ICCS

International Conference
on Chemical Structures

2025 International Conference on Chemical Structures
June 1st – June 5th , 2025 | Noordwijkerhout, The Netherlands

PROGRAM & ABSTRACTS

MOE

MOEsaic

Web Application for SAR and Ligand Analytics

- Structure-Activity and Property Relationships
- Visualize SAR Data and Trends
- Perform Substructure and Similarity Searches
- Profile and Analyze R-Groups
- Analyze Matched Molecular Pairs (MMPs)
- Design Novel Virtual Ligands
- Document Analysis Results and Collect Notes

Chemical Computing Group

MOEsaic is a web-based application for analyzing SAR data, visualizing trends, exploring new virtual leads, and documenting results. Its streamlined interface features interactive MMP analysis and R-group profiling for fast assessment of property cliffs and SAR transferability of fragments, an integrated sketcher for designing novel structures, and models for property prediction.

chemcomp.com

Preface

Welcome to the 13th International Conference on Chemical Structures. 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 165 abstract submissions which were all subjected to a review process carried out by our Scientific Advisory Board of international reviewers from academia and industry. This allowed us to compile an outstanding scientific program of 35 plenary lectures and 92 posters, 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 social event in the beautiful historic city of Leiden.

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. Val Gillet in recognition of her work in the area of Chem(o)informatics at the University of Sheffield (UK) [2]. Dr. Gillet has agreed to give a keynote lecture on Sunday evening before dinner right after the opening of the conference.

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 Egon Willighagen. Papers can be submitted at any date up to the 1st of December 2025, and authors should mention in their cover letter that the manuscript is intended to be included in the 2025 ICCS article collection. More details online at the journal [3]. 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, molecular modeling, and AI in 2025. Note that in the online program ORCID identifiers and Mastodon, Bluesky, and GitHub 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 page [4].

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,
Pieter Stouten, and Jenke Scheen

1. An overview of all ICCS meetings. https://scholia.toolforge.org/event-series/Q47501052
2. Scholarly output of Dr. Val Gillet: https://scholia.toolforge.org/author/Q42717125
3. Journal of Cheminformatics Special issue: https://www.biomedcentral.com/collections/ICCS25
4. 2025 ICCS page : https://scholia.toolforge.org/event/Q133457282

Contents

Organizing Committee

- Gerard JP van Westen, University of Leiden, The Netherlands
- Willem Jespers, University of Groningen, The Netherlands
- Egon Willighagen, Maastricht University, The Netherlands
- Markus Wagener, Grünenthal, Germany
- Jenke Scheen, CHARM Therapeutics, United Kingdom
- Frank Oellien, AbbVie, Germany
- Pieter Stouten, Freelance Consultancy, Belgium

Scientific Advisory Board

- Andreas Bender, University of Cambridge, UK
- Barabara Zdrazil, EMBL-EBI, UK
- Christoph Steinbeck, Friedrich Schiller University, Germany
- Christos Nicolau, Recursion, USA
- Elif Ozkirimli, Roche, Switzerland
- Ester Kellenberger, University of Strasbourg, France
- Herman van Vlijmen, Johnson and Johnson, BE
- Johan Aqvist, Uppsala University, Sweden
- John Overington, Drug Hunter, USA
- Marwin Segler, Microsoft Research, UK
- Matthias Rarey, University of Hamburg, DE
- Ola Engkvist, AstraZeneca, Sweden
- Pat Walters, Relay Therapeutics, USA
- Peter Ertl, Novartis, Switzerland
- Rajarshi Guha, Vertex Pharmaceuticals, USA
- Sereina Riniker, ETH Zurich, Switzerland
- Teodoro Laino, IBM Zurich, Switzerland
- Val Gillet, University of Sheffield, UK
- Woody Sherman, Psivant Therapeutics

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

Chemical Computing Group

www.chemcomp.com

Platinum Sponsors

Chemaxon

www.chemaxon.com

EXSCALATE

www.exscalate.com

Gold Sponsors

CDD VAULT® Complexity Simplified

www.collaborativedrug.com

NextMove Software

www.nextmovesoftware.com

Silver Sponsors

Alipheron

www.alipheron.com

OpenEye | cadence MOLECULAR SCIENCES

www.eyesopen.com

cresset®

www.cresset-group.com

D E Shaw Research

www.deshawresearch.com

FACCTs

www.faccts.de

Schrödinger

www.schrodinger.com

Symposiumsium Bag Sponsor

Dompé

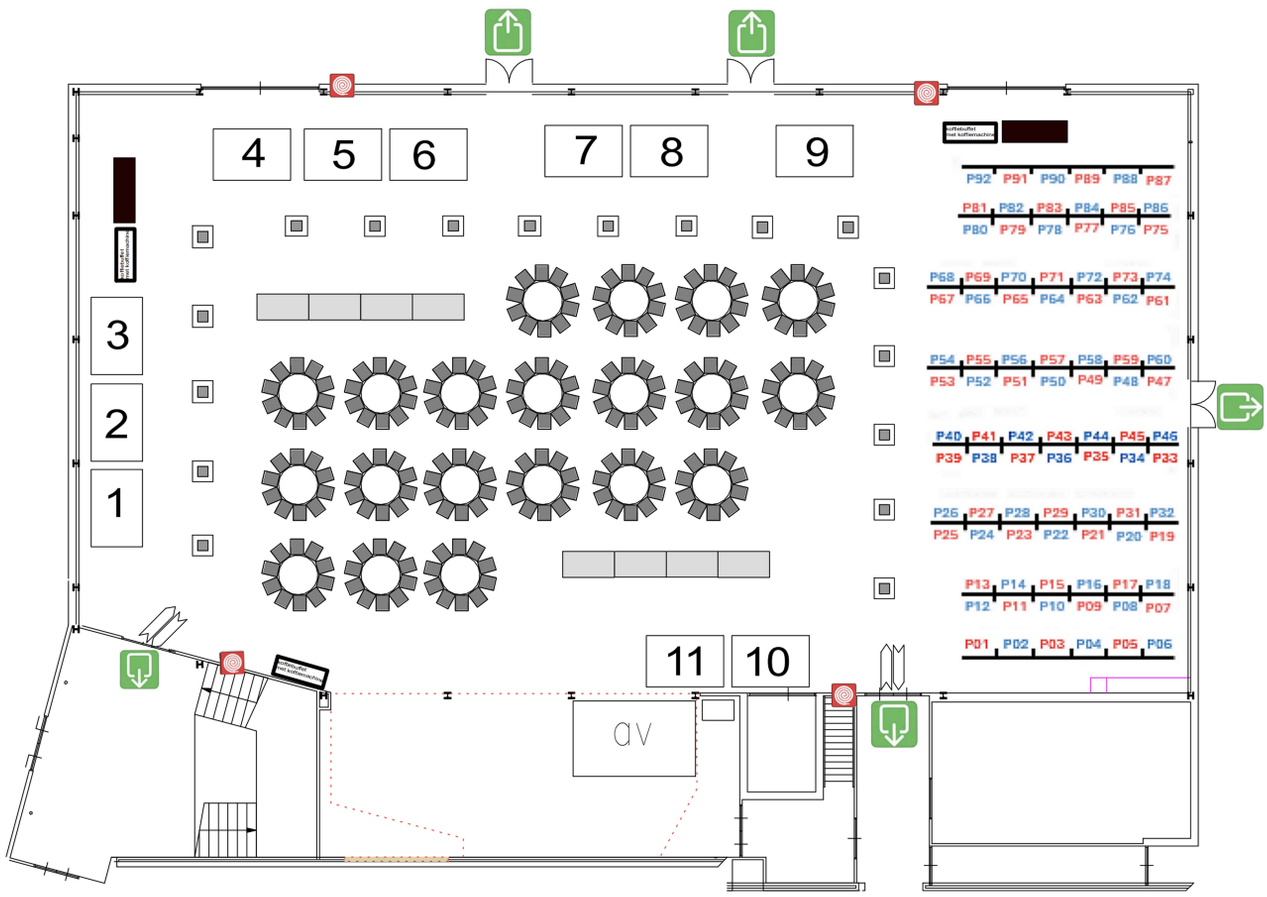
www.dompe.com

Poster Awards Sponsor

evolvus

www.evolvus.com

Exhibition Floor Plan



Exhibitors:

1. Chemical Computing Group
2. Collaborative Drug Discovery
3. D E Shaw Research
4. Cresset
5. Nextmove
6. OpenEye
7. Alipheron
8. Chemaxon
9. FACCTs
10. Qsimulate
11. Simulations Plus

Exposition hours are:

Monday, June 2nd	15.30–19.30
Tuesday, June 3rd	15.30–19.30
Wednesday, June 4th	09.00–13.00

Exhibitors

Chemical Computing Group

www.chemcomp.com

CDD VAULT®
Complexity Simplified

www.collaborativedrug.com

D E Shaw Research

www.deshawresearch.com

cresset®

www.cresset-group.com

NEXTMOVE SOFTWARE

www.nextmovesoftware.com

OpenEye | cadence® MOLECULAR SCIENCES

www.eyesopen.com

Alipheron

www.alipheron.com

Chemaxon

www.chemaxon.com

FACCTs

www.faccts.de

QSIMULATE

www.qsimulate.com

SimulationsPlus

www.simulations-plus.com

Workshops Sunday June 1st | 15:00–17:00 h

Generative design of novel active molecules based on multiple objective criteria

Sarah Witzke and Guido Kirsten, Chemical Computing Group

New lead compounds discovered computationally – or identified by virtual, experimental and/or fragment screening – invariably need to be optimised for activity, ease of synthesis and other pharmacokinetic properties. This workshop explores how computational applications in the Molecular Operating Environment (MOE) software system are used to facilitate both the discovery and optimization processes. These include:

- Using binding pocket information to guide de novo design
- Scaffold Replacement, Fragment Growing and Reaction-based Transformation of starting molecules
- Bioisosteric Replacement
- Multi-objective compound scoring and filtering (descriptors, models, pharmacophores, docking scores)
- Intelligent enumeration of reactions for library design
- Cheminformatics analysis of SAR data sets
- Attendees will therefore gain an awareness of the arsenal of techniques available to address topics in computational drug design.

