Registration**<br>*Atrium Lounge* |
| 15:00 - 17:00 | **Pre-conference workshops** |
| | **De novo design of novel compounds to meet multiple property constraints** Chemical Computing Group |
| | **KNIME Workshop**<br>KNIME |
| 17:00 - 18:00 | **Free Time** |
| 18:00 - 18:15 | **Welcome**<br>*Rotonde* |
| 18:15 - 19:00 | **Keynote Address - CSA Trust Mike Lynch Award**<br><br>**RDKit: where did we come from and where are we going?**<br>Awardee Dr. Greg Landrum, ETH Zurich |
| 19:00 - 20:00 | **Welcome Reception**<br>*Atrium* |
| 20:00 - 22:00 | **Reception Dinner**<br>*Atrium* |

# Monday, June 13

| | |
|---|---|
| 8:30 - 12:00 | **Session A - Analysis of Large Chemical Datasets**<br>*Rotonde* |
| 8:30 - 9:00 | **A-1: 25 years of small molecule optimization at Novartis: A retrospective analysis of chemical series evolution**<br>Maximilian Beckers, Novartis Pharma, Switzerland |
| 9:00 - 9:30 | **A-2: GeoMine: On-The-Fly Geometric Pattern Mining in Binding Sites**<br>Joel Graef, Universität Hamburg, Germany |
| 9:30 - 10:00 | **A-3: Papyrus - A large scale curated dataset aimed at bioactivity predictions**<br>Olivier JM Béquignon, Leiden University, The Netherlands |
| 10:00 - 10:30 | **Coffee Break**<br>*Atrium* |
| 10:30 - 11:00 | **A-4: Improving Torsion Library Patterns with SMARTScompare**<br>Patrick Penner, Roche, Switzerland |
| 11:00 - 11:30 | **A-5: PSnpBind: A database of mutated binding site protein-ligand complexes constructed using a multithreaded virtual screening workflow**<br>Ammar Ammar, Maastricht University, The Netherlands |
| 11:30 - 12:00 | **A-6: Recent Advances in Chemical Search of Ultra-large Databases**<br>Roger Sayle, NextMove Software, UK |
| 12:00 - 13:00 | **Lunch**<br>*Atrium* |
| 13:00 - 15:00 | **Session B - Structure-Activity and Structure-Property Prediction**<br>*Rotonde* |
| 13:00 - 13:30 | **B-1: Chemical feature visualization to interpret neural network models for toxicity prediction**<br>Moritz Walter, University of Sheffield, UK |
| 13:30 - 14:00 | **B-2: The Influence of Nonadditivity on Machine Learning and Deep Learning Models**<br>Eva Nittinger, AstraZeneca, Sweden |
| 14:00 - 14:30 | **B-3: Challenges of tracking SARS Cov-2 M-protease inhibitors from patents**<br>Christopher Southan, Medicines Discovery Catapult, UK |
| 14:30 - 15:00 | **B-4: An innovative approach of Toll-like receptor dynamics exploitation for structure optimization through 3D pharmacophore analysis**<br>Valerij Talagayev, Freie Universität Berlin, Germany |
| 15:00 - 15:30 | **Coffee Break**<br>*Atrium* |
| 15:30 - 19:30 | **Poster Session & Exhibition**<br>*Atrium* |
| 15:30 - 17:30 | **Poster Presentations Red**<br>*Atrium* |
| 18:30 - 19:30 | **Reception**<br>*Atrium* |
| 19:30 - 21:30 | **Dinner**<br>*Atrium* |

## Tuesday, June 14

| | |
|---|---|
| 08:30 - 10:00 | **Session C – Dealing with Biological Complexity**<br>*Rotonde* |
| 08:30 - 09:00 | **C-1: A Systems Biology Workflow to Support the Diagnosis of Pyrimidine and Urea Cycle Disorders**<br>Denise Slenter, Maastricht University, The Netherlands |
| 09:00 - 09:30 | **C-2: Modeling, Proper Validation, and Discovery of Synergistic Drug Combinations**<br>Eugene Muratov, UNC Chapel Hill, US |
| 09:30 - 10:00 | **C-3: Conformational Chirality and Protein Structure Analysis**<br>Inbal Tuvi-Arad, The Open University of Israel, Israel |
| 10:00 - 10:30 | **Coffee Break**<br>*Atrium* |
| 10:30 - 14:30 | **Session D – Structure-based approaches**<br>*Rotonde* |
| 10:30 - 11:00 | **D-1: Describing protein dynamics for proteochemometric bioactivity prediction: 3DDPDs**<br>Marina Gorostiola González , Leiden University, The Netherlands |
| 11:00 - 11:30 | **D-2: Mechanism of passive membrane permeability from weighted ensemble simulations in the cloud**<br>David LeBard, OpenEye Scientific, US |
| 11:30 - 12:00 | **D-3: Integrated Structural Cheminformatics Analysis Tools for Customisable Chemogenomics Driven**<br>Dominique Sydow, Sosei Heptares, UK |
| 12:00 | **GROUP PHOTO** |
| 12:00 - 13:00 | **Lunch**<br>*Atrium* |
| 13:00 - 13:30 | **D-4: A novel antibiotic target: Identifying bacterial ribosomal assembly inhibitors via 3D pharmacophore-based virtual screening**<br>Theresa Noonan, Freie Universität Berlin, Germany |
| 13:30 - 14:00 | **D-5: Dynamic interaction patterns enable characterization of opioid-peptide binding to the atypical chemokine receptor 3**<br>Kristina Sophie Puls, Freie Universität Berlin, Germany |
| 14:00 - 14:30 | **D-6: Development of potent FPR1 antagonists and partial agonists based on structural modelling and a detailed understanding of binding characteristics**<br>Sarah Maskri, University of Münster, Germany |
| 14:30 - 15:00 | **Coffee Break**<br>*Atrium* |
| 15:00 - 19:30 | **Poster Session & Exhibition**<br>*Atrium* |
| 15:00 - 17:00 | **Poster Presentations Blue**<br>*Atrium* |
| 18:30 - 19:30 | **Reception**<br>*Atrium* |
| 19:30 - 21:30 | **Conference Dinner**<br>*Atrium* |

## Wednesday, June 15

| | |
|---|---|
| 08:30 - 12:30 | **Session E - Chemoinformatics Approaches**<br>*Rotonde* |
| 08:30 - 09:00 | **E-1: Chemical Annotation: A new similarity score for automated design and ranking**<br>Baptiste Canault, GlaxoSmithKline, UK |
| 09:00 - 09:30 | **E-2: Conformers Everywhere: Conformer Ensembles, Conformer Energies, 3D-ADMET and Machine Learning Potentials**<br>Andreas Göller, Bayer AG |
| 09:30 - 10:00 | **E-3: NFDI4Chem – The National Research Data Infrastructure for Chemistry**<br>Oliver Koepler, TIB – Leibniz Information Centre for Science and Technology, Germany |
| 10:00 - 10:30 | **Coffee Break**<br>*Atrium* |
| 10:30 - 11:00 | **E-4: Automated Ligand Design meets Synthesis Planning**<br>Hans Briem, Bayer AG, Germany |
| 11:00 - 11:30 | **E-5: De novo design of synthetically accessible molecules using an evolutionary algorithm**<br>Alan Kerstjens, University of Antwerp, Belgium |
| 11:30 - 12:00 | **E-6: Assigning Diastereomers by Comparing Experimental and Theoretical IR Spectra** Sereina Riniker, ETH Zurich, Switzerland |
| 12:00 - 12:30 | **E-7: Tautomerism analyses in preparation of InChI V2**<br>Mark Christian Nicklaus, IUPAC InChI Tautomerism Group, NIPER, India |
| 13:00 | **Lunch Box** |
| 13:00 - 23:00 | **Excursion**<br>Visit to the Zuiderzee Museum and Sailing Cruise on the IJsselmeer.<br>Dinner will be served on board. |

## Thursday, June 16

| | |
|---|---|
| 07:30 - 08:30 | **Hotel Check-Out** |
| 08:30 - 11:30 | **Session F - Artificial Intelligence Approaches**<br>*Rotonde* |
| 08:30 - 09:00 | **F-1: Improved classification of protein function by a localized 3D protein descriptor and deep learning**<br>Karel Johannes van der Weg, Forschungszentrum Jülich, Germany |
| 09:00 - 09:30 | **F-2: Augmented Hill-Climb improves language-based de novo molecule generation as benchmarked via the open source MolScore platform**<br>Morgan Cole Thomas, University of Cambridge, UK |
| 09:30 - 10:00 | **F-3: Explaining and avoiding failure modes of artificial intelligence for small molecule design**<br>Maxime Langevin, Sanofi Aventis R&D, France |
| 10:00 - 10:30 | **F-4: Multi-Instance Learning Approach to Predictive Modeling of Molecular Properties: new or well forgotten old?**<br>Pavel Polishchuk, Palacky University, Czech Republic |
| 10:30 - 11:00 | **Coffee Break & Hotel Check-Out**<br>*Atrium* |
| 11:00 - 11:30 | **F-5: Neural Fingerprints: Generating Domain-specific Molecular Fingerprints Using Neural Networks.**<br>Janosch Menke, University of Münster, Germany |
| 11:30 - 12:00 | **F-6: Ranking generated molecule conformations using deep-learning predicted deviation to target-bound conformations**<br>Benoit Baillif, University of Cambridge, UK |
| 12:00 - 12:30 | **F-7: Digital Chemistry at Syngenta: From academic labs to industrial applications**<br>Arndt Finkelmann, Syngenta Crop Protection AG, Switzerlands |
| 12:30 - 13:00 | **F-8: Translating data to predictive models**<br>Akos Tarcsay, ChemAxon Kft, Hungary |
| 13:00 - 13:15 | **Closing Remarks** |
| 13:15 - 14:00 | **Lunch or Box Lunch** |
| 13:30 - 14:00 | **Shuttle Busses leave for Schiphol Airport** |
| 14:00 - 16:00 | POST-CONFERENCE WORKSHOPS |
| | **Reaction Processing with the CACTVS Toolkit**<br>Xemistry |
| 16:30 - 17:00 | **Shuttle Busses leave for Schiphol Airport** |

# Poster Session RED

## Analysis of Large Chemical Datasets



## Artificial Intelligence Approaches



## Chemoinformatics Approaches

## Dealing with Biological Complexity



## Structure-Activity and Structure-Property Prediction



## Structure-based Approaches

# Poster Session BLUE

## Analysis of Large Chemical Data Sets



## Artificial Intelligence Approaches



## Cheminformatics Approaches



## Dealingwith Biological Complexity

## Structure-Activity and Structure-Property Prediction



## Structure-based Approaches

# Plenary Session Abstracts

# Keynote Address CSA Trust Mike Lynch Award

## RDKit: where did we come from and where are we going?

Gregory A. Landrum[1]

[1]*Institute for Physical Chemistry, ETH Zürich, Vladimir-Prelog-Weg 2, 8093 Zürich, Switzerland*

In this talk I will start with a brief overview of the current status and history of the RDKit project and community. The RDKit was originally developed within a small startup and used for building predictive models for ADME, Tox, and biological activity. In 2006 we shut down the startup, released the RDKit under an open-source (BSD) license, and I moved to Novartis. Since then, open-source development has continued and the project has grown through my time in large pharma (Novartis), a software startup (KNIME), and now in academia (the ETH).

After this overview I'll move on to talk about planned future work on the toolkit. In order to put the planned work into context, I'll spend some time talking about my perspective on chemical registration systems and making chemical data FAIR. This all inevitably leads to a challenge to the computational chemistry and cheminformatics communities to up their game when it comes to keeping track of and reporting their own work.

1. RDKit: Open-source cheminformatics. https://www.rdkit.org

# Session A: INTEGRATION OF CHEMICAL INFORMATION

## A-1: 25 years of small molecule optimization at Novartis: A retrospective analysis of chemical series evolution

M. Beckers[1], N. Fechner[1], N. Stiefl[1]

[1] *Novartis Institutes for BioMedical Research, Novartis Pharma AG, Novartis Campus, 4002 Basel, Switzerland*

In early drug discovery, optimization of small molecules with respect to safety and biological activity is an important task to deliver drug candidates with highest chances of success in subsequent clinical trials. Typically, analyses have focused on the comparison of approved and failed drug candidates or have investigated the association of structural and measured properties with clinical outcomes. However, the actual optimization process is barely characterized, mainly due to missing annotations about chemical series that have been worked on in past projects.
In this contribution, we report a reconstruction of ~3000 chemical series from our Novartis in-house compound database. We present modifications made to the previously published protocol for automated chemical series identification[1,2], which allowed application to datasets with more than 100k compounds. Based on our reconstruction we characterize the determined series and their connections with each other.
Using the registration dates of the compounds, we further characterized the evolution of chemical properties over time. Determination of active optimization phases allowed us to trace both structural and ADMET properties during optimization of the molecules. Our analysis revealed multiple patterns, which are repeatedly observed in the reconstructed series. We investigate the influence of the chemists on the observed trends and quantify the extent to which the respective ADMET properties can be improved over time.

1. F. Kruger *et al.* Automated Identification of Chemical Series: Classifying like a Medicinal Chemist. *J. Chem. Inf. Model. 2020, 60, 6, 2888–2902*
2. M. Beckers *et al*. *manuscript in preparation*, 2022

## A-2: GeoMine: On-The-Fly Geometric Pattern Mining in Binding Sites

J. Graef[1], C. Ehrt[1], K. Diedrich[1], M. Poppinga[1,2], N. Ritter[2], M. Rarey[1]

[1]*Universität Hamburg, ZBH - Center for Bioinformatics, Bundesstraße 43, 20146 Hamburg, Germany.*
[2] *Universität Hamburg, Department of Informatics, Databases and Information Systems Group, Vogt-Kölln-Straße 30, 22527 Hamburg, Germany.*

Structural investigations and comparisons of binding sites help in understanding a protein's function, discovering ligand off-targets, or analyzing interaction geometries with several applications in drug design. The Protein Data Bank (PDB)[1] provides a substantial amount of data to make use of but the need for a tool with fast and comprehensive search capabilities persists. With GeoMine[2,3], we have developed a database-driven search engine for DoGSite[4]-predicted and ligand-based binding sites in the PDB that allows in-depth structural investigations. Queries are any combinations of point-based patterns, and a multitude of textual and numerical filters. For the query points (either atoms or aromatic centers), several properties can be defined, e.g., nucleic acid, protein, ligand, water, or metal, the residue type of a protein atom, its solvent exposure, the secondary structure type, or the corresponding functional group can be described. Points can be connected via distance constraints or interactions, e.g., hydrogen bonds, pi-pi, or ionic interactions, and angle ranges between these can be defined. Therefore, virtually any 3D pattern can be created and efficiently searched for in the PDB or user-defined subsets thereof thanks to the database design which was carefully fine-tuned based on the query requirements.
GeoMine is publicly available on the Proteins*Plus*[5] web server and currently allows structural investigations in 308,507 ligand-based and 759,011 predicted binding sites. Based on its capabilities it has evolved into a multi-purpose pattern mining tool uniting the capabilities of several methods designed for binding site comparison, the analysis of geometrical preferences of interaction patterns, the filtering of large binding site databases based on protein-, ligand-, and site-based descriptors, the detection of structural residue motifs, or the search for typical protein-ligand interaction patterns.

1. Berman, H.M., et al., The Protein Data Bank. Nucleic Acids Res., **2000**, 28, 235-242
2. Diedrich, K., et al., GeoMine: Interactive Pattern Mining of Protein-Ligand Interfaces in the Protein Data Bank. Bioinformatics, **2020**, 37, 3, 424-425
3. Graef, J., et al., Searching Geometric Patterns in Protein Binding Sites and Their Application to Data Mining in Protein Kinase Structures. J. Med. Chem., **2022**, 65, 2 1384-1395
4. Volkamer, A et al., Analyzing the Topology of Active Sites: On the Prediction of Pockets and Subpockets. J. Chem. Inf. Model., **2010**, 50, 11, 2041-2052
5. Schöning-Stierand, K., et al., ProteinsPlus: Interactive Analysis of Protein–Ligand Binding Interfaces. Nucleic Acids Res., **2020**, 48, W48-W53

## A-3: Papyrus - A large scale curated dataset aimed at bioactivity predictions

O.J.M. Béquignon O.J.M.[1], B.J. Bongers[1], W. Jespers[1], A.P. IJzerman[1], B. van de Water[1], G.J.P. van Westen[1]

[1]*Division of Drug Discovery and Safety, LACDR, Leiden University, Leiden, The Netherlands*

With the recent advancements in machine learning and more specifically deep learning, the ability of algorithms to converge to a stable state and optimal solution to predict ligand-protein bioactivity data is challenged, specifically when considering small datasets. Additionally, the trove of bioactivity data that can be used in this regard suffers from the use of different standardisation rules applied to molecular structures, bioactivity measurements and units.

This study aims at combining and standardising multiple sources of bioactivity data while annotating the quality of each datapoint. In a second step, it aims at evaluating the diversity of the aggregated data, named Papyrus dataset, not only in terms of chemical space but also in terms of target and bioactivity spaces.

The ChEMBL[1] and ExCAPE-DB[2] large public datasets were combined with that of four other articles 3–6. Molecular structures were standardised using the ChEMBL structure pipeline while determining canonical ionization and tautomerization states. Targets were annotated with ChEMBL's protein family classification. Finally, protein-compound interactions were categorised as low-, medium- or high-quality data based on (i) the ChEMBL curators' confidence in the assay and (ii) that the correct protein targets were assigned and on (iii) the precision associated with the bioactivity measurement: censored and binary data being associated with low quality.

Subsets of the Papyrus dataset relating to adenosine, C-C chemokine and monoamine receptors, kinases and members of the solute carrier 6 (SLC6) transport family were created and bioactivities were modelled using extreme gradient boosted trees.

Finally, the chemical, protein target and bioactivity spaces were evaluated as functions of chemical environments, sequences and fold similarities.

The Papyrus dataset consisting of 59,763,781 compound-protein pairs was aggregated and its data normalised. Quantitative structure-activity relationship and proteochemometrics models were fitted with mean balanced accuracy and Maxwell correlation coefficient of 0.69 and 0.39 and Pearson r and root-mean-square error of 0.55 and 0.94 respectively. Additionally, the Papyrus dataset shows increased chemical environment diversity as well as less sparse bioactivity matrix. It is anticipated that the Papyrus dataset can be exploited in a myriad of ways and filtered or altered for specific research questions.

1. Gaulton, A. et al. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res. 40, D1100-7 (2012).
2. Sun, J. et al. ExCAPE-DB: An integrated large scale dataset facilitating Big Data analysis in chemogenomics. J. Cheminform. 9, 1–9 (2017).
3. Sharma, R., Schürer, S. C. & Muskal, S. M. High quality, small molecule-activity datasets for kinase research. F1000Research 5, 1366 (2016).
4. Christmann-Franck, S. et al. 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. 56, 1654–1675 (2016).
5. Merget, B., Turk, S., Eid, S., Rippmann, F. & Fulle, S. Profiling Prediction of Kinase Inhibitors: Toward the Virtual Assay. J. Med. Chem. 60, 474–485 (2017).
6. Klaeger, S. et al. The target landscape of clinical kinase drugs. Science (80). 358, (2017).

## A-4: Improving Torsion Library Patterns with SMARTScompare

P. Penner[1], W. Guba[2], R. Schmidt[1,3], A. Meyder[1,2], M. Stahl[2], M. Rarey[1]

*[1]Universität Hamburg ZBH- Center for Bioinformatics, Bundesstr. 43 20146 Hamburg, Germany*
*[2] Roche Pharma Research & Early Development, Roche Innovation Center Basel, F.Hoffmann-La Roche Ltd. CH-4070 Basel, Switzerland*
*[3] BioSolveIT GmbH, An der Ziegelei 79 53757 Sankt Augustin, Germany*

Molecular Geometry is not only defined by bond lengths and bond angles but also by 4-point dihedral angles. Torsions are subject to many nuanced effects that lead to multiple angles/local minima they can occupy. The chemical environment determines the torsional potential energy curve. Associating torsion angle data from known structures with a description of their chemical environment is a way to describe how likely particular angles are in a molecule.
The Torsion Library[1][2] is a collection of torsion angle statistics from crystal structure data associated with chemical environments encoded as SMARTS. The frequency of torsion angles in crystal structures, which can be considered low energy states, are used to detect strain in chemical environments matched by the SMARTS expression. The Torsion Library is expert-curated and every SMARTS is written with the intention to express meaningful chemistry. The SMARTS syntax, however, can be very complicated and full of implications. Recent advances in comparing SMARTS expressions can be used to automatically support expert curation.
We have adapted SMARTScompare[3] to work with the specifications of the Torsion Library. This can be used to rapidly detect inconsistencies in SMARTS. Fixing these inconsistencies has very tangible effects on matching behavior and ultimately the classification of molecule conformations into relaxed or strained.[4] We have furthermore reworked large parts of the Torsion Library ecosystem, including a new public webserver at https://torsions.zbh.uni-hamburg.de.

1. Schärfer, C. et al., Torsion Angle Preferences in Druglike Chemical Space: A Comprehensive Guide. *J. Med. Chem.*, **2013**, 56, 5, 2016-2028
2. Guba, W. et al., Torsion Library Reloaded: A New Version of Expert-Derived SMARTS Rules for Assessing Conformations of Small Molecules. *J. Chem. Inf. Model.*, **2016**, 56, 1, 1-5
3. Schmidt, R. et al., Comparing Molecular Patterns Using the Example of SMARTS: Theory and Algorithms. *J. Chem. Inf. Model.* **2019**, 59, 6, 2560-2571
4. Penner, P. et al., The Torsion Library: Semi-automated Improvement of Torsion Rules with SMARTScompare. *J. Chem. Inf. Model. Under Review*

## A-5: PSnpBind: A database of mutated binding site protein-ligand complexes constructed using a multithreaded virtual screening workflow

A. Ammar[1], C. Evelo[1], R. Cavill[2], E. Willighagen[1]

*[1]Department of Bioinformatics—BiGCaT, NUTRIM, Maastricht University, The Netherlands 2*
*[2]Department of Data Science and Knowledge Engineering, Maastricht University, The Netherlands*

Over the last 50 years, pharmacogenomics has studied the genetic basis for inter-individual drug response variability [1]. Many factors are involved in patient-drug response, in particular, natural genetic variants that can affect the protein structure and stability and alter ligand-binding affinity [2]. Studies have shown that 80% of patients carry at least one functional variant in the drug targets of the top 100 most commonly prescribed drugs in the United States [3]. These effects have been reported in literature on very limited and small datasets that are not suitable for machine learning prediction models. Ideally, a large dataset of binding affinity changes due to binding site single-nucleotide polymorphisms (SNPs) is needed to build a machine learning (ML) model predicting these effects. However, to the best of our knowledge, such a dataset did not exist yet. Having a large database of protein-ligand complexes covering a wide range of binding pocket mutations and a large small molecules' landscape is of great importance for several types of studies. For example, developing machine learning models to predict protein-ligand affinity or a SNP effect on it requires an extensive amount of data. In this work, we present PSnpBind: A large database of 0.6 million mutated binding site protein-ligand complexes constructed using a multithreaded molecular docking approach. A 7-step workflow was developed (Figure 1) to integrate, preprocess, prepare and dock protein-ligand complexes using a distributed Kubernetes-based infrastructure. PSnpBind provides a web interface to explore and visualize the protein-

ligand complexes and a REST API to programmatically access the different aspects of the database contents. PSnpBind is open source and freely available at https://psnpbind.org.

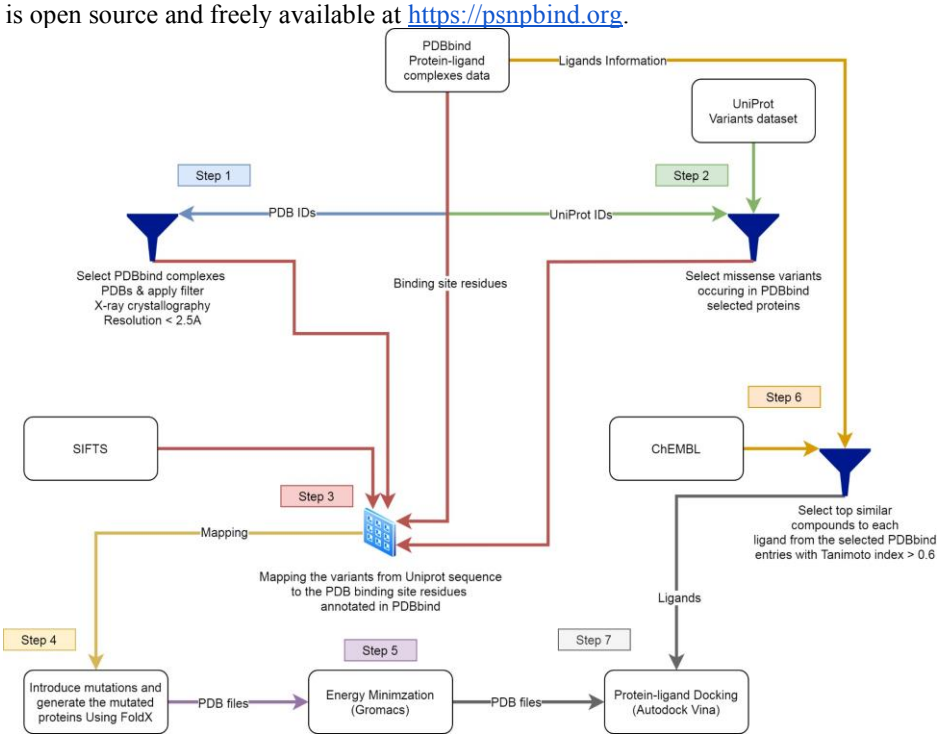



**Figure 1**: Methodology workflow. Steps 1, 2 and 3 filter the data from the main sources and map them together. Step 4 and 5 prepare the selected protein PDBs and their mutated versions for docking. Step 6 prepares the ligands. Step 7 performs the docking.


1. Daly A. Pharmacogenetics and human genetic polymorphisms. Biochemical Journal. 2010;429(3):435–449.
2. Wilke RA, Dolan ME. Genetics and Variable Drug Response. JAMA. 2011;306(3).
3. Schärfe CPI, Tremmel R, Schwab M, Kohlbacher O, Marks DS. Genetic variation in human drug-related genes. Genome Medicine. 2017;9(1).
4. Ammar, A., Cavill, R., Evelo, C. et al. PSnpBind: a database of mutated binding site protein–ligand complexes constructed using a multithreaded virtual screening workflow. J. Cheminform 14, 8 (2022). https://doi.org/10.1186/s13321-021-00573-5


## A-6: Recent Advances in Chemical Search of Ultra-large Databases


R.A. Sayle[1], J.W. Mayfield[1]

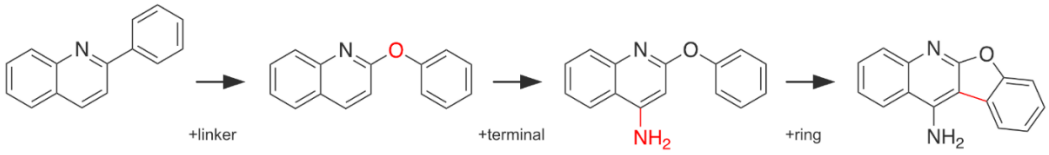
[1] *NextMove Software, Cambridge, United Kingdom*


The prolific growth of chemical databases can be seen by comparing abstracts from previous ICCS conferences. At the last ICCS in 2018, Enamine's "make-on-demand" database of molecules available for purchase had 647 million compounds, today (in February 2022) it contains over 22 billion. Even with the benefit of Moore's law, today's computer hardware isn't 35 times faster than it was four years ago, and already trillion compound databases are on the horizon.

Keeping on top of this exponential growth represents an on-going challenge to the field of cheminformatics, requiring continual innovation and advances in the algorithms and techniques for working with large data sets. This presentation will discuss several of these advanced techniques, from dealing with modern Non-Uniform Memory Access (NUMA) hardware and latest CPU architectures, through efficient federation and distribution of chemical searches over multiple servers, to file format and data compression improvements to reduce the total I/O required for each search.

This last category is particularly relevant to state-of-the-art chemical similarity measures based on Graph Edit Distance (GED) and large graph databases. These "deep sublinear" approaches allow search times to remain about the same (and even decrease) compared to 2018, despite the rapid increases in the underlying data sets, but at the storage expense of a large, precomputed index. Back in 2018, the graph index of chemical space used by the authors contained 80 billion nodes and required 6 terabytes of disk space. Today's graph indices contain over 675 billion nodes and require 34 terabytes of disk space. Efficient representation of these huge

graphs [1,2,3,4,5] is critical to keeping the disk space requirements manageable and providing interactive performance.





1. Irwin J.J., Tang K.G., Young J., Dandarchuluun C., Wong B.R., Khurelbaatar M., Moroz Y.S, Mayfield J. and Sayle R.A. "ZINC20 – A Free Ultralarge-scale Chemical Database for Ligand Discovery", Journal of Chemical Information and Modeling (JCIM), **2020**, vol. 60, issue 12, pp. 6065-6073.
2. Grabowski S., and Bieniecki W., Tight and Simple Web Graph Compression for Forward and Reverse Neighbor Queries. Discrete Applied Mathematics, **2014**, vol. 163, part 4, pp. 298-306.
3. Lemire D., and Boytsov L., Decoding Billions of Integers per second through Vectorization, Software: Practice and Experience, **2015**, vol. 45, issue 1, pp. 1-29.
4. Sayle R. and Delany J., SMILES Multigram Compression, Presented at Daylight User Group Meeting (MUG01), 2001, Santa Fe, New Mexico.
5. Baeza-Yates R.A., A Fast Set Intersection Algorithm for Sorted Sequences. In Proceedings of the 15th Annual Symposium on Combinatorial Pattern Matching (CPM), **2004**, Lecture Notes in Computer Science (LNCS), vol. 3109, pp. 400-408.

# Session B: STRUCTURE-ACTIVITY AND STRUCTURE-PROPERTY PREDICTION

## B-1: Chemical feature visualization to interpret neural network models for toxicity prediction

M. Walter[1], S.J. Webb[2], V.J. Gillet[1]

[1]*Information School, University of Sheffield, Regent Court, 211 Portobello, Sheffield S1 4DP, UK*
[2]*Lhasa Limited, Granary Wharf House, 2 Canal Wharf, Leeds LS11 5PS, UK*

Deep Neural Network (DNN) models have become a popular machine learning technique for bioactivity prediction of chemicals. Due to their complex structure, it is difficult to understand predictions made by these models which limits confidence. Current approaches to tackle this problem such as SHAP or integrated gradients provide insights by attributing importance to input features of individual compounds.[1,2] While these methods have produced promising results, they do not shed light on representations of compounds in hidden layers. Feature visualization has emerged as a popular tool to understand which features are detected by hidden layer neurons of DNN models in image classification models.[3]
The present study focuses on feedforward neural networks with RDKit's Morgan fingerprints as input. Inspired by feature visualization techniques, a novel method was developed to automatically extract chemical features responsible for activation of hidden neurons. This method leverages both information about training compounds strongly activating hidden neurons and learned model parameters. Using Ames mutagenicity as a well-understood toxicity endpoint, the method was able to extract known toxicophores. Moreover, extracted substructures can be mapped onto test compounds to obtain model explanations incorporating hidden layer representations of compounds. Using toxicophores from the Derek expert system[4] as ground truth, the explanatory capability of the approach was evaluated using attribution AUCs as metric[5]. For a large number of compounds, explanations match ground truth with an AUC above 0.8.
The proposed method may be used to extract novel toxicophores by leveraging chemical features encoded in DNN models. Furthermore, understanding of model predictions is increased by providing explanations complementary to those obtained with established attribution methods. While not explored in the present study, the proposed method could be adapted to other DNN architectures such as graph-convolutional neural networks.

1. Rodriguez-Perez, R., et al., Interpretation of Compound Activity Predictions from Complex Machine Learning Models Using Local Approximations and Shapley Values. Journal of Medicinal Chemistry., **2020,** 63, 16, 8761-8777
2. Preuer, K., et al., Interpretable Deep Learning in Drug Discovery. arXiv:1903:02788v2., **2019**
3. Olah, C., et al., Feature Visualization, Distill. **2017**
4. Marchant, C., et al., In Silico Tools for Sharing Data and Knowledge on Toxicity and Metabolism: Derek for Windows, Meteor, and Vitic. Toxicology Mechanisms and Methods., **2008**, 18, 2-3, 177-187
5. McCloskey, K. et al., Using Attribution to Decode Binding Mechanism in Neural Network Models for Chemistry., PNAS, **2019**, 116, 24, 11624-11629

## B-2: The Influence of Nonadditivity on Machine Learning and Deep Learning Models

E. Nittinger*[1], K. Kwapien[1], J. He[2], C. Margreitter[2], A. Voronov[2], C. Tyrchan[1]

[1] *Medicinal Chemistry, Research and Early Development, Respiratory and Immunology (R&I), BioPharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden*

[2]*Computational Chemistry, Discovery Sciences, R&D, AstraZeneca, Gothenburg, Sweden*

Matched molecular pairs (MMPs) is nowadays a commonly applied concept and is used in many computational tools for structure activity relationship analysis, biological activity prediction or optimization of physicochemical properties. Up to date, it has not been rigorously shown that MMPs, i.e. changing only one substituent between two molecules, can be predicted with higher accuracy and precision in contrast to any other chemical compound pair. Here, we evaluate the predictability of four classical physico-chemical parameters – logD, solubility, permeability, and clearance – in combination with different machine learning algorithms. It is expected that any model should be able to predict the small defined change with high accuracy

and reasonable precision.
The nonadditivity analysis code from Christian Kramer[1] was used to classify inhouse AstraZeneca data into additive and nonadditive. The mmpdb package[2] was used to obtain all MMPs. Thus, four different datasets were generated – (1) all data, (2) MMPs, (3) additive MMPs, and (4) nonadditive MMPs and used for training and evaluation of ML models – PLS, RF, SVR, XGBoost, and DNN.
Our study confirms that MMPs are easier to predict than using all data. In agreement to our previous ChEMBL study[3], additive data is the easiest to predict, while nonadditive data is most difficult even with deep learning models (Figure 1), which should be better suited to model non-linear events. This highlights the importance of recognizing nonadditivity events, which can reveal critical changes in SAR, and leaves the field with a still standing challenge.

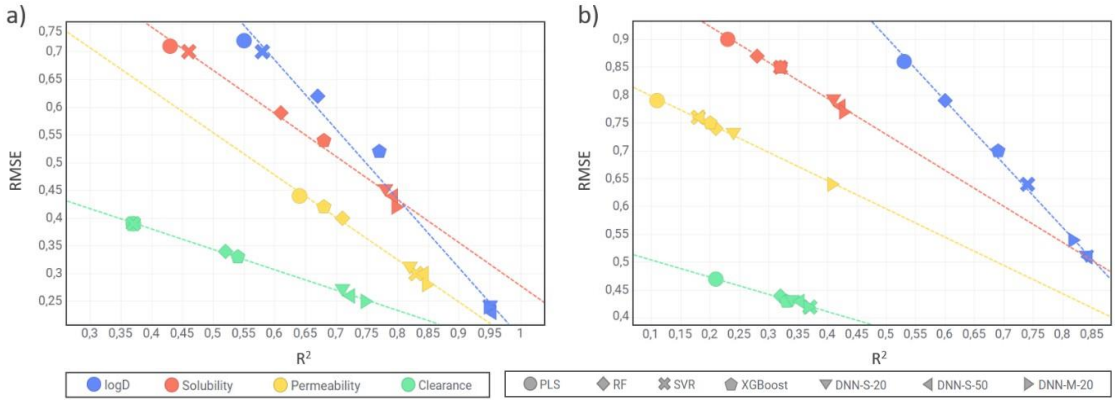



**Figure 1.** Comparison of different models and endpoints; $R^2$ against RMSE for test A) only additive data and B) only nonadditive data.

1. Kramer, C., Nonadditivity Analysis. J Chem Inf Model, **2019**, 59, 4034–4042.
2. Dalke, A., *et al.*, mmpdb: An Open-Source Matched Molecular Pair Platform for Large Multiproperty Data Sets. J Chem Inf Model, **2018**, 58, 902–910.
3. Gogishvili, D., *et al.*, Nonadditivity in public and inhouse data: implications for drug design. J Cheminform.

## B-3: Challenges of tracking SARS Cov-2 M-protease inhibitors from patents

C. Southan

*Data Sciences, Medicines Discovery Catapult, SK10 4ZF, UK*

Despite the success of COVID-19 vaccines there remains an urgent need for small-molecule antivirals. The recent orally effective M-protease inhibitor PF-07321332 thus represents a breakthrough. While Pfizer first declared the structure at an ACS meeting in April 2021 their patent WO2021250648 "Nitrile Containing Antiviral Compounds" published on the 16th of Dec 2021 followed closely by their paper on the 24th (PMID: 34726479). The problem is that the extraction of structures and activity data proceeds at different speeds by different commercial and public sources. For SARS Cov-2 targets published lead structures are curated by the Guide to Pharmacology and complemented by full SAR data sets curated from patents by BindingDB. Both resources promptly submit to PubChem. ChEMBL also extracts from papers, but the release cycle is long. This work will look at the timings affecting the flow of post-publication structures and data into PubChem as well as SciFinder. While some chemistry has been automatically extracted from WO2021250648 by Google Patents and WIPO Patentscope, SureChEMBL has not yet subsumed the chemistry into their database. None of these patent extractions has yet fed through to PubChem where the structures would be usefully merged at the compound level. SciFinder has indexed the substances but not the activity values. It also turns out that a new set of Pfizer 2021 patents include a new M-protease inhibitor series, possibly as PF07321332 back-ups. Tracking these through the system illustrates the technical challenges for automated pipelines to extract the correct structures from PDFs. Methods using open tools such as OPSIN and OSRA for manual patent curation will be exemplified. These can improve extraction fidelity for key compounds but are obviously more difficult to scale. There are additional recently declared clinical candidates including SH-879 from Sosei Heptares and S-217622 from Shinogi. The former is still blinded (i.e., no structure in the public domain) but the latter is specified in a preprint and has been curated by BindingDB (CID162533924). However, neither of these two have surfaced from patents so far (February). An academic team from Stanford also claims to have filed on their ML1000 lead compound, CID155925840. We can expect that more M-

protease patents from companies, as well as academic groups, will be published in 2022. The open science COVID Moonshot efforts have just nominated CID156906151 as their clinical candidate. It is thus important that all quality data, both open and commercial, can be extracted and tracked quickly into resources that are FAIR. Large sets of analogue activity data from patents are particularly suitable for classical pharmacophore modelling or AI/ML approaches. Extracting SAR from an expanding range of patents will thus enhance the development of further improved clinical M-protease inhibitors for battling the pandemic.

## B-4: An innovative approach of Toll-like receptor dynamics exploitation for structure optimization through 3D pharmacophore analysis

V. Talagayev[1], A. Dolsak[2], D. Sribar[1], G. Wolber[1], M. Sova[2], G. Weindl[3]

[1] *Molecular Design Lab, Institute of Pharmacy, Freie Universität Berlin, Königin-Luisestr.+ 4, 14195 Berlin, Germany*

[2] *Chair of Pharmaceutical Chemistry, Faculty of Pharmacy, University of Ljubljana, Askerceva c. 7, SI-1000, Ljubljana, Slovenia*

[3] *Pharmacology and Toxicology Section, Pharmaceutical Institute, University of Bonn, Gerhard-Domagk-Str. 3, 53121, Bonn, Germany*

Toll-like receptors (TLRs) are pattern recognition receptors responsible for the recognition of pathogen-associated molecular patterns of viruses, fungi, bacteria and parasites and therefore play an important role in the human immune response against infections. The involvement of Toll-like receptor 8 (TLR8) in multiple diseases has been reported in recent decades. Excessive activation of TLR8 leads to inflammation and autoimmune diseases, prompting a requirement for the discovery of novel selective TLR8 modulators. Through the use of 3D pharmacophore-based virtual screening, subsequent molecular docking studies and further visual selection small molecule 6-(trifluoromethyl)pyrimidin-2-amine-based TLR8 inhibitors with a novel core pyrimidine scaffold were discovered, synthesized and experimentally validated.[1,2]

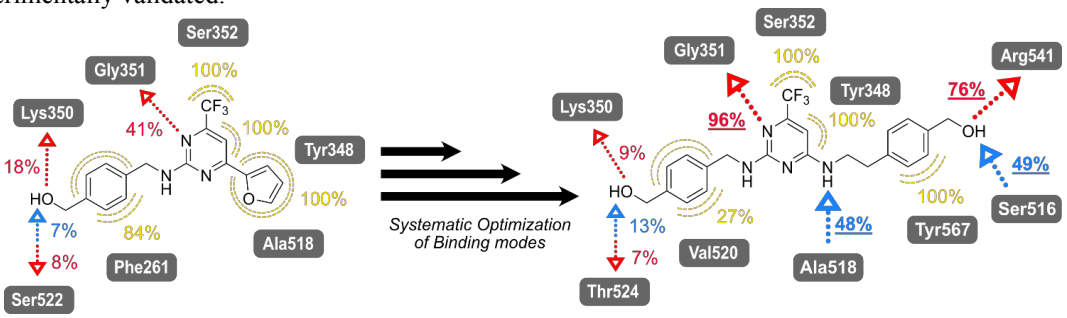



**Figure 1:** Optimization of TLR8 antagonists with the application of Dynophores. The optimized compound on the right displays additional hydrogen bonding interactions indicated by red and blue arrows, while lipophilic contacts by yellow dotted lines.

We present an innovative optimization strategy for novel 6-(trifluoromethyl)pyrimidin-2-amine-based TLR8 inhibitors based on the application of Dynophores3, a novel method of optimization and structure-activity relationship investigation through the analysis of 3D pharmacophores4 over the course of molecular dynamics simulations (Figure 1). This allows the systematic prediction of the stability of the ligand binding as well as of the interactions between the ligand and the receptor, thus facilitating the design of compounds with increased inhibitory potency.

1. Šribar, D., et al., Identification and characterization of a novel chemotype for human TLR8 inhibitors., European journal of medicinal chemistry, **2019**, 179, 744-752.
2. Dolšak, A., et al., Further hit optimization of 6-(trifluoromethyl) pyrimidin-2-amine based TLR8 modulators: Synthesis, biological evaluation and structure–activity relationships, European Journal of Medicinal Chemistry, **2021**, 225, 113809.
3. Bock, A., et al., Ligand binding ensembles determine graded agonist efficacies at a G protein coupled receptor, Journal of Biological Chemistry, **2016**, 291.31,16375-16389.
4. Schaller, D., et al., Next generation 3D pharmacophore modeling, Wiley Interdisciplinary Reviews: Computational Molecular Science, **2020**, 10.4, e1468.

# Session C: DEALING WITH BIOLOGICAL COMPLEXITY

## C-1: A Systems Biology Workflow to Support the Diagnosis of Pyrimidine and Urea Cycle Disorders

D.N. Slenter[a], I.M.G.M. Hemel[a,b], C.T. Evelo[a,b], J. Bierau[c,d], E.L. Willighagen[a], L.K.M. Steinbusch[c]

*[a] Department of Bioinformatics (BiGCaT), NUTRIM, Maastricht University, Maastricht, The Netherlands [b] Maastricht Centre for Systems Biology (MaCSBio), Maastricht University, Maastricht, The Netherlands [c] Department of Clinical Genetics, Maastricht University Medical Center, Maastricht, The Netherlands [d] Department of Clinical Genetics, Erasmus Medical Center, Rotterdam, The Netherlands*

Pyrimidine (and purine) metabolism provides essential high-energy vehicles which serve as fuel and building blocks as well as being the messenger molecules that steer these processes. The pyrimidine pathway overlaps with the urea cycle through a common metabolite (carbamoyl phosphate); the urea cycle being responsible for the production of several amino acids and removing ammonia. Various Inherited Metabolic Disorders (IMDs) disrupt these pathways and the biomarkers for these disorders overlap substantially between the IMDs. We used these well-known overlapping pathways as a proof-of-concept for the development of a framework that combines clinical and theoretical biomarkers with pathway models through network approaches and semantic web technologies.

We integrated literature and expert knowledge into machine-readable pathway models for pyrimidine and urea cycle disorders, including disease information and relevant downstream biomarkers. The theoretical change for each biomarker per disease was compiled based on a manual database search. Data of 16 previously diagnosed patients with various pyrimidine and urea cycle disorders were analysed with our framework. The top three pathways of interest were retrieved through semantic web technologies, by selecting pathways that covered most unique markers as well as showing overlap with the theoretical marker data. These pathways, and the corresponding clinical data, were visualised through network analysis. Two expert laboratory scientists evaluated our approach.

The number of relevant biomarkers for each patient varies greatly (five to 48), and likewise the pathways covering most unique biomarkers differ for equivalent disorders. The two experts reached similar conclusions with our proposed framework as with their current workflow. More specifically for the framework described here, they reached similar conclusions regarding the diagnosis of nine patient samples without knowledge about clinical symptoms or sex. For the remaining seven cases, four interpretations pointed in the direction of a subset of disorders, which could be prioritised for further investigation. Three cases were found to be undiagnosable with the data available.

The presented workflow supports the diagnosis of several IMDs of pyrimidine metabolism and the urea cycle, by directly linking biological pathway knowledge and theoretical biomarker data to clinical cases. This workflow is adaptable to analyse different types of IMDs, difficult patient cases and functional assays in the future. Furthermore, the pathway models can be used as a basis to perform various other types of (omics) data analysis, e.g. transcriptomics, metabolomics, fluxomics.

## C-2: Modeling, Proper Validation, and Discovery of Synergistic Drug Combinations

E.N. Muratov[1]

*[1]Laboratory for Molecular Modeling, UNC Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599, USA.*

In this presentation, we will discuss most recent theoretical developments and applications in the modeling, proper model validation, and discovery of chemical mixtures with desired effects. The QSAR modeling of organic mixtures requires the use of specific descriptors to characterize the different chemicals involved, taking into account their stoichiometry. We will present the system of descriptors developed for modeling chemical combinations and will discuss their advantages and disadvantages. Motivated by increasing interest in QSAR modeling of mixtures and the lack of techniques for assessing the true performance modeling of

mixtures more complex than the binary, we present a collection of statistical validation strategies for models built on N-ary mixtures, each strategy applicable to a different modeling goal. For our purposes, a model's goal is related to the composition of external mixtures of interest (binary, ternary, …, n-ary) and reflected as the proportion of mixture constituents in the intended external test data not found in the training dataset. Each such goal is characterized by a different degree of statistical dependence between training and test sets, and our validation strategies are designed to account for this dependence when assessing a real model's performance. We contend that validation of QSAR models of mixture datasets without regard for this dependence will likely lead to an unrealistic notion of model performance and, therefore, a high chance of model failure upon deployment.

We will present case studies on discovery of synergistic drug combinations against SARS-CoV-2 and pancreatic cancer. In the first case study, we hypothesized that combining drugs with independent mechanisms of action could result in synergy against SARS-CoV-2, thus generating better antiviral efficacy. Using in silico approaches, we prioritized 73 combinations of 32 drugs with potential activity against SARS-CoV-2 and then tested them in vitro. Sixteen synergistic and eight antagonistic combinations were identified; among 16 synergistic cases, combinations of the FDA-approved drug nitazoxanide with remdesivir, amodiaquine, or umifenovir were most notable, all exhibiting significant synergy against SARS-CoV-2 in a cell model. However, the combination of remdesivir and lysosomotropic drugs, such as hydroxychloroquine, demonstrated strong antagonism. Overall, these results highlight the utility of drug repurposing and preclinical testing of drug combinations for discovering potential therapies to treat COVID-19.

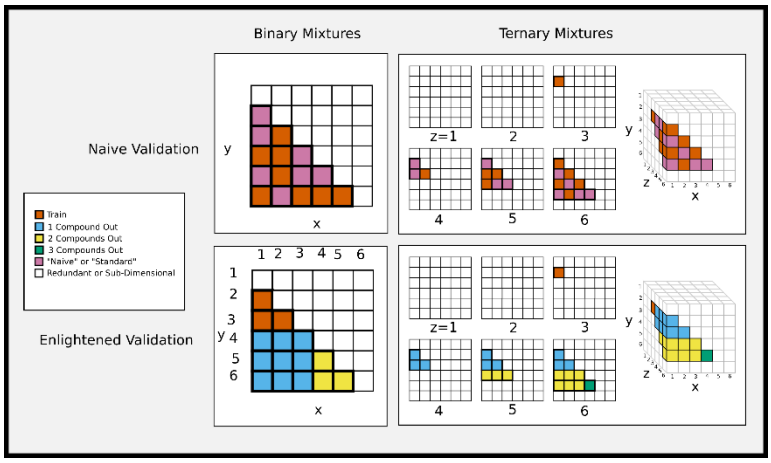



**Figure 1.** A visual comparison between naive validation and rational validation on small binary and ternary datasets. The colors of the squares designate how they would be used in a modeling task. Orange squares are always the mixtures used for training, and the other colors are labeled with their relationship to the training set. For example, in the naive validation case, training and test mixtures are chosen randomly, which results in a random scattering of orange and purple squares. The ternary mixture visualization is shown as both a cube (right), and slices of that same cube (left) to show the squares that are occluded in the cube visualization. Note: white squares are either redundant or show mixtures that are below the dimension of the visualization (e.g., binary mixtures in the ternary case). Different squares could be highlighted in each case, but they are chosen here to be as condensed as possible.

## C-3: Conformational Chirality and Protein Structure Analysis

Inbal Tuvi-Arad

*Department of Natural Sciences, The Open University of Israel, Raanana, Israel*

One of the most challenging frontiers of structural biology involves the analysis of the enormously rich library of the conformers of the building blocks of proteins. Quantitative analysis of these conformations requires a global and highly sensitive geometrical descriptor. We show that the continuous chirality measure, that quantifies the distance of a given structure from its nearest achiral conformer, is a suitable parameter for this purpose. Based on this parameter, we have developed three analysis tools: 1) The *chiral Ramachandran plot* (CRP), in which each point in the traditional Ramachandran plot is colored according to the chirality level of the relevant residue. The CRP identifies natural levels of chirality per residue type in the various secondary structure segments and can explore general trends of conformational changes. 2) The *protein chirality spectrum*, in which the chirality level per residue is plotted along the protein sequences, and upon comparing with general trends, can highlight special transitional points such as α-helix kinks, β-strand twists, and junctions that connect different secondary structure segments. 3) The *conformational similarity plot*, where residues' chirality of one peptide are plotted against the chirality level of their corresponding residues in other peptides of a protein homomer in order to highlight regions of high distortion. Analysis of crystallographic data from hundreds of proteins by these tools demonstrate, by visual and quantitative means, the role played

by side chains in dictating the conformational flexibility of proteins, and enrich our understanding of the complexity of protein structure.

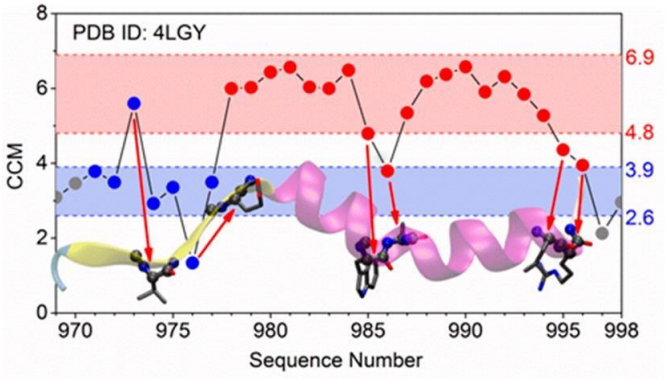



**Figure 1**: A segment from a protein chirality spectrum showing α-helix kinks and β-strand twists

1. Baruch-Shpigler Y., et al., Chiral Ramachandran plots I: Glycine. Biochemistry., **2017**, 56, 5635-5643
2. Wang H., et al., Chiral Ramachandran Plots II: General trends and proteins chirality spectra. Biochemistry., **2018**, 57, 6395–6403
3. Shalit Y. and Tuvi-Arad I., Side chain flexibility and the symmetry of protein homodimers. PLOS ONE., **2020**, 15(7): e0235863

# Session D: STRUCTURE-BASED APPROACHES

# D-1: Describing protein dynamics for proteochemometric bioactivity prediction: 3DDPDs

M. Gorostiola González[1,2], R.L. van den Broek[1], T.G.M. Braun[1], M. Chatzopolou[1], A.P. IJzerman[1], L.H. Heitman[1,2], G.J.P van Westen[1]

[1]*Division of Drug Discovery and Safety, Leiden Academic Centre for Drug Research, Leiden University, The Netherlands.*
[2]*Oncode Institute, The Netherlands.*

Proteochemometric (PCM) modelling is a powerful computational drug discovery tool used in bioactivity prediction of potential drug candidates.[1] In PCM, features are computed to describe small molecules and proteins, which directly impact the quality of the predictive models.[2,3] State-of-the-art protein descriptors are based on physicochemical, electrostatic, or topological properties.[4] However, these are calculated from the protein sequence and neglect the dynamic nature of proteins, which computationally can be simulated with molecular dynamics (MD). In small molecules, fingerprints calculated from MD simulations have been shown competitive performance to state-of-the-art descriptors in solvation free-energy predictions and substrate classification tasks.[5,6]

Here, we designed novel 3D dynamic protein descriptors (3DDPDs) to be applied in bioactivity prediction tasks with PCM models. We started by leveraging publicly available G protein-coupled receptor (GPCR) MD data from GPCRmd.[7] GPCRs exist in different conformational states that allow transmission of diverse signals and that can be modified by ligand interactions, among other factors.[8] To translate the MD-encoded protein dynamics, two types of 3DDPDs were considered: residue-specific (RS) and protein-specific (PS) 3DDPDs. The descriptors were developed by calculating distributions of trajectory coordinates and partial charges, applying dimensionality reduction, and subsequently condensing them into (fixed-length) fingerprints per residue or protein, respectively.

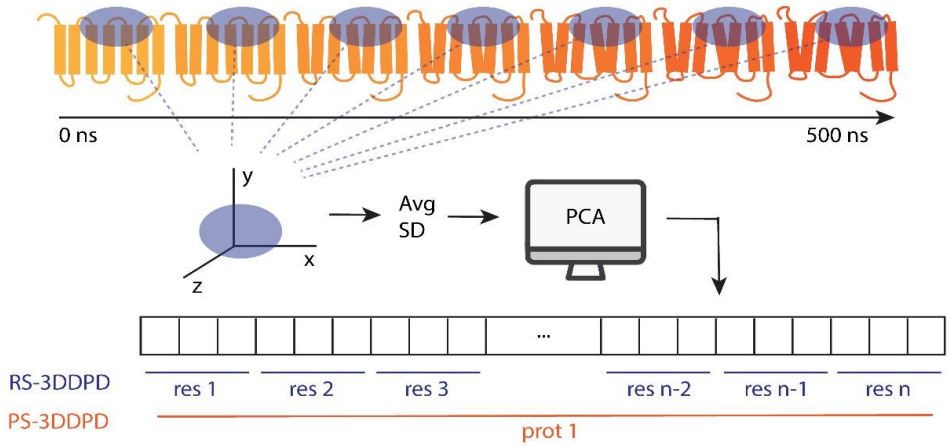



**Figure 1.** Visual abstract: Schematic representation of the development of 3DDPDs.

To evaluate the performance of the 3DDPDs, they were benchmarked on a regression PCM model against the state-of-the-art protein descriptors. The performance of our RS and PS 3DDPDs was equivalent to that of the best-performing established protein descriptor, z-scales (3) + z-scales (avg). Combinations of classic descriptors with 3DDPDs were also explored, often increasing the performance of the former.

These results highlight 3DDPDs as a steppingstone for further research on protein descriptors used for predicting drug-target interactions based on protein dynamics. This work could be very relevant in assessing differences in bioactivity driven by dynamic alterations caused by, for example, cancer related mutations.

1. Van Westen, G. J. P., Wegner, J. K., Ijzerman, A. P., Van Vlijmen, H. W. T. & Bender, A. Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets. *Medchemcomm* **2**, 16–30 (2011).
2. Van Westen, G. J. P. *et al.* Benchmarking of protein descriptor sets in proteochemometric modeling (part 2): Modeling performance of 13 amino acid descriptor sets. *J. Cheminform.* **5**, 42 (2013).
3. Van Westen, G. J. P. *et al.* Benchmarking of protein descriptor sets in proteochemometric modeling (part 1): Comparative study of 13 amino acid descriptor sets. *J. Cheminform.* **5**, 41 (2013).
4. Bongers, B. J., IJzerman, A. P. & Van Westen, G. J. P. Proteochemometrics – recent developments in bioactivity and selectivity modeling. *Drug Discov. Today Technol.* **32–33**, 89– 98 (2019).

5. Riniker, S. Molecular Dynamics Fingerprints (MDFP): Machine Learning from MD Data to Predict Free-Energy Differences. *J. Chem. Inf. Model.* **57**, 726–741 (2017).
6. Gebhardt, J., Kiesel, M., Riniker, S. & Hansen, N. Combining molecular dynamics and machine learning to predict self-solvation free energies and limiting activity coefficients. *J. Chem. Inf. Model.* **60**, 5319–5330 (2020).
7. Rodriguez-Espigares, I. *et al.* GPCRmd uncovers the dynamics of the 3D-GPCRome. *Nat. Methods* **17**, 777–787 (2020).
8. Orgován, Z., Ferenczy, G. G. & Keserű, G. M. The role of water and protein flexibility in the structure-based virtual screening of allosteric GPCR modulators: an mGlu5 receptor case study. *J. Comput. Aided. Mol. Des.* **33**, 787–797 (2019).

## D-2: Mechanism of passive membrane permeability from weighted ensemble simulations in the cloud

D. LeBard

*OpenEye Scientific, Santa Fe, NM 87508, USA*

Despite high *in vitro* potency, if a drug-like molecule cannot reach its biological target, it will be unable to perform its designed therapeutic function. Although active transport can be responsible for charged molecules entering certain cells, it is widely believed that passive permeation is the dominant membrane transit mechanism for most neutral drug-like molecules. Indeed, on the way to their target, all therapeutics must cross many biological barriers – whether epithelial cells lining the gastrointestinal tract, brain cells gating the central nervous system, or cell membranes for intracellular targets – yet a general understanding of membrane permeation remains a grand challenge of pharmaceutical development. Experimental assays made from artificial membranes or genetically engineered cell lines are often used to measure membrane permeability in a pharmaceutical setting. However, such empirical techniques can only provide an estimate of the permeability coefficient and have no information on the microscopic details of the permeation process to guide the design of new molecules. Permeability predictors like qualitative structure-permeability models can infer permeability coefficients of new compounds through statistical analysis of pre-existing experimental data, but they also provide no mechanistic understanding to help guide rational drug design. Computational methods based on the inhomogeneous solubility-diffusion (ISD) model have been developed that rely on thermodynamic parameters taken from molecular dynamics trajectory data. Unfortunately, even when the ISD model offers an accurate estimate for the permeability coefficient, it is still incapable of providing direct information about the permeation mechanism because it contains no information about the permeation pathways. To bridge the gap between permeation mechanism and permeability coefficient prediction, we have developed a new kinetic model of passive permeability. This method combines the weighted ensemble path sampling strategy with elastic cloud computing in the Orion platform for an efficient permeability prediction workflow. The output of the workflow includes an estimate of permeability coefficient, the top-weighted reactive permeation pathways, as well as analysis of the rate-limiting steps to gain insight into the permeation mechanism. We compare our predicted permeability coefficients to experimental methods for a set of traditional drug-like molecules that obey Lipinski's Rule of 5 (Ro5), and several molecules that fall outside the Ro5 as well. Implications for the use of this method as a computational assay will also be discussed.

## D-3: Integrated Structural Cheminformatics Analysis Tools for Customisable Chemogenomics Driven Kinase and GPCR Drug Design

D. Sydow[1,3,*], N.M. O'Boyle[1], A.J. Kooistra[2], A. Volkamer[3], C. de Graaf[1]

*[1]Sosei Heptares, Steinmetz Building, Granta Park, Cambridge CB21 6DG, United Kingdom*
*[2]Department of Drug Design and Pharmacology, University of Copenhagen, Universitetsparken 2,*
*2100 Copenhagen, Denmark*
*[3]In Silico Toxicology and Structural Bioinformatics, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Augustenburger Platz 1, 13353 Berlin, Germany*; *[*]Current affiliation: Sosei Heptares*

Drug discovery is a complex and iterative process, which involves many manual steps and individual decisions. Transferring knowledge from one project to another is increasingly useful with the growing amount

of data for drug targets such as kinases and GPCRs. This allows the detection of otherwise hidden patterns in the GPCRome and kinome and the guidance of the next experimental steps based on past experiences. We present here structural cheminformatics tools and analyses integrating chemical, pharmacological, and structural data for the development of customisable Computer-Aided Drug Discovery across the kinome and GPCRome.

Kinases are well-studied drug targets for decades, resulting in ~12,000 monomeric structures in the PDB [1] representing over 300 of the roughly 500 human kinases and enabling a detailed cartography of functionally relevant kinase subpockets. These datasets are comprehensively annotated in the KLIFS database [2, 3]. In the KinFragLib project [4, 5], kinase-bound co-crystallised ligands were fragmented and recombined with respect to the subpockets that they occupy. Such subpocket fragment pools are a relevant inspiration for hit expansion for kinases and beyond. In the KiSSim project [6, 7], physicochemical and spatial properties of kinase subpockets were encoded as fingerprint in order to detect (dis)similar kinase pockets. Such an assessment can guide early off-target detection and opportunities for selectivity and polypharmacology studies. Both methods are available as Python packages and are executable as pipelines in Jupyter notebooks [5, 7].

Subpocket-focused investigations of binding sites is as important for kinases as it is for GPCRs. The diversity of GPCR ligand binding sites in the combined structural GPCRome – covered by 400+ unique GPCR-ligand complexes in the PDB, covering over 110 different GPCRs, and another 370+ X-ray crystallographic and cryo-EM structures, and 75 different GPCR StaRs from Sosei Heptares' proprietary StaR® technology platform – can be exploited for GPCR Structure-Based Drug Discovery [8, 9, 10, 11].

We discuss how to rationalise ligand binding modes and mutation effects based on detailed subpocket annotations, to guide ligand repurposing by identifying orthosteric/allosteric subpocket similarity, and to select protein modelling templates by combining sequence and ligand similarity assessments. These structural chemogenomics tools are available as KNIME workflows as an extension of the 3D-e-Chem project [12, 13] and building on the GPCRdb [10, 11]. It is noteworthy, that data integration and pipeline automatization is crucial to assess the vast amount of data. However, manual curation remains an integral part of the drug discovery process.


1. Berman, H. M., et al., The Protein Data Bank. NAR, 2000, 28, 235–242
2. Kanev, G. K., et al., KLIFS: an overhaul after the first 5 years of supporting kinase research. NAR, **2020**, 49, D1, D562–D569
3. KLIFS database: https://klifs.net/
4. Sydow, D. and Schmiel P., et al., KinFragLib: Exploring the kinase inhibitor space using subpocket-focused fragmentation and recombination. JCIM, **2020**, 60, 12, 6081–6094
5. KinFragLib GitHub repository: https://github.com/volkamerlab/kinfraglib
6. Sydow, D., et al., KiSSim: Predicting off-targets from structural similarities in the kinome. ChemRxiv, **2021**
7. KiSSim GitHub repository: https://github.com/volkamerlab/kissim
8. Congreve, M., et al., Impact of GPCR structures on drug discovery. Cell, **2020**, 181, 1, 81-91
9. Vass, M., et al., Chemical diversity in the G protein-coupled receptor superfamily. Trends Pharmacol Sci, 2018, 39, 5, 494-512
10. Kooistra, A. J., et al., GPCRdb in 2021: Integrating GPCR sequence, structure and function. NAR, **2021**, 49, D335–D343
11. GPCRdb website: https://gpcrdb.org/
12. Kooistra, A. J., et al., (2018). 3D-e-Chem: Structural cheminformatics workflows for computer-aided drug discovery. ChemMedChem, 2018, 13, 6, 614–626
13. 3D-e-Chem GitHub repository: https://github.com/3D-e-Chem/workflows


## D-4: A novel antibiotic target: identifying bacterial ribosomal assembly inhibitors via 3D pharmacophore-based virtual screening


T. Noonan[1], D. Schaller[2], R. Nikolay[3], C. Spahn[3], M. Bermudez[4], G. Wolber[1]

*[1]Institute of Pharmacy, Freie Universität Berlin, Königin-Luise-Straße 2+4, 14195, Berlin, Germany; [2]Institute of Physiology, Charité Berlin, Charitéplatz 1, 10117 Berlin, Germany [3]Institute of Medical Physics and Biophysics, Charité Berlin, Charitéplatz 1, 10117 Berlin, Germany; [4]Department of Pharmaceutical and Medicinal Chemistry, Westfälische Wilhelms-Universität, Corrensstr. 48, 48149 Münster, Germany*

The 50S ribosomal subunit is the large subunit of the bacterial 70S ribosome, and its crucial role in bacterial protein synthesis renders it a popular target for existing ribosome-inhibiting antibiotics. However, in the face of increasing bacterial resistance, there is an unmet need for novel antibiotic classes as well as targets. The aim of this computer-aided drug design project is to go beyond targeting the function of the static ribosome, to identify antibiotics which disrupt the dynamic process of ribosomal assembly (Fig. 1). This approach is based on cryo-electron microscopy models of the 50S subunit at various stages during its assembly, which comprises the sequential association of ribosomal RNA (rRNA) and ribosomal proteins (r-proteins)[1]. These atomistic structures enable the identification of small molecules that bind r-proteins at the protein-rRNA interface to prevent their successful integration into the 50S complex, thus interrupting assembly. We developed a computational filtering process to identify ribosomal protein L17 as a suitable target. In the absence of known ligands, we identified a binding site on L17 and subsequently created a 3D pharmacophore from scratch for use in virtual screening. We achieved this by carefully analyzing both static and dynamic potential interaction patterns. This involved the use of PyRod[2], which uses water molecules as probes during molecular dynamics simulations to generate 3D pharmacophore features. Structure-based methods including molecular docking were applied to filter the initial hit list down to a selection of compounds, and their binding to L17 was experimentally validated via a biophysical assay using bio-layer interferometry. The compound displaying the highest affinity for L17 was chosen as a starting point for further optimization. A shape-based search using ROCS v.3.2.9 (OpenEye Scientific Software, Santa Fe, NM)[3] for analogues of this hit identified further compounds capable of binding L17. Thus, this study showcases modern computational methods in combination with biophysical validation for the design of ligands for a truly novel antibiotic target.

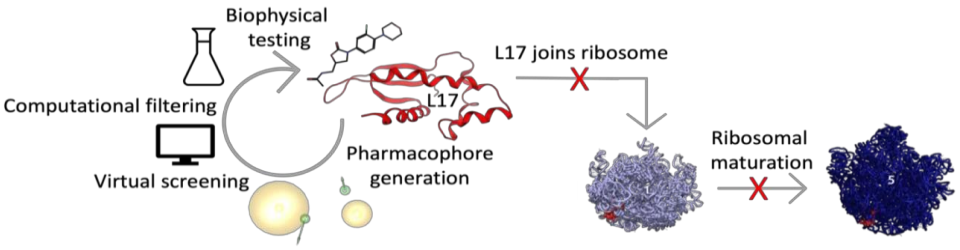



**Figure 1:** Iterative workflow for the identification and optimization of ligands for L17 (red).

1. Nikolay, R., et al., Structural Visualization of the Formation and Activation of the 50S Ribosomal Subunit during In Vitro Reconstitution. Mol. Cell., **2018**, 70(5), 881-93.
2. Schaller, D., et al., PyRod: Tracing Water Molecules in Molecular Dynamics Simulations. J. Chem. Inf. Model., **2019**, 59, 2818−2829.
3. Hawkins, P., et al., Comparison of Shape-Matching and Docking as Virtual Screening Tools. J. Med. Chem. **2007**, 50(1), 74-82.

## D-5: Dynamic interaction patterns enable characterization of opioid-peptide binding to the atypical chemokine receptor 3

K. Puls[1], S. Pach[1], G. Wolber[1], M. Bermudez[2]

[1]*Department of Pharmaceutical Chemistry, Institute of Pharmacy, Freie Universität Berlin, Königin-Luise-Str. 2+4, D-14195 Berlin, Germany*

[2]*Department of Pharmaceutical and Medicinal Chemistry, Westfälische Wilhelms-Universität, Corrensstr. 48, 48149 Münster, Germany*

The atypical chemokine receptor 3 (ACKR3) is an opioid scavenger receptor but the binding characteristics of opioid peptides remain elusive[1]. Computational approaches are further hindered by the lack of an experimentally solved structure of the ACKR3. Our research focuses on the elucidation of structural determinants involved in opioid peptide binding to the ACKR3. For this purpose, we used the AlphaFold2 prediction of the human ACKR3 model and performed a docking-based structure-activity relationship (SAR) study of the opioid peptide adrenorphin (YGGFMRRV) and a series of its analogs. The SAR analysis revealed several previously unknown factors important for opioid peptide binding at the ACKR3 and rationalized the experimentally determined activity differences of the studied ligands[1]. The obtained binding modes of adrenorphin and its potent analog LIH383 (FGGFMRRK) were further investigated in molecular dynamics (MD) simulations and analyzed via fully automatically generated dynamic pharmacophores (*Dynophores*[2]). Dynophores consist of chemical feature density clouds that represent the occurrence frequency of interactions over the simulation time facilitating detection of time-dependent protein-ligand interaction changes. Analysis of the occurrence frequency of ionic interactions and their respective interaction distances rationalizes the

higher affinity of LIH383 for the ACKR3 compared to adrenorphin (Figure 1).

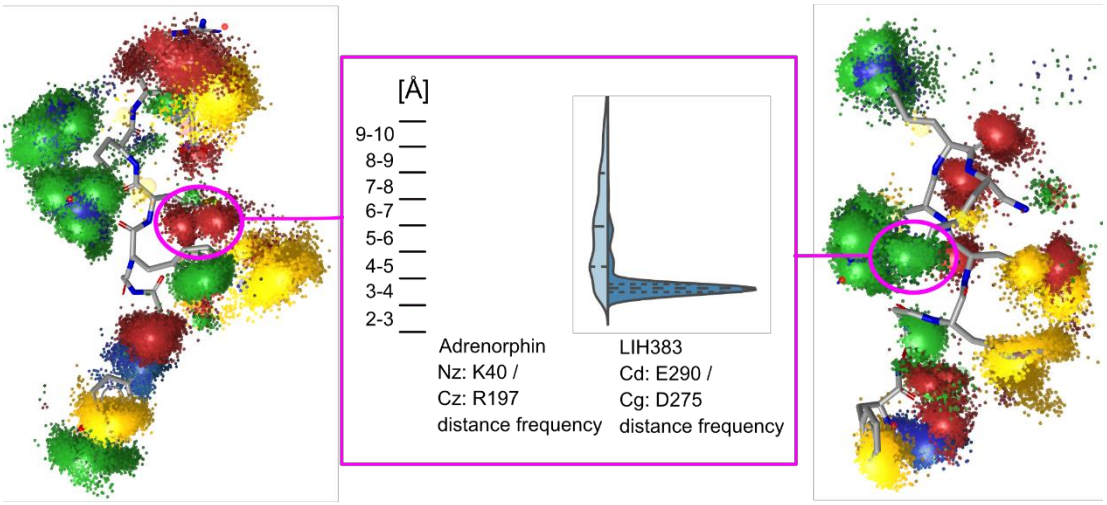



**Figure 1**: Dynophore of adrenorphin (left) and LIH383 (right). Feature clouds refer to interactions occurring over simulation time. Hydrophobic contacts are depicted in yellow, ionic interactions in blue and red, hydrogen bond donators in green. Stronger ionic interactions in the LIH383 complex due to shorter interaction distances over MD simulations (middle).