Additionally, we will demonstrate the exploration and exploitation of chemical space using the PNN based REINVENT approach. The QSAR, pharmacophore, fingerprint and docking scoring functions in MOE can be used in an interface to filter the output of the generative model efficiently. The trained PNN will be used to sample a manageable number of interesting compounds.

Chemaxon Workshop: Machine Learning Assisted Molecular Design with Chemaxon's Python Toolkit and Design Hub

Mark Szabo, Chemaxon

Design Hub is a compound design and tracking platform for drug discovery teams and their external collaborators that connects scientific hypotheses, candidate compound selection, and computational capabilities. With the rise of machine learning (ML) in drug discovery, platforms like Design Hub, combined with ML-driven methods, offer increased value for rational compound design. By leveraging Chemaxon's freshly released native Python libraries, researchers can directly access molecular fingerprints, descriptors, and physico-chemical property calculations, making the design process more data-driven and efficient than ever before.

During this workshop, the participants will have the opportunity to build a machine learning workflow using open-source libraries such as scikit-learn while leveraging Chemaxon's Python libraries for feature generation. A lightweight predictive model will be trained and wrapped as a Python-based plugin, which will be callable from Design Hub. The second part of the workshop will demonstrate how this custom plugin can be integrated into Design Hub. We will build a workflow where structures designed within the platform will be sent to the ML plugin, and the returned predictions — e.g., activity, synthetic feasibility, or other calculated properties — will guide compound prioritization within an interactive design generation workflow.

By the end of the session, participants will have seen a concrete example of how open-source ML tools, powered by Chemaxon descriptors, can be effectively combined with Design Hub's production-grade cheminformatics infrastructure to support real-world molecular design workflows

A hands-on training is possible for participants using laptops and arriving 30 minutes before the workshop starts for installation and setup. Participants without coding experience are welcome too! For more information, please contact Mark Szabo: mszabo21@chemaxon.com

Social Media, Photos, and Code of Conduct

Rules of Engagement

The 2025 ICCS 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. Any form of abuse should be reported at the conference office.

Photos

Participants are allowed to take photos during the meeting, but make sure you have permission to reproduce presented results. 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 Chemische Congressen VI* reserves the right to use any photograph taken by the conference photographer at the 2025 International Conference on Chemical Structures, without the expressed written permission of those included within the photograph. SCC 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.

X / BlueSky / Mastodon

The official hashtag for this meeting is #ICCS2025. 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/xSjv9f56 The same rules of engagement apply online as they apply in person.

Social Event

Schedule

Time	
12.45	Departure by Bus
13.15–13.45	Arrival and Welcome with coffee
13.45–14.00	Briefing and group formation
14.00–16.00	‘Standsganzebord in Leiden’
16.00–17.00	Drinks sponsored by D E Shaw Research
17.00–19.00	Pubquiz
19.00–22.00	Walking dinner
21.00 / 22.00 / 23.00	Busses back

The City of Leiden

Leiden is a city and municipality in the province of South Holland, Netherlands. The municipality of Leiden has a population of 127,046 but the city forms one densely connected agglomeration with its suburbs Oegstgeest, Leiderdorp, Voorschoten and Zoeterwoude with 215,602 inhabitants. Leiden is located on the Oude Rijn, at a distance of some 20 km (12 mi) from The Hague to its south and some 40 km (25 mi) from Amsterdam to its north.

A university city since 1575, Leiden has been one of Europe‘s most prominent scientific centres for more than four centuries. University buildings are scattered throughout the city and the many students from all over the world give the city a bustling, vivid and international atmosphere. Many important scientific discoveries have been made here, giving rise to Leiden‘s motto: ‚City of Discoveries‘. Leiden University is one of Europe‘s top universities, with thirteen Nobel Prize winners. It is a member of the League of European Research Universities and positioned highly in all international academic rankings. It is twinned with Oxford, the location of the United Kingdom‘s oldest university.

Leiden is a city with a rich cultural heritage, not only in science, but also in the arts. The painter Rembrandt was born and educated in Leiden. Other Leiden painters include Lucas van Leyden, Jan van Goyen and Jan Steen.

Sources: https://en.wikipedia.org/wiki/Leiden
https://picryl.com/media/plattegrond-van-leiden-met-stadsgezicht-3c0c2f

Scientific Program

Sunday June 1st

13.00-18.00	**Registration** *Conference Desk*
15.00-17.00	**Pre-conference workshops**
	Generative design of novel active molecules based on multiple objective criteria *Chemical Computing Group*
	Machine Learning Assisted Molecular Design with Chemaxon's Python Toolkit and Design Hub *Chemaxon*
17.00-18.00	**Break**
18.00-18.15	**Welcome** *Rotonde*
18.15-19.00	**Keynote address – CSA Trust Mike Lynch Award** **A Chemoinformatics Journey: An Evolution of Methods** *Awardee: Dr. Val Gillet , University of Sheffield*
19.00-20.00	**Welcome Reception** *Atrium*
20.00-22.00	**Reception Dinner** *Atrium*

Scientific Program

Monday June 2nd

08.30-15.00	**Session A - Artificial Intelligence, Machine Learning, and QSAR** *Rotonde*		
08.30-09.00	A-01	**GENEOnet: Revolutionizing Drug Discovery with the Most Accurate Protein Binding Pocket Detection Using GENEOs** *Carmine Talarico*	*Dompé farmaceutici S.p.A.*
09.00-09.30	A-02	**The future of computational chemistry: AI Target-Ligand Co-Folding?** *Christian Tyrchan*	*AstraZeneca*
09.30-10.00	A-03	**Drugging the undruggable: A highly accurate method for detecting and ranking cryptic pockets** *David LeBard*	*OpenEye, Cadence Molecular Sciences*
10.00-10.30	**Coffee Break** *Atrium*		
10.30-11.00	A-04	**Improving Target-Adverse Event Association Prediction by Mitigating Topological Imbalance in Knowledge Graphs** *Terence Egbelo*	*University of Sheffield*
11.00-11.30	A-05	**Refined ADME Profiles for ATC Drug Classes** *Raquel Parrondo-Pizarro*	*Chemotargets, S.L.*
11.30-12.00	A-06	**ADMET modelling with a quantum chemically pretrained Graphormer. Beyond benchmarking results** *Kostiantyn Chernichenko*	*J&J Innovative Medicine*
12.00-12.30	**Lunch** *Atrium*		
13.00-13.30	A-07	**Towards experiment-aware bioactivity model(er)s** *Linde Schoenmaker*	*Leiden University (LACDR)*
13.30-14.00	A-08	**OpenMMDL - Simplifying the Complex: Building, Simulating, and Analyzing Protein–Ligand Systems in OpenMM** *Valerij Talagayev*	*Freie Universitaet Berlin*
14.00-14.30	A-09	**High-accuracy QM in life sciences: From drug properties to binding modes** *Christoph Riplinger*	*FACCTs GmbH*
14.30-15.00	A-10	**Quantifying the Unknown: A Comparative Study of Deep Learning-Based Uncertainty Quantification Methods for Bioactivity Assessment** *Bola Khalil*	*Leiden University and Johnson&Johnson*
15.00-15.30	**Coffee Break** *Atrium*		
15.30-17.30	**Poster Session RED & Exhibition** *Atrium*		
18.30-19.30	**Reception** *Atrium*		
19.30-21.30	**Dinner** *Atrium*		

Scientific Program

Tuesday June 3rd

08.30-12.00	**Session A - Artificial Intelligence, Machine Learning, and QSAR** *Rotonde*		
08.30-09.00	A-11	**Exploration of Synthesis Space by Application of Evolutionary Strategies** *Emma Sarah Armstrong*	*University of Sheffield*
09.00-09.30	A-12	**Visualization and Clustering of Ultra-Large Chemical Space** *Johannes Kaminski*	*University of Münster*
09.30-10.00	A-13	**CACHE Challenge #1: Searching for Hit Molecules in Ultra-Large Chemical Libraries Guided By De Novo Design** *Pavel Polishchuk*	*Palacky University*
10.00-10.30	**Coffee Break** *Atrium*		
10.30-11.00	A-14	**Discovery of Novel CYP19A1 Inhibitors Using Machine Learning-Driven Virtual Screening and Structure-Based Approaches** *Sijie Liu*	*Freie Universität Berlin*
11.00-11.30	A-15	**SpectruMS: A Multi-modal Foundation Model for Better Generalizability on Tandem MS2 Data** *Aya Abdelbaky*	*Pangea Bio*
11.30-12.00	A-16	**Leveraging institutional data to improve LLM performance** *Valery Tkachenko*	*Science Data Experts*
12.00-12.30	**Lunch** *Atrium*		
13.00-15.00	**Session B - New Modalities and Large Chemical Data Sets** *Rotonde*		
13.00-13.30	B-01	**A Workflow Pairing Rational and Computational PROTACs Design Yields a Novel BRD4 Degrader** *Olga Tarkhanova*	*Chemspace LLC*
13.30-14.00	B-02	**Benchmarking Searching in Combinatorial Spaces with the Approved Drug Space** *Modest von Korff*	*Alipheron AG*
14.00-14.30	B-03	**COCONUT 2.0: A Comprehensive Improved Open Database for Natural Products Research** *Christoph Steinbeck*	*Friedrich-Schiller-University Jena*
14.30-15.00	B-04	**StrAcTable – Combining Structural and Bioactivity Data with Atomic Precision for Protein-Ligand Complex Datasets** *Torben Gutermuth*	*University of Hamburg*
15.00-15.30	**Coffee Break** *Atrium*		
15.30-17.30	**Poster Session BLUE & Exhibition** *Atrium*		
18.30-19.30	**Reception** *Atrium*		
19.30-21.30	**Dinner** *Atrium*		