Our study shows for the first time how opioid peptides bind to the ACKR3 and explains differences in ligands potency measured experimentally. These results serve as a starting point for further studies on mechanistic understanding of ACKR3 that serves as an alternative therapeutic target for OR-related disorders such as depression and addiction.

1. Meyrath, M., et al. The atypical chemokine receptor ACKR3/CXCR7 is a broad-spectrum scavenger for opioid peptides. *Nat. Commun.* **2020**, *11*, 3033.
2. Schaller, D., et al. Next generation 3D pharmacophore modeling. WIREs Computational Molecular Science **2020**, 10.

## D-6: Development of potent FPR1 antagonists and partial agonists based on structural modelling and a detailed understanding of binding characteristics

S. Maskri[1], D. Pajonczyk[2], C. A. Raabe[2], U. Rescher[2], O. Koch[1]

[1]*Institute of Pharmaceutical and Medicinal Chemistry, University of Münster, correnstrasse 48, 48149 Münster, Germany*
[2]*Institute of Medical Biochemistry, University of Münster, Waldeyerstraße 15, 48149 Münster, Germany*

Formyl peptide receptors (FPRs) are chemotactic G-protein-coupled receptors (GPCRs) that recognize bacterial and mitochondria-derived formylated peptides. While a wide range of *N*-formylated peptides (as formyl-MLF) are potent FPR agonists, Boc-MLF or Boc-FLFLFL are FPR antagonists [1]. We used the recently published agonist-bound structure of the human FPR family member FPR2 complexed to the synthetic peptide agonist WKYMVm (pdb:6LW5), together with the published cryo-EM structure in complex with Gi (pdb:6OMM) as a starting point for an in-depth analysis of the agonist binding mode of the FPR1 homolog using molecular dynamics simulations.
FPR1 activity is discussed in the context of inflammatory diseases and might serve as a target for therapeutic interventions. Within the human FPR family, FPR2 is evolutionary closest to FPR1. Based on the overall sequence similarity of ~70% of FPR1 and FPR2, we built a homology model for FPR1 apo and holo bound to WKYMVm. Subsequently, a combination of docking and molecular dynamics simulations were used for a detailed analysis of several agonists and antagonists in complex with FPR1. Especially, the ability to turn N-term formylated agonists into antagonists just by inserting a tert-butyloxycarbonyl (Boc) group was of high interest. Our computational analysis revealed that although the binding mode of N-formylated agonistic and Boc-modified antagonistic peptides is identical, the sterically demanding Boc group is inserted between transmembrane helices three (TM3) and five (TM5), thus preventing the conformational change required for FPR activation. As this seems to be a generic feature, other known agonists were chosen as starting points for

the development of potent antagonists by a transformation into Boc and Fmoc-protected peptides, and their experimental validation revealed new antagonists and partial agonists of high interest [2]. A theory about the partial agonism could also be derived based on molecular dynamics simulations.
The retrieved binding modes and the underlying conformational changes for (partial) activation will be discussed in detail.

1. Raabe, C.A., Gröper, J, Rescher, U. Biased perspectives on formyl peptide receptors. *Biochim Biophys Acta Mol Cell Res*. **2019**, 1866:305-316.
2. Koch, O., Maskri, S., Rescher, U., Raabe, C.A., Pajonczyk, D. Development of potent antagonists and partial agonists through structural modelling of binding characteristics of potent formylated FPR1 agonists, EP 21214029.7 filed 13 December 2021

# Session E: CHEMOINFORMATICS APPROACHES

## E-1: Chemical Annotation: A new similarity score for automated design and ranking

B. Canault[1], P. Pogany[1], S. Pickett[1]

[1]*Molecular Design, Data and Computational Sciences, GlaxoSmithKline, Gunnels Wood Rd, Stevenage SG1 2NY, United-Kingdom*

In the context of accelerating drug design to propose innovative new medicines, GSK has implemented a small molecule automated design platform called BRADSHAW [1]. This automated platform is able to facilitate the generation of new chemical ideas, the modelling and prediction of crucial properties, filtering, ranking and selection of novel compounds. The final step is the annotation of results for discussion and dissemination of results. In the application of BRADSHAW we have identified several challenges in the calculation and application of appropriate chemical space metrics, which go beyond standard similarity based approaches. Given a set of lead molecules how far is too far? Can we distinguish compounds that are space filling from those that are extrapolating? How do we identify compounds for active learning? In this context, the Chemical Annotation score has been implemented. Chemical Annotation combines different cheminformatic views of "similarity" (fingerprints, maximum common substructure, reduced graph, edit distance) into a consistent metric applicable to both Hit-to-lead and Lead Optimisation stages. The scores provide a framework for understanding and interpreting compound selections in an active learning pipeline. In this presentation we will describe the current status of BRADSHAW with several examples of its successful use in projects and discuss the development and validation of Chemical Annotation within that context.

1. Green, D.V.S., Pickett, S., Luscombe, C. et al. BRADSHAW: a system for automated molecular design. J Comput Aided Mol Des 34, 747–765 (2020)

## E-2: Conformers Everywhere: Conformer Ensembles, Conformer Energies, 3D-ADMET and Machine Learning Potentials

A.H. Göller[1]

[1]*Bayer AG, Computational Molecular Design, Aprather Weg 18a, 42096 Wuppertal, Germany*

Conformers are everywhere. The identification of relevant low-energy conformers is important for target binding and thus for pharmacophore alignments, docking, or free energy perturbation based on such poses. Moreover, all physicochemical and ADMET properties are determined by interactions with off-targets, membranes or the medium. Finally, many spectroscopic properties can only be computed if complete low-energy ensembles are taken into account.

The talk will provide insights from a thorough benchmark study [1] of various force field, semiempirical and quantum-mechanical methods for relative conformer energies performed for 100 drug molecules, identifying OPLS 3 and GFN1-xTB as reliable low-cost methods.

Having identified reliable methods for energy calculation, I will then focus on ensemble complete-ness [2], i.e. the set of minimum and accessible non-minimum conformers co-existing in solvents, relevant for crossing membranes or entering binding pockets. It turns out that three conformer generator methods investigated were not able to cover the conformer space identified by molecular dynamics simulations, and even more important, that the population differences in different solvents could not be described by doing post optimization in continuum solvation. Thus, by missing most of the accessible conformer space any follow-up computation is expected to fail.

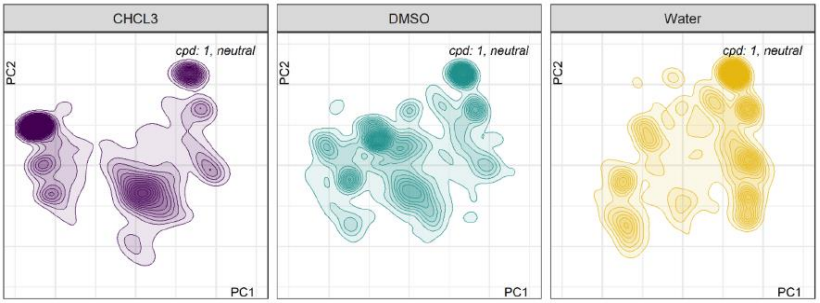



**Figure 1**: Conformer ensemble maps for a macrocycle from MD simulations in 3 solvents

Conformer ensembles are supposed to play a major role for ADMET properties. Membrane permeability for instance requires major changes in the ensemble populations while crossing the barrier. Results on first 3D-ADMET machine learning models for Caco-2 and for logD are presented.

Finally, I will give an outlook on the development of machine learning potentials for conformer ensembles with QM quality in a Bayer AG wide research project [3] that will further amplify our efforts in the areas of 3D-ADMET, calculations of experimental spectra, or molecular dynamics.

1. Cavasin, A.T., et al., Reliable and Performant Identification of Low-Energy Conformers in the Gas Phase and Water. J. Chem. Inf. Mod., **2018**, 58, 1005-1020
2. Seep, L., et al., Ensemble completeness in conformer sampling: the case of small macrocycles, **2021**, 13, 55
3. https://www.linkedin.com/feed/update/urn:li:activity:6891744599526703104/

## E-3: NFDI4Chem - The National Research Data Infrastructure for Chemistry

O. Koepler[1], F. Bach [2], S. Herres-Pawlis[3], N. Jung[4], J. Liermann[5], S. Neumann[6], M. Razum[2], C. Steinbeck[7]

*[1]Lab Linked Scientific Knowledge, TIB - Leibniz Information Centre for Science and Technology, Welfengarten 1B, 30173 Hannover, Germany:*
*[2] Research Data, E-Research (ER-FD), FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany;*
*[3] Institute for Inorganic Chemistry , RWTH Aachen University, Landoltweg 1A, 52074 Aachen, Germany;*
*[4] Institute of Biological and Chemical Systems (IBCS), Karlsruhe Institute of Technology (KIT), Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany;*
*[5] Department of Chemistry, Johannes Gutenberg University Mainz, Duesbergweg 10-14, 55128 Mainz, Germany; [6] Bioinformatics and Scientific Data, Leibniz Institute of Plant Biochemistry, Weinberg 3, 06120 Halle, Germany: [7] Institute for Inorganic and Analytical Chemistry, Friedrich-Schiller-University Jena, Lessingstr. 1, 07743 Jena*

The German National Research Data Infrastructure for Chemistry (NFDI4Chem) aims to develop a federation of data repositories, digital workflows and data standards to support researchers in all stages of the research data management (RDM) life cycle[1]. At the very heart of NFDI4Chem's vision is the Smartlab and the idea of capturing research data from the earliest possible point in time in rich, semantically annotated form in the lab using electronic laboratory notebooks and analytical device integration. From there data can easily be submitted to specialized and generic repositories either provided by the NFDI or by international organizations making it available to the community. NFDI4Chem operates within the context of the National Research Data Infrastructure (NFDI), which will eventually include 30 collaborating consortia from all fields of science. We expect that initiatives like the NFDI will substantially improve the data availability for data intensive sciences such as cheminformatics. While structural information is available in large numbers today, rich collections of experimentally determined properties of chemical compounds and materials are still rare. NFDI4Chem will provide such experimental data with associated structure and contextual information from experiment documentation. To achieve this, a whole spectrum of activities in different directions are necessary.


**Figure 1**: NFDI4Chem Activities and Services

In this presentation we will outline the work and achievements after more than one year of development NFDI4Chem. This will include showcasing NFDI4Chem services and the work on metadata standards and terminologies to annotate data, federating data repositories, considerations for a legally compliant provision of data, and implementing research data into the publication process. We will furthermore suggest routes towards a paradigm shift leading to a wide-spread RDM adoption in the academic communities.

1. Steinbeck C, et al., Research Ideas and Outcomes **2020**, 6:e55852.

## E-4: Automated Ligand Design meets Synthesis Planning

H. Briem[1], G. Mogk[1,2], M. Schimeczek[2]

[1]*Bayer AG, Computational Molecular Design, Berlin, Germany*
[2] *Bayer AG, Applied Mathematics Innovation Projects, Leverkusen, Germany*
[3] *Bayer AG, Technology & Operations, Wuppertal, Germany*

Computational tools for automated molecular design have recently gained much attention.[1,2] These tools can combine on-the-fly generation of huge virtual compound libraries, high-quality predictions of a large number of relevant properties and multiparameter-based selection of the most interesting candidates to be made in the lab. In the course of establishing these methods at Bayer Pharma, we realized that synthetic feasibility of the proposed molecules is often the bottleneck for quickly executing multiple consecutive "Design-Make-Test-Analyze" (DMTA) learning cycles required for fast compound optimization. Thus, we recently took the next logical step and combined our automated virtual design approach, called AIOLI ("AI-based Optimization of Ligands), with the powerful retrosynthesis planning tool CHAI ("Chemistry using AI") developed inhouse. This combination allows us to select those virtual molecules that not only feature superior molecular properties but in addition promise to be synthetically accessible with readily available chemicals and a low number of synthetic steps.

1. Besnard, J., et al., Automated design of ligands to polypharmacological profiles. Nature, **2012**, 492, 7428, 215-220
2. Green, D.V.S, et al., BRADSHAW: a system for automated molecular design. Journal of Computer-Aided Molecular Design, **2020**, 34, 747-765

## E-5: De novo design of synthetically accessible molecules using an evolutionary algorithm

A. Kerstjens[1], H. De Winter[1]

[1] *Department of Pharmaceutical Sciences, Faculty of Pharmaceutical, Biomedical and Veterinary Sciences, University of Antwerp, Universiteitsplein 1A, 2610 Wilrijk, Belgium*

Given an objective function that predicts a molecular property of interest, typically biological activity, *de novo* molecular design is a useful technique to identify molecules that maximize or minimize said function. However, when applied carelessly, a common drawback of these methods is that they tend to design synthetically unfeasible molecules.
LEADD (Lamarckian Evolutionary Algorithm for *de novo* Drug Design) [1] is an algorithm and software that optimizes the fitness of molecules while preserving their synthesizability by imitating the molecular connectivity of reference synthetically accessible molecules.
Molecules in a reference virtual library are assigned atom types and fragmented. Broken bonds result in typed fragment connectors. The resulting fragments and their frequencies are recorded. Atom types involved in broken bonds are assumed to be compatible. LEADD designs molecules as graphs of molecular fragments, with bonds being formed solely between compatible connectors.
A population of molecules is optimized in an evolutionary algorithm. Mutations alter the individuals stochastically and the objective function exerts selective pressure. Fragment frequencies are used to bias the outcome of the mutations. A novel set of genetic operators ensures that the compatibility rules are respected in a computationally efficient manner.
LEADD was compared to an alternative evolutionary algorithm [2] and a virtual screen in a standardized benchmark [3]. Both evolutionary algorithms were able to find fitter molecules than a virtual screen and did so more efficiently. However, LEADD found slightly fitter and substantially easier to synthesize molecules than the comparable algorithm. The major factor accounting for LEADD's improved synthesizability was identified as the atom typing scheme's degree of degeneracy, with more exhaustive atom typing schemes leading to better synthesizability but worse optimization power.

1. Kerstjens, A., De Winter, H., LEADD: Lamarckian evolutionary algorithm for de novo drug design. Journal of Cheminformatics. **2022**, Volume 14, Article 3
2. Jensen, J., A graph-based genetic algorithm and generative model/Monte Carlo tree search for the exploration of chemical space. Chemical Science, **2019**, Volume 10, pp. 3567-3572

3. Brown, N. et al., GuacaMol: Benchmarking Models for de novo Molecular Design. Journal of Chemical Information and Modeling, **2019**, Volume 59, pp. 1096-1108

## E-6: Assigning Diastereomers by Comparing Experimental and Theoretical IR Spectra

S. Riniker[1]

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

The relative stereochemistry and isomeric substitution pattern of organic molecules is typically determined using nuclear magnetic resonance spectroscopy (NMR). However, NMR spectra are sometimes nonconclusive, e.g., if spectra are extremely crowded, coupling patterns are not resolved, or if symmetry reasons preclude interpretation. Infrared spectroscopy (IR) can provide additional information in such cases, because IR represents a molecule comprehensively by depiction of the complete set of ist normal vibrations. The challenge is thereby that diastereomers and substitution isomers often give rise to highly similar IR spectra, and visual distinction is insufficient and may be biased. For this purpose, we have recently developed the IR spectra alignment (IRSA) algorithm[1] for automated optimal alignment. IRSA provides a set of quantitative metrics to identify the candidate structure that agrees best with the experimental spectrum. We will first present the basic idea and procedure of IRSA, followed by a discussion of the most recent improvements for handling strongly overlapping peaks in the IR spectrum and for aligning multiple spectra from different sources (e.g., IR and VCD or Raman). In addition to a performance assessment on rigid to flexible small molecules, the potential of the IRSA approach is demonstrated with the application to the natural product mutanobactin D,[2] a cyclic peptide of the human microbiome.

1. Böselt, L., et al., Determining the Regiochemistry and Relative Stereochemistry of Small and Druglike Molecules Using an Alignment Algorithm for Infrared Spectra. Anal. Chem., **2020**, 92, 9124

2. Pultar, F., et al., Mutanobactin D from the Human Microbiome: Total Synthesis, Configurational Assignment, and Biological Evaluation. J. Am. Chem. Soc., **2021**, 143, 10389

## E-7: Tautomerism analyses in preparation of InChI V2

D. Dhaked[1], M. Nicklaus[2]

[1]*Dept. of Pharmacoinformatics, NIPER, Kolkata, India*
[2] *InChI Tautomerism Working Group, IUPAC, RTC, NC, USA*

The InChI Project "Redesign of Handling of Tautomerism for InChI V2" (Project No.: 2012-023-2-800, https://iupac.org/project/2012-023-2-800) was created to address the shortcomings of the current InChI (version 1) algorithm by establishing requirements and guidelines for improving the handling of tautomerism in the next generation of InChI. We will present an overview of the work of this task group. More than 100 tautomeric transforms have been identified mostly from experimental literature but also from reviews and text books, other software and databases. We report on a collection of such rules,[1] a tautomer database of results extracted from experimental literature,[2] an online tool to calculate tautomers for user-submitted structures based on more than 80 rules,[3] and a recent analysis of tautomeric conflicts in 40 databases,[4] "tautomeric conflict" being defined as an occurrence of two or more structures within a data set identified by the tautomeric rules applied as being tautomers of each other. We also report on analyses of a small subset of these rules that could be integrated into the InChI V1 algorithm. We present an outlook of how all these results may influence the development of an InChI V2 algorithm.

1. Dhaked, D.K. *et al*. Toward a Comprehensive Treatment of Tautomerism in Chemoinformatics Including in InChI V2. *J. Chem. Inf. Model.* **2020,** 60: 1253–1275.
2. Dhaked, D.K. *et al.* Tautomer Database: A Comprehensive Resource for Tautomerism Analyses. *J. Chem. Inf. Model.* **2020,** 60: 1090–1100.
3. Tautomerizer - https://cactus.nci.nih.gov/tautomerizer/.
4. Dhaked, D.K., and M. Nicklaus. Tautomeric Conflicts in Forty Small-Molecule Databases. **2021.** ChemRxiv 10.26434/chemrxiv.14779254.v1.

# Session F: ARTIFICIAL INTELLIGENCE APPROACHES

## F-1: Improved classification of protein function by a localized 3D protein descriptor and deep learning

K. van der Weg[1,2], E. Merdivan[3], M. Piraud[3], H. Gohlke[1,2,4]

[1]*Computational Biophysical Chemistry, Jülich Supercomputing Centre, Wilhelm-Johnen-Straße 52428 Jülich, Germany*
[2]*Institute of Bio- and Geosciences: Bioinformatics (IBG-4), Forschungszentrum Jülich, Wilhelm-Johnen-Straße 52428 Jülich, Germany*
[3]*Helmholtz AI, Helmholtz Zentrum München, Ingolstädter Landstraße 1, D-85764 Neuherberg, Germany*
[4]*Computational Pharmaceutical Chemistry, Heinrich-Heine-University, Universitätsstr. 1, 40225 Düsseldorf, Germany*

It is essential to understand target enzyme function for applications in biomedicine and biotechnology. A good method to predict the function of new enzymes is the classification through (deep) neural networks in combination with large structural datasets. The best performing structural function predictors use the backbone[1] or $C_\alpha$ atom[2] locations in connection with a Graph Convolutional Network (GCN). However, that way, information from side-chain atoms that often carry out enzymatic function is omitted. Furthermore, to reduce the computational requirement needed for these large systems, a more sophisticated representation of an enzyme than the sequence or fold is needed.
Here, we show that we can improve enzyme function prediction compared to established methods in the field by creating a localized 3D enzyme descriptor consisting of 30 different atom types and applying it in a 3D-Convolutional Neural Network and a 3D-GCN. We generated a database consisting of 9039 structural models of enzymes with TopModel[3] to supplement structures obtained from the Protein Data Bank. Testing our descriptor on this database improves the F1-score up to 17% in enzyme classification tasks compared to fold representation methods. Furthermore, we implemented better GCNs, SchNett[3] and DimeNetPP[4], for atom classification. This increases the performance by 13% and 16% on the enzyme classification task.
Our results demonstrate that a localized 3D descriptor is the better alternative to current reduced structure representations used in enzyme prediction networks. We anticipate that the localized 3D descriptor can be used in other protein prediction tasks, e.g., in ligand binding site detection and protein-ligand binding affinity prediction. Moreover, we show that current methods can improve their performance by implementing SchNett and DimenetPP for atom prediction tasks.

1. A. Amidi, et al. EnzyNet: enzyme classification using 3D convolutional neural networks on spatial representation, ArXiv, 1707.06017
2. V. Gligorijević, et al. Structure-based protein function prediction using graph convolutional networks. Nat Commun, doi: 10.1038/s41467-021-23303-9
3. D. Mulnaes, et al. TopModel: Template-Based Protein Structure Prediction at Low Sequence Identity Using Top-Down Consensus and Deep Neural Networks, J. Chem. Theory Comput., doi: 10.1021/acs.jctc.9b00825
4. K.T. Schütt, et al. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions, arXiv, 1706.08566
5. J. Klicpera, et al. Fast and Uncertainty-Aware Directional Message Passing for Non-Equilibrium Molecules, NeurIPS-W, 2020

## F-2: Augmented Hill-Climb improves language-based *de novo* molecule generation as benchmarked *via* the open source MolScore platform

M. Thomas[1], N.M. O'Boyle [2], A. Bender A[1,*], C. de Graaf C[2,*]

[1] *Centre for Molecular Informatics, University of Cambridge, Cambridge, UK*

[2] *Computational Chemistry, Sosei Heptares, Cambridge, UK*

Artificial intelligence is now strongly embedded in computational drug discovery and design, such as the

combination of language-based models (such as recurrent neural networks) with reinforcement learning (RL) to condition SMILES generation towards desirable endpoints to aid drug design. However, RL can be a sample-inefficient learning process, sometimes requiring up to $10^5$ molecules to be sampled to optimize more complex objectives [1, 2]. This serves as a practical limitation in particular when using more computationally expensive scoring functions, such as docking, for structure-based drug design.

We will discuss the limitations of current RL strategies used with language-based models and how our proposal – a hybrid strategy we call Augmented Hill-Climb (AHC) – addresses those limitations. This strategy improves optimization ability about 1.5-fold and sample-efficiency about 45-fold compared to REINVENT [1] when conducting *de novo* structure-based design *via* Glide docking on four GPCR targets ($D_2$, μ, $AT_1$ and $OX_1$ receptors). Moreover, benchmarking this strategy against other commonly used RL strategies on six representative tasks of varying difficulty highlights state-of-the-art performance, not only with respect to objective optimization but also the chemistry being generated (which is often overlooked in generative model publications, Figure 1). These improvements enable more computationally expensive scoring functions to be tractable without the need for large compute clusters by reducing sample requirements by two orders of magnitude, to the order of $10^3$.

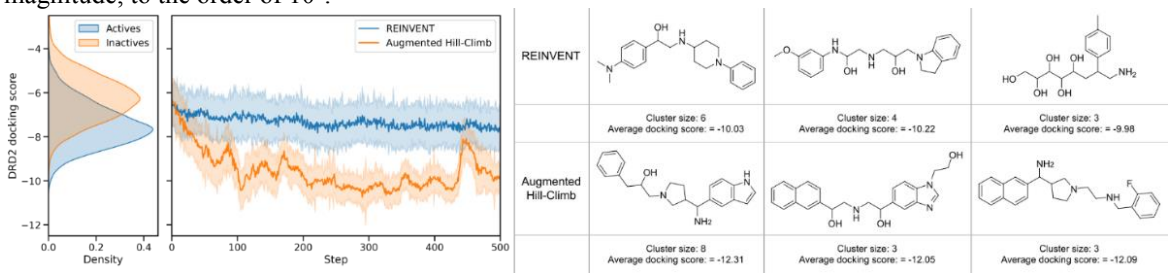



**Figure 1**: Example of improved Dopamine Receptor D2 docking score optimization and the three most common chemotypes generated de novo.

This comparison and benchmarking of goal-directed generative model algorithms and/or scoring functions was conducted by our simple-to-use python framework MolScore [3]. We will discuss how this framework implements a variety of scoring function capabilities, diversity filters, score modifications, performance metrics, apps (to provide a graphical user interface for parameter setup and optimization monitoring) and is easily extendable and implementable (with just three lines of code). We will describe how the MolScore framework has been used for comparing structure- and ligand-based scoring functions for *de novo* molecule generation [2], comparing prior datasets for generative model training, as well as benchmarking molecular representations and QSAR algorithms and their subsequent effect on *de novo* molecule generation.

Olivecrona, M., Blaschke, T., Engkvist, O. and Chen, H., Molecular de-novo design through deep reinforcement learning. J Cheminform 9, 48 (2017)
Thomas, M., Smith, R.T., O'Boyle, N.M., de Graaf, C. and Bender, A., Comparison of structure- and ligand-based scoring functions for deep generative models: a GPCR case study. J Cheminform 13, 39 (2021)
Thomas, M., MolScore: An automated scoring function to facilitate and standardize evaluation of goal-directed generative models for de novo molecular design. https://github.com/MorganCThomas/MolScore

## F-3: Explaining and avoiding failure modes of artificial intelligence for small molecule design

M. Langevin[1, 2], R. Vuilleumier [1], M. Bianciotto[2]

[1]*PASTEUR, Département de chimie, Ecole Normale Supérieure, PSL University, Sorbonne Université, CNRS, Paris, France*
[2]*Molecular Design Sciences - Integrated Drug Discovery, Sanofi R&D, Vitry-sur-Seine, France*

Boosted by recent progress in machine learning, Artificial Intelligence (AI) for small molecule design has sparked widespread interest in medicinal chemistry[1]. AI algorithms are especially used to design *in-silico* novel molecular structures, aided by structure-based or machine learning models that guide the AI towards compounds with desired predicted potency and Absorption-DistributionMetabolization-Excretion-Toxicity (ADMET) profile.

Nevertheless, some questions still remain concerning the ability of AI algorithms to perform unbiased exploration of chemical space, and whether they might be exploiting biases of the predictive models that guide them. In a recent study[2], it was shown that AI generates compounds with high scores according to the model used to guide it (the optimization score), but with low scores according to control models (i.e. models trained on the same data and the same endpoint). This problematic behavior has prompted discussions within the scientific community[3, 4] about whether it could hamper the use of AI for drug design.

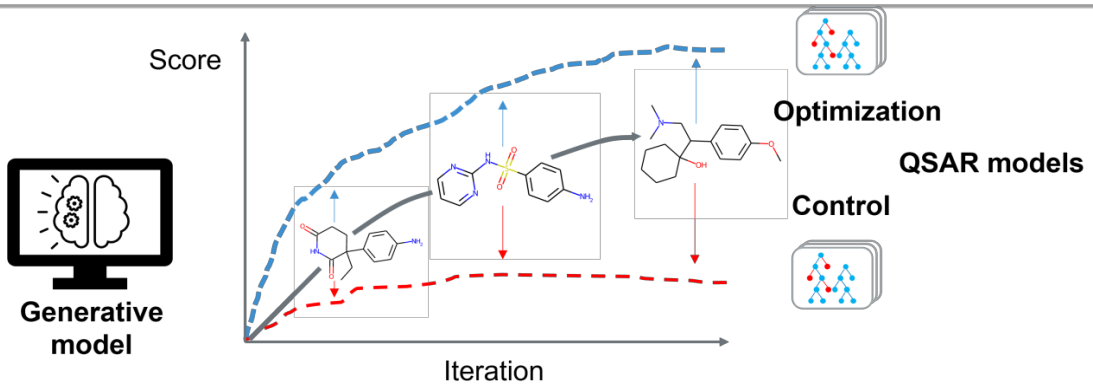



**Figure 1**: As the generative AI algorithm proposes compounds with increasing optimization score (e.g. predicted bioactivity by a QSAR model), control scores given by similar models on the same endpoint remain significantly lower. We investigate in-depth this problematic behavior.

We investigate this problematic behavior of AI algorithms[5]. To understand it, we conduct an in-depth analysis of the datasets used previously to evaluate AI algorithms for molecular design, and show that biases already present in the datasets and predictive machine learning models are causing the failures of AI. To follow through this analysis, we gather datasets devoid of those biases. We then identify simple conditions where the AI algorithms do generate structures that are predicted as favorable by the optimization and the control models.

1. Schneider, P. et al. Rethinking drug design in the artificial intelligence era. Nature Reviews Drug Discovery., **2019**, 19, 353–364.
2. Renz, P., Rompaey, D.V., Wegner, J.K., Hochreiter, S., Klambauer, G. On failure modes in molecule generation and optimization. Drug Discovery Today: Technologies., **2019,** 32-33, 55–63
3. Thomas, M., Smith, R.T., O'Boyle, N.M., de Graaf, C., Bender, A. Comparison of structure- and ligand-based scoring functions for deep generative models: a GPCR case study. Journal of Cheminformatics., **2021** 13(1), 39
4. Walters, W.P., Barzilay, R.: Critical assessment of AI in drug discovery. Expert Opinion on Drug Discovery., **2021**, 16(9), 937–947
5. Langevin, M., Vuilleumier, R., Bianciotto. Explaining and avoiding failure modes in goaldirected generation. Chemrxiv., **2021**, Chemrxiv preprint, 10.26434/chemrxiv-2021-4m6b3v2

## F-4: Multi-Instance Learning Approach to Predictive Modeling of Molecular Properties: new or well forgotten old?

T. Madzhidov[1], D. Zankov[1,2], A. Varnek[2], P. Polishchuk[3]

[1]*Chemoinformatics and Molecular Modeling, Kazan Federal University, Kazan, 29 Kremvlevskaya, Russia*

[2] *Laboratory of Chemoinformatics, University of Strasbourg, Strasbourg, 4 rue B. Pascal, France*

[3] *Institute of Molecular and Translational Medicine, Palacky University Olomouc, Olomouc, Hněvotínská 1333/5, Czech Republic*

Modern structure-property modeling approaches use machine learning algorithms to build predictive models for different chemical properties. The molecule is numerically described with a vector of chemical descriptors, often with 2D descriptors that encode only 2D information of the molecule. Meanwhile, 3D molecular information can be valuable for modeling many of the molecular properties, but 3D modeling approaches are affected by the problem of the proper choice of the molecule conformations for model building. 4D-QSAR was proposed to solve this problem as a technique that operates on averaged conformation ensembles descriptors. Multi-instance machine learning approaches were proposed for working with objects that could have many different representations. We review possible approaches of multi-instance learning and show that 4D and nD QSAR approaches utilize naïve multi-instance learning (MIL) approaches.
Possible applications of multi-instance machine learning for QSAR are described. We will show how recent

deep learning-based approaches of MIL can be applied for 3D QSAR taking into account multiple conformations of the molecule. Multi-instance learning (MIL) can be considered as the development of ordinary single-instance learning (SIL), where the molecule is represented by a single conformation, often by the lowest-energy one. We prepared the first comprehensive comparison of MIL approaches with traditional QSAR approaches based on 2D and 3D descriptors [1] in the task of modeling the biological activity of compounds. The results show that 3D MIL models outperform single-instance 3D QSAR models (built using the lowest-energy conformation), and in many cases MIL outperforms traditional QSAR models based on 2D descriptors. Moreover, we show that attention-based MIL approaches can correctly highlight bioactive conformation.

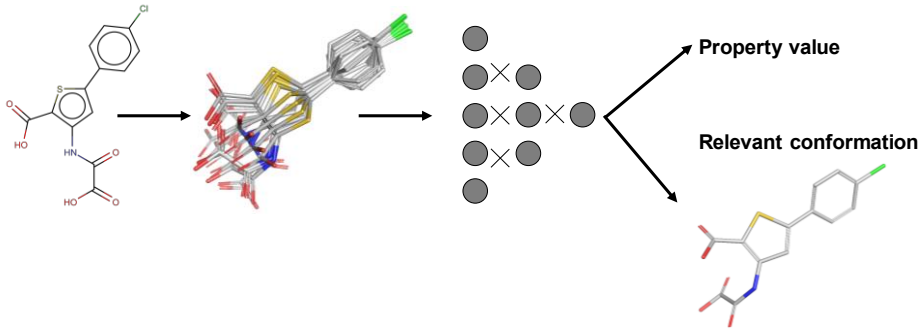



**Figure 1**: Multi-instance learning approach in molecular modeling

We also tested the 3D MIL approach on the task of predicting the catalyst enantioselectivity [2] in asymmetric organic synthesis. Our results show that for diverse examples of catalysts and reactions, the 3D MIL approach show better quality than the SIL and the classical 2D models.

1. Zankov et al., QSAR modeling based on conformation ensembles using a multi-instance learning approach. JCIM, **2021**, 61, 10, 4913-4923
2. Zankov et al., Multi-instance learning approach to predictive modeling of catalysts enantioselectivity. Synlett, **2021**, 32, 18, 1833-1836

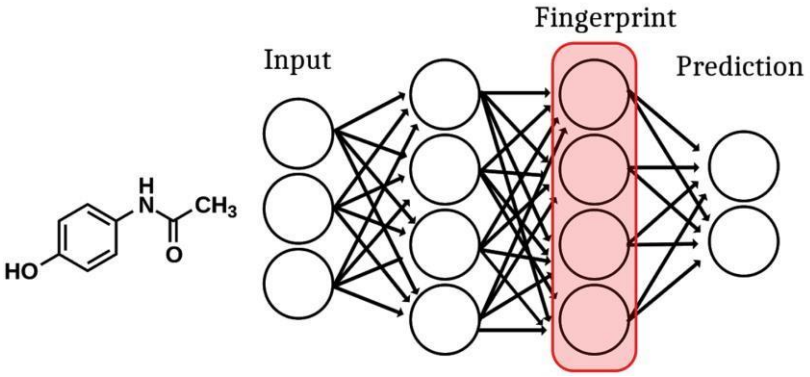
*This work was supported by subsidy allocated to Kazan Federal University for the state assignment in the sphere of scientific activities (agreement No 075-03-2021-299/6).*

## F-5: Neural Fingerprints: Generating Domain-specific Molecular Fingerprints Using Neural Networks

J. Menke[1], O. Koch[1]

[1] *University of Münster, Corrensstraße 48, 48149 Münster*

Similarity-based virtual screening remains an important technique in the early stages of the drug discovery process. Amongst other things, the success relies on the appropriate choice of the underlying molecular representation, the molecular fingerprint. Our work focuses on improving these molecular representations to encapsulate more domain-relevant information with the help of neural networks. This approach works by extracting activations of the last hidden layer of a trained neural network as a novel neural network fingerprint representation for similarity-based virtual screening.




For a thorough validation of this strategy, different architectures were trained on a variety of predictive tasks. Initially, we focused on kinase inhibitors [1] and the creation of an activity-sensitive molecular representation using a dataset of around 50,000 molecules that were tested on 160 kinases. Later, we developed natural

product-specific neural fingerprints as a structure-sensitive molecular representation [2]. For both, the kinase inhibitor, and the natural products traditional feed-forward networks (MLP) were compared to Graph Neural Networks (GNNs) in their neural fingerprint-based virtual screening performance. The evaluation of these fingerprints was done through multiple similarity searches, for which the classification quality of the found molecules was analyzed. The neural fingerprints were compared to well-established fingerprints like the ECFP [3] or other fingerprints like the autoencoder-based CDDD[4].
For both, the kinase inhibitors, as well as natural products, the neural fingerprints outperform other fingerprints in similarity search, by providing overall more active hits than any other. We could show that it is possible to generate domain-specific neural fingerprints as a structure- and activity-sensitive molecular representation through the usage of supervised training for neural networks. Interestingly, we found that GNNs, compared to MLPs, created worse neural fingerprints when trained on the same tasks. Additionally, we were able to extract a Natural Product Likeness Score[2], as an alternative measure of assessing how likely a molecule is a natural product.

1. Menke, J., Koch, O., Using Domain-Specific Fingerprints Generated Through Neural Networks to Enhance Ligand-Based Virtual Screening. J. Chem. Inf. Model. **2021**, 61(2): 664-675.
2. Menke, J., Massa, J., Koch, O., Natural product scores and fingerprints extracted from artificial neural networks. Comput. Struct. Biotechnol. J. **2021**, 19. 4593-4602
3. Rogers, D., Hahn ,M., Extended-Connectivity Fingerprints. J. Chem. Inf. Model. **2010**, 50(5): 742-754.
4. Winter, J. Montanari, F., et al., Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations Chem. Sci. **2019**, 10, 1692-1701

## F-6: Ranking generated molecule conformations using deep-learning predicted deviation to target-bound conformations

B. Baillif[1], A. Bender[1], J. Cole[2], I. Giangreco[2]

*[1]Yusuf Hamied Department of Chemistry, University of Cambridge, Lensfield Rd, CB2 1EW, Cambridge, United Kingdom*
*[2] Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, United Kingdom*

Conformation generation is an essential process in drug discovery to identify potential 3D structures of small molecule drug candidates in the contexts of docking or pharmacophore searching. While state-of-the-art generators attempt to produce diverse low-energy (likely) conformations, they are not able to retrieve target-bound poses of all small molecules [1], which could be of higher energy (i.e., by adopting elongated conformations). A method to identify the "bioactiveness" of a conformation could hence guide existing methods to generate or emphasize those conformations that could represent likely target-bound poses. In this regard, using 11000 ligand poses in the PDBBind dataset and up to 100 conformations generated with the CSD conformer generator for each molecule, we trained a deep learning model based on SchNet [2] to predict the atomic root-mean-square deviation (ARMSD) of an input molecule conformation to its closest known bioactive conformation. On an external dataset (Platinum [1]), ARMSD predictions achieved a mean root-mean-squared error of 0.60 ± 0.03, allowing to rank sets of conformations of a molecule in a fit-for-purpose mode: the model ranks first a bioactive conformation among generated conformations for 30% of the Platinum molecules. Moreover, the model enriches the 10% closest generated conformations to bioactives in the 20% top ranked for a molecule with an average Enrichment Factor ($EF_{20\%}$) of 1.73 ± 0.07, thereby outperforming CSD conformer generator ordering which achieved an average $EF_{20\%}$ of 1.17. Furthermore, short-listing conformations for input in rigid-ligand docking experiments using GOLD allows to reach similar docking power (retrieving the correct pose) to flexible-ligand docking for a significantly lower runtime compared to using all generated conformations. Hence, the approach presented here could lead to improved screening results (and potentially also reduced computational expense) in drug design applications requiring input conformations.

1. Friedrich, N.-O., et al., High-Quality Dataset of Protein-Bound Ligand Conformations and Its Application to Benchmarking Conformer Ensemble Generators. *Journal of Chemical Information and Modeling* **2017**, *57* (3), 529–539.

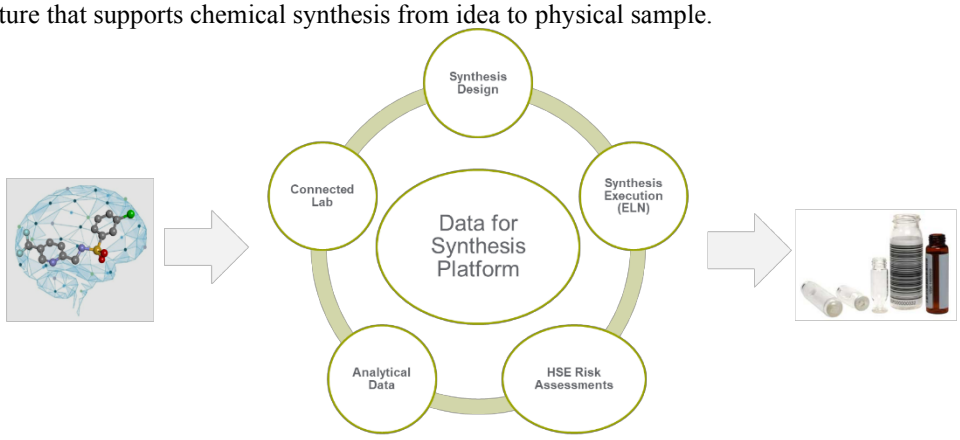
2. Schütt, K. T., et al., SchNet – A Deep Learning Architecture for Molecules and Materials. J. Chem. Phys. 2018, 148 (24), 241722.

## F-7: Digital Chemistry at Syngenta: From academic labs to industrial applications

A.R. Finkelmann[1]

[1]*Syngenta Crop Protection AG, R&D IT, Stein, Switzerland*

Modern agrochemicals must strike the right balance across a large panel of target properties from biological efficacy, environmental impact, resistance management, and cost of goods. This is arguably one of the most complex optimization tasks in the chemical industry. Recent breakthroughs in inverse design and generative chemistry enable to rethink this optimization approach.[1,2] Successful adoption of inverse design as research strategy requires high quality data to build accurate models for relevant target properties. Most importantly, compounds need to be designed that can be readily synthesized. To address these challenges, Syngenta Crop Protection Research Chemistry has initiated an ambitious program to overhaul the whole software infrastructure that supports chemical synthesis from idea to physical sample.




In this presentation we will describe the main concepts and philosophy that went into the design of the platform and how it enables to integrate recent cutting-edge technology in a production environment that will ultimately serve hundreds of chemists worldwide. We will highlight the underlying modeling of chemical information and incorporation of large-scale reaction data for reaction prediction and mapping of synthesis targets and routes against the network of known organic reactions. Several challenges that are subject to current research will be touched.

1. Vanhaelen, Q.; Lin, Y.-C.; Zhavoronkov, A. The Advent of Generative Chemistry. *ACS Med. Chem. Lett.* **2020**, *11* (8), 1496–1505
2. Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering. *Science (80).* **2018**, *361* (6400), 360–365

## F-8: Translating data to predictive models

A. Tarcsay[1], L. Antal[1], E. Andras[2], V. Jobbagy[1], D. Szisz[1]

[1]*Chemaxon Kft., Budapest Zahony str. 7, Hungary*

Biological, chemical and physical properties of molecules are encoded in their molecular structure. The challenge lies in discovering the relationships between the molecular graphs and the measured activity. Where data is measured, collected and curated for a series of compounds there is an opportunity to find the hidden relationships.

Chemical structures come in various shapes and sizes, depending on the scientists or even algorithms that create them. Though variability may sometimes seem subtle to a trained chemist's eyes, these can introduce inconsistencies that impair chemical search algorithms or model building. Structure normalization is a key component of any cheminformatics workflow with an often underestimated significance. Finding relationships between chemical structures and their measured properties primarily relies on the representation of the

chemical matter. Variability of the calculated features and descriptors for these representations can influence data analysis and accuracy of the predictions. During the first part of the presentation we will present the effect of chemical normalization on investigating correlations and building predictive models.
The second part of the talk will incorporate the results of model building on 163 ChEMBL targets extracted from the bioactivity benchmark set[1]. Results with different descriptor generation methods including ECFP fingerprints, MACCS key, structural properties, geometry properties and phy-chem properties will be discussed in detail. This part focuses on summarizing the results of more than 3000 Random Forest models.
Finally model development for ADMET targets will be highlighted including hERG cardiotoxicity prediction, permeability and blood brain barrier penetration. We will describe how these models can be built, analyzed, optimized and deployed using our new machine learning platform.

1. Eelke B, et al., Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. Journal of Cheminformatics., **2017**, 9.

# Poster Session Abstracts RED

## P-01: HASTENing structure-based virtual screening of large chemical libraries

T. Kalliokoski[1] and Ainoleena Turku[1]

[1] *Orion Pharma, Orionintie 1A, 02101 Espoo, Finland*

Virtual screening using molecular docking is one of the standard *in silico*-tools in early drug discovery. The process consumes a lot of computational resources, especially when docking a great number of molecules or using multiple conformations of the docking target (ensemble docking). A machine learning (ML) methodology "macHine leArning booSTEd dockiNg" (HASTEN) was developed to increase the throughput of structure-based virtual screening[1]. HASTEN employs an iterative process that is run until enough many high-scoring compounds have been retrieved (Figure 1). HASTEN was validated using 12 datasets from literature, together with one in-house dataset. Validation studies demonstrated that by docking 10% of a three to four million compounds database, HASTEN could retrieve approximately 80% of the top scoring compounds. The methodology is independent of the docking software and enables also usage of different ML algorithms (Chemprop[2] is deployed by default). HASTEN is freely available from https://github.com/TuomoKalliokoski/HASTEN.

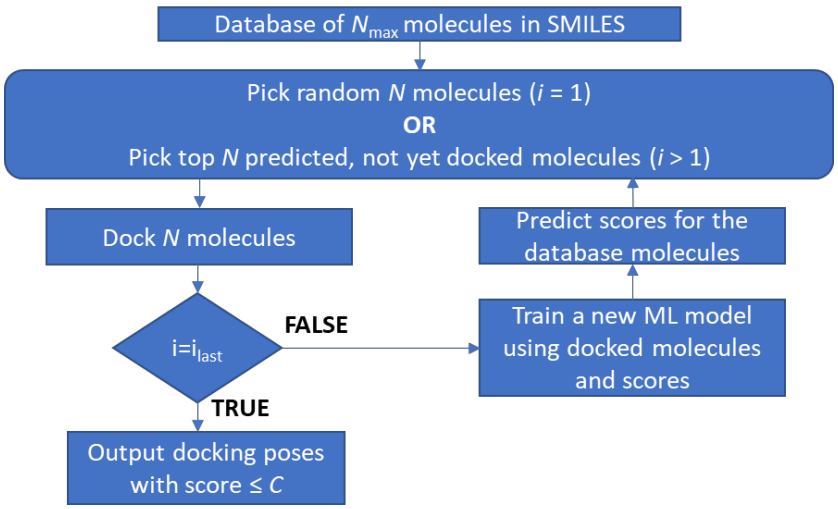



**Figure 1**: Virtual screening with the iterative HASTEN process. As predicting the docking scores for molecules with ML is much faster than docking these molecules, significant computational resources are saved when only N * i molecules are docked. (N * i << $N_{max}$, $N_{max}$ > 1e6). N, number of compounds; i, iteration; C, cut-off.

Since the publication of HASTEN, we have been using the software intensively in-house to accelerate our virtual screening campaigns. Experiences from several screens will be discussed, together with practical advice on how to run virtual screens in an efficient manner on databases of different sizes.

1. Kalliokoski T, Machine Learning Boosted Docking (HASTEN): An Open-source Tool to Accelerate Structure-based Virtual Screening Campaigns. Mol. Inform., **2021**, 40, 210089.
2. Yang K et al. Analyzing Learned Molecular Representations for Property Prediction. J. Chem. Inf. Model **2019**, 59, 3370-3388.

## P-03: DEL design at Ryvu

M. Król, K. Baczyński, M. Potocki, I. Mames, A. Zarębski, A. Sabiniarz

*Ryvu Therapeutics, R&D Center for Innovative Drugs*
*Leona Henryka Sternbacha 2, 30-394 Kraków, Poland*

The identification of hits, compounds that show promising bioactivity, remains one of the major challenges in drug discovery. Traditionally, the main sources of bioactive chemical matter are known active compounds and high-throughput screening (HTS)[1,2]. However, HTS libraries are limited in size and contain only a minute

fraction of the available chemical space[3,4]. Additionally, as HTS is a random technique, resources are required to perform a screen scale linearly with library size.
Selection-based techniques, such as DNA Encoded Libraries (DELs), where all library members are simultaneously screened against a target of interest, can be much more efficient in hit identification. DEL approach, inspired by phage display technology, includes split-and-pool synthesis, oligonucleotide compound tagging, affinity-based selection and PCR-based identification. Briefly, in consecutive cycles of DEL generation oligonucleotide tag ligations and block couplings are performed in turns to synthesize a pool of uniquely tagged combinatorial chemistry based compounds.
Here we will discuss early assessment of DEL design and implementation at Ryvu, with the comparison of chemical and physicochemical space covered by current Ryvu HTS libraries and proposed DELs. We will show optimization of DELs in the physicochemical space and briefly touch on the design of databases and computational analytical tools to mine DEL generated data.

1. Brown, DG., Boström, J., Where Do Recent Small Molecule Clinical Development Candidates Come From? Journal of Medicinal Chemistry 2018 61 (21), 9442-9468
2. Dragovich, PS., Haap, W., Mulvihill, MM., Plancher, JM., Stepan, AF., Small-Molecule Lead-Finding Trends across the Roche and Genentech Research Organizations. Journal of Medicinal Chemistry 2022 65 (4), 3606-3615
3. Reymond, JL., The Chemical Space Project. Accounts of Chemical Research 2015 48 (3), 722-730
4. Volochnyuk, DM., Ryabukhin, SV., Moroz, YS., Savych, O., Chuprina, A., Horvath, D., Zabolotna, Y., Varnek, A., Judd, DB., Evolution of commercially available compounds for HTS. Drug Discovery Today 2019 24(2), 390-402

## P-05: 40 million PubChem structures from patents: both treasure trove and junk yard

C. Southan
*Data Sciences, Medicines Discovery Catapult, SK10 4ZF, UK*

Compared to the literature, the patent corpus has both pros and cons for chemistry data mining The latter include being a) a “Cinderella” source that is difficult to get to grips with, b) massively redundant document corpus from patent families and kind codes and c) include various degrees of deliberate obfuscation to impede data mining. Pros include a) paradoxically, compared to restricted access to the literature, they are completely open for text mining and entity extraction, b) they contain ~ 3x to ~5x more medicinal chemistry SAR than published papers, c) include discloses of new drug targets and chemotypes years ahead of papers d) consitute a rich source of executed synthesis protocols and experimental chemistry property data e) within the last few years open automated chemical named entity recognition (CNER) has broken the monopoly of commercial chemistry curation. Because Medicines Discovery Catapult needs to keep up with developments in both commercial and open sources this work was undertaken to update our overview of patent extractions in general and the expanding integration within PubChem in particular. The four largest PubChem sources, SureChEMBL, Google Patents, WIPO, and IBM, use similar CNER pipelines that include name look-ups, IUPAC conversions and image-to-struc extractions. Their compound (CID) counts are 21.5, 17.9, 17.7 and 10.7 million, respectively, and together with small sources such as NextMove Software synthetic pathway extractions at 1.8 million, the CNER sources add up to just under 40 million from the PubChem March 2022 total of 111 million. The “treasure trove“ aspects that will be presented includes a) expert curation of SAR from patents by BindingDB with 400K compounds from 5.4K US patents and data points covering 2,197 target proteins b) extensive coverage of the ~5 million exemplified compounds from all C07 and A61 patent classified filings relevant to medicinal chemistry c) the ability to track back to exact example numbers in documents via SureChEMBL and WIPO. However, this presentation will also outline the “junkyard“ aspects. These include a) beyond the ~ 5 million structures linkable to data how much of a junk yard the other 35 million represent b) CNER produces artifactual structures from broken IUPAC strings and mixture extractions of various sorts, c) all the large extraction sources diverge significantly in exactly what chemistry their own pipelines pull out and d) the 28 million patent document to chemistry links represent significant massive over-mapping (but reasons for this will be discussed). All things considered however, the PubChem team are congratulated on their efforts not only in wrangling and integrating these sources but also linking and search-indexing the chemistry linked to the patent documents they were extracted from.

## P-07: Artificial Intelligence for Compound Design and Automation of DMTA Cycles

Sauer S.[1], Matter H.[1], Hessler G. [1], Grebner C.[1]

[1]*Synthetic Molecular Design, Integrated Drug Discovery, Sanofi-Aventis Deutschland GmbH, 65926 Frankfurt am Main, Germany*

The development of novel drugs is a multiparameter optimization progress, which requires several iterations of designing, making, testing, and analyzing compounds (also known as DMTA-cycles). Recent applications of artificial intelligence (AI) show very promising results on how AI can improve this process.
AI-based de-novo design of new compounds requires a reward function estimating the suitability of a molecule against a given target.[1] This reward function can be defined by physicochemical properties, 2D or 3D similarity to reference molecules[1], or machine learning approaches, if sufficient amount of activity data is available to train a predictive model.[2]
Moreover, a 3D structure of the target protein can also be used to compute the reward by docking the generated molecules into the protein pocket.[3] The generated pose can be evaluated in different ways to estimate the free binding energy or the binding affinity. Here, we developed a scoring function that predicts the affinity for a given molecule not only from the Glide-XP docking pose but using also general data from the PDBBind database. This scoring function is incorporated into our corporate de-novo design workflow consisting of several state-of-the-art design engines.
Another goal of our work is the acceleration of the DMTA cycle in cooperation with automated combinatorial chemistry. Here, fragment-based approaches are employed to design new compounds which can be easily synthesized from available reagents. One possibility is the enumeration of virtual libraries, followed by several filtering steps.[4] Alternatively, we are working on the adaption of fragment-based de-novo AI-engines to generate molecules using standard reactions suitable for automated synthesis.[5]
Selected case studies will illustrate the potential for AI-based de novo design.

1. Grebner, C., et al., Automated De Novo Design in Medicinal Chemistry: Which Types of Chemistry Does a Generative Neural Network Learn?. J. Med. Chem., **2020**, 63, 8809-8823.
2. Grebner, C, et al., Application of Deep Neural Network Models in Drug Discovery Programs. Chem. Med. Chem., **2021**, 16, 3772.
3. Guo, J., et al., DockStream: a docking wrapper to enhance de novo molecular design., Journal of Chemoinformatics, **2021**, 13, 1, 89.
4. Grebner, C., et al, Virtual Screening in the Cloud: How Big Is Big Enough?. J. Med. Chem., **2020**, 60, 9, 4274-4282.
5. Ståhl, N., et al., Deep Reinforcement Learning for Multiparameter Optimization in de novo Drug Design, Journal of Chemical Information and Modeling, **2019**, 59, 7, 3166-3176.

## P-09: Multi-target uncertainty quantification for *de novo* drug design

S.I.M. Luukkonen[1], E.B. Lenselink[2], M.T.M. Emmerich[3], P.F.W. Stouten[2], G.J.P. van Westen[1]

[1] *Drug Discovery and Safety, Leiden Academic Centre for Drug Research, Einsteinweg 55, Leiden, The Netherlands*
[2] *Galapagos, Generaal De Wittelaan L11, A3, 2800 Mechelen, Belgium*
[3] *Leiden Institute of Advanced Computer Science, Niels Bohrweg 1, Leiden, The Netherlands*

In the recent past, deep learning (DL) has been pivotal in many breakthroughs in the field of artificial intelligence. In image recognition and natural language processing DL-based models even surpassed human abilities. In drug discovery, DL has been used to construct quantitative structure-activity relationship (QSAR) models that allow for the estimation (typically referred to as prediction) of properties of chemical compounds, such as affinities to specific targets[1,2]. In a related context, DL can also be used in *de novo* generation of chemical structures.
Most QSAR models are "black boxes" that produce just a numerical estimate of a property without any indication of uncertainty of the estimate. They are difficult to interpret and lack guarantees of robustness. Model predictions are used for decision making and uncertainty quantification (UQ) is therefore essential. However, if at all, UQ is often not done rigorously and systematically for QSAR models. Drug molecules may interact with more than one target, which can have desired (polypharmacology) or undesired (toxicity) effects, and often these interactions with (similar) targets are correlated. To benefit from these correlations, so-called "multi-task" QSAR models can be developed and often they significantly outperform "single-task"

models.
For this reason, we aim to develop multi-task QSAR models with UQ to improve the quality and applicability of affinity predictions. Furthermore, we are coupling these UQ-QSAR models to DrugEx[3] – a multi-objective, *de novo* molecule generator developed in van Westen's group – to study the effect of incorporating uncertainty in both explorative and exploitative *de novo* generation.
Different methods were applied to publicly available Adenosine receptor family (A1, A2A, A2B and A3) (ant)agonist activity data[4]. Preliminary results of Gaussian Process[5] and (evidential) message-passing neural network-based[6] UQ approaches show that the calculated mean values of the UQ-models are equivalent to single-value QSAR models. Using an approach developed by Galapagos that results in balanced splits (train-validation-test sets: 80%-10%-10%) for all isoform activities, while ensuring optimum chemical separation between sets[7], pChEMBL-values were estimated with RMSE between 0.85 and 1.10, depending on the target. Based on these results, generating uncertainty estimates that are well-correlated with true prediction errors is challenging. Even though they are well-calibrated, the predicted uncertainty distributions are often narrow, making it hard to separate low- and high-confidence predictions. We are engaged in additional efforts to improve these uncertainty estimates.
This research received funding from the Dutch Research Council (NWO) in the framework of the Science PPP Fund for the top sectors.

1. Lenselink, E.B., et al., Beyond the Hype: Deep Neural Networks Outperform Established Methods Using a ChEMBL Bioactivity Benchmark Set. J. Cheminform., **2017**, 9, 45
2. Yang, K., et al., Analyzing Learned Molecular Representations for Property Prediction. J. Chem. Inf. Model., **2019**, 59, 8, 3370
3. Liu, X., et al., *DrugEx* v2: de novo design of drug molecules by Pareto-based multi-objective reinforcement learning in polypharmacology. J. Cheminform., **2021**, 13, 1, 85
4. Béquignon, O.J.M., et al., Papyrus - A large scale curated dataset aimed at bioactivity predictions. ChemRxiv, **2021**
5. Rasmussen, C.E. and Williams, K.I., Gaussian Processes for Machine Learning. MIT Press, **2005**
6. Soleimany, A.P., et al., Evidential Deep Learning for Guided Molecular Property Prediction and Discovery. ACS Cent. Sci., **2021**, 7, 8, 1356
7. Tricarico, G.A., et al., **2022,** manuscript in preparation

## P-11: Planning of chemical synthesis of focused libraries of similars to a given compound

A.A. Fatykhova[1], R.I. Nugmanov[1] , R.N. Mukhametgaleev [1] , T.I. Madzhidov [1], A. Varnek [2]

*[1]Kazan Federal University, Kazan, Russia*
*[2] Strasbourg University, Strasbourg, France*

Currently, the planning of the chemical synthesis of any compound, in particular, medicinal drugs, is an extremely important task. Modern methods of drug development, such as virtual screening, docking, de novo design, allow the generation of many molecules with the necessary biological activity and other properties. One of the problems is that some of the promising molecules cannot move to the next stage of development due to the problem of their synthesis. In this work, we present an approach to planning synthesis of a library of compounds. The approach is based on application of forward chemical synthesis (from reagents to products) technique, combining the Monte Carlo tree search method for guiding optimization and deep learning methods. In contrast to the traditional retrosynthetic approach, the forward synthesis technique optimizes the synthetic path to synthesize the set of molecules, most similar to target ones.
The developed tool consists of the following main blocks: a database of molecules and reaction rules, modules for virtual reactions generation, and heuristic algorithms for fast search based on similarity metrics. The developed approach uses commercially available chemical compounds as initial reagents and rules of reaction transformations to generate new products. Generation of reactions proceed using the Virtual Reactor, which allows the generation of chemically correct structures. Monte Carlo tree search methods applied to effectively navigate the vast space of chemical compounds. Using deep neural networks algorithm quickly selects reagents that are required for obtaining the product molecule as similar to the target one as possible.

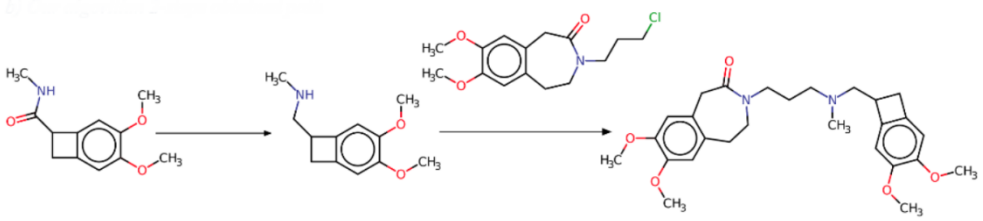


**Figure 1:** Predicted synthesis path for bevantolol by developed algorithm.

The developed approach was validated based on reference pathways of drug molecules, extracted from USPTO database. One predicted path presented in Figure 1.

This work was supported by the Russian Science Foundation (19-73-10137)

## P-13: MoleculeACE: a benchmark for machine learning with activity cliffs

D. van Tilborg[1,2], A. Alenicheva[3], F. Grisoni[1,2]

[1]*Eindhoven University of Technology, Institute for Complex Molecular Systems and Dept. Biomedical Engineering, Eindhoven, Netherlands.*

[2]*Centre for Living Technologies, Alliance TU/e, WUR, UU, UMC Utrecht, Utrecht, The Netherlands.*

[3]*JetBrains Research. Saint Petersburg, Russia.*

Machine learning is accelerating molecule discovery[1], with quantitative-structure activity relationship (QSAR) approaches constituting essential tools in the chemical sciences. QSAR has reached a high level of accuracy for bioactivity prediction. However, activity cliffs – molecules that are highly similar in structure but exhibit large differences in potency[2] (Fig. 1) – are often mispredicted[3,4]. Although large potency differences attributed to subtle structural changes hold key insights for medicinal chemists[4], they might hinder the accuracy of machine learning models. Arguably, models that are well-equipped to predict potency differences in activity cliffs are better at capturing the overall structure-activity landscape. Therefore, assessing a model's predictive performance on activity cliff compounds can give meaningful insights. Despite this, best practices for molecular machine learning in the presence of activity cliffs are currently unknown.

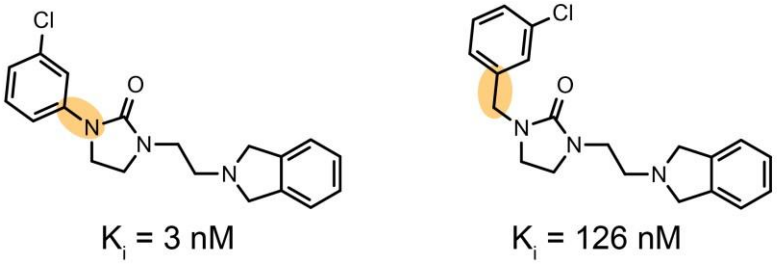



**Figure 1**: Example of an activity cliff on the Dopamine D3 receptor, D3R. Two molecules with highly similar structures show more than a 10-fold difference in their respective inhibitory constant ($K_i$).

In this systematic study, we compared 16 commonly used machine and deep learning strategies for their performance in the presence of activity cliffs. We collected and curated molecular bioactivity data of 30 pharmacologically relevant drug targets from ChEMBL[5] and identified activity cliffs with a consensus approach considering substructure similarity, scaffold similarity, and similarity of SMILES strings. We explored four classical machine learning strategies using four types of molecular descriptors and deep neural networks using two types of "raw" molecular representations, *i.e.*, graphs and SMILES strings.
Our systematic analysis of a total of 690 models revealed that approaches based on human-engineered molecular descriptors outperformed more complex deep learning methods based on SMILES or graphs in their performance on activity cliffs. Importantly, our results highlight that a low overall prediction error

does not guarantee a low prediction error on activity cliff compounds. This aspect highlights the relevance of assessing the performance on activity cliffs alongside traditional performance evaluation strategies, especially when the models have to be applied in a prospective setting (e.g., molecule optimization or virtual screening). To facilitate this, all the data and methods have been collected in an-open access benchmark tool, named MoleculeACE (Activity Cliff Estimation). MoleculeACE allows assessing the predictive performance of machine learning models in the presence of activity cliffs and aims to steer the community towards addressing a current limitation of QSAR methods. The benchmark is available on GitHub at URL: https://github.com/molML/MoleculeACE.

1. Chen, H., Engkvist, O., Wang, Y., Olivecrona, M. & Blaschke, T, The rise of deep learning in drug discovery. *Drug Discov. Today,* (***2018***) 23, 1241–1250.
2. Maggiora GM., *On outliers and activity cliffs – why QSAR often disappoints. J Chem Inf Model*, (**2006**) 46, 1535
3. Stumpfe D, Hu H, Bajorath J., Evolving concept of activity cliffs. *ACS Omega*, (**2019**) 4, 14360–14368
4. Stumpfe, D., Hu, H. & Bajorath, J. Advances in exploring activity cliffs. *J. Comput. Aided Mol. Des, (**2020**)* 34, 929–942
5. Gaulton, A. *et al.,* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res,* (***2012***) 40, D1100–7.

## P-15: The chemistry puppeteer: enhancing the diversity of single-step retrosynthesis

A. Toniato[1], A.C. Vaucher[1], P. Schwaller[2,1] and T. Laino[1]

*[1]IBM Research Europe, Saumerstrassse 4, 8803 Rueschlikon, Switzerland*
*[2] Current address: Laboratory of Artificial Chemical Intelligence, EPFL, RTE Cantonale, 1015 Lausanne, Switzerland*

Retrosynthesis planning is the task of recursively identifying reactions able to decompose a complex molecule into simpler, commercial structures. In order to achieve the goal, most modern AI-based approaches rely on a Deep Learning single-step retrosynthesis model coupled with a search algorithm [1]. One of the main issues is that, usually, the proposed disconnection strategies lack diversity. When the goal is to find a suitable set of precursors for a given target molecule, the generated precursors typically fall in the same chemical macro class (ex. all protection, deprotection or same C-C bond formation with a slightly different set of reagents) and the automatic synthesis planning tools might get stuck. Most of the existing approaches do not allow a machine learning model to provide multiple diverse alternatives to explore and are focused on the top single-step predictions. Truth is that instead there might be multiple ways in which a molecule can be synthesized. In our approach we have developed a text-based Transformer model (the chemistry puppeteer) to increase the diversity of the predictions, by concatenating a token for the reaction class of the molecule with the SMILES. The learned embeddings of the given sample partly codify some characteristics of the reactions belonging to that class. At test time, the use of these tokens allows us to steer the model towards different kinds of disconnection strategies. We show with results on the PISTACHIO [2] dataset that the diversity of the predictions can improve consistently. While the use of excessively specific groupings can decrease the model performances in terms of valid proposed set of precursors, the use of chemically relevant policies (e.g., reaction fingerprints [3]) to construct smaller macro groups allows to recover the quality of the predictions without the loss of the found diversity.

1. Schwaller, P., et al., Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. Chem. Sci. **2020**, 11, 3316–3325
2. Nextmove Software Pistachio. **2021**, http://www.nextmovesoftware.com/pistachio.html, Accessed 2021
3. Schwaller, P., et al., Mapping the space of chemical reactions using attention-based neural networks. Nat. Mach. Intelligence., **2021**, 3, 2, 144-152

## P-17: GenUI: interactive and extensible open source software platform for de novo molecular generation and cheminformatics (updates and perspective)

M. Šícho[1,2], X. Liu[1], D. Svozil[2], G.J.P. van Westen[1]

[1]*Computational Drug Discovery, Drug Discovery and Safety, Leiden Academic Centre for Drug Research, Einsteinweg 55, Leiden, The Netherlands*

[2] *CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Department of Informatics and Chemistry, Faculty of Chemical Technology, University of Chemistry and Technology Prague, Technická 5, 166 28, Prague, Czech Republic*

In recent years, cheminformatics has seen a surge of novel tools for de novo drug design and many novel techniques for automatic generation of molecules with predefined properties have been developed. However, the widespread adoption of these new generative techniques has still not been achieved. Many of the novel approaches are based on complex algorithms such as deep neural networks and require considerable expertise to apply and validate. Therefore, these tools still remain difficult to use and opaque for people who could benefit from them the most. Medicinal chemists and pharmacologists are usually not trained in programming, machine learning and data science so they cannot use these tools directly. However, it would be beneficial if these tools can be used more routinely by experimentalists that can provide valuable prospective validation and feedback. With GenUI [1] we aim to give generative tools a simple and easy-to-use graphical user interface (GUI), but also focus on the development of convenient extensibility features to motivate cheminformatics researches to enhance the tools they develop with a GUI. From a web browser, GenUI can load and standardize structural and activity data from multiple sources, explore the imported data on an interactive chemical space map and manage inputs and outputs to create various predictors and generators. In this contribution, the GenUI platform and its main features will be introduced with special focus on those added in the latest release. We will also add perspective on future development and discuss the potential of such tools to accelerate drug discovery.

## P-19: Applying machine learning for virtual drug discovery and development of adenosine A2A ligands combining in silico medicinal chemistry and quantitative systems pharmacology

H.W. van den Maagdenberg[1,2], J.G.C. van Hasselt[2], P.H. van der Graaf[2,3], G.J.P. van Westen[1]

[1]*Drug Discovery and Safety, Leiden Academic Centre for Drug Research, Einsteinweg 55, 2333 CC Leiden, The Netherlands.*

[2] *Division of Systems Biomedicine and Pharmacology, Leiden Academic Centre for Drug Research, Leiden University, Einsteinweg 55, 2333 CC Leiden, The Netherlands.*

[3] *Certara, University Road, Canterbury Innovation Centre, Unit 43, CT2 7FG Canterbury, Kent, UK.*

Many promising machine learning techniques have been developed and successfully applied to optimize target affinity of molecules for the discovery of novel drug candidates. However, clinical efficacy of a potential drug is dependent on more than just affinity. Quantitative systems pharmacology models can describe the relationship between receptor activation and biomarkers for efficacy and toxicity. The aim of this study is to explore the integration of systems pharmacology models in the virtual drug discovery pipeline for discovering new therapeutics in a case study focused on targeting the adenosine $A_{2a}$ receptor ($A_{2a}R$).

Immuno-oncology agents, such as adenosine $A_{2a}R$ inhibitors[1], are promising new cancer therapeutics, but they do not have sufficient effect in many patients. Immuno-oncology is complex and the mechanistic link between target and effect is often not well-understood. Therefore, targeting the $A_{2a}R$ will be the case study used to test the proposed combination of systems pharmacology and medicinal chemistry.

Novel ligands were generated for the adenosine $A_{2a}R$ using DrugEx[2], which is a multi-objective de novo generator using recurrent neural network(RNN)-based reinforcement learning (RL). First, quantitative structure activity models (QSAR) were trained on a dataset of activity data for the $A_1$, $A_{2a}$, $A_{2b}$ and $A_3$ receptors. During the RL process, activity of the model output is predicted by the QSAR models as feedback

for improving the RNN. The solutions are scored using pareto ranking and a Tanimoto-based crowding distance algorithm, prioritizing chemically diverse molecules with activity on the $A_{2a}$ and $A_{2b}$ and no activity for $A_1$ and $A_3$ receptors.
Different machine learning QSAR models were compared, including a deep neural network, support vector machines and XGboost. Random forest models had the highest value for the AUC of the ROC curves for all four targets and were therefore used as predictors in the reinforcement learning. The trained model was used to predict 10,000 ligand, of those 99% were valid and 88% had the desired activity profile.
The results show that DrugEx can successfully predict $A_{2a}R$ ligands for continuation of this research. The characteristics of the resulting set of molecules will be used to predict the clearance and volume of distribution. Several approaches will be compared from a recent overview of machine learning-based predictive models for human pharmacokinetics[3] and the optimal approach will be used for parameter prediction. The resulting parameter estimates will be used in a systems pharmacology model[4] to compare tumour inhibition efficacy of the predicted inhibitors. The workflow predictions will be validated by a limited test set of known $A_{2a}R$ inhibitors with data available on in-vivo tumour inhibition. In conclusion, a virtual drug discovery pipeline with integrated systems pharmacology will be created to improve in silico drug prediction.