Scientific Program

Wednesday June 4th

08.30-12.30	**Session C - Advanced Cheminformatics Techniques** *Rotonde*		
08.30-09.00	C-01	**Transformers for molecular property prediction: Domain adaptation efficiently improves performance** *Afnan Sultan*	*Saarland university*
09.00-09.30	C-02	**Navigating Synthon Space: Property-Driven Molecular Optimization for Pharmacokinetics** *Rafał Adam Bachorz*	*Simulations Plus*
09.30-10.00	C-03	**Scaffold Hopping with Generative Reinforcement Learning** *Luke Rossen*	*Eindhoven University of Technology*
10.00-10.30	**Coffee Break** *Atrium*		
10.30-11.00	C-04	**Honey, I shrunk the database: Making multi-billion compound libraries as small as possible** *John Wilkinson Mayfield*	*NextMove Software*
11.00-11.30	C-05	**Docking-based geometric graph models for kinase-ligand affinity prediction** *Andrius Bernatavicius*	*Leiden University*
11.30-12.00	C-06	**Validating the prediction of lowest-energy tautomers and conformers against experimental techniques** *Bernardo de Souza*	***F****ACCTs GmbH*
12.00-12.30	C-07	**Modernising the Reaction Vector Framework: From Legacy Code to Validated Synthesis** *James Webster*	*University of Dundee*
12.30-12.45	**Break to get ready (Box Lunch)**		
12.45-23.00	**Social Event** *Leiden*		

Scientific Program

Thursday June 5th

08.30-13.00	**Session D - Integrative Structure-Based Drug Design** *Rotonde*		
08.30-09.00	D-01	**DockM8: All-in-One Open-Source Platform for Consensus Virtual Drug Screening** *Antoine Michel Lauder Lacour*	*Saarland University*
09.00-09.30	D-02	**Computational Challenges in Modeling Metal-Binding Sites in Proteins: A Multiscale Approach to Copper-Ligand Interactions in the Plant Receptor ETR1** *Lisa Sophie Kersten*	*Heinrich Heine University*
09.30-10.00	D-03	**AI and MD-aided computational docking pipeline to elucidate TGF-beta type I receptor & signaling mediator interaction.** *Leon Moritz Obendorf*	*Freie Universität Berlin*
10.00-10.30	D-04	**Advancing free-energy calculations by combining multiscale modeling and multistate enhanced sampling** Domen Pregeljc	ETH Zurich
10.30-11.00	**Coffee Break** *Atrium Boulevard*		
11.00-11.30	D-05	**How useful are protein folding tools for drug design?** *Henriette Willems*	*University of Cambridge*
11.30-12.00	D-06	**Quantum Mechanics for Ligand Design, Binding Affinities, Toxicity Risk and Process Chemistry: Successes and Obstacles** *Andreas Göller*	*Bayer AG*
12.00-12.30	D-07	**MDPath: Unraveling Allosteric Communication Pathways through Molecular Dynamics Simulations** *Niklas Piet Doering*	*Freie Universität Berlin*
12.30-13.00	D-08	**How unsociable is the fragment space, and can we do better?** *Philipp Janssen*	*Saarland University*
13.00-13.15	**Lunch** *Box or sit down*		
13.15-14.00	**Break to get ready**		
13.30	**Shuttle Bus to Schiphol**		
14.30	**Shuttle Bus to Schiphol**		
16.15	**Shuttle Bus to Schiphol**		

Poster Session RED

Advanced Cheminformatics Techniques

P01	**USING CHEMBL TO IMPROVE MOLECULE GENERATION** Eloy Félix, Noel M. O'Boyle *Chemical Biology Services, EMBL's European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, UK*
P03	**CHEMICAL PROBES IN THE SCIENTIFIC LITERATURE: ILLUMINATING NOVEL TARGET-DISEASE ASSOCIATIONS AND STRENGTHENING THE EXISTING EVIDENCE** Melissa F. Adasme[1], David Ochoa[2], Irene Lopez[2], Hoang-My-Anh Do[1], Ellie MacDonagh[2], Andrew Leach[1], Noel O'Boyle[1], Barbara Zdrazil[1] *[1] Chemical Biology Services, EMBL's European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, UK* *[2] Open Targets, Wellcome Genome Campus, Hinxton Cambridgeshire UK*
P05	**FEDERATED REPRESENTATION LEARNING USING KNOWLEDGE DISTILLATION** Thierry Hanser, Jeffrey Plante, Rob Thomas Stephane Werner *Lhasa Limited, Molecular Informatics and AI team, Leeds, UK*
P07	**TAMING TAUTOMERS: EXTENDING ECFP FOR TAUTOMER INVARIANCE** Pirie R. M. E.[1] Mayfield J. W., Sayle R. A. *NextMove Software, Cambridge, UK*
P09	**PRACTICAL MOLECULAR PROPERTY PREDICTION WITH MOLPIPELINE – A SCIENTIFIC INDUSTRY PERSPECTIVE** Conrad Stork, Christian W. Feldmann, Jennifer Hemmerich, Jochen Sieg, Frederik Sandfort, Philipp Eiden, Miriam Mathea *BASF SE, Ludwigshafen, Germany*
P11	**INTEGRATING IN SILICO ENRICHMENT PREDICTION WITH DE NOVO BUILDING BLOCK SELECTION FOR EFFICIENT COMBINATORIAL LIBRARY DESIGN** Remco L. van den Broek[1], A. Nandkeolyar[2], E. A. van der Nol[1], S. M. McKenna[1], M. Šícho[1,3], S. Pomplun[1], D. L. Mobley[2], G. J. P. van Westen[1], and W. Jespers[1] *[1] Division of Medicinal Chemistry, Leiden Academic Centre for Drug Research (LACDR), Leiden University, Leiden, The Netherlands* *[2] Department of Pharmaceutical Sciences, University of California Irvine, Irvine, California, United States of America* *[3] Laboratory of Informatics and Chemistry, University of Chemistry and Technology, Prague, Czech Republic*
P13	**MOLMEDB – MOLECULES ON MEMBRANES DATABASE** Storchmannová K[1], Juračka J[1,2], Martinát D[1], Bazgier V[1], Galgonek J[3], Türková A[4], Zdrazil B[5], Berka K[1] *[1] Department of Physical Chemistry, Palacky University in Olomouc, Czech Republic* *[2] Department of Computer Science, Palacky University in Olomouc, Czech Republic* *[3] Institute of Organic Chemistry and Biochemistry of the Czech Academy of Sciences, Prague, Czech Republic* *[4] AI|ffinity s.r.o., Brno-Medlánky, Czech Republic* *[5] European Molecular Biology Lab., European Bioinformatics Institute (EMBL-EBI), Hinxton United Kingdom*
P15	**THE NEXT GENERATION OF THE IUPAC INTERNATIONAL CHEMICAL IDENTIFIER (INCHI)** Gerd Blanke[1], Andrey Yerin[2], Clare Tovee[3], Ian Bruno[3], Jonathan Goodman[4], Richard Hartshorn[5], Ulrich Schatzschneider[6], Djordje Baljozovic[7], Felix Bänsch[8], Frank Lange[7], Jan Brammer[7], Nauman Ullah Khan[7], Sonja Herres-Pawlis[7] *[1] StructurePendium GmbH, Essen, Germany* *[2] ACD/Labs, Porto, Portugal* *[3] Cambridge Crystallographic Data Centre, Cambridge, UK* *[4] Yusuf Hamied Department of Chemistry, University of Cambridge, Lensfield Road, UK* *[5] School of Physical and Chemical Sciences, University of Canterbury, Christchurch, New Zealand* *[6] Institut für Anorganische Chemie, Julius-Maximilians-Universität Würzburg, Germany* *[7] Institut für Anorganische Chemie, RWTH Aachen, Germany* *[8] Beilstein-Institut zur Förderung der Chemischen Wissenschaften, Frankfurt am Main, Germany*

Advanced Cheminformatics Techniques cont.

P17	**PROQSAR: AN END-TO-END FRAMEWORK FOR THE AUTOMATED CONSTRUCTION OF PREDICTIVE MODELS** Tuyet-Minh Phan[1], Tieu-Long Phan*[,2,3], Tuyen Ngoc Truong[1] *[1] Falcuty of Pharmacy, University of Medicine and Pharmacy at Ho Chi Minh City, Ho Chi Minh City, Vietnam* *[2] Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, Universität Leipzig, Germany* *[3] Department of Mathematics and Computer Science, University of Southern Denmark, Odense M, Denmark*
P19	**PREFIX-BASED DECISION TREE FOR FASTER GENERATION OF SMARTS-BASED FINGERPRINTS** Andrew Dalke *Andrew Dalke Scientific AB, Sweden*
P21	**A COMPREHENSIVE MAPPING OF CHEMICAL AND PHARMACOLOGICAL SPACES FOR DRUG DISCOVERY AND REPURPOSING: INSIGHTS FROM MARKETED DRUGS AND DRUG CANDIDATES** Candida Manelfi[1], Valerio Tazzari[1], Carmen Cerchia[2], Pieter F. W. Stouten[1,3], Andrea Rosario Beccari[1] *[1] EXSCALATE, Dompé Farmaceutici SpA, Napoli, Italy* *[2] Department of Pharmacy, University of Naples "Federico II", Napoli, Italy* *[3] Stouten Pharma Consultancy BV, Sint-Katelijne-Waver, Belgium*
P23	**TANGIBLE CHEMICAL SPACE: EASY ACCESS TO A FULLY ANNOTATED DATABASE OF COMPOUNDS BY ENUMERATION OF TWO- OR THREE-COMPONENT REACTIONS** Valerio Tazzari[1], Candida Manelfi[1], Carmen Cerchia[2] , Anna Fava[1], and Andrea Rosario Beccari[1] *[1] EXSCALATE, Dompé Farmaceutici SpA, Naples, Italy* *[2] Department of Pharmacy, University of Naples "Federico II", Naples, Italy*