1. Augustin, R. C., et al., Next steps for clinical translation of adenosine pathway inhibition in cancer immunotherapy. Journal for ImmunoTherapy of Cancer. **2022**, 10, 2, e004089
2. Liu, X., et al., DrugEx v2: De Novo Design of Drug Molecule by Pareto-based Multi-Objective Reinforcement Learning in Polypharmacology. Journal of Cheminformatics. **2021**, 13, 1, 85
3. Danishuddin, V., et al., A decade of machine learning-based predictive models for human pharmacokinetics: Advances and challenges. Drug discovery today. **2022**, 27, 2, 529-537
4. Voronova, V., et al., Evaluation of Combination Strategies for the A2AR Inhibitor AZD4635 Across Tumor Microenvironment Conditions via a Systems Pharmacology Model. **2021**, 12

## P-21: Combining shape and electrostatics in a spectral geometry-based 3D molecular descriptor

J. Middleton[1], G, Ghiandoni[2], M. Packer[3], M. Zhuang[1], V. J. Gillet[1]

*[1]Information School, University of Sheffield, Regent Court, 211 Portobello, Sheffield, S1 4DP*
*[2]AstraZeneca R&D IT, Melbourn Science Park, Royston SG8 6EE*
*[3]AstraZeneca Early Oncology R&D, Alderley Park, Macclesfield, SK10 4TG*

It has been well established that shape complementarity plays an important role in molecular recognition. However, shape is not the only important property involved in this process. Electrostatic complementarity has also long been known to play a critical role in the binding process between a drug molecule and a therapeutic target (Weiner et al., 1982). Consequently, there are various molecular representation methods that capture both shape and electrostatic features in order to maximize the discriminative information and isolate molecules of interest. For instance, ElectroShape (Armstrong et al., 2010) which is built upon the Ultrafast Shape Recognition (USFR) descriptor, incorporates charge information as a ratio of charge per unit distance. This descriptor avoids the costly alignment process associated with alignment-based methods and computes molecular similarity based on the moments of 1D distributions. Literature has shown that although alignment-invariant methods typically offer inferior performance to alignment-based 3D approaches (Cleves et al., 2019), these methods offer a more appropriate alternative for large databases. This signifies that there is a notable gap for a molecular descriptor that can retain the efficiency of alignment-invariant approaches as well as offer more competitive application performance relative to established alignment-based methods.
Seddon et al. (2019) developed an alignment-invariant molecular descriptor known as MOLSG which is based on the concepts of spectral geometry. The MOLSG descriptor captures shape information through the application of the Laplace Beltrami Operator (LBO) to a 2D molecular surface embedded in 3D space. This shape information is obtained through an eigendecomposition of the LBO which is then subsequently fed as input into a local geometry descriptor method to present a more refined view of the local shape information. To enable comparisons between molecules, a global geometry descriptor is employed. This is achieved using the bag of features method which attempts to extract a set of meaningful geometric words through an unsupervised clustering method. These geometric words are then stored in a codebook which can then be applied to unseen molecules in order to derive a description of molecular shape based on the representative geometric words. The benefit of this codebook approach is that since they are generated ad hoc using a

training set of molecules, it presents an opportunity to potentially maximize the quality of information captured for a given application. This could provide the MOLSG descriptor an advantage in application performance.
Here, the MOLSG descriptor workflow has been modified to include electrostatic information whilst retaining the beneficial properties of the original descriptor such as rotational and translational invariance. This has been achieved by leveraging a graph convolutional neural network developed by Rathi et al. (2020) which has been shown to produce electrostatic potential (ESP) surfaces of comparable quality to computationally expensive DFT ESP surfaces whilst taking a fraction of the time (Rathi et al., 2020). The MOLSG approach has been extended to incorporate the electrostatic information produced by this neural network to form MOLSG-Electro. This investigation aims to determine the optimal encoding of electrostatic information within the previously developed MOLSG workflow. MOLSG-Electro is then compared to established descriptors in virtual screening applications using the DUD-E dataset.

Armstrong, M. S., Morris, G. M., Finn, P. W., Sharma, R., Moretti, L., Cooper, R. I., & Richards, W. G. (2010). ElectroShape: fast molecular similarity calculations incorporating shape, chirality and electrostatics. *Journal of computer-aided molecular design*, 24(9), 789–801. https://doi.org/10.1007/s10822-010-9374-0

Cleves, A.E., Johnson, S.R. & Jain, A.N. Electrostatic-field and surface-shape similarity for virtual screening and pose prediction. *Journal of Computer-Aided Molecular Design*, 33, 865–886 (2019). https://doi.org/10.1007/s10822-019-00236-6

Rathi, P. C., Ludlow, R. F., & Verdonk, M. L. (2020). Practical High-Quality Electrostatic Potential Surfaces for Drug Discovery Using a Graph-Convolutional Deep Neural Network. *Journal of Medicinal Chemistry*, *63*(16), 8778–8790. https://doi.org/10.1021/acs.jmedchem.9b01129

Seddon, M. P., Cosgrove, D. A., Packer, M. J., & Gillet, V. J. (2019). Alignment-Free Molecular Shape Comparison Using Spectral Geometry: The Framework. *Journal of Chemical Information and Modeling*, *59*(1), 98–116. https://doi.org/10.1021/acs.jcim.8b00676

Weiner, P. K., Langridge, R., Blaney, J. M., Schaefer, R., & Kollman, P. A. (1982). Electrostatic potential molecular surfaces. *Proceedings of the National Academy of Sciences of the United States of America*, 79(12), 3754–3758. https://doi.org/10.1073/pnas.79.12.3754

## P-23: Using Matched Molecular Pairs for CoreDesign®

J. Stacey[1], A. Dossetter[1], E. Griffen[1], A. Leach[1], L. Reid[1], P. de Sousa[1], B. Khan[1], B. Kwan[1]

[1]*MedChemica Ltd. MedChemica Ltd, Alderley Park, Macclesfield, Cheshire, UK, SK10 4TG*

In drug discovery an active core is typically selected and altered to improve properties of interest. Scaffold hopping is an approach used to explore new chemistry around this core (/scaffold) of interest. Typically, scaffolds are a ring system or systems that are connected by linkers. Several studies have already been undertaken to examine the potential ring chemical space and replacements[1-3].

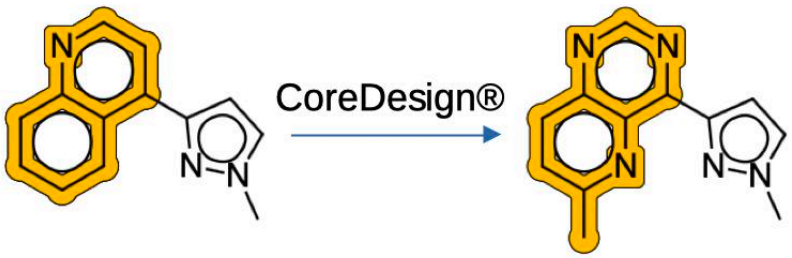



**Figure 1**: An example of a CoreDesign® transformation, where the core is highlighted in yellow.

MedChemica's technology currently has the capability to perform focussed small chemical transformations to improve properties using Matched Molecular Pairs (MMPs) via the program RuleDesign®. In order to apply a MMP within RuleDesign® certain statistics must be obeyed so that the transformation has a high potential of improving the property of interest. CoreDesign® is complementary to RuleDesign® as it relaxes

the probability of success in accordance to improving the property of interest by including all transformations which are focussed on core and linker changes. A key benefit of CoreDesign is prioritising core exchanges where there is good synthetic precedent.

1. Pitt, W., et al., Heteroaromatic Rings of the Future., *J. Med. Chem.*, **2009**, 52, 2952-2963
2. Tu, M., et al., Exploring Aromatic Chemical Space with NEAT: Novel and Electronically Equivalent Aromatic Template., *J. Chem. Inf. Model*., **2012**, 52, 1114-1123
3. Ertl, P. Magic Rings: Navigation in the Ring Chemical Space Guided by the Bioactive Rings., *J. Chem. Inf. Model*., **2021**, DOI: 10.1021/acs.jcim.1c00761

## P-25: The Future of InChI

G. Blanke[1], J. Goodman[2], R. Potenzone[3], S. Heller[4]

*[1]StructurePendium Technologies GmbH, Essen, Germany and Technical Director of the InChI Trust, Cambridge UK*
*[2]University of Cambridge, Department of Chemistry, Cambridge, UK and Secretary of the subcommittee on the IUPAC International Chemical Identifier*
*[3]Community Outreach Director of the InChI Trust;*
*[4]Project Director of the InChI Trust and chair of the subcommittee on the IUPAC International Chemical Identifier;*

Nowadays InChI standard covers most of the organic chemistry as Jonathan M. Goodman, Igor Pletnev, Paul Thiessen, Evan Bolton, and Stephen R. Heller pointed out in their article "InChI version 1.06: now more than 99.99% reliable" [1]. In the past few years, the interest in the areas of organometallics, enhanced stereochemistry, and improved tautomer recognition has increased. There are also groups looking more broadly at extending InChI to be able to handle Markush structures, mixtures, nanomaterials, and biologics. All of them lead to considerations to adapt the InChI core code. These code developments might become a change of tyres without stopping the vehicle and may lead to changes to the InChI notation and the InChI Key.
This talk will provide an overview of the current and potential future developments including the move of the InChI development to an open distributed version control system like GitHub that will allow the participation of the community in the further programming.

1. J Cheminform 13 40, 2021, https://www.doi.org/10.1186/s13321-021-00517-z

## P-27: PKD- KG: A drug repurposing knowledge graph for Autosomal Dominant Polycystic Kidney Disease (ADPKD)

B. Khalil [1,2], D. Araripe [2,3], J.-M. Neefs [1], H. van Vlijmen [1], N.Dyubankova [1], G.J.P. van Westen [2]

*[1] Janssen Research & Development, LLC, In-Silico Discovery, and External Innovation (ISD&EI), B-2340 Beerse, Belgium*
*[2] Division of Drug Discovery and Safety, Leiden Academic Centre for Drug Research (LACDR), Leiden University, P.O. Box 9502, 2300 RA, Leiden, The Netherlands*
*[3] Department of Human Genetics, Postzone S-04-P, Leiden University Medical Centre (LUMC), P.O. Box 9600, 2300 RC Leiden, Netherlands.*

ADPKD is an inherited disorder primarily caused by mutations in PKD1 and PKD2, encoding for polycystin-1 (PC1) and polycystin-2 (PC2), respectively. It is characterized by formation of renal cysts leading to enlargement of kidneys and eventually renal failure. ADPKD is the most prevalent genetic kidney disease with an estimated incidence of 9.3 cases per 10,000 people [1] and only one FDAapproved drug, Tolvaptan. [2] According to a recent study [3], ADPKD phenotypic features are reversible, providing hope for a cure discovery. In recent years, machine learning technologies have helped to improve the effectiveness and speed of several phases of the drug discovery pipeline. Of those, Biomedical knowledge graphs (KG) can help with the understanding and modelling of complex biological systems and diseases, and many other tasks, such as drug repurposing and target gene-disease prioritisation [4]. Existing KG research often suffers from issues such as

sparse and noisy datasets, insufficient modelling methods and non-uniform evaluation metrics [5].
In this work, we introduce PKD-KG, a multi-relational, attributed biomedical KG, with a focus on PKD domain-specific information, incorporating multiple types of entities (nodes) and relationships (edges). The source database is a multi-dimensional relational PostgreSQL database, extracted from multi-omics data, and text-mined from literature. Two commercially available knowledge acquisition tools were utilized: Euretos Knowledge Platform (www.euretos.com) and Causaly (www.causaly.com/). They captured: 295 and 276 genes/targets, 502 and 174 chemical compounds, 99 and 75 pathways related to PKD, respectively, which only includes entities referenced for at least three times or more. The data is then filtered and combined with the publicly available and internal databases (Figure 1). Machine learning algorithms use this input to learn a knowledge graph embedding (KGE), which is then used to generate hypotheses suggesting a list of prioritized targets, novel disease-target, and disease drug (repurposing) connections or to analyse and visualise the complexities of the disease model, allowing the PKD community to interpret findings directly from the KG.

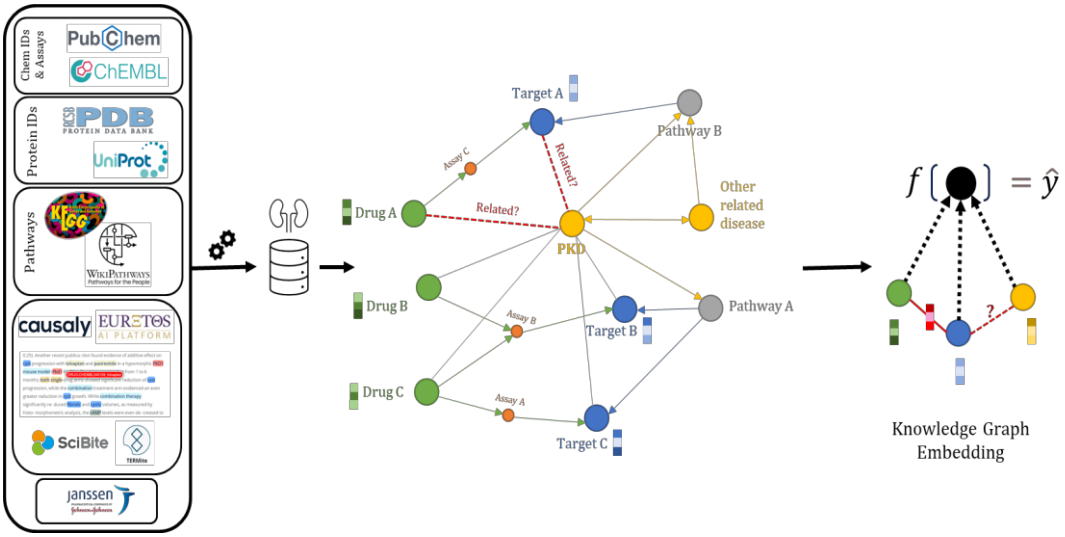



**Figure 1**: A graphical Abstract of the PKD-KG workflow showing some of the database resources, the graph with entities representing drugs, targets, assays, pathways, and diseases, each with an embedded feature vector of relevant descriptors.


1. Lanktree, M. B. *et al.* Prevalence Estimates of Polycystic Kidney and Liver Disease by Population Sequencing. *J Am Soc Nephrol* **29**, 2593–2600 (2018).
2. Gattone, V. H., Wang, X., Harris, P. C. & Torres, V. E. Inhibition of renal cystic disease development and progression by a vasopressin V2 receptor antagonist. *Nat Med* **9**, 1323–1326 (2003).
3. Dong, K. *et al.* Renal plasticity revealed through reversal of polycystic kidney disease in mice. *Nat Genet* **53**, 1649–1663 (2021).
4. Bonner, S. *et al.* A Review of Biomedical Datasets Relating to Drug Discovery: A Knowledge Graph Perspective. *Arxiv* (2021).
5. Zheng, S. *et al.* PharmKG: a dedicated knowledge graph benchmark for biomedical data mining. *Brief Bioinform* **22**, (2020).


## P-29: Molecular dynamics-based elucidation of Flap endonuclease 1 flexibility for DNA cleavage


Z. Hosni,[a] V. D'annibale,[b] A.N. Nardi,[b] G. Chen,[b] E. Brudenell,[c] M. D'abramo,[b] J. Sayers,[c] V.J. Gillet [a]

[a] *Information School, Regent Court (IS), 211 Portobello, Sheffield, UK S1 4DP*

[b] *Dipartimento di Chimica, Sapienza Università di Roma, Italy, P. le A. Moro, n° 5, 00185 Roma*

[c] *Department of Infection, Immunity and Cardiovascular Disease, L Floor, The Medical School, Beech Hill Road, Sheffield, UK, S10 2RX*

Flap endonucleases (FENs) are highly conserved structure-specific metalloenzymes that catalyse a specific incision to remove 5' flaps in double-stranded DNA substrates. These proteins have crucial roles in various cellular processes, such as DNA replication. Mutations that compromise Fen1 (a mammalian flap endonuclease) expression levels or activity have severe pathological consequences.1 It is claimed that the DNA is bound in a bent conformer in the region of the active site of Fen1 enzyme.2,3 However, the structure around the active site is one of the most variable parts of the superfamily, and is primarily responsible for determining the substrate specificity of a given nuclease in the superfamily. This makes these nucleases particularly challenging to study as a means to elucidate the role of order and disorder in protein function.4 Molecular dynamics (MD) is a very suitable approach to study protein flexibility as atoms and molecules are allowed to interact for a period of time by approximations of known physics, giving a view of the motion of the particles. We have applied MD to investigate the conformational variation in the disordered region of the FEN1 enzyme with the aim of gaining a better understanding of the catalyzed DNA strand cleavage.

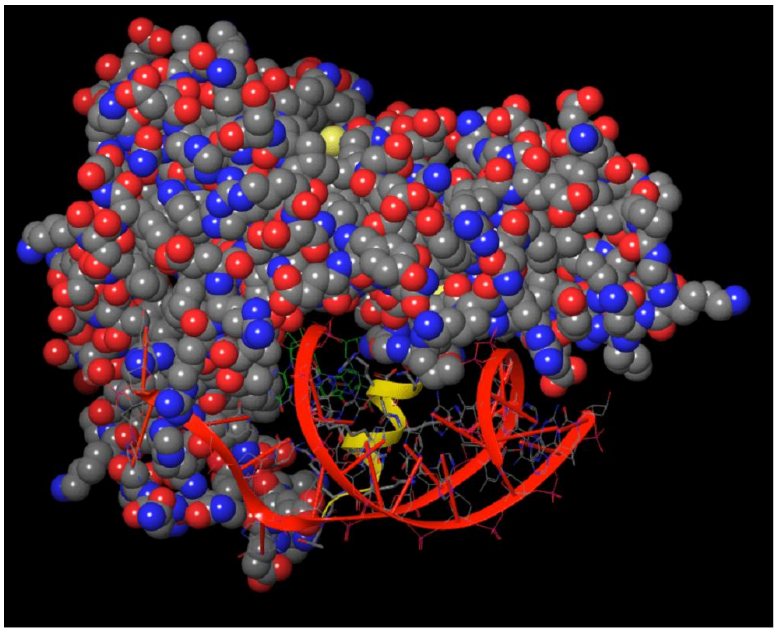


*Figure 1:3D structure of the DNA penetration within the gap of the FEN1 nuclease.*

We initially elucidated the conformational difference between wild/mutated TaqDK-DNA by measuring the RMSD and applying the principal component analysis referring to the whole protein and the flexible loop located between the residues 70 and 90 in order to spot groups of conformations that belong the mutant or the wild variant of the enzyme. We also studied the correlation between the residues by computing the fluctuation of different residues. The analyses of the interactions between the DNA strands and the enzyme were conducted by the monitoring of the distance between the atoms of interest and by the application of a temperature gradient, respectively. It has been revealed that the MD simulation can give very promising insights to better elucidate the mechanism of the DNA cleavage and to help the crystallographer to further refine the outcome of the X-ray diffraction.

1. Kucherlapati, M., Yang, K., Kuraguchi, M., Zhao, J., Lia, M., Heyer, J., Kane, M.F., Fan, K., Russell, R., Brown, A.M. and Kneitz, B., Haploinsufficiency of Flap endonuclease (Fen1) leads to rapid tumor progression. *Proceedings of the National Academy of Sciences*, **2002**. 99(15), pp.9924-9929.
2. Orans, J., McSweeney, E.A., Iyer, R.R., Hast, M.A., Hellinga, H.W., Modrich, P. and Beese, L.S., Structures of human exonuclease 1 DNA complexes suggest a unified mechanism for nuclease family. *Cell*, **2011**. 145(2), pp.212-223.
3. Dervan, J.J., Feng, M., Patel, D., Grasby, J.A., Artymiuk, P.J., Ceska, T.A. and Sayers, J.R., Interactions of mutant and wild-type flap endonucleases with oligonucleotide substrates suggest an alternative model of DNA binding. *Proceedings of the National Academy of Sciences*, **2002**. 99(13), pp.8542-8547.
4. AlMalki, F.A., Flemming, C.S., Zhang, J., Feng, M., Sedelnikova, S.E., Ceska, T., Rafferty, J.B., Sayers, J.R. and Artymiuk, P.J., Direct observation of DNA threading in flap endonuclease complexes. *Nature structural & molecular biology*, **2016**. *23*(7), pp.640-646.

## P-31: Testing the limits of prediction in QSPR models considering their applicability domain

M. von Korff[1], T. Sander[1]

[1]*Idorsia Pharmaceuticals Ltd, Hegenheimer Mattweg 91, 4123 Allschwil, Switzerland*

In drug discovery, molecules are optimized towards desired properties. In this context, machine learning is frequently used for extrapolation in drug discovery projects. Hence, models used for extrapolation are at the border of their applicability domain. Despite the frequent usage, any systematic analysis of the effectiveness of extrapolation in drug discovery has not yet been performed. In response, this study examined the capabilities of six machine learning algorithms to extrapolate from 243 datasets. To guarantee the coherence of the applicability domain, the data sets were constructed by degradation of three blockbuster drugs. The response values calculated from the molecules in the datasets were molecular weight, cLogP, and the number of sp3-atoms. Three experimental setups were chosen for response values. Shuffled data were used for interpolation, whereas data for extrapolation were sorted from high to low values, and the reverse. Extrapolation with sorted data resulted in much larger prediction errors than extrapolation with shuffled data. The error was correlated with the distance measures in the applicability domain. Additionally, this study demonstrated that linear machine learning methods are preferable for extrapolation.

## P-33: Predictive-based selection of drug candidates for Autosomal Dominant Polycystic Kidney Disease (ADPKD)

D. Araripe[1,2], B. Khalil[1,3], H. Bange[4], L. Price[4], D.J.M. Peters[2], G.J.P. van Westen[1]

[1]*Division of Drug Discovery and Safety, Leiden Academic Centre for Drug Research, Leiden University, P.O. Box 9502, 2300 RA, Leiden, The Netherlands*
[2]*Department of Human Genetics, Postzone S-04-P, Leiden University Medical Centre (LUMC), P.O. Box 9600, 2300 RC Leiden, The Netherlands.*
[3]*Janssen Research & Development, LLC, In-Silico Discovery and External Innovation (ISD&EI), B-2340 Beerse, Belgium*
[4]*Crown Bioscience, BioPartner Center, 2333 CH Leiden, The Netherlands.*

Autosomal polycystic kidney disease (ADPKD) is an inherited disorder with an incidence of 9.3 in every 10,000 people[1] and is the main cause of end-stage kidney failure. Tolvaptan, an arginine vasopressin receptor 2 (AVPR2) antagonist,[2] is the only FDA-approved small molecule for ADPKD. It delays kidney function decline, but owing to its side effects, it cannot be administered to all patients. Therefore, ADPKD urges the discovery and development of novel or repurposed therapies. In recent years, In-silico technologies like machine learning have been used to improve the effectiveness and speed of several phases of the drug discovery pipeline. In this work, we propose a data-driven pipeline for evaluating the potential of approved drugs for being repurposed for ADPKD as well as proposing new compounds with activity for this condition. Our approach is based on Crown Bioscience's in-house compound screening data, which measures cAMP-dependent cyst swelling on a 3D Murine Pkd1-/- cell line model.[3,4] Our current dataset contains 2498 compounds screened in duplicates, where the phenotypical readout consists of the total Rhodamine-Phalloidin staining signal of an image stack, representing the three-dimensional cyst size. This continuous readout was separated by a threshold calculated on the control conditions, resulting in 159 active and 2339 inactive compounds. These readouts are used for cystinhibition QSAR modeling based on the extended connectivity fingerprints of radius 2 (ECFP4) of our compounds. The active and inactive compounds underwent a stratified random split into 70% train and 30% test set 5 times for cross-validation, where our baseline random forest classification model displayed a ROC AUC of 0.78 ± 0.05. Furthermore, by querying public databases with our active compounds, we identified 123 of our active compounds with annotated bioactivity values for a total of 624 targets that can be potentially explored for designing active compounds against ADPKD using structure-based methods. These targets will be then assessed whether they have a known connection with the disease based on the current literature.
If approved, this work will showcase the most relevant compounds and targets identified as well as a discussion on the limitations of the proposed techniques.

1. Lanktree, M. B. *et al.* Prevalence estimates of polycystic kidney and liver disease by population sequencing. *Journal of the American Society of Nephrology* **29**, 2593–2600 (2018).
2. Chebib, F. T. *et al.* A practical guide for treatment of rapidly progressive ADPKD with tolvaptan. *Journal of the American Society of Nephrology* **29**, 2458–2470 (2018).
3. Booij, T. H. *et al.* High-Throughput Phenotypic Screening of Kinase Inhibitors to Identify Drug Targets for Polycystic Kidney Disease. *SLAS Discovery* **22**, 974–984 (2017).
4. Booij, T. H. *et al.* In vitro 3D phenotypic drug screen identifies celastrol as an effective in vivo inhibitor of polycystic kidney disease. *Journal of Molecular Cell Biology* **12**, 644– 653 (2020).

## P-35: Virtual Distillation of Naphthas Using Molecular Property Prediction Algorithms

M.R. Dobbelaere[1], Y. Ureel[1], F.H. Vermeire[1], C.V. Stevens[2], K.M. Van Geem[1]

*[1]Laboratory for Chemical Technology, Ghent University, Belgium*
*[2] SynBioC Research Group, Ghent University, Belgium*

Naphtha is a fraction of fossil or renewable oils which boils between approximately 20 °C and 200 °C. It is a complex mixture of hydrocarbons, mainly n-paraffins, iso-paraffins, olefins, naphthenes, and aromatics (PIONA). Distillation curves are among the main characterization methods for oil fractions and they can be used for the calculation of other bulk properties. However, naphthas from plastic waste pyrolysis are typically rich in olefins which complicates the experimental determination of the distillation curve. Therefore, it is important to have an accurate computational method that is able to predict boiling points of naphthas at different percentages of evaporation starting from a lumped composition.
Here, we will present how distillation curves can be predicted using molecular property prediction algorithms, which are originally designed for pure components. The first step of the algorithm consists of a rule-based method which converts a lumped composition into a composition on molecular level.[1] The composition at any point of the distillation curve is defined by the fractions of the molecules with a boiling point above the corresponding distillation temperature. This complex mixture is then represented via a linear combination of the fractions and the molecular representation vectors. The resulting vector has the same dimension of a molecular representation vector, so that a chemical mixture can be considered as a pseudo-molecule. The model is trained on boiling points of both pure components and mixtures and a distillation curve is reconstructed from the predicted boiling points. Trained on only ~1000 experimental pure component normal boiling temperatures, a mean absolute error on the test set is achieved of 2.5 K.[1] Typically, pure component properties are available in higher numbers than mixture properties, so that a wide range of chemical mixture properties can be predicted simply by using existing molecular property prediction tools.

. Dobbelaere, M.R.; Ureel, Y; Vermeire, F.H.; Tomme, L.; Stevens, C.V.; Van Geem, K.M. Machine Learning for Physicochemical Property Prediction of Complex Hydrocarbon Mixtures. *Ind. Eng. Chem. Res.* **2022**, submitted.

## P-37: Use of semi-quantitative (censored) data for QSAR modeling of hERG inhibitory potency

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

*[1] VšĮ "Aukštieji algoritmai", A. Mickevičiaus 29, LT-08117 Vilnius, Lithuania*
*[2] ACD/Labs, Inc., 8 King Street East, Suite 107, Toronto, Ontario, M5C 1B5, Canada*

In a previous publication, we presented a probabilistic classification model based on literature data of hERG inhibitory potential for >6600 drug-like compounds.[1] The main goal of the current study was to bring those predictions onto a quantitative scale by deriving a QSAR model capable of estimating the actual inhibitory potencies.
Experimental data were represented as hERG $IC_{50}$ constants, recorded as either exact values, or open-ended intervals, such as $IC_{50} < 1$ μM or $IC_{50} > 30$ μM (censored data points). The need to account for censored observations arose since many experimental studies perform precise measurements only near the practical classification threshold (around 10 μM), and report all other results in semi-quantitative manner, expressed

as intervals. Model development was performed using Gradient Boosting Machine (GBM) statistical method with a custom optimization function adapted for censored regression objective. Only a minimal set of readily interpretable physicochemical descriptors was used, including LogP, acid and base pKa, molecular size and topology parameters.
During this study the experimental database was further expanded to an overall size of >8800 molecules. According to the preliminary modeling results, when $pIC_{50}$ predictions by the quantitative model are converted back to binary classification at 10 μM threshold, the external validation set compounds can be classified with >75% overall accuracy. The quantitative model also shows a tendency of slightly outperforming the analogous logistic model and consistently achieves a better balance between sensitivity and specificity metrics. A major advantage of this type of model is that its output is much easier to interpret and allows the user not only to discern potential hERG inhibitors from non-inhibitors, but also to rank the compounds by their inhibitory potential.

1. Didziapetris, R., Lanevskij, K. Compilation and physicochemical classification analysis of a diverse hERG inhibition database. J. Comput. Aided Mol. Des., **2016**, 30, 1175-1188.

## P-39: DFT and ML modeling of peptide properties for cytotoxicity prediction

A. Markovnikova[1], A. Novikov[2], M. Kurushkin[1]

*[1]Chemistry Education Research and Practice Laboratory, SCAMT Institute, ITMO University, 9 Lomonosova Str., Saint Petersburg, Russian Federation, 191002*
*[2] Infochemistry Scientific Center, ITMO University, 9 Lomonosova Str., Saint Petersburg, Russian Federation, 191002*

The prediction of cytotoxicity for various chemical substances (including peptides) is an innovative topic in modern machine learning (ML). The DFT calculations could be used for generation of initial datasets for training of artificial neural networks. This report is focused on the presentation of our attempts in modeling of peptide properties for cytotoxicity prediction based on combined DFT and ML approaches (Figure 1). Our project consisted of two parts. The first part is calculations of various properties of peptides – we found and constructed several valid ML models that allow to predict following parameters: ΔG, ΔH, ΔS, ε, μ, ν, ρ, ζ. We presented an extensive and diverse database of peptide conformational energies. Our database contains five different classes of model geometries: dipeptides, tripeptides, and disulfide-bridged, bioactive, and cyclic peptides. All the reference energies have been calculated at the LC-ωPBE-XDM/aug-cc-pVTZ level of theory, which is shown to yield conformational energies with an accuracy in the order of tenths of a kcal/mol when compared to complete-basis-set coupled-cluster reference data.[1] The second part deals with correlations between the physical and chemical properties of peptides and parameters of real substances useful for medical industry and pharmacy. The preliminary calculations were carried out using interactive molecular dynamics in virtual reality open-source multi-person framework NarupaXR[2]. The ORCA program package[3] was additionally used for future advanced DFT calculations. The Weka software[4] was used for machine learning. Results of this work would be useful as a fundamental basis for treatment of inflammatory and autoimmune desires as well as for creation of anti-cancer innovative drugs.

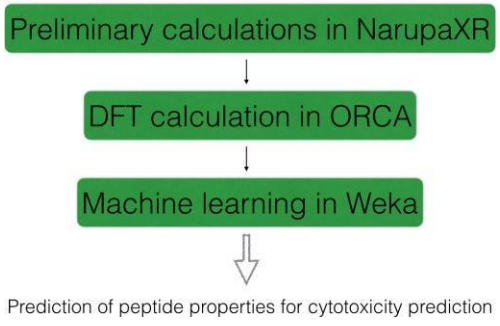



**Figure 1**: Road map for our project.

. Prasad, V., et al., PEPCONF, a diverse data set of peptide conformational energies. Scientific data., **2019**, 6, 1, 1-9

. Jamieson-Binnie, A., et al., Visual Continuity of Protein Secondary Structure Rendering: Application to SARS-CoV-2 Mpro in Virtual Reality. Frontiers in Computer Science., **2021**, 63

. Neese, F. The ORCA program system. Wiley Interdisciplinary Reviews: Computational Molecular Science., **2012**, 2, 1, 73-78.

. Frank, E., et al., The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, **2016**.

## P-41: Conservation Analysis of anti-TB Target DnaE1 and Identification of Potential Interactions of DnaE1 Inhibitor Nargenicin on the Human Proteome

R.C.M Kuin[1], T.H.W Bäck[2], M.H. Lamers [3], G.J.P. van Westen[1]

*[1]Leiden Academic Centre for Drug Research (LACDR), Einsteinweg 55, 2333 CC Leiden, The Netherlands*

*[2]Leiden Institute of Advanced Computer Science (LIACS), Niels Bohrweg 1, 2333 CA Leiden, The Netherlands*

*[3] Department of Cell & Chemical Biology, Leiden University Medical Center (LUMC), Einthovenweg 20, 2333 ZC Leiden, The Netherlands*

Tuberculosis (TB) is a bacterial infection caused by Mycobacterium tuberculosis (Mtb) and is the leading cause of death from a single infectious agent worldwide.[1] Presently, the emergence of resistant forms of TB and in particular multidrug- and extensively drug-resistant TB (MDR/XDR-TB) form a growing threat to global health. Hence, there is an urgent need for new antibiotics that inhibit novel targets.

Here, the replicative DNA polymerase DnaE1 from Mtb, which is responsible for DNA synthesis in this bacterium, is explored as an anti-TB target using computational techniques. Recent studies have shown that the natural product nargenicin inhibits DnaE1 and inhibits growth in Mtb.[2] In this study the sequence conservation of DnaE1 was calculated to identify structurally and functionally important residues that are not likely to play a role in developing resistance against nargenicin, as mutating these residues is likely to be destructive for the protein. For this, an multi-sequence alignment (MSA) consisting of 17 bacterial replicative DNA polymerase sequences from different species was used. Using this MSA, the Shannon Entropy was calculated per residue position. Several unique residues were identified (I220, R332, V865 and H978) that were present in the replicative DNA polymerase, but not conserved in the other polymerases; see **Figure 1**. This suggests that these residues might be crucial for DnaE1 functioning, but follow-up experiments are needed to elucidate their roles.

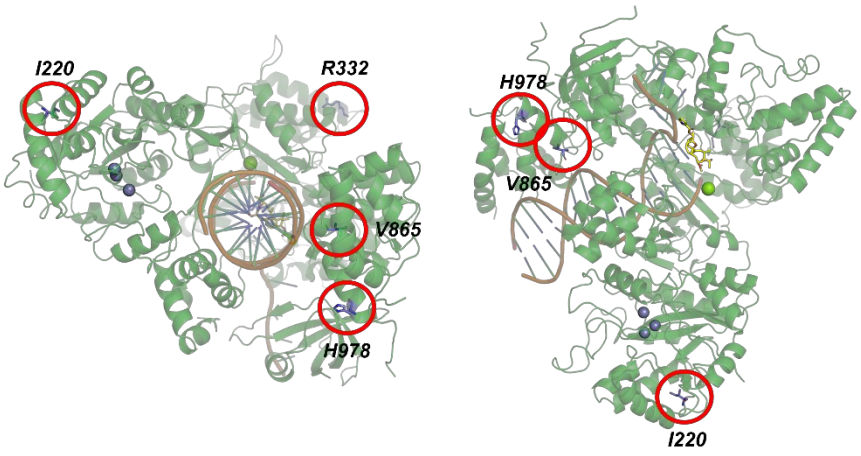



**Figure 1:** Location of highly conserved residues in Mycobacterial DnaE1. Positions of the conserved residues are marked by red circles. DNA is shown in orange and blue sticks and the inhibitor nargenicin is shown in yellow sticks. The cryo-EM structure of Mtb DnaE1 bound to DNA and nargenicin was obtained from Chengalroyen et al.[2]

Furthermore, in this study potential off-target interactions of nargenicin with the human proteome were identified using computational docking experiments. For this, a set of high-confidence AlphaFold protein structures[3] was selected, after which each of these structures was used to generate a nargenicin-protein complex using docking in AutoDock Vina[4]. High-affinity complexes were redocked using Molsoft ICM[5].

Analysis of results led to identification of four proteins that potentially interact with nargenicin. Currently, additional in vitro and in silico experiments are underway to validate these preliminary results.

To summarize, we have identified conserved residues for different bacterial replicative DNA polymerase genes and identified proteins to which nargenicin potentially binds. We are currently following these findings up with experimental validation.


1. Fukunaga et al, "Epidemiology of Tuberculosis and Progress Toward Meeting Global Targets - Worldwide, **2019**."
2. Chengalroyen et al., "DNA-Dependent Binding of Nargenicin to DnaE1 Inhibits Replication in Mycobacterium Tuberculosis", ACS Infectious Diseases., **2022**, Article ASAP
3. Jumper et al., "Highly Accurate Protein Structure Prediction with AlphaFold", Nature., **2021**, 596, 7873, 583-589
4. Trott and Olson, Trott and Olson, "AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading", Journal of computational chemistry., **2010**, 31, 2, 455-461
5. Neves, Totrov, and Abagyan, Neves, Totrov, and Abagyan, "Docking and Scoring with ICM: The Benchmarking Results and Strategies for Improvement", Journal of computer-aided molecular design, **2012**, 26, 6, 675


## P-45: Structural Investigations of Protein Kinases with GeoMine


C. Ehrt[1], J. Graef[1], K. Diedrich[1], M. Poppinga[1,2], N. Ritter[2], M. Rarey[1]

*[1]Universität Hamburg, ZBH - Center for Bioinformatics, Research Group for Computational Molecular Design, Bundesstraße 43, 20146 Hamburg, Germany*

*[2] Universität Hamburg, Department of Informatics, Databases and Information Systems Group, Vogt-Kölln-Straße 30, 22527 Hamburg, Germany*


GeoMine[1] on the Proteins*Plus*[2] web server enables textual, numerical, and geometrical searches in 1,067,518 ligand-based and predicted binding sites in the Protein Data Bank (PDB).[2] Given a protein binding site of interest, individual user-defined patterns can be designed for geometric searching in binding sites on the atomic level. This opens new opportunities for typical challenges, e.g., in the field of protein kinase research.[4]

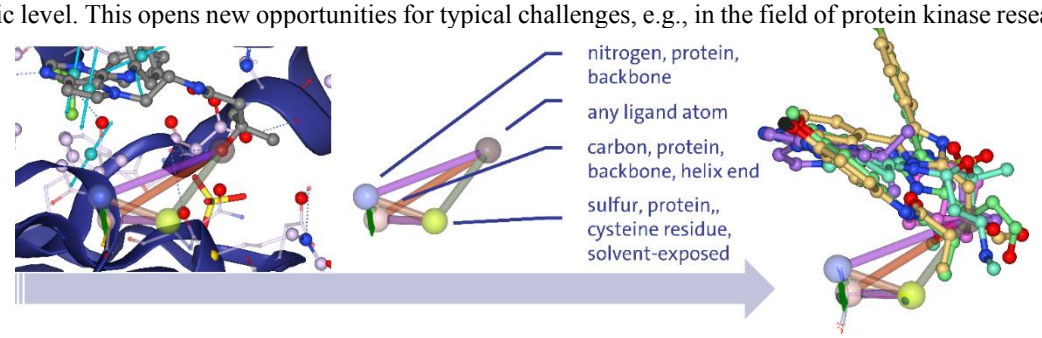



**Figure 1**: GeoMine Query to Screen for Reactive Cysteine Residues in the Proximity of Ligands

In this contribution, we show how GeoMine can be used to screen for rare interactions and exploit these, e.g., for the design of selective inhibitors. Additionally, we designed queries based on known protein-ligand interaction patterns of highly active, but unselective compounds to find which kinase binding sites in the Kinase Ligand Interaction Fingerprints (KLIFS)[5] database match these interaction patterns. Subsequently, we explore whether additional ligand interaction anchors might lead to more selective inhibitors. In a final example, we perform a screening for reactive cysteine residues in protein kinase structures to identify them and demonstrate its potential for differentiating between cysteine positions that are frequently occurring in protein kinases and cysteine positions that are specific for kinase families. All these applications of the method show that GeoMine can serve as a highly flexible and comprehensive tool to assist in drug design processes, not only for protein kinases but for a multitude of pharmaceutically interesting targets in the PDB.


1. Diedrich, K., et al., GeoMine: Interactive Pattern Mining of Protein-Ligand Interfaces in the Protein Data Bank. Bioinformatics, **2020**, 37, 3, 424-425
2. Schöning-Stierand, K., et al., ProteinsPlus: Interactive Analysis of Protein–Ligand Binding Interfaces. Nucleic Acids Res., **2020**, 48, W48-W53
3. Berman, H.M., et al., The Protein Data Bank. Nucleic Acids Res., **2000**, 28, 235-242
4. Graef, J., et al., Searching Geometric Patterns in Protein Binding Sites and Their Application to Data Mining in Protein Kinase Structures, J. Med. Chem., **2022**, 65, 2 1384-1395

5. Kanev, G.K., et al., KLIFS: An Overhaul After the First 5 Years of Supporting Kinase Research, Nucleic Acids Res., **2021**, 49, D1, D562-D569

## P46: The application of the MM/GBSA method in the binding pose prediction of FGFR inhibitors

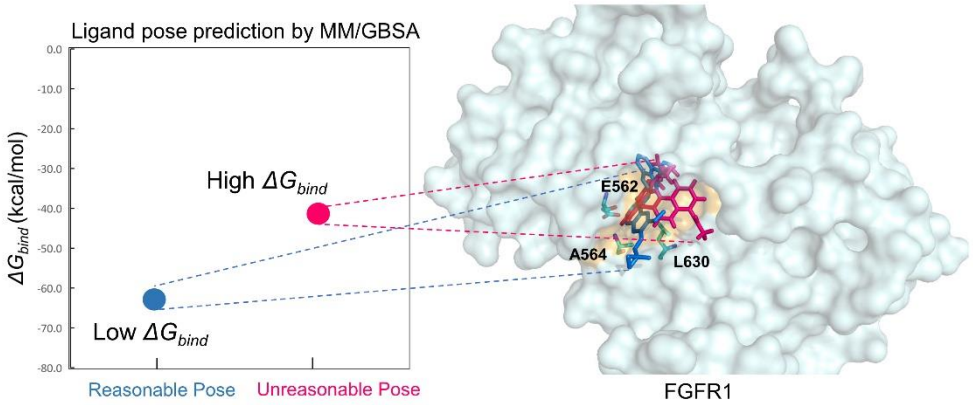
Y. Chen[1], Q. T. Wang[2]

[1]*Institute of Pharmacy, Freie Universität Berlin, Berlin 14195, Germany*

[2]*West China School of Pharmacy, Sichuan University, Chengdu 610041, China*

The success of structure-based drug design, or more specifically lead optimization, is highly dependent on a known binding pose of the protein-ligand system. However, this is not always available to many groups. Therefore, a reliable and cost-effective alternative approach is of great interest. In this work[1], we set out to explore the applicability of the popular and easy-to-use MD-based MM/GBSA[2], method to determine the binding poses. Although this method has been introduced and widely used for a long time, much effort was made to explore its performance to estimate binding affinity between different ligands in previous studies. However, its performance was not satisfactory in this regard. This is why we want to emphasize the application of MM/GBSA in ligand pose prediction, which might be a more appropriate application scenario for MM/GBSA.

This work is trying to answer two major scientific questions: 1) which is the best way to determine the binding pose of a ligand using MD simulation and MM/GBSA calculation; 2) is longer MD simulation useful for pose prediction? And how long would be good enough? Given the amount of known co-crystal structures and the importance of kinases as drug targets, we chose FGFR as an example. A total of 28 ligands of FGFR, including 10 with co-crystal structures, were studied. For each ligand, 2 to 5 poses were generated, and each was simulated for more than 100 ns. It was found that MM/GBSA combined with MD simulation significantly improved the success rate of docking methods from 30-40% to 70%. This work demonstrates a way that the MM/GBSA method can be more accurate in ligand pose prediction than it is in ligand affinity ranking, filling a gap in structure-based drug discovery when the binding pose is unknown.




***Graphic Abstract***. *MM/GBSA calculation based on long MD simulations distinguished the reasonable binding pose from the unreasonable pose of the ligands. Lower binding free energy ($\Delta G_{bind}$) was predicted for the correct binding pose for each FGFR inhibitor.*

**Acknowledgments**
Reproduced from Ref. 1 with permission from the PCCP Owner Societies.

**References**

1. Chen, Y.; Zheng, Y.; Fong, P., *et al*, The application of the MM/GBSA method in the binding pose prediction of FGFR inhibitors. *Physical chemistry chemical physics : PCCP* **2020,** *22* (17), 9656-9663.
2. Kollman, P. A.; Massova, I.; Reyes, C., *et al,* Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Accounts of Chemical Research* **2000,** *33* (12), 889-897.

## P-47: Ultralarge Virtual Screening Identifies SARS-CoV-2 Main Protease Inhibitors with Broad-Spectrum Activity against Coronaviruses

Luttens A.[1], Gullberg H.[2], Abdurakhmanov E.[1], Vo Duc D.[1], Akaberi D.[1], Talibov O. V.[3], Nekhotiaeva N.[2], Vangeel L.[4], De Jonghe S.[4], Jochmans D.[4], Krambrich J.[1], Tas A.[5], Lundgren B.[2], Gravenfors Y.[2], Craig J. A.[1], Atilaw Y.[1], Sandström A.[1], Moodie W.K. L.[1], Lundkvist Å.[1], van Hemert J. M.[5], Neyts J.[4], Lennerstrand J.[1], Kihlberg J.[1], Sandberg K.[1], Danielson H.[1], Carlsson J.[1]

[1]*Uppsala University, Uppsala, Sweden*
[2]*Stockholm University, Stockholm, Sweden*
[3]*MAX IV Laboratory, Lund University, Lund, Sweden*
[4]*Rega Institute, KULeuven, Leuven, Belgium*
[5]*Leiden University Medical Center, Leiden, The Netherlands*

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has caused the greatest health crisis of this generation and already led to >5 million deaths world-wide. Analogous to common cold viruses, SARS-CoV-2 is expected to continue to circulate and pose a significant threat to our society. Despite promising vaccination and booster programs against COVID-19, antiviral drugs will likely be crucial to control the future outbreaks of coronaviruses. Among the proteins encoded by the SARS-CoV-2 genome, the main protease ($M^{pro}$) has emerged as a promising target. Inhibition of $M^{pro}$ blocks the processing of polyproteins produced by translation of the viral RNA, which is an essential step in SARS-CoV-2 replication.

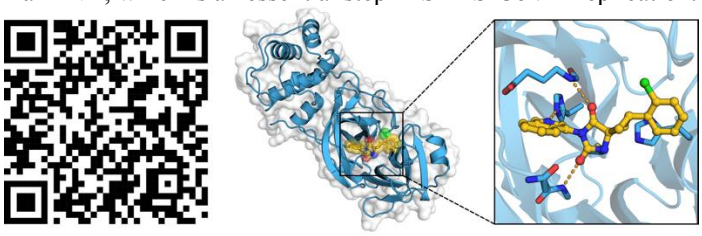


**Figure 1**: SARS-CoV-2 $M^{pro}$ complexed with a novel broad-spectrum inhibitor

The determination of high-resolution crystal structures of SARS-CoV-2 proteins has enabled virtual screening campaigns to identify hits that can be developed into antiviral drugs.[1] Structure-based docking algorithms can sample and score binding poses in seconds, making it possible to evaluate large libraries and this approach is not restricted to compounds that are physically available.[2] The size of libraries with commercially available compounds is growing rapidly and >20 billion make-on-demand molecules are currently available from chemical suppliers. These libraries provide opportunities to identify potential therapeutic agents that can readily be synthesized and tested for activity, but require development of effective strategies for navigation in this enormous chemical space. We present two different strategies to search for $M^{pro}$ inhibitors in ultra-large chemical libraries using structure-based docking.[3] Synergy between molecular modeling, protein crystallography and organic synthesis led to a novel broad-spectrum inhibitor of coronaviruses.

1. Douangamath A., et al., Crystallographic and electrophilic fragment screening of the SARS-CoV-2 main protease. *Nat. Commun.*, **2020**, 11, 5047
2. Bender B., Gahbauer S., Luttens A., et al., A practical guide to large-scale docking. *Nat. Protoc.*, **2021**, 16, 4799–4832
3. Luttens A., et al., Ultralarge Virtual Screening Identifies SARS-CoV-2 Main Protease Inhibitors with Broad-Spectrum Activity against Coronaviruses. *J. Am. Chem. Soc*., **2022**, 144, 7, 2905–2920

## P-49: GenCReM: de novo generation of synthetically feasible compounds based on genetic algorithm

A. Ivanová[1], G. Minibaeva[1], P. Polishchuk.[1]

[1]*Institute of Molecular and Translational Medicine, Faculty of Medicine and Dentistry, Palacký University, Hněvotínská 1333/5, 779 00 Olomouc, Czech Republic*

De novo generation approaches provide an alternative to virtual screening. They suggest a more effective way to adaptively explore the chemical space, which is extremely huge and is not affordable for exhaustive enumeration and screening. Synthetic feasibility of de novo generated compounds is still the main issue of many currently available de novo generation approaches. In our work we used the recently developed CReM[1] approach which allows to implicitly control synthetic feasibility of generated compounds and the genetic algorithm to efficiently explore chemical space. The main goal of the developed tool is to generate sets of diverse compounds satisfying given criteria represented by a custom scoring function. The large number of structurally diverse output hits was a particular goal because it gives a possibility to a researcher to choose which molecular series to explore further.
In our pilot study we used docking as a major component of the objective function. Other component of the objective function was diversity of molecules composing a chromosome. Thus, the main goal was to maximize docking score and diversity of the generated molecules. Optionally, other important parameters can be controlled, e.g. lipophilicity, the number of rotatable bonds, drug-likeness, ligand-protein interaction fingerprints to preserve important protein-ligand interactions, etc. The implemented pipeline has two working modes: i) scaffold decoration, where selected atoms are protected from mutations and ii) unrestricted de novo generation, where molecules are generated completely from scratch.
The approach was evaluated on a range of targets using different setups of the CReM generator (different fragment databases and context radiuses). For generated compounds we calculated synthetic accessibility score and showed that in all cases high-scored synthetically feasible sets of compounds were obtained. The estimated synthetic accessibility predictably improved if we chose a greater radius and/or a more restricted fragment database obtained by fragmentation of more synthetically feasible ChEMBL compounds. Thus, there is no need to implicitly include a synthetic accessibility estimate in the objective function. The developed approach based on a genetic algorithm can be easily extended to other approaches and models besides docking, e.g. pharmacophore or machine learning models. It can be also applied for multi-objective optimization that is planned in future studies.
This work was funded by the INTER-EXCELLENCE LTARF18013 project (MEYS), the European Regional Development Fund - Project ENOCH (No. CZ.02.1.01/0.0/0.0/16_019/0000868) and ELIXIR CZ research infrastructure project (MEYS Grant No: LM2018131).

1. Polishchuk, P.,CReM: Chemically reasonable mutations framework for structure generation, Journal of Cheminformatics, 2020, 12, 1, 1-18.

## P-51: MD pharmacophore-based search for novel MARK4 inhibitors

Kutlushina A.[1], Mokshyna O.[1], Hruba L.[1], Gurska S.[1], Dzubak P.[1], Hajduch M.[1], Polishchuk P[1]

[1]*Institute of Molecular and Translational Medicine, Faculty of Medicine and Dentistry, Palacky University, Hnevotinska 5, Olomouc, Czech Republic*

Microtubule affinity-regulating kinase 4 (MARK4) is a Ser/Thr protein kinase, which affects microtubules through phosphorylation of MAP2, MAP4 and tau proteins. Since microtubules are involved in many biological processes, MARK4 can be a potential target for the treatment of Alzheimer's disease, cancer, atherosclerosis and type II diabetes. In our study we developed and applied a special workflow which includes MD pharmacophores and docking. The goal of the study was to find MARK4 inhibitors with novel scaffolds. First, we selected 3 ligands from our previous studies which demonstrated high inhibitory activity. They were docked to MARK4 (PDB 5ES1) and further 100 ns molecular dynamic (MD) simulations were performed of these three complexes. From each frame of these MD trajectories we extracted 3D pharmacophores by the previously developed pharMD tool[1]. To reduce redundancy identical or very similar pharmacophores were identified and removed by their 3D pharmacophore hashes which were calculated by pmapper[2]. The remaining pharmacophore models were used to screen the Enamine database consisting of more than 2 million compounds. Compounds were ranked by their consensus scoring. 1000 top scored compounds were docked to MARK4 using Autodock Vina. Finally 24 compounds were selected and purchased from Enamine. Two compounds demonstrated activity in primary screening and $IC_{50}$ values were measured for them. One of them demonstrates moderate activity ($IC_{50}$ = 12 uM). Activity of the other compound was high ($IC_{50}$ = 30 nM) and comparable to activity of reference compounds used on the first stage of the study. However, the scaffold of the newly identified inhibitor was completely different from those reference structures. Further we plan to

measure selectivity of these identified hits to other subtypes of MARK and perform preliminary ADME studies.

This work was funded by the INTER-EXCELLENCE LTARF18013 project (MEYS), the European Regional Development Fund - Project ENOCH (No. CZ.02.1.01/0.0/0.0/16_019/0000868) and ELIXIR CZ research infrastructure project (MEYS Grant No: LM2018131).

1. Polishchuk, P.; Kutlushina, A.; Bashirova, D.; Mokshyna, O.; Madzhidov, T., Virtual Screening Using Pharmacophore Models Retrieved from Molecular Dynamic Simulations, International Journal of Molecular Sciences, **2019**, 20, 5834.
2. Kutlushina, A.; Khakimova, A.; Madzhidov, T.; Polishchuk, P., Ligand-Based Pharmacophore Modeling Using Novel 3D Pharmacophore Signatures, Molecules, 2018, 23, 3094.

# Poster Session Abstracts BLUE

## P-02: Ring systems in natural products: structural diversity, physicochemical properties, and coverage by synthetic compounds

Y. Chen[1], C. Rosenkranz[2], S. Hirte[1], J. Kirchmair[1]

[1]*Department of Pharmaceutical Sciences, Division of Pharmaceutical Chemistry, Faculty of Life Sciences, University of Vienna, 1090 Vienna, Austria*

[2] *Center for Bioinformatics (ZBH), Universität Hamburg, 20146 Hamburg, Germany*

The majority of modern small-molecule drugs is inspired, to different extents, by natural products (NPs).1 Much of the significance of NPs can be attributed to their ring systems, which form the structural core of many drugs. However, in spite of their importance, our understanding of NP ring systems and how their full potential can be harnessed in drug discovery and design is still limited.
Here we present a comprehensive cheminformatic analysis of more than 35,000 NP ring systems with regard to their structural and physicochemical properties, and compare them with those of ring systems found in readily purchasable, synthetic compounds and approved drugs. The data sets were carefully curated to obtain clean sets of NPs and synthetic compounds. In addition to key 2D physicochemical properties such as molecular weight and number of hydrogen bond donors/acceptors, 3D shape and electrostatic properties were explored.
In NPs research, stereochemical information is important as they contribute largely to the structural complexity and bioactivity activities. However, this information is often incomplete in the databases (and sometimes even wrong).[2] Therefore, most cheminformatics studies disregard stereochemical information. To maximize the usage of available structures even when the available stereochemical information is incomplete and also keep the accuracy, we took stereochemistry into account by following an evidence-based logic.
This study shows that structures of NP ring systems are much more diverse than those of ring systems observed in synthetic compounds. In particular a large number of macrocycles are represented by NPs but not among synthetic compounds. Approximately half of the NP ring systems are represented by ring systems with identical or related 3D shape and electrostatic properties. Meanwhile, only about 2% of the NP ring systems are observed in approved drugs, leaving a huge number of potential ring systems to be explored in small-molecule drug discovery.

1. Newman, D. J., Cragg, G. M. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. *J. Nat. Prod.* **2020**, *83* (3), 770–803.
2. Kramer, C., Podewitz, M., Ertl, P., Liedl, K. R. Unique Macrocycles in the Taiwan Traditional Chinese Medicine Database. *Planta Med.* **2015**, *81* (6), 459–466.

## P-04: Utilizing the semantic web and network tools to integrate pharmacokinetic, -dynamic, and OMICS data with metabolic (disease) pathways

D. Slenter[1], E. Willighagen[1]

[1]*Dept of Bioinformatics - BiGCaT, NUTRIM, Maastricht University, The Netherlands*

Large portions of biomedical data and knowledge are captured in various databases and research papers, having limited capabilities to interact with each other. Integrating information from these resources could be useful for repurposing existing drugs for (rare) metabolic diseases and investigating the synergies of drug combinations, by in silico models supported by the appropriate kinetic and drug (response) data. Unfortunately, this data often needs to be manually scavenged from various sources which are incompatible with other tools. This project describes our approach to apply the semantic web technology Resource Description Framework (RDF) to increase the interoperability of said data and create pharmacological compatible pathways for metabolic disorders on the fly. First, to test our approach, six machine readable pathways were created with PathVisio and uploaded to WikiPathways, converting the pathway to the RDF format. Second, four kinetics databases and literature were examined to identify relevant kinetic parameters for each individual pathway. Third, an RDF model of the kinetic data, compatible with the pathway model and other databases, was created. Fourth, three drug-target databases were investigated to find applicable

inhibitors and their respective dynamic data. Fifth, the data was collapsed onto the pathways through Cytoscape, allowing for the integration with various types of OMICS data. This approach led to pathway models supported by available kinetic and drug-target information for various inherited metabolic disorders. By capturing this data in a semantic model, researchers can easily assess which interactions are missing data, shortening wet-lab time. Furthermore, adding additional data is user-friendly, allowing others to utilize and extend our method for other pathways of interest.

## P-06: The DECIMER (Deep lEarning for Chemical ImagE Recognition) project

K. Rajan[1], H.-O. B.[1], A. Zielesny[2], C. Steinbeck[1]

[1]*Institute for Inorganic and Analytical Chemistry, Friedrich-Schiller-University Jena, Lessingstr. 8, 07743 Jena, Germany*

[2]*Institute for Bioinformatics and Chemoinformatics, Westphalian University of Applied Sciences, August-Schmidt-Ring 10, D-45665 Recklinghausen, Germany*

Significant amounts of information on chemical compounds, their structures, and their properties have been published in scientific articles. Only a fraction of this knowledge is available in open databases. Retrospectively curating open data from books and journals automatically or semi-automatically, therefore, is a timely challenge [1]. Towards this end, tools for extracting chemical structure depictions and converting them into computer-readable formats are needed. Such an Optical Chemical Structure Recognition (OCSR) tool translates the image of a chemical structure into a machine-readable representation. With DECIMER (Deep lEarning for Chemical ImagE Recognition) [2] an open-source automated software solution has been developed to address the OCSR problem through deep learning for image segmentation and recognition. DECIMER includes a deep-learning-based segmentation algorithm (DECIMER-Segmentation) [3] for automated recognition and segmentation of chemical structures from the scientific literature, as well as an OCSR engine (DECIMER-Image transformer) [4] with a deep learning model based on CNN + Transformer networks which can predict SMILES with over 90% accuracy from depictions of chemical structures.
DECIMER can be applied to older articles before vector images were introduced to PDFs since it uses bitmap images of journal pages. Segmented images can be directly fed into the DECIMER-Image transformer to be converted into SMILES strings. Benchmark results on the OCSR benchmark datasets show that the DECIMER - Image transformer outperforms all currently available open-source algorithms for OCSR. DECIMER's code is open source and the trained models are also openly available.

1. Rajan K, Brinkhaus HO, Zielesny A, Steinbeck C (2020) A review of optical chemical structure recognition tools. J Cheminform 12:60
2. Rajan K, Zielesny A, Steinbeck C (2020) DECIMER: towards deep learning for chemical image recognition. J Cheminform 12:65
3. Rajan K, Brinkhaus HO, Sorokina M, Zielesny A, Steinbeck C (2021) DECIMER-Segmentation: Automated extraction of chemical structure depictions from scientific literature. J Cheminform 13:20
4. Rajan K, Zielesny A, Steinbeck C (2021) DECIMER 1.0: deep learning for chemical image recognition using transformers. J Cheminform 13:61

## P-08: New approaches for antimicrobial peptides prediction using Machine-Learning

C. Bournez[1], M. Riool[2], L. de Boer[2], R.A. Cordfunke[3], L. de Best[4], R. van Leeuwen[4], J.W. Drijfhout[3], S.A.J. Zaat[2], G.J.P. van Westen[1]

[1]*Division of Medicinal Chemistry, Drug Discovery and Safety, Leiden Academic Centre for Drug Research, Leiden University, P.O. Box 9502, 2300 RA, Leiden, Netherlands*

[2]*Department of Medical Microbiology and Infection Prevention, Amsterdam institute for Infection and Immunity, Amsterdam UMC, University of Amsterdam, 1105 AZ Amsterdam, Netherlands*

[3]*Department Immunology , Leiden University Medical Center, 2300 RC Leiden, Netherlands*

[4]*Madam Therapeutics B.V., Pivot Park Life Sciences Community, Kloosterstraat 9, 5349AB Oss, Netherlands*

The existing risk that we enter in a « post-antibiotic era », as stated by The World Health Organization (WHO), makes urgent the finding of new drugs and alternatives to classical antibiotics, which may no longer be effective against common infections [1]. One solution could arise from antimicrobial peptides (AMPs), natural innate host defense molecules produced by all forms of life [2]. Indeed, unlike most traditional small molecule antibiotics, AMPs do not have a specific structural target but directly act on the membrane of the microorganism causing its lysis [3]. Hence, it is important to create fast and accurate methods to discover and design potent AMPs as alternative therapeutics. However, finding novel AMPs through classical wet-lab screening is both time and money consuming. Therefore, various *in silico* approaches were developed over the last years and the keen interest in the field is growing [4]. Still some improvements can occur as the majority of such tools do not consider bacterial species, or at least membrane structure, differences in their application and they do not use a negative dataset based on experimental data.

Here we present an innovative AMP prediction tool based on a machine learning algorithm able to predict activity against either Gram-positive, Gram-negative bacteria or fungi. The input dataset was collected from either previously obtained experimental data from partners or public databases. Moreover, the peptides were labelled AMP (≤15 µM) or Non-AMP (>25 µM) based on these experimental data so that all peptides present proven experimental results. Different algorithms were assessed and created, ensemble tree-based algorithms presenting best outcomes. A primary selection of >50 peptides, predicted as active or inactive per our models, were synthesized and tested against both Gram-positive and Gram-negative bacteria showing promising results.

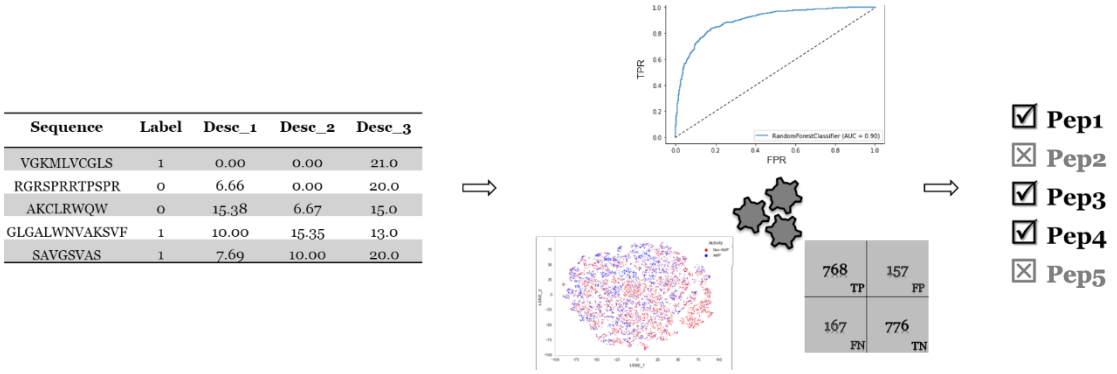



**Figure 1**: Illustration of the different steps (creation to prediction) of our model

1. Reardon, S. WHO Warns against "post-Antibiotic" Era. *Nature* **2014**, doi:10.1038/nature.2014.15135.
2. Ageitos, J.M.; Sánchez-Pérez, A.; Calo-Mata, P.; Villa, T.G. Antimicrobial Peptides (AMPs): Ancient Compounds That Represent Novel Weapons in the Fight against Bacteria. *Biochemical Pharmacology* **2017**, *133*, 117–138, doi:10.1016/j.bcp.2016.09.018.
3. Huan, Y.; Kong, Q.; Mou, H.; Yi, H. Antimicrobial Peptides: Classification, Design, Application and Research Progress in Multiple Fields. *Frontiers in Microbiology* **2020**, *11*, 2559, doi:10.3389/fmicb.2020.582779.
4. Wu, Q.; Ke, H.; Li, D.; Wang, Q.; Fang, J.; Zhou, J. Recent Progress in Machine Learning-Based Prediction of Peptide Activity for Drug Discovery. *Current Topics in Medicinal Chemistry* **2019**, *19*, 4–16, doi:10.2174/1568026619666190122151634.

## P-10: Application of DeepSMILES to machine-learning of chemical structures

O'Boyle NM[1], Dalke A[2], Thomas M[3], Bender A[3], de Graaf C[1]

[1] *Computational Chemistry, Sosei Heptares, Cambridge, UK*
[2] *Andrew Dalke Scientific AB, Trollhättan, Sweden*
[3] *Centre for Molecular Informatics, University of Cambridge, Cambridge, UK*

DeepSMILES [1, 2] is a SMILES-like syntax designed to be more suited than SMILES to machine-learning and manipulation of chemical structures. When machine-learning models are applied to the generation of SMILES strings (as in the GPCR case study [3]), syntactically invalid SMILES strings are often generated due to the mismatch of particular components (parentheses, ring closure numbers) that must occur in pairs. The syntax of DeepSMILES avoids these problems by using only close parentheses (and indicating the branch length), and only a single symbol for ring closure (indicating the ring size).

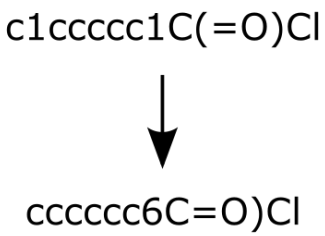



**Figure 1:** A SMILES string (top) and its equivalent DeepSMILES string (bottom)

Here we show the results of a comparison of SMILES to DeepSMILES in the context of *de novo* molecule generation by a recurrent neural network (RNN). The model was trained either on the one-hot encoded representations of the SMILES or DeepSMILES string representation of molecules in ChEMBL28. Once trained with equal training parameters and epochs, *de novo* molecules were sampled in the form of the respective string. The ratio of syntactically valid to invalid strings was identified as well as the nature of any syntax errors in order to understand how well the RNN has learnt the respective syntax. Furthermore, we measured generative model performance by a suite of metrics on de novo molecules to indicate any performance benefit from using syntactically simplified DeepSMILES. We also compare the results to those for SELFIES [4], a string representation of molecules designed to only represent valid molecules.
Beyond its application to machine learning, the DeepSMILES syntax has also found applications in other areas of cheminformatics for the manipulation of chemical information. For example, we show how it has been applied to the problem of fuzz testing cheminformatics software to identify bugs in parsers.

1. O'Boyle NM, Dalke A. DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures. ChemRxiv, 2018. (Preprint)
2. DeepSMILES GitHub repository. https://github.com/baoilleach/deepsmiles
3. Thomas M, Smith RT, O'Boyle NM, de Graaf C, Bender A. Comparison of structure- and ligand-based scoring functions for deep generative models: a GPCR case study. J. Cheminf. 2021, 13, 39.
4. Krenn M, Häse F, Nigam A, Friederich P, Aspuru-Guzik A. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. Mach. Learn. Sci. Technol. 2020, 1, 045024.

## P-12: Towards Predicting Enzyme Activity by Traversing Biomedical Knowledge Graphs

T. Egbelo[1], V. Sykora[2], M. Bodkin[2], Z. Zhang[1], V. Gillet[1]

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

[2]*Research Informatics and In Silico R&D, Evotec (UK) Ltd, 114 Park Drive, Abingdon OX14 4RZ, United Kingdom*

New drug discovery remains central to the aspiration of improving health care. Nevertheless, the drug discovery process is complex and stubbornly resource intensive. The conceptualisation of biological systems as networks along with their representation using suitable graph data models has opened the door for the adaptation of a great diversity of machine learning methods that exploit the relational nature of such data. For drug discovery, a particularly rich avenue for network-based knowledge discovery has been to cast compound property prediction as a knowledge graph completion, or link prediction, problem. In a biomedical knowledge graph, which is in essence a heterogeneous network integrating the relationships between entities such as genes, proteins, compounds and diseases, a variety of interesting properties of a given entity are encoded as direct links to other entities, and these properties may correlate with more complex patterns within the graph. Identifying and exploiting such associations lies at the heart of drug discovery when framed as a knowledge graph completion problem. This poster shall summarise the results of the initial efforts to explore and tackle it in this form.
As a first step, a knowledge graph was created by integrating public data sources from the areas of proteomics, chemistry and pharmacology. Data from Ensembl (Gene and Protein nodes), ChEMBL (Compound, Assay and Measurement nodes), the Experimental Factor Ontology (Disease nodes) and the Gene Ontology (Biological Process nodes), among others, have been merged into a single graph (Figure 1).

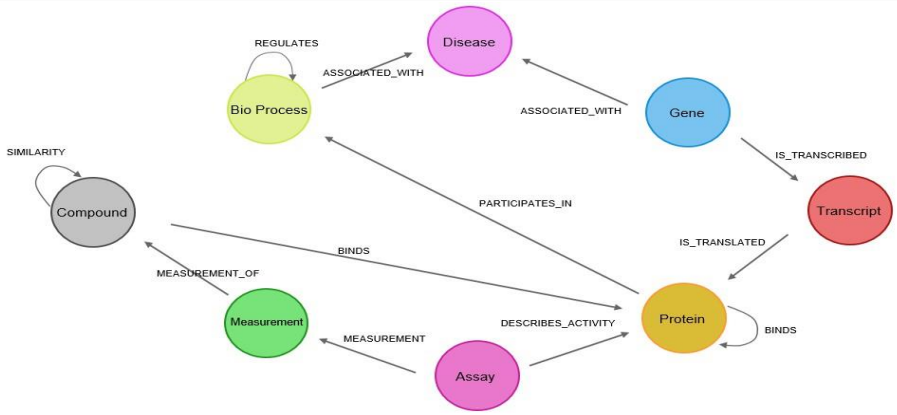



**Figure 1**: Schema of the knowledge graph used in this work.

The resulting resource therefore displays a rich connectivity between key entities relevant to the drug discovery process. The central hypothesis of this work is that this connectivity encodes consistent (and as yet unknown) patterns that allow the inference of untested compounds' activity in assays of interest.