Artificial Intelligence, Machine Learning, and QSAR

P25	**THE COMPOUND MAPPER: BRIDGING PRACTITIONERS AND BIOACTIVITY DATA WITH AUTOMATED QUALITY CONTROL** David Alencar Araripe[1,2], Linde Schoenmaker[1], Olivier Béquignon[1.3,4], Gerard J.P. van Westen[1] *[1] Division of Medicinal Chemistry, Leiden Academic Centre for Drug Research, Leiden University, The Netherlands* *[2] Department of Human Genetics, Leiden University Medical Centre (LUMC), The Netherlands* *[3] Amsterdam UMC location Vrije Universiteit Amsterdam, Neurosurgery, The Netherlands* *[4] Cancer Center Amsterdam, Cancer Biology and Immunology, Amsterdam, The Netherlands*
P27	**DISCOVERING NOVEL BETA-LACTAMASE INHIBITORS WITH AN AI-BASED VIRTUAL PIPELINE** H.W. van den Maagdenberg[1,2], B.J. Bongers[1], P.H. van der Graaf[2,3], J.G.C. van Hasselt[2], G.J.P. van Westen[1] *[1]Medicinal Chemistry, Leiden Academic Centre for Drug Research, The Netherlands* *[2] Systems Pharmacology and Pharmacy, Leiden Academic Centre for Drug Research, Leiden University, The Netherlands* *[3] Certara, University Road, Canterbury Innovation Centre, Unit 43, Kent, UK*
P29	**BALANCING COMPLEXITY AND EFFICIENCY: SCALABLE MACHINE LEARNING APPROACHES FOR REACTION YIELD PREDICTION** Idil Ismail & Sereina Riniker *Department of Chemistry and Applied Biosciences, ETH Zürich, Switzerland*
P31	**BIO-ISOSTERE GUIDED MOLECULAR PROPERTY PREDICTION** Anatol Ehrlich[1], Nils M. Kriege[1], Christoph Flamm[2] *[1] Faculty of Computer Science, University of Vienna, Austria* *[2] Department of Theoretical Chemistry, University of Vienna, Austria*
P33	**PREDICTION OF IN VIVO PK PROFILES FROM CHEMICAL STRUCTURES AND IN VITRO ADME EXPERIMENTS** Moritz Walter[1], Bettina Gerner[2], Hermann Rapp[2], Hannes Wendelin[3], Christofer Tautermann[1], Miha Skalic[1], Jens Markus Borghardt[2], Lina Humbeck[1] *[1 2] Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach, Germany* *[1] Medicinal Chemistry [2] Drug Discovery Sciences* *[3] Boehringer Ingelheim RCV GmbH & Co KG, Cancer Research DDS ADME, Wien, Austria*
P35	**CENSORED LOSS: INCLUSION OF CENSORED DATA IN MOLECULAR AFFINITY MODELLING** Marc A. Boef*[1], R. L. van den Broek[1], G. J. P. van Westen[1] *[1] Division of Medicinal Chemistry, Leiden Academic Center for Drug Research, The Netherlands*

P37	**EVALUATING MACHINE LEARNING MODELS FOR MOLECULAR PROPERTY PREDICTION: PERFORMANCE AND ROBUSTNESS ON OUT-OF-DISTRIBUTION DATA** Hosein Fooladi[1,2,3], Thi Ngoc Lan Vu[1,2,3], and Johannes Kirchmair[1,2] *[1]Christian Doppler Laboratory for Molecular Informatics in the Biosciences, Department for Pharmaceutical Sciences, University of Vienna, Austria* *[2]Department of Pharmaceutical Sciences, Division of Pharmaceutical Chemistry, University of Vienna, Austria* *[3]Vienna Doctoral School of Pharmaceutical, Nutritional and Sport Sciences (PhaNuSpo), University of Vienna, Austria*
P39	**P39: Trialblazer: A Chemistry-Focused Predictor of Toxicity Risks in Late-Stage Drug Development** Huanni Zhang[1,2,3], Matthias Welsch[1,2,3], William Schueller[1], Johannes Kirchmair*[1,2,3] *[1]Department of Pharmaceutical Sciences, Division of Pharmaceutical Chemistry, Faculty of Life Sciences, University of Vienna, Austria* *[2]Christian Doppler Laboratory for Molecular Informatics in the Biosciences, Department for Pharmaceutical Sciences, University of Vienna, Austria* *[3]Vienna Doctoral School of Pharmaceutical, Nutritional and Sport Sciences (PhaNuSpo), University of Vienna, Austria*
P41	**FAST AND SCALABLE 3D PHARMACOPHORE SCREENING WITH PHARMACOMATCH** Daniel Rose[1,2,3], Oliver Wieder[1,2], Thomas Seidel[1,2], Thierry Lager[1,2] *[1] Department of Pharmaceutical Sciences, Division of Pharmaceutical Chemistry, Faculty of Life Sciences, University of Vienna, Austria* *[2] Christian Doppler Laboratory for Molecular Informatics in the Biosciences, Department of Pharmaceutical Sciences, University of Vienna, Austria* *[3] Vienna Doctoral School of Pharmaceutical, Nutritional and Sport Sciences, University of Vienna, Austria*
P43	**UNLOCKING THE POTENTIAL OF C2-CARBOXYLATED 1,3-AZOLES: A COMPUTATIONAL DESIGN-MAKE-TEST-ANALYZE (DMTA) APPROACH** Kerrin Janssen[1], Johannes Kirchmair[2,3], Jonny Proppe[1] *[1] Institute of Physical and Theoretical Chemistry, TU Braunschweig, Germany* *[2] Department of Pharmaceutical Sciences, University of Vienna, Austria* *[3] Christian Doppler Laboratory for Molecular Informatics in the Biosciences, Department for Pharmaceutical Sciences, University of Vienna, Austria*
P45	**APPLICATION OF DFT TO ASSESS NITROSAMINE FORMATION RISKS: TERTIARY AMINES AREN'T A RISK, EXCEPT WHEN THEY ARE** E. Pye, M. Kawamura, M. Burns, C. Barber *Lhasa Limited, Granary Wharf House, Leeds, UK*
P47	**CLOSING THE GENERATIVE AI SBDD LOOP: FROM GPCR STRUCTURE TO REINFORCEMENT LEARNING GUIDED DE NOVO LIGAND DESIGN AND BACK AGAIN** Chris de Graaf *Structure Therapeutics, USA / China*
P49	**CONFORMAL CALIBRATION OF QSAR CLASSIFIERS** Sébastien Guesné, Stéphane Werner and Thierry Hanser *Lhasa Limited, Granary Wharf House, Leeds, United Kingdom*
P51	**ANNalog – GENERATION OF MEDCHEM-SIMILAR MOLECULES** Wei Dai[1], Jonathan D. Tyzack[2], Chris de Graaf[3], Arianna Fornili[1], Noel M.O'Boyle[4] *[1] School of Physical and Chemical Science, Queen Mary University of London, United Kingdom* *[2] Nxera Pharma, Steinmetz Building, Granta Park, Great Abington, Cambridge, United Kingdom* *[3] Structural Therapeutics, South San Francisco, USA* *[4] EMBL's European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire, United Kingdom*
P53	**IMPROVING AFFINITY OF EPCAM-BINDING PEPTIDES USING ALPHAFOLD2 FOR IN VIVO TUMOR IMAGING** Nada Badr[1*], Nicky Janssen[3], Mark Fonteyne[1], Jochem R. van der Burgt[1], Taryn March[1], Lysanne D.A.N. de Muynck[1], Shadhvi Bhairosingh[1], Robbert Kim[2], Alexander L. Vahrmeijer[1], Gerard J.P. van Westen[3], Peter J. K. Kuppen[1] *[1] Department of Surgery, Leiden University Medical Center, Leiden, The Netherlands* *[2] Department of Cell and Chemical Biology, Leiden University Medical Center, Leiden, The Netherlands* *[3] Medicinal Chemistry, Leiden Academic Centre for Drug Research, Leiden, The Netherlands* *[*] Presenting author*

Artificial Intelligence, Machine Learning, and QSAR cont.

P55	**INTEGRATION OF STEREOCHEMISTRY WITHIN DRUGEX FOR BETTER SAMPLE EFFICIENCY** Chiel Jespers[1], Martin Sicho[1,3], Mike Preuss[2], Gerard van Westen[1] *[1] Systems Pharmacology and Pharmacy, LACDR, Leiden University, The Netherlands* *[2] Leiden Institute of Advanced Computer Science, Leiden, The Netherlands* *[3] National Infrastructure for Chemical Biology, Department of Informatics and Chemistry, Faculty of Chemical Technology, University of Chemistry and Technology Prague, Czech Republic*
P57	**MULTI-TASK IS WHAT YOU NEED! MULTI-TASK MACHINE LEARNING MODELS FOR MOLECULAR PROPERTY PREDICTION** Eelke Bart Lenselink[1], Giovanni A. Tricarico[1], Marie-Pierre Dréanic[2], Johan Hofmans[1], Kenneth Goosens[1], Stephane de Cesco[1] *[1] Galapagos NV, Mechelen, Belgium* *[2]Galapagos SASU, Romainville, France*
P59	**FANTASTIC SMILES AUGMENTATION METHODS AND WHERE TO FIND THEM** H. Brinkmann[1], A. Argante[1], H. ter Steege[1], F. Grisoni[1,2] *[1] Institute for Complex Molecular Systems (ICMS), Eindhoven AI Systems Institute (EAISI), Department of Biomedical Engineering, Eindhoven University of Technology, The Netherlands* *[2] Centre for Living Technologies, Alliance TU/e, WUR, UU, UMC Utrecht, The Netherlands*
P61	**CONSTRAINED GENERATION OF MOLECULES USING DIFFUSION MODELS** Cristian Pop[1], James Longden[2], Aniket Ausekar[2], Andreas Bender[1] *[1] Iuliu Hatieganu University of Medicine and Pharmacy, Faculty of Pharmacy, Cluj-Napoca, Romania* *[2] Evolvus Inc., Frankfurt am Main, Germany*
P63	**A NOVEL STATISTICAL MACHINE LEARNING FRAMEWORK FOR ENHANCED DRUG SAFETY PREDICTION IN ZEBRAFISH ASSAYS** Filippo Lunghini[1], Christian Cortes Campos[2], Vincenzo Pisapia[3], Gentzane Sánchez Elexpuru[2], Sylvia Dyballa[2], Francesco Sacco[3], Daniela Iaconis[1], Vincenzo Di Donato[3], Andrea Beccari[1] *[1] EXSCALATE, Dompé Farmaceutici SpA, Italy* *[2] ZeClinics SL, Barcelona* *[3] Professional Service Department, SAS Institute, Milan, Italy*
P65	**PREDICTING THE DISSIPATION KINETICS OF AGROCHEMICALS IN SOIL** Vincent-Alexander Scholz,[1,2] Richard Marchese-Robinson,[3] Sevil Payvandi,[3] Timothy J. C. O'Riordan[3] and Johannes Kirchmair[1,4] *[1] Department of Pharmaceutical Sciences, Division of Pharmaceutical Chemistry, University of Vienna, Austria* *[2] Vienna Doctoral School of Pharmaceutical, Nutritional and Sport Sciences (PhaNuSpo), University of Vienna, Austria* *[3] Syngenta UK, Jealott's Hill International Research Centre, Bracknell, Berkshire, United Kingdom* *[4] Christian Doppler Laboratory for Molecular Informatics in the Biosciences, Department for Pharmaceutical Sciences, University of Vienna, Austria*
P67	**NLP-INSPIRED OPERATORS FOR DE NOVO DESIGN AUGMENTATION** Hanz Tantiangco[1], James Webster[2], Beining Chen[3], Val Gillet[1] *[1] Information School, The University of Sheffield, U.K.* *[2] Drug Discovery Unit, The University of Dundee, U.K.* *[3] Department of Chemistry, The University of Sheffield, U.K.*