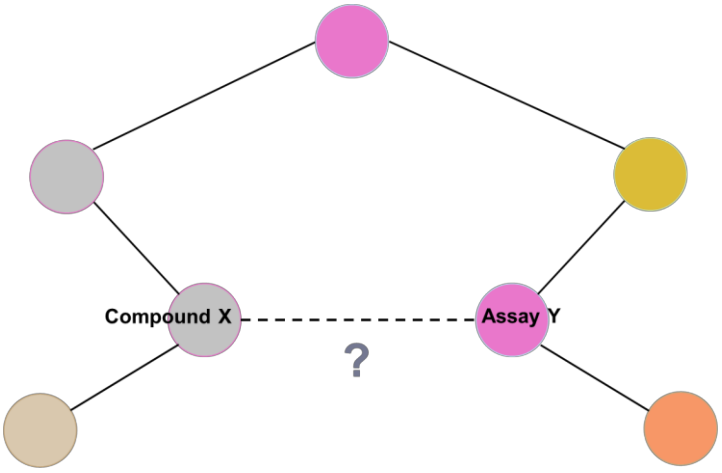



**Figure 2**: Framing the prediction of compound activity in an assay as knowledge graph completion. Node colours code for different classes of entities in the knowledge graph.

Whereas much, if not most, of the recent literature on biomedical knowledge graph completion utilises deep graph embedding models, this poster will highlight an approach to the task based on traversing the observable graph. Tackling the task in this manner serves to both build familiarity with the underlying data as well as achieve inference model transparency and explicability – properties that are of great value in drug discovery. Inspired by previous research with close links to logical inference (Lao et al 2011, Mitchell and Gardner 2015), the approach used in this work leverages observable, rather than latent, knowledge graph topological properties to enable the inference of compound activity in a set of kinase assays. The method relies on the characteristics of the paths within the knowledge graph between a given candidate compound and a target assay to predict the likelihood of a direct connection between the two, which would signify that the compound would demonstrate activity in the assay if tested in a laboratory experiment (Figure 2). Follow-on work may further investigate the suitability of knowledge graph Horn rule mining approaches as detailed by Galárraga et al (2013/2015).
It is intended that the lessons from the exploration summarised in the poster will inform the development of methodologies that combine the transparency of graph traversal-based techniques with the learning potential of deep embedding techniques later on in the first author's PhD project.

1. Lao, N., Mitchell, T., & Cohen, W. (2011, July). Random walk inference and learning in a large scale knowledge base. In Proceedings of the 2011 conference on empirical methods in natural language processing (pp. 529-539).
2. Gardner, M., & Mitchell, T. (2015, September). Efficient and expressive knowledge base completion using subgraph feature extraction. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (pp. 1488-1498).
3. Galárraga, L. A., Teflioudi, C., Hose, K., & Suchanek, F. (2013, May). AMIE: association rule mining under incomplete evidence in ontological knowledge bases. In Proceedings of the 22nd international conference on World Wide Web (pp. 413-422).

4. Galárraga, L., Teflioudi, C., Hose, K., & Suchanek, F. M. (2015). Fast rule mining in ontological knowledge bases with AMIE+. The VLDB Journal, 24(6), 707-730.
5. Horn, A. (1951). On sentences which are true of direct unions of algebras. The Journal of Symbolic Logic, 16(1), 14-21.

## P-14: TERP: a machine learning approach for predicting and prioritizing specialized metabolite tailoring enzyme products

D. Meijer[1], M. Medema[1], J. van der Hooft[1]

[1] *Bioinformatics Group, Wageningen University, The Netherlands*

Biosynthetic gene clusters (BGCs) are local clusters of genes that encode for enzymatic machinery that produce a secondary metabolite. These metabolites are of high scientific and commercial interest because of their potent bioactive properties, for example as antibiotics [1]. Although genome mining can readily identify novel BGCs from (meta)genomic data [2], prediction of the final product is often indecisive. Compound scaffold elucidation of non-ribosomal peptide (NRP), polyketide type 1 (PK1), and NRP-PK1 hybrid natural products from their respective BGCs can be inferred through the use of rule-based mechanisms due to their modular build-up. However, from the BGC coding sequences alone it is not always possible to directly infer the regioselectivity of the encoded tailoring enzymes (e.g., for methylation, glycosylation, or halogenation). Existing tools solve this by creating permutation libraries of the possible product space, resulting in a diverse set of putative but not always plausible products [3].

In this work we will present a novel data-driven approach called TERP (Tailoring Enzymes Regioselectivity Predictor) for prioritizing tailored BGC products. TERP uses tiny graph neural networks with a custom node-labeling strategy in order to perform edge prediction between a tailoring moiety and the scaffold structure. Based on domain presence of known tailoring enzymes in the mined BGCs, multiple tailoring reactions can be chained until no valid edges are predicted and a final tailored product is emitted. TERP mutates simplified heterogeneous graphs of molecular structures based on mined substructures from a chemical compound class-based subset of known natural product chemical space. We use data augmentation in the form of simplified heterogeneous graphs to improve model training in low data regimes.

As bioactivity predictors need complete chemical structures for making predictions, prioritizing tailored BGC products significantly aids in coupling putative BGCs to the most likely bioactivities of its compounds. Not only will this help in isolating BGCs coding for the biosynthesis of compounds with specific pharmaceutical properties, it will also help with inferring microbe-microbe and microbe-plant interactions from genomic sequence data alone.

1. Medema, M. H., de Rond, T., & Moore, B. S. (2021). Mining genomes to illuminate the specialized chemistry of life. In *Nature Reviews Genetics* (Vol. 22, Issue 9, pp. 553-571). Nature Research. https://doi.org/10.1038/s41576-021-00363-7
2. Blin, K., Shaw, S., Kloosterman, A. M., Charlop-Powers, Z., van Wezel, G. P., Medema, M. H., & Weber, T. (2021). AntiSMASH 6.0: Improving cluster detection and comparison capabilities. *Nucleic Acids Research*, *49*(W1),
3. W29-W35. https://doi.org/10.1093/nar/gkab335
4. Skinnider, M. A., Johnston, C. W., Gunabalasingam, M., Merwin, N. J., Kieliszek, A. M., MacLellan, R. J., Li, H., Ranieri, M. R. M., Webster, A. L. H., Cao, M. P. T., Pfeifle, A., Spencer, N., To, Q. H., Wallace, D. P., Dejong, C. A., & Magarvey, N. A. (2020). Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-19986-1

## P-16: Enzeptional: enzyme optimization via a generative language modeling-based evolutionary algorithm

Y.G. Nana Teukam[1,2], M. Manica[1], F. Grisoni[2], T. Laino[1]

[1]*IBM Research Europe, CH-8803, Rüschlikon, Switzerland*
[2]*Eindhoven University of Technology, Institute for Complex Molecular Systems and Dept. Biomedical Engineering, Eindhoven, Netherlands.*

Enzymes are molecular machines optimized by nature to allow otherwise impossible chemical processes to occur. Besides the increased reaction rates, they present remarkable characteristics to enable more sustainable reactions: mild conditions, less toxic solvents, and reduced waste. Billion years of evolution have made enzymes extremely efficient. However wide adoption in industrial processes requires faster design using in-silico methodologies, a daunting task far from being solved. The majority of methods operate by introducing mutations in an existing amino acid (AA) sequence using a variety of assumptions and strategies to introduce variants in the original sequence. More recently, machine learning and deep generative networks have gained popularity in the field of protein engineering by leveraging prior knowledge on protein binders, their physicochemical properties, or the 3D structure. Here, we cast the problem of enzyme optimization as an evolutionary algorithm where mutations are modeled via generative language modeling. Relying on pretrained language models trained on AA sequences, we apply transfer learning and train a scoring model on a dataset of biocatalysed chemical reactions that is used to drive the optimization process. Our methodology allows designing enzymes with higher biocatalytic activity, emulating the evolutionary process occurring in nature by sampling optimal sequences modeling the underlying proteomic language.

1. Chapman, Jordan Ismail, Ahmed Dinu, Cerasela. (2018). Industrial Applications of Enzymes: Recent Advances, Techniques, and Outlooks. Catalysts. 8. 238. 10.3390/catal8060238.
2. Bloom, Jesse Labthavikul, Sy Otey, Christopher Arnold, Frances. (2006). Protein stability promotes evolvability. Proceedings of the National Academy of Sciences of the United States of America. 103. 5869- 74. 10.1073/pnas.0510098103.
3. Poole, Alan Ranganathan, Rama. (2006). Knowledge-based potentials in protein design. Current opinion in structural biology. 16. 508-13. 10.1016/j.sbi.2006.06.013
4. Baek, Minkyung et al. (2021). Accurate prediction of protein structures and interactions using a three track neural network. Science. Vol 373, Issue 6557, pp. 871-876
5. Schwaller, Philippe Hoover, Benjamin Reymond, Jean-Louis Strobelt, Hendrik Laino, Teodoro. (2021). Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. Science Advances. 7. eabe4166. 10.1126/sciadv.abe4166
6. Probst D, Manica M, Nana Teukam YG, Castrogiovanni A, Paratore F, Laino T. Biocatalysed synthesis planning using data-driven learning. Nat Commun. 2022 Feb 18;13(1):964. doi: 10.1038/s41467-022-28536-w. PMID: 35181654; PMCID: PMC8857209.

## P-18: Algorithmic Advances in Diverse Fingerprint Selection

A. Dalke

*Andrew Dalke Scientific, Storgatan 50, Trollhättan, Sweden*

MaxMin and sphere exclusion methods are used to select diverse subsets from a chemical database[1]. These methods may take minutes or hours to select a diverse subset of 10 million binary fingerprints. Recent work in chemfp applies experience from high-performance similarity search to triple the performance of these methods, relative to the widely-used RDKit implementations, when using Tanimoto similarity.

In particular, BitBound search ordering[2] plays a vital role in improving MaxMin performance by an order of magnitude when selecting diverse candidates which must also be diverse from a set of references (eg, to select compounds from vendor catalog which are also dissimilar from an in-house collection).

MaxMin is a fast approximate solution to selecting the most diverse subset. It requires an initial pick, often chosen at random or with one of several heuristics. Chemfp's novel "heapsweep" algorithm gives an exact solution. It combines the heap-based approach of its MaxMin implementation with the sweep family of algorithms[3] to iteratively pick the globally most diverse fingerprints. While significantly slower than MaxMin, heapsweep identifies the most diverse fingerprint from the 2.1 million of ChEMBL 29 in about 6 seconds, which is fast enough to use heapsweep to identify the first pick for MaxMin.

The chemfp implementations are single-threaded. Further optimization and parallelism may improve the performance several-fold. The presentation will end with a discussion of ongoing research to scale diversity selection to larger data sets using distributed computing, and present a novel approach to diversity selection using set coverage.

1. Snarey, M., et al., Comparison of algorithms for dissimilarity-based compound selection. Journal of Molecular Graphics and Modelling, **1997**, 15, 6, pp372-385
2. Swamidass, S. J., et al., Bounds and Algorithms for Fast Exact Searches of Chemical Fingerprints in Linear and Sublinear Time. J. Chem. Inf. Model., **2007**, *47*, 2, 302–317

3. Borassi, M., et al., Fast diameter and radius BFS-based computation in (weakly connected) real-world graphs: With an application to the six degrees of separation games. Theoretical Computer Science, **2015**, 586, 59–80

## P-20: Human Pharmacokinetic Prediction using Predicted Animal Pharmacokinetic Parameters and Computed Physicochemical Properties

S. Seal[1], K. Handa[1,2], A. Bender A[1]

[1] *Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK*

[2] *Toxicology & DMPK Research Department, Teijin Institute for Bio-medical Research, Teijin Pharma Limited, 4-3-2 Asahigaoka, Hino-shi, Tokyo 191-8512, Japan*

Understanding human pharmacokinetics (PK) parameters that affect the blood concentration profile of a drug, such as the steady-state volume of distribution ($VD_{ss}$), and total body clearance (CL), is critical in clinical trials. While animal PK parameters are considered most predictive for modelling human PK parameters, only limited animal PK data is available in the public domain. In this work, we integrated a combination of observed and predicted animal PK data to model the human PK parameters $VD_{ss}$, and CL for 1,335 unique compounds in a two-step process.

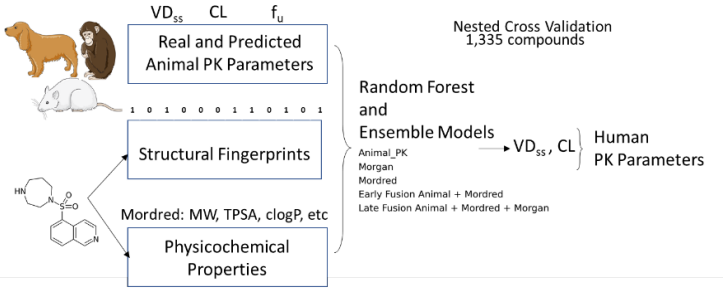



**Figure 1**: Human Pharmacokinetic Prediction using Predicted Animal Pharmacokinetic Parameters and Computed Physicochemical Properties

Firstly, we were using Animal PK parameters [$VD_{ss}$, CL, fraction unbound in plasma (fu)] of rats, dogs, and monkeys for 399 unique compounds where this data was available directly. Using a Random Forest algorithm and Mordred physicochemical descriptors, we predicted animal PK parameters for compounds where animal PK data was unavailable. Secondly, we used Morgan fingerprints, Mordred descriptors and either available or predicted animal PK parameters in a Random Forest algorithm to predict human CL and $VD_{ss}$, where the model was validated using repeated nested cross-validation. We found that human $VD_{ss}$ was best predicted using only Mordred descriptors ($R^2$=0.53, Geometric Mean Fold Error, GMFE=2.13). However, for human CL, higher predictive accuracies of ($R^2$=0.31 and GMFE=2.53) were observed for models combining Morgan fingerprints, Mordred descriptors and Animal PK parameters while using these features separately, the R2 dropped by 19.2%, 10.7% and 29.2% respectively. For best performing models in the prediction of human $VD_{ss}$, 57% compounds and for human CL, 50% compounds were within a 2-fold geometric mean-fold error of the experimental values. These results suggest that $VD_{ss}$, where mechanisms are based on drug binding with tissue components, can be suitably predicted using only physicochemical properties, while all feature spaces are required by the best performing model to predict CL, where complex mechanisms such as metabolism and excretion via multiple pathways are involved. Although not directly comparable due to different datasets and validation methods, our results perform better than previous models which used features such as rat PK data. [1,2] In conclusion, integrating animal PK features from across a range of species can be used for fit-for-purpose and improved PK prediction in drug discovery.

1. Miljković F., et al., Machine Learning Models for Human In Vivo Pharmacokinetic Parameters with In-House Validation. Mol Pharm. 2021; 18(12)
2. Iwata H., et al., Prediction of Total Drug Clearance in Humans Using Animal Data: Proposal of a Multimodal Learning Method Based on Deep Learning. J Pharm Sci. 2021; 110 (4)

## P-22: Prediction of new active ligands for the Vitamin D Receptor

M.I. Agea[1], W. Dehaen[1], M. Šícho[1], D. Sedlák[2], J. Kolla[2], J. Kirchmair[3], D. Svozil[1,2]

*[1]Department of Informatics and Chemistry & CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Faculty of Chemical Technology, University of Chemistry and Technology, Prague, Czech Republic.*
*[2]CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Institute of Molecular Genetics, AS CR v.v.i., Prague, Czech Republic.*
*[3]Universität Hamburg, Center for Bioinformatics (ZBH), Hamburg, Germany; Department of Pharmaceutical Sciences, Division of Pharmaceutical Chemistry, Faculty of Life Sciences, University of Vienna, Vienna, Austria.*

Vitamin D receptor (VDR), a well-known nuclear receptor, is involved in the regulation of calcium homeostasis and cancer chemoprevention. Many researchers have tried and still try to find new VDR active ligands, both computationally and experimentally, with no luck. In this work, we have developed a virtual screening workflow based on a VDR pharmacophore and a Exponential Consensus Ranking (ECR)[1] docking approach. The pharmacophore is very useful to assure important ligand-protein interactions, which for VDR are mainly three H bonds, and it acts as a huge initial filter. The ECR consensus docking has demonstrated to improve individual docking results in terms of AUC, accuracy and Enrichment Factor (EF). This workflow has been applied to two libraries. The first one, our own VDR morphing library enriched by active compounds generated by Molpher[2]. The second one, the 'in-stock' ZINC database. We have discovered new VDR ligands while evidencing, once more, Molpher's capability to explore the active chemical space of a particular target.

1. Palacio-Rodríguez, K., et al., Exponential consensus ranking improves the outcome in docking and receptor ensemble docking. Scientific Reports. **2019,** 9, 1, 1-14.
2. Hoksza, D., et al., Molpher: a software framework for systematic chemical space exploration. Journal of Cheminformatics., **2014**, 6, 7, 1-13.

## P-24: Reaction InChI: Present and Future

[1]G. Blanke, [2]G. Grethe, [3]G. Gygli,[4] H. Kraut,[5] I. Öri,[6] J.H. Jensen,[7] J.M. Goodman,[8] N. Davis

*[1]StructurePendium Technologies GmbH, Essen, Germany, [2]Poway, CA 92064, US, [3]Rheinfelden, Schweiz, [4]InfoChem Gesellschaft für chemische Information mbH, München, Germany, [5]ChemAxon Kft., Budapest, Hungary, [6]Biochemfusion ApS, Ølsted, Denmark, [7]University of Cambridge, Department of Chemistry, Cambridge, CB2 1EW, UK, [8]California Section, American Chemical Society, US*

Since its first release in 2017, the International Chemical Identifier for Reactions (Reaction InChI, RInChI) provides a vendor-neutral, machine-readable string identifying chemical reactions. RInChI is used in databases and cheminformatics software packages like drawing tools that provide the calculations of RInChIs from reaction depictions, publishers are integrating the RInChI into their applications.
So, what is new with the next release: In the upcoming release additional auxiliary layers will be introduced to the RInChI that include the ability to assign atom-atom mapping information and chemical process details, such as reaction temperature and yield. The introduction of No-structures identifiers will allow the description of enzyme reactions as well as any other reaction component that cannot be represented by a unique chemical structure, such as natural products or biologics. Last but not least, programming of the next version will be moved to GitHub to achieve a better participation by the user community for the further development.

## P-28: Building classifiers to link hepatic transcriptomic profile in humans with varying degree of hepatic fibrosis

M. A. González Hernández[1], L. Verschuren[3], M.C. Morrison[2], J. Venhorst[2], M. Wioleta[4], B. Coornaert[4], S. Wink[4], R. Hanemaaijer[2], G.J.P. van Westen[1]

*[1]Drug Discovery and Safety, Leiden Academic Centre for Drug Research, Einsteinweg 55, Leiden, The Netherlands*
*[2] Department of Metabolic Health Research, The Netherlands Organization for Applied Scientific Research (TNO), 2333 CK Leiden, The Netherlands*

[3] *Department of Microbiology and Systems Biology, TNO, 3704 HE Zeist, The Netherlands*
[4] *Galapagos Leiden, Zernikedreef 16, 2333 CL Leiden, The Netherlands*

Hepatic fibrosis can develop in response to various etiologies, including metabolic dysregulation, ethanol consumption and viral infections. Here, we focus on metabolic dysregulation that results from a high fat diet intake and increases the prevalence of non-alcoholic fatty liver disease (NAFLD) in metabolic pathologies such as obesity and metabolic syndrome. Of importance, liver fibrosis is one of the leading causes of hepatocellular carcinoma and cirrhosis. Liver fibrosis can be understood as a complex pathology that is characterized by high deposition of extracellular matrix components such as type 1 collagen that can increase tissue rigidity and alter metabolic processes [1]. Currently, there is a lack of understanding in the heterogeneity in underlying disease mechanisms that drive fibrosis progression *in vivo* as well as *in vitro* hepatic models that recapitulate *in vivo* complexity.

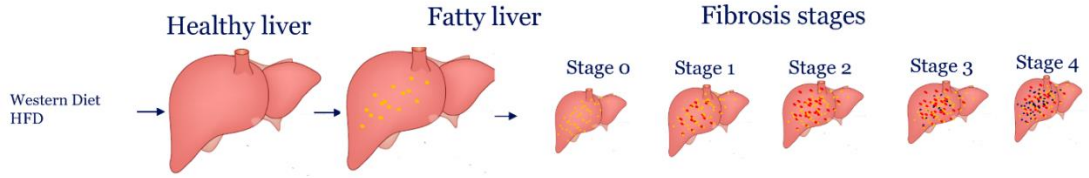



***Figure 1**: The progression of liver disease*

Therefore, we focused on publicly available datasets from human hepatic transcriptomics data in the GEO repository. Our objective is to identify patient-specific patterns in the gene expression profile that characterize subgroups of patients based on clusters of genes and/or specific pathways as well as the relationship of the patient-specific patterns with disease pathology. Transcriptomics data came from various technologies including RNA-seq, microarray platforms as well as scRNAseq. First, we employed various techniques (t-SNE, PCA, MDS) for dimensionality reduction and visualization. Secondly, through unsupervised learning methods, we investigated the structure of the data to investigate subgroups of individuals and their relevant regulatory genes and disease pathway patterns. Furthermore, we fit various machine learning classification algorithms (support vector machines, random forest, logistic regression and k-nearest neighbors, XGBoost) and artificial neural networks algorithms to investigate the gene signature prediction of fibrosis stages ($0-4$).The models were validated using a 5-fold cross validation approach and will be discussed on their performance in independent datasets.
Of relevance, this approach can provide biological insight into the regulatory genes and pathways that can be taken into consideration to develop hepatic *in vitro* models for fibrosis and in the future test various compounds for drug discovery.

1. Brenner, David et al, Molecular and cellular mechanisms of liver fibrosis and its regression. Nature reviews Gastroenterology and Hepatology, **2021,** volume 18, 18, 151-166
2. Arun J. Sanyal et al., Gene Expression predicts histological severity and reveals distinct molecular profiles of nonalcoholic fatty liver disease. Scientific reports. **2019**, volume 9, issue, pages

## P-30: An automated workflow to expand AOP-Wiki Stressor chemical knowledge and identify potential activators of Adverse Outcome Pathways

M. Martens[1], C.T. Evelo[1, 2], E.L. Willighagen[1]

[1]*Department of Bioinformatics - BiGCaT, NUTRIM, Maastricht University, Maastricht, The Netherlands*
[2]*Maastricht Centre for Systems Biology (MaCSBio), Maastricht University, Maastricht, The Netherlands*

The purpose of Adverse Outcome Pathways (AOPs) is to organize mechanistic knowledge on toxicological processes upon exposure to a (chemical) stressor leading to an Adverse Outcome through a series of Key Events (KEs) by the activation of Molecular Initiating Events (MIEs). [1] The implementation of this concept for risk assessments is aimed to facilitate the replacement, reduction, refinement (3Rs) of animal testing. [2] Qualitative descriptions of AOPs are generally stored in the public AOP-Wiki. With the recent development of the AOP-Wiki RDF, its contents were made more FAIR by providing the data in RDF and allowing computational access. [3.4] The majority of stressors describe chemicals, though the AOP-Wiki lacks

chemical-related knowledge and capabilities to identify potential activators of MIEs. Our goal was to expand the current knowledge on chemicals that exists in the AOP-Wiki and identify additional potential activators of MIEs based on structural and functional similarity.
This was done by developing a Jupyter notebook that initiates with the extraction of AOP-Wiki content using the AOP-Wiki RDF through executing SPARQL queries (see Figure 1) with the SPARQLwrapper Python library. This was followed by the generation of chemical compound identifiers using the Chemistry Development Kit (CDK) Python Wrapper [5], and BridgeDb [6] for a range of alternative chemical database identifiers. Furthermore, visualisations were generated, chemical characteristics were extracted from Wikidata using SPARQL, and functionally and structurally similar chemicals were identified using CDK. Overall, this workflow not only extends Stressor chemical knowledge from AOP-Wiki but also provides potential activators of MIEs.

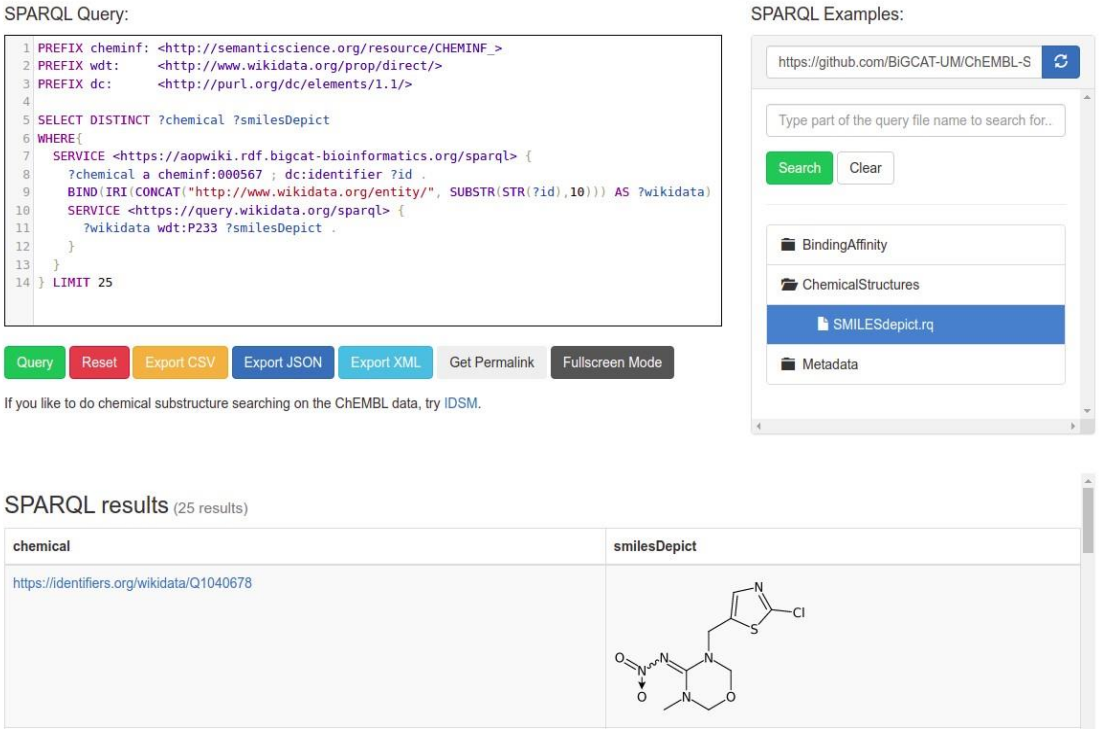



**Figure 1**: Using a SPARQL query to extract chemicals and chemical structures from the AOP-Wiki


1. Ankley, G.T., et al., Adverse outcome pathways: A conceptual framework to support ecotoxicology research and risk assessment. Environmental Toxicology and Chemistry, **2010**, 29, 730–741
2. Burden, N., et al., Aligning the 3Rs with new paradigms in the safety assessment of chemicals. Toxicology, **2015**, 330, 62-66
3. Martens, M., et al., Providing Adverse Outcome Pathways from the AOP-Wiki in a Semantic Web Format to Increase Usability and Accessibility of the Content. Applied In Vitro Toxicology, **2022**
4. Wilkinson, M., et al., he FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, **2016**, 3, 160018
5. Willighagen, E.L., et al., The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. Journal of Cheminformatics, **2017**, 9


## P-32: Exploring aspartic protease inhibitor binding to design selective antimalarials


R. Bobrovs[1], E. E. Basens[1], L. Drunka[1], I. Kanepe[1], S. Matisone[1], K.K. Velins[1], V. Andrianov[1], G. Leitis[1], D. Zelencova-Gopejenko[1], D. Rasina[1], A. Jirgensons[1], K. Jaudzems[1,2]

[1] *Latvian Institute of Organic Synthesis, Aizkraukles iela 21, Riga, Latvia*

[2] *Faculty of Chemistry, University of Latvia, Jelgavas iela 1, Riga, Latvia*


Selectivity is a major issue in the development of drugs targeting pathogen aspartic proteases. Here we explore the selectivity determining factors by studying specifically designed malaria aspartic protease (plasmepsin)

open-flap inhibitors. 2-Aminoquinazolin-4(3H)-one based plasmepsin inhibitors with various flap pocket substituents are synthesized and their potencies against several aspartic proteases are determined. Metadynamics simulations are used to uncover the complex binding/unbinding pathways of these inhibitors, and describe the critical transition states in atomistic resolution. Our findings demonstrate that plasmepsin inhibitor selectivity can be achieved by targeting the flap loop with hydrophobic substituents that enable ligand binding under the flap loop, as such behaviour is not observed for several other aspartic proteases. The ability to estimate compound selectivity before they are synthesized is of great importance in drug design, therefore, we expect that our approach will be useful in selective inhibitor design not only against aspartic proteases, but other enzyme classes as well.

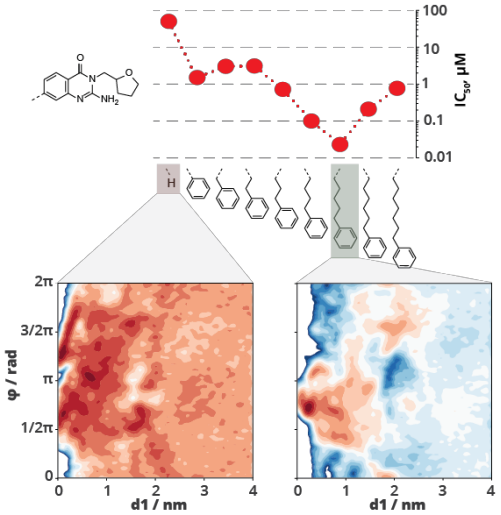



**Figure 1**: Binding free energy surfaces of 2-aminoquinazolin-4(3H)-ones designed, synthesized and enzymatically tested for potency against plasmepsin II

This work was supported by the Latvian Council of Science, project No. lzp-2020/2-0012. RB acknowledges European Regional Development Fund project No. 1.1.1.2/VIAA/2/18/379 for financial support.

## P-34: Proteochemometric modeling identifies chemically diverse norepinephrine transporter inhibitors

B.J. Bongers[1]†, H.J. Sijben [1]†, P.B.R. Hartog[1], A.P. IJzerman[1], L.H. Heitman[1], G.J.P van Westen[1].

[1] *Division of Drug Discovery and Safety, Leiden Academic Centre for Drug Research, Leiden University, The Netherlands*

† *These authors contributed equally*

Solute carriers (SLCs) are a divergent class of transporters and compared to some of the other major receptor families, such as kinases and G protein-coupled receptors (GPCRs), they are understudied[1]. Yet SLCs can play a critical role in complex diseases and as such several SLCs are currently regarded as drug targets[2–4]. From a drug discovery perspective, it is challenging to design family-wide studies to find new ligands that interact with SLCs. Instead, the focus lies on single subfamilies, or even a single SLC, to identify novel compounds.

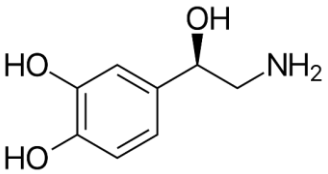



**Figure 1**: The molecular structure of Norepinephrine

The norepinephrine transporter (NET / SLC6A2) is involved in the rapid re-uptake of the neurotransmitter norepinephrine (NE) from the synaptic clefts of noradrenergic neurons in the peripheral and central nervous system[5]. As one of the most well-characterized transporters, NET is an established drug target for depression, chronic pain and narcolepsy, with several marketed drugs available. Despite the abundance of pharmacological data on NET ligand binding, there is a need for the development of novel inhibitors with improved affinity and selectivity over other monoamine transporters[6].

Here, we aimed to find new NET inhibitors using computational modeling. We applied multiple optimization

steps during dataset creation, including similarity networks and stepwise feature selection, to end up with an optimal training set for our model, which was created by using proteochemometrics and stacking of several machine learning techniques. The model was applied to a large virtual database of Enamine, from which 22,000 of the 600 million predicted compounds were clustered to end up with 46 chemically diverse candidates. 32 of these candidates were synthesized and tested using an impedance-based assay. We identified five hit compounds with sub-micromolar inhibitory potencies towards NET, which are promising for follow-up optimizations. This study demonstrates a comprehensive computational pipeline to predict new potential ligands, which could be applied to any protein that has enough interaction data available.


1. César-Razquin, A. *et al.* A Call for Systematic Research on Solute Carriers. *Cell* **162**, 478–487 (2015).
2. Rask-Andersen, M., Masuram, S., Fredriksson, R. & Schiöth, H. B. Solute carriers as drug targets: Current use, clinical trials and prospective. *Molecular Aspects of Medicine* **34**, 702–710 (2013).
3. Girardi, E. *et al.* A widespread role for SLC transmembrane transporters in resistance to cytotoxic drugs. *Nature Chemical Biology 2020 16:4* **16**, 469–478 (2020).
4. Okabe, M. *et al.* Profiling SLCO and SLC22 genes in the NCI-60 cancer cell lines to identify drug uptake transporters. *Molecular Cancer Therapeutics* **7**, 3081–3091 (2008).
5. Bönisch, H. & Brüss, M. The Norepinephrine Transporter in Physiology and Disease. *Handbook of Experimental Pharmacology* **175**, 485–524 (2006).
6. Xue, W. *et al.* Recent Advances and Challenges of the Drugs Acting on Monoamine Transporters. *Current Medicinal Chemistry* **27**, 3830–3876 (2018).


## P-36: Multi-Instance Learning Approach to Predictive Modeling of Catalyst Enantioselectivity


T. Madzhidov[1], D. Zankov[1,2], A. Varnek[2]

[1]*Chemoinformatics and Molecular Modeling, Kazan Federal University, Kazan, 29 Kremvlevskaya, Russia*

[2] *Laboratory of Chemoinformatics, University of Strasbourg, Strasbourg, 4 rue B. Pascal, France*


The production of enantiomerically pure organic compounds is a hot topic in modern organic chemistry. Enantioselective catalysis is a powerful technology for the synthesis of enantiomerically pure compounds using special organic catalysts. Chemoinformatics is an appealing technology aiming to empower experimentalists in the quest for developing new catalysts. Preliminary theoretical research enables the identification of the most promising catalysts before their experimental testing, reducing the time and overheads needed to find an appropriate catalyst.

We developed a new chemoinformatics-based protocol for constructing accurate models for the prediction of catalyst enantioselectivity. The catalysts were represented by multiple conformations, which were encoded with special 3D descriptors developed in our group and probed in predicting the biological activity of molecules. Models were constructed with multi-instance neural networks. Multi-instance (MI) machine learning algorithms can be applied to process the multiple conformations (instances) of a catalyst. In the multi-instance approach, a molecule (catalyst) is presented by a bag of instances (i.e., a set of conformations), and a label (a selectivity value) is available only for a bag (a molecule), but not for individual instances (conformations). The multi-conformation models were compared with single-conformation models constructed with the lowest-energy catalyst conformation.

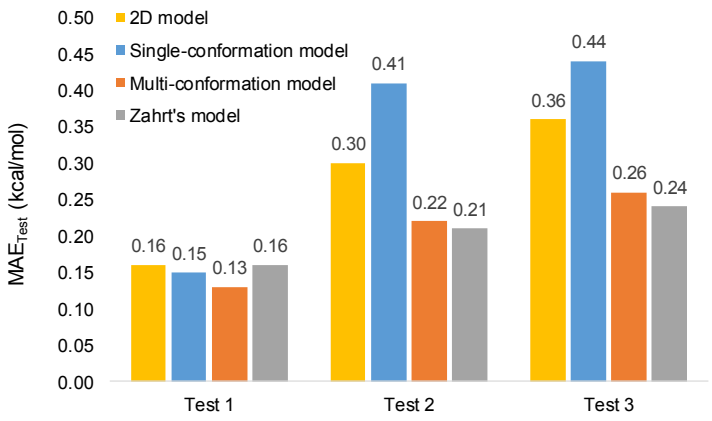



**Figure 1**: Mean Absolute Error (MAE, kcal/mol) obtained for test sets 1-3.

The 2D, single- and multi-conformation models were built on the training set of 384 data points resulting from a combination of 24 catalysts with 16 reactions. The models were validated on three test sets selected according to different scenarios: (a) new reactions with known catalysts, (b) known reactions with new catalysts, and (c) new reactions with new catalysts. Thus, Test set 1 contained 216 instances resulting from a combination of 24 catalysts from the training set with 9 new reactions, Test set 2 included 314 instances (19 new catalysts / 16 training reactions) and Test set 3 contained 171 instances (19 new catalysts / 9 new reactions). Performances of 2D, single-conformation and multi-conformation models (Mean Absolute Error, MAE) in comparison with those of the model by Zahrt et al. [1] are given in Figure 1. These results demonstrate the importance of accounting for all representative catalyst conformations in predictive modeling.