Integrative Structure-Based Drug Design

P69	**EXTENDING ENVELOPING DISTRIBUTION SAMPLING TOWARDS NE-EDS: A NON-EQUILIBRIUM APPROACH TO FREE-ENERGY ESTIMATION** Shu-Yu Chen[1,†], Enrico Ruijsenaars[1,†], Philippe H. Hünenberger[1], Sereina Riniker[1] *[1] Department of Chemistry and Applied Biosciences, ETH Zurich, Switzerland* *[?] S.-Y.C. and E.R. contributed equally to this work*
P71	**IMERGE-FEP 2.0: GENERATING INTERMEDIATE R-GROUPS FOR CHALLENGING FREE ENERGY PERTURBATIONS** Daan A. Jiskoot[1,2], Linde Schoenmaker[2], Jeroen L.A. Pennings[3], Willie J. G. M. Peijnenburg[1,4], David L. Mobley[5], Pim N.H. Wassenaar[1], Gerard J.P. van Westen[2], Willem Jespers[2,6] *[1] Centre for Safety of Substances and Products, National Institute for Public Health and the Environment (RIVM), Bilthoven, the Netherlands* *[2] Div. Drug Discovery and Safety, Leiden Academic Centre for Drug Res., Leiden University, The Netherlands* *[3] Centre for Health Protection, Nat. Inst. for Public Health and the Env. (RIVM), Bilthoven, The Netherlands* *[4] Institute of Environmental Sciences, Leiden University, The Netherlands* *[5] Dept. of Pharmaceutical Sciences, University of California, Irvine, United States* *[6] Dept. of Med. Chemistry, Photopharmacology and Imaging, Groningen Research Institute of Pharmacy, Groningen, The Netherlands*
P73	**ASSESSING THE ROLE OF MACHINE LEARNING-BASED POSE SAMPLING IN VIRTUAL SCREENING** Thi Ngoc Lan Vu[1,2,3], Hosein Fooladi[1,2,3], Johannes Kirchmair[1,2] *[1] Christian Doppler Laboratory for Molecular Informatics in the Biosciences, Department for Pharmaceutical Sciences, University of Vienna, Austria* *[2] Department of Pharmaceutical Sciences, Division of Pharmaceutical Chemistry, University of Vienna, Austria* *[3] Vienna Doctoral School of Pharmaceutical, Nutritional and Sport Sciences (PhaNuSpo), University of Vienna, Austria*
P75	**ULTRALARGE TAILORED LIBRARY OF AMINO ACID DERIVATIVES DESIGNED TO TARGET PEPTIDE GPCRS** Pach S.[1], Carlsson J.[1] *[1] Science for Life Laboratory, Department of Cell and Molecular Biology-BMC, Uppsala University, Sweden*
P77	**OXYTOCIN-SIGNALING-INSPIRED ALLOSTERIC MODULATOR DESIGN FOR THE μ-OPIOID RECEPTOR** Marvin Taterra[1], Marcel Bermudez[1], *[1] Universität Münster, Institute of Pharmaceutical and Medicinal Chemistry, Germany*
P79	**IMPROVED PROTEIN-LIGAND STRUCTURE MODELING USING CONSERVED SCAFFOLD PLACEMENT** Jonathan Pletzer-Zelgert[1], Matthias Rarey[1], Bernd Kuhn[2] *[1] ZBH - Center for Bioinformatics, University of Hamburg, Hamburg, Germany* *[2] Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd., Basel, Switzerland*
P81	**STRUCTURE-BASED PHARMACOPHORE SCREENING OF TREM2** Kevin Lam[1], Gerhard Wolber[1], Moustafa Gabr[2] *[1] Molecular Design Group, Institute of Pharmacy, Department of Biology, Chemistry & Pharmacy, Freie Universität Berlin, Germany* *[2] Moustafa Gabr Laboratory, Department of Radiology, Weill Cornell Medical College, New York, US*
P83	**IDENTIFICATION, SYNTHESIS AND BIOLOGICAL ASSESSMENT OF ALLOSTERIC LIGANDS FOR THE C-C CHEMOKINE RECEPTOR 5** Kian Noorman van der Dussen[1], Martin Šícho[1,2*‡], Khaled Essa[1,3‡], Yao Yao[1], Benthe Bleijs[1], Laura Heitman[1,3], Gerard van Westen[1,3*], Daan van der Es[1], Willem Jespers[1] *[1] Division of Medicinal Chemistry, Leiden Academic Centre for Drug Research (LACDR), Leiden University, The Netherlands.* *[2] CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Department of Informatics and Chemistry, University of Chemistry and Technology Prague, Czech Republic* *[3] Oncode Institute, Leiden, The Netherlands*

New Modalities and Large Chemical Data Sets

P85	**GENERATION OF CUSTOM SYNTHETICALLY ACCESSIBLE COMBINATORIAL CHEMICAL SPACES USING MACHINE LEARNING-BASED REAGENT FILTERING – DESIGN OF FREEDOM SPACE 4.0** Anna Kapeliukha[1,3], Serhii Hlotov[1,3], Marina Vasylchuk[1], Mykola Protopopov[1,3], Olga Tarkhanova[1], Yurii Moroz[2,3] *[1] Chemspace LLC, Kyiv, Ukraine* *[2] Enamine Ltd., Kyiv, Ukraine* *[3] Taras Shevchenko National University of Kyiv, Ukraine*

Open Science, Omics, and Natural Products

P87	**IDENTIFYING OFF-TARGET DRUG INTERACTIONS MEDIATED VIA DNA METHYLATION** Delaney A. Smith[1], Russ B. Altman[2] *[1] Department of Biochemistry, Stanford University Medical School, Stanford CA, USA* *[2] Departments of Genetics, Bioengineering, and Bioinformatics, Stanford University Medical School, Stanford CA, USA*
P89	**VHP4SAFETY COMPOUND WIKI: AN OPEN SCIENCE APPROACH TO COLLECT DOMAIN SPECIFIC KNOWLEDGE** Willighagen E[1], Zare Jeddi, M[2], Sinke L[3] *[1] Dept of Translational Genomics, Maastricht University, The Netherlands* *[2] Shell Global Solutions International BV, The Netherlands* *[3] Leiden Academic Centre for Drug Research, University of Leiden, The Netherlands*
P91	**A MUTATOR EFFECT CAUSED BY TWO AMINO ACID CHANGES IN THE DNA BINDING REGION OF M. SMEGMATIS DNAE1: NOVEL INSIGHTS INTO DNA POLYMERASE FIDELITY USING IN SILICO AND IN VIVO APPROACHES** R.C.M. Kuin[1,2], M.H. Lamers[1], G.J.P. van Westen[2] *[1] Leiden Academic Centre for Drug Research (LACDR), Leiden, The Netherlands* *[2] Department of Cell & Chemical Biology, Leiden University Medical Center (LUMC), The Netherlands*

Poster Session BLUE

Advanced Cheminformatics Techniques

P02	**STELLAR: DEVELOPING AND OPTIMIZING A NOVEL ADVANCED DOCKING PROTOCOL FOR PROCESSING LARGE PEPTIDES WITHOUT AI ASSISTANCE. A CANCER CONTEXT APPLICATION** Alejandro Rodríguez-Martínez, Jochem Nelen, Miguel Carmena-Bargueño, Carlos Martínez-Cortés, Horacio Pérez-Sánchez *Structural Bioinformatics and High Performance Computing Research Group (BIO-HPC), UCAM Universidad Católica de Murcia (UCAM), Murcia, Spain*
P04	**TOWARDS MORE RELIABLE DISTANCE GEOMETRY-BASED CONFORMER GENERATION** Niels Maeder, Gregory A. Landrum, Sereina Riniker *Department of Chemistry and Applied Biosciences, ETH Zürich, Switzerland*
P06	**INVESTIGATING STRUCTURAL INFORMATION IN GRAPH-BASED NEURAL FINGERPRINTS FOR SIMILARITY SEARCHES** S. Homberg[1], M. Modlich[2], B. Risse[2], O. Koch[1] *[1] Institute of Pharmaceutical and Medicinal Chemistry, University of Münster, Germany* *[2] Computer Vision and Machine Learning Systems, University of Münster, Germany*
P08	**AUTOMATIC ANNOTATION OF SITES OF METABOLISM FROM SUBSTRATE-METABOLITE PAIRS** R. A. Jacob[1,2,3], J. Kirchmair[1,2] *[1] Department of Pharmaceutical Sciences, Division of Pharmaceutical Chemistry, University of Vienna, Austria* *[2] Christian Doppler Laboratory for Molecular Informatics in the Biosciences, Department of Pharmaceutical Sciences, University of Vienna, Austria* *[3] Vienna Doctoral School of Pharmaceutical, Nutritional and Sport Sciences (PhaNuSpo), University of Vienna, Austria*
P10	**BAYESIAN ILLUMINATION: INFERENCE AND QUALITY-DIVERSITY ACCELERATE GENERATIVE MOLECULAR MODELS** Jonas Verhellen[1,2] *[1] University of Copenhagen, Department of Drug Design and Pharmacology, Denmark* *[2] University of Oslo, Centre for Integrative Neuroplasticity, Norway*
P12	**THE MOLECULE FRAGMENTATION FRAMEWORK (MORTAR): A RICH CLIENT APPLICATION FOR ALGORITHMIC SUBSTRUCTURE EXTRACTION** F. Bänsch[1], C. Steinbeck[2], A. Zielesny[1], J. Schaub[2] *[1] Institute for Bioinformatics and Chemoinformatics, Westphalian University of Applied Sciences, Recklinghausen, Germany* *[2] Institute for Inorganic and Analytical Chemistry, Friedrich Schiller University Jena, Germany*
P14	**RETROSPECTIVE EVALUATION OF MOLECULE ENUMERATION METHODS** Pierre-Yves Libouban[1], David Hahn[1], Natalia Dyubankova[1], Dries Van Rompaey[1], Gary Tresadern[1] *[1] In-Silico Discovery, Johnson & Johnson, Beerse, Belgium*
P16	**DECONVOLUTION OF “ORIGIN-OF-LIFE" REACTION NETWORKS** Nico Domschke[1], Richard Golnik[1], Chen Wang[2], Ales Charvat[2], Thomas Gatter[1], Bernd Abel[2], Peter F. Stadler[1] *[1] Bioinformatics, Leipzig University, Germany* *[2] Wilhelm-Ostwald-Institut für Physikalische und Theoretische Chemie, Leipzig University, Germany*
P18	**DECIPHERING MOLECULAR EMBEDDINGS WITH CENTERED KERNEL ALIGNMENT** Matthias Welsch[1,2,3], Steffen Hirte[1,3], Johannes Kirchmair[*,1,2] *[1] Department of Pharmaceutical Sciences, Division of Pharmaceutical Chemistry, Faculty of Life Sciences, University of Vienna, Austria* *[2] Christian Doppler Laboratory for Molecular Informatics in the Biosciences, Department for Pharmaceutical Sciences, University of Vienna, Austria* *[3] Vienna Doctoral School of Pharmaceutical, Nutritional and Sport Sciences (PhaNuSpo), University of Vienna, Austria*