1. Zahrt et al., Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. Science, **2019**, 363, 6424

## P-38: VHP4Safety: building a virtual human for safety assessment

L. Schoenmaker[1], G.J.P. van Westen[1]

[1]*Division of Drug Discovery and Safety, Leiden, The Netherlands*

While animal testing plays a pivotal role in the safety assessment of new chemicals, this traditional "gold standard" has significant shortcomings. A major disadvantage of animal models is that they do not perfectly reflect toxicity in humans [1]. With the growing understanding of the mechanisms of toxicity and the advances in computational modeling, *in silico* methods can play an important role in addressing this issue.
We are establishing the Virtual Human Platform for Safety Assessment (VHP4Safety) in order to reduce the need for animal testing [2]. To this end, our consortium is creating a platform consisting of *in silico* models based on relevant human data. This data takes the form of existing human data, *in vitro* models, and clinical data related to three case studies; kidney disease, neurodegenerative disease, and thyroid-mediated developmental neurotoxicity. These case studies will be used as the basis to build, evaluate and improve the platform. The platform, in turn, consists of a combination of toxicokinetic and toxicodynamic models [3]. Specifically, we are working toward predicting exposure in the body and linking this to toxic effects using adverse outcome pathways [4,5].
In doing so, the VHP4Safety is building a flexible platform and making a first step towards creating a virtual human for safety assessment.
The VHP4Safety project NWA 1292.19.272 is part of the NWA research program 'Research along Routes by Consortia (ORC)' and is funded by the Netherlands Organization for Scientific Research (NWO) and coordinated by Utrecht University, Utrecht University of Applied Sciences, and RIVM.

1. Piersma A., et al., Validation redefined., Toxicology in Vitro, **2018**, 46:163-165. doi:10.1016/j.tiv.2017.10.013
2. https://vhp4safety.nl/
3. Zhang Q., et al., Bridging the Data Gap From in vitro Toxicity Testing to Chemical Safety Assessment Through Computational Modeling., Front Public Health, **2018**, 6. doi:10.3389/fpubh.2018.00261
4. Vamathevan J., et al., Applications of machine learning in drug discovery and development., Nature Reviews Drug Discovery, **2019**, 18(6), 463-477. doi:10.1038/s41573-019-0024-5
5. Pittman, M., et al., AOP-DB: A database resource for the exploration of Adverse Outcome Pathways through integrated association networks., Toxicology And Applied Pharmacology, **2018**. 343, 71-83. doi: 10.1016/j.taap.2018.02.006

## P-40: *In silico* identification of dual targeting potential BACE1 and GSK-3β inhibitors for Alzheimer's disease

N.G. Bajad[1], S. Kumar Singh[1]

[1]*Department of Pharmaceutical Engineering & Technology, Indian Institute of Technology (Banaras Hindu University), Varanasi- 221005, India.*

Alzheimer's disease (AD) is a progressive neurodegenerative disorder characterized by the extracellular deposition of amyloid-β (Aβ) peptides as diffused and neuritic plaques and hyper-phosphorylation of tau (p-tau) protein accumulated intracellularly as neurofibrillary tangles (NFTs). The progression of AD can be slowed down with the designing of disease-modifying therapeutic agents that are supposed to interfere with

the pathogenic steps. The enzymes BACE 1 and GSK-3β are involved in the initiation of Aβ production through the cleavage of the extracellular domain of APP and phosphorylation of various substrates, respectively, leading to cognitive deficiencies in AD. Designing multi-target-directed drugs hitting more than one target against multifactorial diseases, like AD, is one of the worthwhile approaches in drug discovery. Targeting two enzymes, BACE 1 and GSK-3β, involved in distinct pathological conditions with a single inhibitor, could be a conducive approach. The *in silico* approach has been implemented to identify dual targeting inhibitors. Two pharmacophore models were constructed based on the reported potential ligands with co-crystallized protein structure, and common pharmacophore features of both were identified. The pharmacophore models were further used to screen the zinc database to obtain ligands that converge the two pharmacophore models simultaneously. The potential ligands, ZINC043166082 and ZINC225533551 inhibiting both BACE1 and GSK3β, have been identified with the studies including, Structure-based virtual screening, molecular docking, drug-likeness, PAINS filtering, ADME properties prediction, toxicity risk assessment analysis, and molecular dynamics studies. The obtained ligands are expected to be good leads against BACE1 and GSK3β on experimental evaluation.

## P-42: Atomistic insight into substrate activity of SARS-CoV-2 papain-like protease and human casein kinase 1

L. Tesmer[1], A. R. Mehdipour[1,2], G. Hummer[1,3]

[1] *Department of Theoretical Biophysics, Max-Planck Institute of Biophysics, Frankfurt, Germany*
[2] *Center for Molecular Modeling, Ghent University, Ghent, Belgium*
[3] *Institute of Biophysics, Goethe University Frankfurt, Frankfurt, Germany*

We used molecular dynamics (MD) to study enzyme-substrate interactions in two drug targets: the papain-like protease (PLpro) of SARS-CoV-2 [1] and the human casein kinase 1 (CK1) [2-3].

PLpro plays a critical role in coronavirus replication. In addition, PLpro can suppress the innate immune response by preferentially cleaving ISG15 compared with ubiquitin [1]. With MD simulations, we could confirm that PLpro from SARS-CoV-2 interacts more tightly with ISG15. While ISG15 remained bound in three independent MD runs of 3.2 μs, the distal ubiquitin of di-ubiquitin separated from PLpro in four out of six runs on a microsecond timescale. We observed a water-mediated dissociation mechanism for ubiquitin and identified L75T of PLpro as a key mutation distinguishing the earlier SARS-CoV and new SARS-CoV-2 coronavirus, as it weakens the hydrophobic cluster within the binding interface. Hence, pharmacological inhibition of PLpro in SARS-CoV-2 not only blocks viral replication but also simultaneously boosts the antiviral immune response [1].

CK1 regulates a variety of important cellular pathways, including DNA repair. In oocytes after chemotherapy-induced DNA damage, CK1 is associated with activating a cell-death program that leads to infertility in women. A key step in this process is the third phosphorylation of p63, which converts p63 into an active conformation. With MD simulations, we could trace the slow kinetics of this decisive step ─ compared to the other three phosphorylation events ─ to an unusual enzyme-substrate interaction. The simulations identified the stabilizing interactions between CK1 and p63 as persistent salt bridges and tight hydrophobic contacts in a form unfavorable for phospho-transfer. When inhibiting CK1 in mice, the oocytes remained intact, even under the influence of chemotherapeutic agents [2]. Our experimental collaborators found that CK1 can be autophosphorylated [3], resulting in reduced enzyme activity. Using MD simulations, we observed that the phosphorylated form exhibited greater plasticity than the native form. In particular, the integrity of the substrate binding site in the phosphorylated form was altered, which explains the reduced activity [3].

In summary, MD simulations allowed us to investigate enzyme-substrate interaction in full atomic detail. We could contribute to the understanding of disease mechanisms, in particular COVID-19 and infertility in women after chemotherapy.

1. Shin, D., et al., Papain-like protease regulates SARS-CoV-2 viral spread and innate immunity. Nature, **2020**, 587, 657–662.
2. Gebel, J., et al., p63 uses a switch-like mechanism to set the threshold for induction of apoptosis. Nature Chemical Biology, **2020,** 16(10), 1078-1086.
3. Cullati, S, et al., Autophosphorylation of the CK1 kinase domain regulates enzyme activity and substrate specificity. **2022**, submitted.

# P-44: Extracting 3D pharmacophores from molecular dynamics simulations: a case study

S. Pach[1], D. Schaller[2], G. Wolber[1]

[1]*Pharmaceutical and Medicinal Chemistry, Institute of Pharmacy, Freie Universität Berlin, KöniginLuise-Str. 2–4, 14195 Berlin, Germany*

[2] *In Silico Toxicology, Institute of Physiology, Charité Berlin, Charitéplatz 1, 10117 Berlin, Germany*

Flaviviral infections are associated with an increased risk of neurological complications or hemorrhage. The flaviviral serine protease NS2B-NS3 is involved in the processing of a viral polyprotein into functional components of viral particles during replication. Therefore, it represents a promising target to combat flaviviral infections efficiently. Our goal is to develop computational models for the identification of competitive small molecule inhibitors of NS2B-NS3 proteases. Here, we present the first prospective usage of *PyRod*[1], a novel computational method, to identify crucial interaction points in the substrate binding pocket of NS2B-NS3 (Figure 1A)[2]. The interaction points were detected by tracing water molecules in the environment of the flaviviral proteases over the course of molecular dynamics (MD) simulations. The most geometrically favorable and durable water-protein interactions were automatically converted into 3D pharmacophore models, allowing us to perform a virtual screening campaign. The rationally designed small molecule pan-flaviviral protease inhibitors showed inhibitory activity in the low micromolar range[2]. The use of *PyRod* allowed us to overcome two challenging properties of the substrate-binding site of NS2B-NS3: shallowness and high hydrophilicity[3]. In order to rationally characterize the binding modes of our inhibitors and explain observed activity differences, we extracted dynamic 3D pharmacophore patterns (*Dynophores*, Figure 1B)[4] from inhibitor-protease MD simulations. Careful visual inspection of MD trajectories and statistical evaluation of the *Dynophores* allowed us to identify polymorphic mutations in the flaviviral proteases that might explain the activity differences found for our inhibitors.

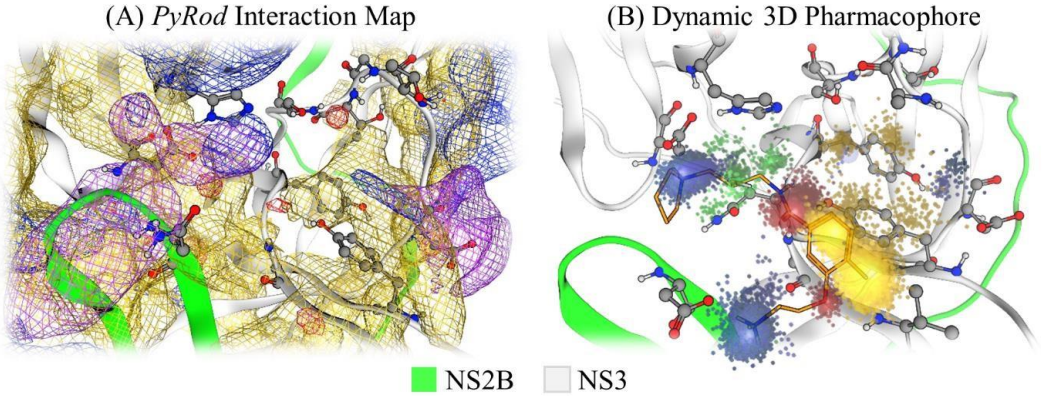



**Figure 1**: The interaction maps obtained from tracing (A) water molecules and (B) inhibitor in NS2B-NS3 protease environment during molecular dynamics simulations.

1. Schaller, D., et al., PyRod: Tracing Water Molecules in Molecular Dynamics Simulations. J. Chem. Inf. Model., **2019**, 59, 2818−2829.
2. Pach, S., et al., Catching a Moving Target: Comparative Modeling of Flaviviral NS2B-NS3 Reveals Small Molecule Zika Protease Inhibitors. ACS Med. Chem. Lett., **2020**, 11, 514-520.
3. Nitsche, C. Proteases from dengue, West Nile and Zika viruses as drug targets. Biophys. Rev., **2019**, 11, 157−165.
4. Bock, A., et al., Ligand Binding Ensembles Determine Graded Agonist Efficacies at a G Protein-Coupled Receptor. J. Biol. Chem., **2016**, 291, 16375-89.

## P-46: The application of the MM/GBSA method in the binding pose prediction of FGFR inhibitors

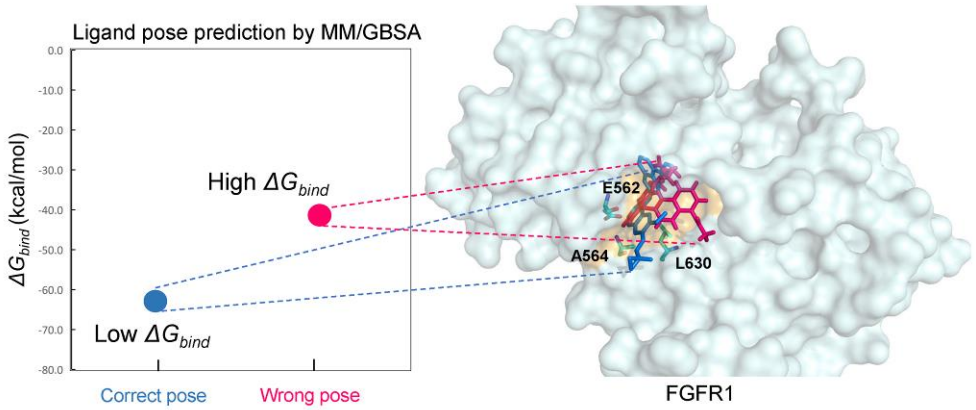
Y. Chen[1], Q. T. Wang[2]

[1]*Institute of Pharmacy, Freie Universität Berlin, Berlin 14195, Germany*

[2]*West China School of Pharmacy, Sichuan University, Chengdu 610041, China*

The success of structure-based drug design, or more specifically lead optimization, is highly dependent on a known binding pose of the protein-ligand system. However, this is not always available to many groups. Therefore, a reliable and cost-effective alternative approach is of great interest. In this work[1], we set out to explore the applicability of the popular and easy-to-use MD-based MM/GBSA[2], method to determine the binding poses. Although this method has been introduced and widely used for a long time, much effort was made to explore its performance to estimate binding affinity between different ligands in previous studies. However, its performance was not satisfactory in this regard. This is why we want to emphasize the application of MM/GBSA in ligand pose prediction, which might be a more appropriate application scenario for MM/GBSA.

This work is trying to answer two major scientific questions: 1) which is the best way to determine the binding pose of a ligand using MD simulation and MM/GBSA calculation; 2) is longer MD simulation useful for pose prediction? And how long would be good enough? Given the amount of known co-crystal structures and the importance of kinases as drug targets, we chose FGFR as an example. A total of 28 ligands of FGFR, including 10 with co-crystal structures, were studied. For each ligand, 2 to 5 poses were generated, and each was simulated for more than 100 ns. It was found that MM/GBSA combined with MD simulation significantly improved the success rate of docking methods from 30-40% to 70%. This work demonstrates a way that the MM/GBSA method can be more accurate in ligand pose prediction than it is in ligand affinity ranking, filling a gap in structure-based drug discovery when the binding pose is unknown.




**Graphic Abstract**. MM/GBSA calculation based on long MD simulations distinguished the correct binding pose from the wrong pose of the ligands. Lower binding free energy (ΔGbind) was predicted for the correct binding pose for each FGFR inhibitor.

**Acknowledgments**
Reproduced from Ref. 1 with permission from the PCCP Owner Societies.

**References**

1. Chen, Y.; Zheng, Y.; Fong, P., *et al*, The application of the MM/GBSA method in the binding pose prediction of FGFR inhibitors. *Physical chemistry chemical physics : PCCP* **2020,** *22* (17), 9656-9663.
2. Kollman, P. A.; Massova, I.; Reyes, C., *et al,* Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Accounts of Chemical Research* **2000,** *33* (12), 889-897.

## P-48: Automated design of synthetically accessible compounds

P. Polishchuk[1], A. Ivanova[1], G. Minibaeva[1], J. Pecha[1], A. Kutlushina[1]

[1]*Institute of Molecular and Translational Medicine, Faculty of Medicine and Dentistry, Palacky University, Hnevotinska 5, Olomouc, Czech Republic*

The chemical space is enormous and estimated to consist of ~$10^{33}$ molecules. Many approaches were developed to rationally navigate within this space. Despite many successes in structure generation using conventional or deep learning approaches, synthetic accessibility of generated molecules remains the major issue. We developed a framework of Chemically Reasonable Mutations (CReM)[1] which naturally solves this issue to the large extent. The basic idea is that fragments occurring in the same local context in existing molecules are interchangeable and their replacements should result in synthetically accessible compounds. We create a database of interchangeable fragments by exhaustive fragmentation of known compounds and annotate each fragment by the corresponding local chemical context of a given radius. Later, we use this database of interchangeable fragments to perform modifications of compound structures to iteratively search for molecules with better properties.

We developed several tools for fully automatic exploration of the chemical space based on the CReM approach. These tools can address different project goals: i) scaffold decoration; ii) enumeration of analogue series; iii) hit/lead optimization; iv) fragment growing within the protein cavity; v) de novo design of diverse sets of promising hits. Some tools can be used for unsupervised enumeration of chemical space (an example is at https://crem.imtm.cz), others perform exploration guided by molecular docking or pharmacophore models. In the latter case we developed an approach to generate structures fitting a given 3D pharmacophore, which has only few analogs. All created tools are part of the CReM Suite for de novo design and structure optimization.

The key feature of these approaches is that synthetic accessibility of generated compounds is controlled indirectly and explicit inclusion of synthetic accessibility to the objective function is not necessary: i) database of fragments can be enumerated from synthetically more accessible molecules that improves synthetic accessibility of generated compounds, ii) choosing the greater radius of considered local context results in more synthetically feasible structures, iii) in-house compound libraries can be fragmented to create a custom fragment database that will result in generation of compounds more synthetically accessible for a particular research group. The estimated coverage of the chemical space depends on the chosen setup, but it is comparable to the analogous spaces enumerated using other approaches (e.g. MASSIV, AZSpace, EVOspace, PGVL, etc).

We demonstrated the ability to generate synthetically accessible compounds on a number of benchmark tasks (including e.g. Guacamol[2]) and also applied the developed tools in the ongoing projects on i) optimization of new tubulin inhibitors, ii) fragment-based design of inhibitors of SARS-CoV2 main protease, iii) de novo design of inhibitors of *Mycobacterium tuberculosis* virulence factor Zmp1 and iv) inhibitors of mycobacterial ATP synthase.

This work is funded by the INTER-EXCELLENCE LTARF18013 project (MEYS), the European Regional Development Fund - Project ENOCH (No. CZ.02.1.01/0.0/0.0/16_019/0000868) and ELIXIR CZ research infrastructure project (MEYS Grant No: LM2018131).

1. Polishchuk, P., CReM: chemically reasonable mutations framework for structure generation. Journal of Cheminformatics, **2020,** 12, 28
2. P. Polishchuk, Control of Synthetic Feasibility of Compounds Generated with CReM, Journal of Chemical Information and Modeling **2020**, 60, 6074-6080.

## P-50: Structure-based generation of synthetically feasible molecules

G. Minibaeva[1], A. Ivanová[1], P. Polishchuk[1]

[1]*Institute of Molecular and Translational Medicine, Faculty of Medicine and Dentistry, Palacký University, Hněvotínská 1333/5, 779 00 Olomouc, Czech Republic*

Identifying molecules that selectively interact with a biological target is a key step in drug discovery. Nowadays, computer-aided molecular design plays an important role in the development of new drugs. In particular, de novo approaches are increasingly used to search for new biologically active molecules. In this case, new compounds with desired pharmacological properties are assembled in the target cavity guided by the general principles of intermolecular interaction. One of the problems of de novo design tools is difficulty to control synthetic feasibility of generated compounds.

In this work, a tool for the design of drug-like compounds inside protein binding sites was developed. This tool includes the use of the CReM method [1] to generate ligand structures and molecular docking by AutoDock Vina [2] to assess their binding to a target protein. The CReM method is intended for generating

new compounds based on interchangeable fragment databases. The main idea of the CReM method is to consider the nearest chemical environment of a fragment when performing a replacement: if the nearest environment (chemical contexts) of two fragments is the same, the fragments are interchangeable.
The developed tool has two modes: i) iteratively growing of a fragment co-crystallized with a protein preserving the position of the parent part of the molecule and ii) de novo compound generation. In the latter case we use a preliminary created set of starting fragments from ChEMBL compounds. Those fragments have from 8 to 15 heavy atoms, from 1 to 5 distinct hydrogen-bond donor/acceptors centers, logP < 2, TPSA > $25A^2$, at most one halogen atom, at most two rotatable bonds and no structural alerts. This starting set of fragments is docked and iteratively grown. We implemented several strategies to select molecules on each iteration: greedy, Pareto or clustering-based selection.
The developed tool was used to grow small ligands co-crystallized with 3C-like protease of SARS-CoV-2 (5RGX, 5RH2) and to de novo generation of ligands form CDK2, dopamine D2 and other targets frequently used in benchmarking studies. During testing of the tool, it was studied how the choice of the following parameters such as fragment databases and context radius affected the structural diversity and synthetic accessibility of the generated compounds. Based on obtained results, we concluded, the synthetic complexity of the generated structures decreases with increasing radius, as well as with using a base of fragments obtained from synthetically more accessible compounds.
This work was funded by the INTER-EXCELLENCE LTARF18013 project (MEYS), the European Regional Development Fund - Project ENOCH (No. CZ.02.1.01/0.0/0.0/16_019/0000868) and ELIXIR CZ research infrastructure project (MEYS Grant No: LM2018131).

1. Polishchuk, P.,CReM :Chemically reasonable mutations framework for structure generation, Journal of Cheminformatics, **2020**, 12, 1, 1-18.
2. Trott, O., Olson, A.J., AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading, Journal of Computational Chemistry, **2010**, 31, 2, 455-461.

## P-52: Automated determination of optimal λ schedules for free energy calculations

H. Loeffler[1], M. Mackey[1]

*Cresset, New Cambridge House, Bassingbourn Road, Litlington, Cambridgeshire, SG8 0SS, UK*

Relative free energy calculations are being increasingly used in drug discovery due to the combination of improved algorithms, faster hardware and better usability. These calculations generally involve the use of alchemical intermediates, where one ligand is transformed into another one in a non-physical way. The choice of exactly how to do this transformation can have a strong effect on the stability and accuracy of the free energy calculation. Traditionally the transformation coordinate λ is either varied linearly between the endpoints, or a sigmoidal function is used to provide extra resolution near the endpoints. However, the choice of which λ values to use is not always clear: if too many values are used then computational resources are wasted, while if too large a gap is left between adjacent λ values accuracy can suffer. In this talk we present a new method for automatically determining optimal λ schedules for a given transformation. A short pre-calculation is used to obtain an estimate of the phase space overlap matrix. Analysis of this matrix can reveal regions where the λ values are too far apart for good convergence, in which case an iterative procedure can be applied to introduce additional λ values and repeat the calculation. The output is a set of optimised λ values which ensure a good phase space overlap between all adjacent λ values while minimising the total number of λ windows that need to be simulated. Validation on standard data sets indicates that the simulation time can be reduced by 30% with no loss of accuracy.

# List of Participants

| *Name* | *Affiliation* | *Country* |
|---|---|---|
| Ms. Maria Isabel Agea Lorente | University of Chemistry and Technology Prague | Czech Republic |
| Dr. Ben Allen | MoA Technology | United Kingdom |
| Mr. Ammar Ammar | Maastricht University | Netherlands |
| Mr. André Asanoski | Technische Universität Braunschweig | Germany |
| Mr. Aniket Suresh Ausekar | Evolvus | India |
| Dr. Jon Christian Baber | Janssen | United Kingdom |
| Mr. Benoit Baillif | University of Cambridge | United Kingdom |
| Dr. Jose Batista | Openeye Scientific Software GmbH | Germany |
| Dr. Maximilian Beckers | Novartis Pharma | Switzerland |
| Prof. Andreas Bender | University of Cambridge | United Kingdom |
| Mr. Olivier Jacques Maurice Béquignon | Leiden University | Netherlands |
| Dr. Marc Bianciotto | Sanofi | France |
| Mr. Michel Blakey | University of Southampton | United Kingdom |
| Dr. Gerd Blanke | StructurePendium Technologies GmbH | Germany |
| Dr. Raitis Bobrovs | Latvian Institute of Organic Synthesis | Latvia |
| Mr. Brandon J. Bongers | Leiden University | Netherlands |
| Dr. Colin Titouan Bournez | Leiden University | France |
| Mr. Guillaume Bret | CNRS | France |
| Dr. Hans Briem | Bayer AG | Germany |
| Dr. Baptiste Canault | GlaxoSmithKline | United Kingdom |
| Mr. Joao Carvalho | VIB | Belgium |
| Dr. Ya Chen | University of Vienna | Austria |
| Ms. Yu Chen | Frei Universität Berlin | Germany |
| Ms. Rachel Cliffe | Cresset | United Kingdom |
| Mr. Andrew Dalke | Andrew Dalke Scientific | Sweden |
| Ms. Deepti Niranjan Dandekar | Evolvus | India |
| Dr. Thomas Doorner | Discovery Informatics Information | Germany |
| Mr. Thomas-Martin Dutschmann | TU Braunschweig | Germany |
| Dr. Lukas Eberlein | OpenEye Scientific | Germany |
| Mr. Terence Egbelo | University of Sheffield | United Kingdom |
| Dr. Christiane Ehrt | Universität Hamburg | Germany |
| Dr. Valentina Eigner-Pitto | CAS - Chemical Abstracts Services | Germany |
| Dr. Thomas Engel | LMU Munich | Germany |
| Dr. Peter Ertl | Novartis | Switzerland |
| Dr. Ricardo J Ferreira | Red Glead Discovery AB | Sweden |
| Mr. David Figueiredo Vidal | Leiden University | Netherlands |
| Dr. Arndt Finkelmann | Syngenta Crop Protection AG | Switzerland |

| *Name* | *Affiliation* | *Country* |
|---|---|---|
| Dr. Mahad Gatti iou | Benevolent AI | United Kingdom |
| Dr. Gian Marco Ghiandoni | AstraZeneca | United Kingdom |
| Prof. Masoud Giahi | Islamic Azad University | Iran |
| Prof. Val Gillet | University of Sheffield | United Kingdom |
| Dr. Daria Goldmann | KNIME GmbH | Germany |
| Dr. Andreas Göller | Bayer AG | Germany |
| Mr. Manuel Alejandro Gonzalez Hernandez | Leiden University | Netherlands |
| Ms. Marina Gorostiola González | Leiden University | Netherlands |
| Mr. Joel Graef | Universität Hamburg | Germany |
| Dr. Michael Green | Desupervised | Denmark |
| Dr. Francesca Grisoni | Eindhoven university of technology | Netherlands |
| Dr. Laura Guasch | F. Hoffmann-La Roche Ltd | Switzerland |
| Dr. Judith Günther | Bayer AG | Germany |
| Ms. Archana Chintamani Hajare | Evolvus | India |
| Dr. Richard John Hall | Astex Pharmaceuticals | United Kingdom |
| Mr. Alan Kai Hassen | Leiden University | Netherlands |
| Mr. Andrew Hall Henry | Chemical Computing Group | United Kingdom |
| Mr. Marc Hoffstedt | Technische Universität Braunschweig | Germany |
| Dr. Zied Hosni | Sheffield University | United Kingdom |
| Ms. Jiahui Huang | University of Vienna | Austria |
| Dr. Lina Humbeck | Boehringer Ingelheim Pharma GmbH & Co. KG | Germany |
| Dr. Wolf Ihlenfeldt | Xemistry GmbH | Germany |
| Ms. Aleksandra Ivanová | Institute of Molecular and Translational Medicine | Czech Republic |
| Dr. Célien Jacquemard | University of Strasbourg | France |
| Dr. Zuzana Jandova | Boehringer Ingelheim RCV GmbH & co KG | Austria |
| Dr. Willem Jespers | Leiden University (LACDR) | Netherlands |
| Dr. Tuomo Sakari Kalliokoski | Orion Pharma | Finland |
| Mr. Alan Kerstjens | University of Antwerp | Belgium |
| Mr. Bola Khalil | Janssen Pharmaceutica AND Leiden University | Belgium |
| Mr. Lennart Kinzel | TU Braunschweig | Germany |
| Prof. Johannes Kirchmair | University of Vienna | Austria |
| Prof. Oliver Koch | University of Münster | Germany |
| Mr. Frans Koeman | KNCV | Netherlands |
| Dr. Oliver Koepler | TIB - Leibniz Information Centre for Science and Technology | Germany |
| Dr. Marcin A. Król | Ryvu Therapeutics | Poland |
| Ms. Rosan Kuin | Universiteit Leiden | Netherlands |
| Ms. Alina Kutlushina | Institut Molecular and Translational Medicine | Czech Republic |

| *Name* | *Affiliation* | *Country* |
|---|---|---|
| Dr. Ingvar Lagerstedt | NextMove Software Limited | United Kingdom |
| Ms. Adelene Lai Shuen Lyn | Friedrich-Schiller-University | Luxembourg |
| Dr. Gregory Landrum | ETH Zurich | Switzerland |
| Dr. Udo Ernst Walter Lange | AbbVie | Germany |
| Mr. Maxime Langevin | Sanofi Aventis R&D | France |
| Dr. David N LeBard | OpenEye Scientific | United States |
| Ms. Sijie Liu | Freie Universität Berlin | Germany |
| Dr. Hannes Loeffler | Cresset | United Kingdom |
| Dr. Sohvi Iiri Maria Luukkonen | Leiden University | Netherlands |
| Mr. Marvin Martens | Maastricht University | Netherlands |
| Ms. Sarah Maskri | University of Muenster | Germany |
| Mr. David Meijer | Wageningen University | Netherlands |
| Ms. Janosch Menke | University of Münster | Germany |
| Mr. James Alexander Middleton | University of Sheffield | United Kingdom |
| Ms. Guzel Minibaeva | Palacký University Olomouc | Czech Republic |
| Dr. Ngai Yi Mok | BenevolentAI | United Kingdom |
| Dr. Wijnand Mooij | Dotmatics Limited | Netherlands |
| Dr. Joerg Muehlbacher | Novartis | Switzerland |
| Dr. Daniel Mulnaes | OpenEye Scientific GmbH | Germany |
| Prof. Eugene Muratov | UNC Chapel Hill | United States |
| Dr. Mikko Muuronen | Janssen Pharmaceutica NV | Belgium |
| Mr. Yves Gaetan Nana Teukam | IBM Research Europe - Zurich | Switzerland |
| Dr. Marc Christian Nicklaus | IUPAC InChI Tautomerism Group | United States |
| Dr. Christos Nicolaou | Recursion Pharmaceuticals | United States |
| Dr. Eva Nittinger | AstraZeneca | Sweden |
| Ms. Theresa Noonan | Freie Universität Berlin | Germany |
| Dr. Noel Michael O'Boyle | Sosei Heptares | United Kingdom |
| Dr. Frank Oellien | AbbVie | Germany |
| Dr. Martin Ott | Lhasa limited | United Kingdom |
| Mr. Szymon Pach | Freie Universität Berlin | Germany |
| Mr. Vincenzo Palmacci | University of Vienna / Bayer AG | Germany |
| Mr. Patrick Penner | Universität Hamburg | Germany |
| Dr. Pavlo Polishchuk | Palacky University | Czech Republic |
| Mr. Savins Puertas Martin | University of Sheffield | United Kingdom |
| Ms. Kristina Sophie Puls | Freie Universität Berlin | Germany |
| Dr. Kohulan Rajan | Friedrich Schiller University | Germany |
| Prof. Matthias Rarey | Universität Hamburg | Germany |

| *Name* | *Affiliation* | *Country* |
|---|---|---|
| Mr. Ben Retamal | Collaborative Drug Discovery | United Kingdom |
| Prof. Sereina Riniker | ETH Zurich | Switzerland |
| Dr. Thorsten Rohde | ACS international | Germany |
| Dr. Thomas Lothar Sander | Idorsia Pharmaceuticals Ltd. | Switzerland |
| Mr. Norbert Sas | KNIME | Germany |
| Ms. Susanne Sauer | Sanofi-Aventis Deutschland GmbH | Germany |
| Dr. Roger Anthony Sayle | NextMove Software | United Kingdom |
| Dr. Delia Sayle | NextMove Software | United Kingdom |
| Dr. Andrius Sazonovas | VsI Aukstieji Algoritmai | Lithuania |
| Dr. Josef Scheiber | BioVariance GmbH | Germany |
| Dr. Peter Schmidtke | Discngine | France |
| Ms. Linde Schoenmaker | Universiteit Leiden | Netherlands |
| Mr. Srijit Seal | University of Cambridge | United Kingdom |
| Dr. Martin Šícho | Leiden University | Netherlands |
| Ms. Denise Slenter | Bioinformatics (BiGCaT), Maastricht University | Netherlands |
| Dr. Christopher Southan | Medicines Discovery Catapult | Sweden |
| Dr. Jess Stacey | MedChemica Ltd | United Kingdom |
| Dr. Francesca Stanzione | Cambridge Crystallographic Data Centre | United Kingdom |
| Prof. Christoph Steinbeck | Friedrich-Schiller-University Jena | Germany |
| Dr. Pieter Stouten | Galapagos | Belgium |
| Ms. Dominique Sydow | Sosei Heptares | United Kingdom |
| Ms. Zsofia Szabo | Chemaxon | Hungary |
| Mr. Valerij Talagayev | FU Berlin, Molecular Drug design | Germany |
| Mr. Hanz Tantiangco | University of Sheffield | United Kingdom |
| Mr. Akos Tarcsay | ChemAxon Kft | Hungary |
| Ms. Barbara Ruth Terlouw | Wageningen University and Research | Netherlands |
| Ms. Laura Tesmer | AbbVie | Germany |
| Mr. Morgan Cole Thomas | University of Cambridge | United Kingdom |
| Dr. Samuel Toba | OpenEye Scientific | United States |
| Ms. Alessandra Toniato | IBM Research Zurich | Switzerland |
| Mr. Giovanni Alessandro Tricarico | Galapagos NV | Belgium |
| Ms. Martyna Trojgo | Desupervised | Denmark |
| Prof. Inbal Tuvi-Arad | The Open University of Israel | Israel |
| Ms. Helle Willemijn Van Den Maagdenberg | Leiden University | Netherlands |
| Mr. Karel Johannes van der Weg | Forschungszentrum Juelich | Germany |
| Mr. Derek van Tilborg | Eindhoven University of Technology | Netherlands |
| Dr. Herman Van Vlijmen | Janssen | Belgium |

| *Name* | *Affiliation* | *Country* |
|---|---|---|
| Prof. Gerard JP van Westen | Leiden University (LACDR) | Netherlands |
| Dr. Mariana Vaschetto | Collaborative Drug Discovery. CDD VAULT | United Kingdom |
| Mr. Vincent Vivien | OpenEye Scientific | France |
| Mr. Modest von Korff | Idorsia Pharmaceuticals Ltd. | Switzerland |
| Dr. Markus Wagener | Grünenthal GmbH | Germany |
| Mr. Moritz Walter | The University of Sheffield | United Kingdom |
| Dr. Wendy Warr | Wendy Warr & Associates | United Kingdom |
| Dr. Samuel Webb | Lhasa Limited | United Kingdom |
| Dr. Henriette Willems | University of Cambridge | United Kingdom |
| Dr. Egon Willighagen | Maastricht University | Netherlands |
| Dr. Sarah Witzke | Chemical Computing Group | United Kingdom |
| Prof. Gerhard Wolber | Freie Universität Berlin | Germany |
| Prof. Achim Zielesny | Westphalian University of Applied Sciences | Germany |

## 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)