Advanced Cheminformatics Techniques cont.

P20	**PREDICTING OFF-TARGETS FROM CHEMBL DATA USING THE POLYPHARMACOLOGY BROWSER** Maedeh Darsaraee and Jean-Louis Reymond* *Department of Chemistry, Biochemistry and Pharmacy, University of Berne, Switzerland*
P22	**COMPASS: COMPUTATIONAL POCKET ANALYSIS AND SCORING SYSTEM** Akash Deep Biswas[1]*, Emanuela Sabato[2], Serena Vittorio[2], Parisa Aletayeb[2], Alessandro Pedretti[2], Angelica Mazzolari[2], Carmen Gratteri[3], Andrea R. Beccari[1], Giulio Vistoli[2], and Carmine Talarico[1] *[1] Dompé Farmaceutici S.p.A., Napoli, Italy* *[2] Dipartimento di Scienze Farmaceutiche, Università degli Studi di Milano, Italy* *[3] LIGHT s.c.ar.l., Brescia, Italy*
P24	**LACAN: LEVERAGING ADJACENT CO-OCCURRENCE OF ATOMIC NEIGHBORHOODS** Wim Dehaen[1,2] *[1] Department of Informatics and Chemistry, University of Chemistry and Technology Prague, Czech Republic* *[2] Department of Organic Chemistry, University of Chemistry and Technology Prague, Czech Republic*

Artificial Intelligence, Machine Learning, and QSAR

P26	ENRICHING CHEMBL ASSAY DESCRIPTIONS USING NATURAL LANGUAGE PROCESSING Ines A. Smit[1], Melissa F. Adasme[1], Emma Manners[1], Sybilla Corbett[1], Hoang-My-Anh Do[1], Noel O'Boyle[1], Andrew R. Leach[1], Barbara Zdrazil[1] *[1] Chemical Biology Services, European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, United Kingdom*
P28	**MULTIRETRO – A SYNTHESIS PLANNING TOOLKIT** Alan Kai Hassen[1,2], Jacquelyn L. Klug-McLeod[3], Roger M. Howard[3], Jason Mustakis[3], Antonius P. A. Janssen[4], Gerard J.P. van Westen[4], Mike Preuss[2], Djork-Arné Clevert[1] *[1] Machine Learning Research, Pfizer Research and Development, Berlin, Germany* *[2] Leiden Institute of Advanced Computer Science, Leiden University, The Netherlands* *[3] Pfizer Research and Development, Groton, CT, USA* *[4] Leiden Academic Centre for Drug Research, Leiden University, The Netherlands*
P30	**ROBUST PREDICTION OF THE PHARMACOPHORE FIT SCORES WITH ACTIVE LEARNING** D. Goldmann, C. Grebner, G. Hessler *Synthetic Molecular Design, Integrated Drug Discovery, Sanofi, Frankfurt am Main, Germany*
P32	**USING DEEP LEARNING AND MACHINE LEARNING-BASED DOCKING TO INVESTIGATE METALLOENZYME-SUBSTRATE COMPLEXES** Daniil Lepikhov[1,2], Laura Sandner[1], Silke Leimkühler[2], Ariane Nunes-Alves[1] *[1] Theoretical Structural Biology group, Technical University Berlin, Germany* *[2] Molecular Enzymology group, University Potsdam, Germany*
P34	**FROM THEORY TO PRACTICE: REACTION SIMILARITY SEARCH IN REAL-WORLD APPLICATIONS AT ASTEX** I.N. Derbenev, D. Branduardi, R.F. Ludlow *[1] Computational Chemistry & Informatics Department, Astex Pharmaceuticals, Cambridge, United Kingdom*
P36	**PREDICTION OF PHARMACOKINETICS PROFILE AS TIME SERIES** Uday Abu Shehab[1,3], Gerhard F. Ecker[1], Lina Humbeck[2], Miha Skalic[2], Moritz Walter[2], Andreas Bergner[3] *[1] University of Vienna, Department of Pharmaceutical Chemistry, Austria* *[2] Boehringer Ingelheim Pharma GmbH & Co KG, Medicinal Chemistry Department, Biberach an der Riss, Germany* *[3] Boehringer Ingelheim RCV GmbH & Co KG, Drug Discovery Sciences, Vienna, Austria*

P38	**CHEMISTRY-AWARE FOUNDATION MODEL FOR SMALL MOLECULE ADMET AND POLYPHARMACOLOGY PROPERTY ESTIMATION** Pietro Morerio[1], Filippo Lunghini[2], Alessio Del Bue[1], Andrea Beccari[2] *[1] Istituto Italiano di Tecnologia, Italy* *[2] EXSCALATE, Dompé Farmaceutici SpA, Naples, Italy*
P40	**EXPLOITING SARKUSH AND FREE-WILSON ANALYSIS TO ACCELERATE AN ANTIVIRAL DRUG DISCOVERY PROJECT** Jess Stacey[1], Lauren Reid[1], Al Dossetter[1], Ed Griffen[1], Andrew Leach[1], Phillip de Sousa[1], Bashy Khan[1], Dan James[1], and David Cousins[1] *[1] MedChemica Ltd, Motorworks, Macclesfield, United Kingdom*
P42	**EFFICIENT COMPOUND SELECTION STRATEGIES IN LEAD OPTIMIZATION: INSIGHTS FROM RETROSPECTIVE ANALYSIS** Mas Pablo[1,2], Filoche-Rommé Bruno[2], Vuilleumier Rodolphe[1], Bianciotto Marc[2] *[1] PASTEUR Lab, École Normale Supérieure – PSL, Paris, France* *[2] Molecular Design Sciences, Integrated Drug Discovery, Sanofi, Vitry-sur-Seine, France*
P44	**regAL: PYTHON PACKAGE FOR ACTIVE LEARNING OF REGRESSION PROBLEMS** Elizaveta Surzhikova and Jonny Proppe *Institute of Physical and Theoretical Chemistry, Technische Universität Braunschweig, Germany*
P46	**EXPLORATION OF DATA FROM THE PHARMACEUTICAL INDUSTRY FOR SITE-OF-METABOLISM PREDICTION** Ya Chen[1,2] *[1] Department of Pharmaceutical Sciences, Division of Pharmaceutical Chemistry, University of Vienna, Austria* *[2] Medicinal Chemistry, Research and Early Development, Cardiovascular, Renal and Metabolism (CVRM), BioPharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden*
P48	**PySSA: END-USER PROTEIN STRUCTURE PREDICTION AND VISUAL ANALYSIS WITH ColabFold AND PyMOL** H. Kullik[1], M. Urban[1], J. Schaub[2], A. Loidl-Stahlhofen[3], A. Zielesny[1] *[1] Institute for Bioinformatics and Chemoinformatics, Westphalian University of Applied Sciences, Recklinghausen, Germany* *[2] Institute for Inorganic and Analytical Chemistry, Friedrich Schiller University Jena, Germany* *[3] Lab. of Protein Chemistry, Westphalian University of Applied Sciences, Recklinghausen, Germany*
P50	**REST2-AMP/MM: INTEGRATING ENHANCED SAMPLING WITH MACHINE LEARNING POTENTIALS FOR MOLECULAR CONFORMATIONAL SAMPLING** Riccardo Solazzo[1], Igor Gordiy[1], Sereina Riniker[1] *[1]Department of Chemistry and Applied Biosciences, ETH Zürich, Switzerland*
P52	**COOPERATIVE FREE ENERGY: INDUCED PPI, SOLVATION, AND CONFORMATION IN PROTEIN-LIGAND-PROTEIN TERNARY COMPLEXATION** Shu-Yu Chen[1], Riccardo Solazzo[1], Marianne Fouche[2], Hans-Joerg Roth[2], Birger Dittrich[2], Sereina Riniker[1] *[1] Department of Chemistry and Applied Biosciences, ETH Zurich, Switzerland* *[2] Novartis Biomedical Research, Basel, Switzerland*
P54	**A FEATURE-ENGINEERED DELTA-ML APPROACH FOR MOLECULAR STRUCTURE REFINEMENT: BRIDGING EXPLORATION AND EXPLOITATION IN COMPUTATIONAL CHEMISTRY** Federico Lazzari *Scuola Superiore Meridionale, Napoli, Italy*
P56	**MACHINE LEARNING PREDICTIONS OF THE PROTEIN-LIGAND BINDING AFFINITY WITH FINGERPRINTS, SHAPE AND ELECTROSTATICS** K. Stanciakova[1], M. Krier[1], L. Eberlein[1], G. Stahl[1], J. Chen[2], S. Mandal[2], S. Nath[2], M. Geballe[2] *[1] OpenEye, Cadence Molecular Sciences, Cologne Germany* *[2] OpenEye, Cadence Molecular Sciences, Santa Fe, United States*
P58	**BALANCING DATA QUANTITY AND QUALITY: EVALUATING CURATION STRATEGIES FOR BIOACTIVITY PREDICTION MODELS** Carl C.G. Schiebroek, Gregory A. Landrum, and Sereina Riniker *Department of Chemistry and Applied Biosciences, ETH Zurich, Switzerland*
P60	**PROTAC-SPLITTER: AN AI-BASED SYSTEM TO AUTOMATICALLY IDENTIFY PROTAC LIGANDS** Stefano Ribes[1], Anders Källberg[1], Ranxuan Zhang[1], Eva Nittinger[2], Christian Tyrchan[2], and Rocío Mercado[1] *[1] Department of Computer Science and Engineering, Section for Data Science and AI, Chalmers University of Technology and University of Gothenburg, Sweden* *[2] Medicinal Chemistry, Research and Early Development, Respiratory and Immunology (R&I), BioPharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden*

Artificial Intelligence, Machine Learning, and QSAR cont.

P62	**THE PREDICTEAM'S ALL-INCLUSIVE STRATEGY FOR LEVERAGING THE FULL POTENTIAL OF ADMET PREDICTIONS IN DRUG DISCOVERY** Lara Kuhnke, Uschi Dolfus *Bayer AG, Berlin, Germany*
P64	**MOLECULAR DEEP LEARNING AT THE EDGE OF CHEMICAL SPACE** Derek van Tilborg[1,2], Luke Rossen[1], and Francesca Grisoni[1,2*] *[1] Institute for Complex Molecular Systems (ICMS), Dept. of Biomed. Engineering, Eindhoven University of Technology, The Netherlands* *[2] Centre for Living Technologies, Alliance TU/e, WUR, UU, UMC Utrecht, The Netherlands*
P66	**INTEGRATING STRUCTURAL AND MORPHOLOGICAL FINGERPRINTS: UNDERSTANDING INFORMATION FOR PATTERN IDENTIFICATION AND BETTER TOXICITY PREDICTION** Floriane Odje[1], Elena von Coburg[2], Christopher Wolff[3], Jens Peter von Kries[3], Sebastian Dunst[2], Andrea Volkamer[1] *[1] Data Driven Drug Design, Universität des Saarlandes, Saarbrücken, Germany* *[2] German Federal Institute for Risk Assessment (BfR), Berlin, Germany* *[3] Institute for Molecular Pharmacology, Berlin, Germany*
P68	**GUT MICROBIOTA METABOLIC MIMICKING DRUGS FOR AUTOIMMUNE/INFECTIOUS DISEASES** Shayma El-Atawneh, Oliver Koch *Institute of Pharmaceutical and Medicinal Chemistry, Universität Münster, Germany*

Integrative Structure-Based Drug Design

P70	**MOLECULAR DYNAMICS AND EXPERIMENTAL INSIGHTS INTO THE FUNGISTATIC MECHANISM OF MUTANOBACTIN D** Patricia Brandl[1], Lukas Lüthy[2], Felix Pultar[1], Moritz Hansen[2], Erick M. Carreira[2], Sereina Riniker[1] *[1] Department of Chemistry and Applied Biosciences, ETH Zürich, Switzerland* *[2] Laboratory of Organic Chemistry, ETH Zürich, D-CHAB, Switzerland*
P72	**BENCHMARKING STATE-OF-THE-ART IN SILICO PEPTIDE DESIGN METHODS AND EVALUATING PEPTIDE-OPTIMIZATION FOR FLUORESCENT PROBES** Mark Fonteyne[1,3], Peter J.K. Kuppen[1], Alexander L. Vahrmeijer[1], Willem Jespers[2*], Gerard J. P. van Westen[3*] *[1] Department of Surgery, Leiden University Medical Center, The Netherlands* *[2] Department of Pharmacy, University of Groningen, The Netherlands* *[3] Medicinal Chemistry, Leiden University (LACDR), The Netherlands* ** These authors contributed equally as last authors.*
P74	**BENCHMARKING AI-DRIVEN PROTOCOLS FOR CYCLIC PEPTIDE CONFORMER PREDICTION AND RANKING** Rodrigo Ochoa[1], Jovan Damjanovic[2] *[1] Novo Nordisk A/S, Måløv, Denmark* *[2] Novo Nordisk US R&D, Lexington, United States of America*
P76	**ACTIVE LEARNING FEP USING 3D-QSAR FOR PRIORITIZING BIOISOSTERES IN MEDICINAL CHEMISTRY** Venkata K. Ramaswamy, Matthew Habgood and Mark D. Mackey *Cresset, New Cambridge House, Litlington, Cambridgeshire, UK*
P78	**PIEZO1: IN-SILICO INVESTIGATION OF CHANNEL ACTIVATION AND DEACTIVATION** Joana Massa[1], Benedikt Frieg[2], Christian Tyrchan[2], Oliver Koch[1] *[1] Institute of Pharmaceutical and Medicinal Chemistry, University of Münster, Germany* *[2] Medicinal Chemistry, Research and Early Development, Respiratory and Immunology (R&I), BioPharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden*
P80	**PHARMACOPHORE-BASED DISCOVERY OF NOVEL CYTOCHROME P450 (CYP) 4A11 INHIBITORS TO COMBAT NON-ALCOHOLIC FATTY LIVER DISEASE (NAFLD)** Clemens A. Wolf[1], Matthias Bureik[2], Gerhard Wolber[1] *[1] Pharmaceutical and Medicinal Chemistry (Computer-Aided Drug Design), Institute of Pharmacy, Freie Universität Berlin, Germany* *[2] School of Pharmaceutical Science and Technology, Tianjin University, China*
P82	**DATABASE AUTOPH4: PHARMACOPHORE ANALYSIS OF MULTIPLE PROTEIN STRUCTURES** Chris Williams[1], Andrew Henry[1], Steve Maginn[1], Guido Kirsten[1], Markus Kossner[1], Miklos Feher[2] *[1] Chemical Computing Group, Montreal, Canada* *[2] D.E. Shaw Research, New York, USA*

P84	**FROM 45 BILLION SMALL MOLECULES TO POTENTIAL NEW LIGANDS FOR THE INTRACELLULAR BINDING SITE OF CCR2, USING HIGH THROUGHPUT VIRTUAL SCREENING, MOLECULAR DOCKING AND RELATIVE FREE ENERGY PERTURBATION** D.J.M. van Pinxteren[1,2], H. Gutiérrez-de-Terán[2], E. Gibert[3], F. Martin Garcia[3], and W. Jespers[1,2] *[1] Medicinal Chemistry, Photopharmacology and Imaging group, Rijksuniversiteit Groningen, The Netherlands* *[2] MODSIM Pharma AB, Uppsala, Sweden* *[3] Pharmacelera, PCB, Torre R, Barcelona, Spain*

New Modalities and Large Chemical Data Sets

P86	**SCAFFOLD-BASED LIBRARY DESIGN VS. MAKE-ON-DEMAND SPACE: A COMPARATIVE ASSESSMENT OF CHEMICAL CONTENT** Leonard Bui, Teodora Djikic-Stojsic, Guillaume Bret, Esther Kellenberger *Laboratoire d'Innovation Thérapeutique, UMR 7200 CNRS-Université de Strasbourg, Illkirch-Graffenstaden, France*

Open Science, Omics, and Natural Products

P88	**BIOSYNFONI: A BIOSYNFORMATIC MOLECULAR DESCRIPTOR FOR NATURAL PRODUCT RESEARCH** Lucina-May Nollen[1,2], David Meijer[1], Maria Sorokina[3], Justin J.J. van der Hooft[1,4] *[1] Bioinformatics Group, Wageningen University & Research, The Netherlands* *[2] Current address: Leiden Academic Centre of Drug Research, Leiden University, The Netherlands* *[3] DS&AI, Bayer Pharmaceuticals, Germany* *[4] Department of Biochemistry, University of Johannesburg, South Africa*
P90	**DRUG AND CLINICAL CANDIDATE DRUG DATA IN CHEMBL** Fiona Hunter, Harris Ioannidis, Melissa F Adasme, James Blackshaw, Nicolas Bosc, Sybilla Corbett, Marleen de Veij, Eloy Felix, Tevfik Kizilören, Emma Manners, Juan F. Mosquera, Ines Smit, Barbara Zdrazil and Noel O'Boyl *European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge, United Kingdom*

Oral Presentations

Keynote Address CSA Trust Mike Lynch Award

A Chemoinformatics Journey: An Evolution of Methods

Val Gillet

University of Sheffield, The Wave, UK

As a proud recipient of the Mike Lynch Award, I‘ll start with a brief tribute to Mike, who was a pioneer of chemoinformatics and laid the foundations for Sheffield‘s strong contribution to the field. Crucially for me, he was my PhD supervisor, and his inspiration guided me to a career in chemoinformatics. I will then discuss my contributions to some enduring research themes, including chemical representation, de novo design, and applications of evolutionary methods, and will highlight their development from my PhD and early post-doc work to current research.

Session A: Artificial Intelligence, Machine Learning, and QSAR

A01: GENEOnet: Revolutionizing Drug Discovery with the Most Accurate Protein Binding Pocket Detection Using GENEOs

Talarico, Carmine[1]; Bocchi, Giovanni[2]; Gratteri, Carmen[3]; Frosini, Patrizio[4]; Pedretti, Alessandro[2]; Palermo, Gianluca[5]; Gadioli, Davide[5]; Lunghini, Filippo[1]; Biswas, Akash Deep[1]; Stouten, Pieter F.W.[1]; Beccari, Andrea R.[1]; Fava, Anna[1]; Micheletti, Alessandra[2]

[1] *EXSCALATE – Dompé farmaceutici S.p.A., Napoli, Italy*

[2] *Università degli Studi di Milano, Italy*

[3] *LIGHT s.c.a.r.l, Brescia, Italy*

[4] *Università degli Studi di Bologna, Italy*

[5] *Politecnico di Milano, Italy*

Structure-based virtual screening techniques, such as molecular docking, depend on accurately identifying and calculating binding pockets to efficiently search for potential ligands. In this paper, we present GENEOnet, a machine learning model for protein pocket detection, which leverages Group Equivariant Non-Expansive Operators (GENEOs)[1]. These operators improve model transparency and facilitate the integration of domain-specific knowledge. Unlike other approaches, GENEOnet employs fewer parameters while enhancing interpretability. It analyzes the empty spaces within proteins by converting them into a 3D grid of uniform blocks, or "voxels." These potential pockets are then scored and ranked based on their druggability. Experimental results demonstrate that GENEOnet performs robustly, even with relatively small training datasets, and outperforms state-of-the-art methods across several metrics. Notably, GENEOnet achieves an H_1 score of 0.794—indicating the likelihood that the top-ranked pocket is the correct one—compared to 0.728 for the next best performer, P2Rank.[2]

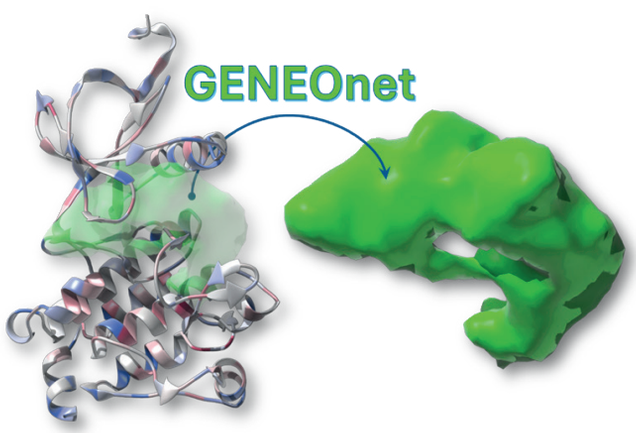


***Figure 1**: GENEOnet pocket generation from proteins' 3D structure*

References

1. Bergomi, M. G., et al., Towards a topological–geometrical theory of group equivariant non-expansive operators for data analysis and machine learning. Nature Machine Intelligence, **2019**, 1, 423–433, https://doi.org/10.1038/s42256-019-0087-3
2. Bocchi, G., et al., GENEOnet: A new machine learning paradigm based on Group Equivariant Non-Expansive Operators. An application to protein pocket detection. arXiv., **2022**, https://doi.org/10.48550/arXiv.2202.00451

A02: The future of computational chemistry: AI Target-Ligand Co-Folding?

Christian Tyrchan[1], Alessandro Tibo[2], Özge Yoluk[3], Gustav Olanders[1], Eva Nittinger[1]

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

[2]Molecular AI, Discovery Sciences, R&D, AstraZeneca, Gothenburg, Sweden

[3]Hit Discovery, Discovery Sciences, R&D, AstraZeneca, Gothenburg, Sweden

Proteins are integral to cellular functions and represent key targets in drug development. The lengthy journey from initial concept to clinical candidate in pre-clinical drug discovery relies heavily on optimizing the Design, Make, Test, and Analyze (DMTA) cycle. This iterative cycle is vital for advancing structure-based design, facilitating the development of hits towards clinical candidates. One major hurdle in this process is determining the binding modes of potential hits. Traditionally, this task has been accomplished through labor-intensive methods such as X-ray crystallography, NMR spectroscopy, and cryo-EM, which involve significant time and resources.

Emerging target-ligand co-folding methods provide a promising alternative [1-3]. These methods, evolved alongside advancements like AlphaFold3, which accommodates ligand placement and non-protein complexes, are set to potentially revolutionize our understanding of dynamic protein-ligand interactions. Unlike standard AI protein folding approaches that often produce a single solution, co-folding methods seem to offer flexibility through conformational ensembles. However, it is crucial to ensure that these ensembles are biologically meaningful, as research suggests that they may not always capture relevant conformations or binding modes [4].

This study evaluates the performance of novel protein-ligand co-folding methods (Boltz-1 [1], NeuralPLexer [2], and RoseTTAFold All-Atom [3]) in predicting and analyzing a dataset of orthosteric-allosteric ligand complexes. We aim to shed light on their predictive capabilities and address challenges highlighted by initiatives like PoseBusters [5], which stress the importance of adhering to physical-chemical constraints.

The findings suggest that while co-folding methods hold promise, their current predictive accuracy largely depends on the training data. Further room for improvement is indicated by capturing binding modes that adhere to essential chemical and physical principles, such as angles, bond length, and non steric clashes among others.

References

1. Wohlwend J, Corso G, Passaro S, Reveiz M, Leidal K, Swiderski W, Portnoi T, Chinn I, Silterra J, Jaakkola T, Barzilay R. Boltz-1 Democratizing Biomolecular Interaction Modeling. bioRxiv [Preprint]
2. Qiao, Z., Nie, W., Vahdat, A., Miller, T. F. & Anandkumar, A. State-specific protein–ligand complex structure prediction with a multiscale deep generative model. Nat Mach Intell 6, 195–208 (2024).
3. Krishna, R. et al. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. Science (1979) 384, (2024)
4. Olanders G, Testa G, Tibo A, Nittinger E, Tyrchan C. Challenge for Deep Learning: Protein Structure Prediction of Ligand-Induced Conformational Changes at Allosteric and Orthosteric Sites. J Chem Inf Model., 64(22), 8481-8494 (2024)
5. Buttenschoen M, Morris GM, Deane CM. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. Chem Sci., 15(9), 3130-3139 (2024)

A03: Drugging the undruggable: A highly accurate method for detecting and ranking cryptic pockets

Neha Vithani, She Zhang, Phu Tang, A. Geoffrey Skillman, David N. LeBard

OpenEye, Cadence Molecular Sciences, Santa Fe, NM, USA

In certain circumstances, proteins involved in the most prolific life-threatening diseases remained elusive as therapeutic targets for decades simply because a binding pocket for a molecular inhibitor could not be found. One classic example is the KRAS protein, which is a GTPase involved in more than 25% of all human cancers, yet persisted as an undruggable target for over 30 years despite immense efforts by academic and industrial researchers to find a viable pocket. A breakthrough for developing KRAS therapies occurred in 2013 from a covalent fragment screen that uncovered the so-called Switch-II pocket as a possible site for therapeutic design. This discovery has since led to two FDA-approved drugs for $KRAS^{G12C}$ and several more compounds in advanced clinical trials for other mutants. To help expand the druggable proteome to include difficult-to-drug targets like KRAS, we present an automated computational workflow that allows anyone to validate a protein for its ligandability. Requiring only a protein structure (X-Ray, Cryo-EM, Al-generated, etc.), our workflow uses Weighted Ensemble path sampling along intrinsic normal modes to generate rare protein conformations that are analyzed with a set of Markov state models to identify residues that cooperatively form pockets. Furthermore, our workflow also ranks detected pockets by their ligandability using a neural network model estimator of the potential affinity of a pocket. The automated workflow is performed entirely with elastic cloud computing inside the Orion® platform and can uncover pockets that form either by conformational selection of rare protein states or through an induced-fit mechanism by a probe molecule. In this work, we present a proof-of-concept study of $KRAS^{G12D}$ to illustrate that our methodology can predict known cryptic pockets, including the Switch-II pocket that remained hidden for decades. Finally, a validation of our automated protein sampling and cryptic pocket detection workflow on a set of 19 proteins that represent 24 unique pocket types will also be presented.

A04: Improving Target-Adverse Event Association Prediction by Mitigating Topological Imbalance in Knowledge Graphs

Terence Egbelo[1], Zeyneb Kurt[1], Charlie Jeynes[2], Mike Bodkin[3], Val Gillet[1]

[1]*Information School, University of Sheffield, UK*

[2]*In silico R&D, Evotec UK*

[3]*School of Life Sciences, University of Dundee, UK*

The drug discovery process yields high rates of failure at the clinical stage in large part because of adverse events (AEs) triggered by modulation of the intended protein targets as well as other proteins. These can force the termination of R&D programmes even after substantial investment.

Establishing connections between molecular targets, test compounds, and clinically observed adverse events (AEs) remains a persistent challenge. It has spurred milestone work such as that by Kuhn et al (2013). They developed a statistical framework for establishing significant associations between targets and AEs from existing evidence. The work provided investigators with the first computational tool for illuminating the safety profiles of candidate therapeutic targets.

A biomedical knowledge graph (KG) integrates domain knowledge that is representable in network form. Core KG components typically include protein interaction networks and biomedical ontologies like the GO. The inference task of predicting new relations in the KG based on its existing content is known as KG completion. The present work tackles the prediction of new target-AE associations as KG completion using a large-scale biomedical knowledge graph.

Inspired by biomedical KG literature precedents (Fu et al 2016, Himmelstein et al 2017), the KG completion approach used in this work leverages interpretable, "metapath"-based predictive patterns, thereby retaining a direct reference to domain semantics. This is in contrast to presently popular "black-box" deep learning-based KG completion methods as reviewed e.g. by Bonner et al (2022a).

This work also showcases a novel approach to handle topological bias in KGs (extensively evidenced by Bonner et al (2022b). This bias arises from the tendency of high-degree nodes in the KG to be overrepresented in existing associations of the type of interest (here target-AE) and leads to poor learning of associations featuring low-degree nodes. In the present problem, such bias can present a major drawback, as low degree indicates a less-studied target where good inference of safety is most needed. Our bias handling method, based on TF-IDF (Spärck Jones 1972) which was originally developed for information retrieval applications, transforms node sparsity into a useful signal when learning AE associations of sparsely connected targets.

Our procedure was found to produce significantly better predictive performance on the most sparsely connected targets (accuracy improvement of ~0.15 on the bottom 15% targets by no. of AE associations) than an existing conventional alternative, the Degree-Weighted Path Count, or DWPC, first reported by Himmelstein et al (2015) and thereafter used e.g. by Himmelstein et al (2017) and Binder et al (2022).

Lastly, we use the metapath-based KG completion approach, as improved by the above modification, to demonstrate prediction interpretability, including in cases where the unmodified and DWPC alternatives make errors.

References

1. Kuhn, M., Al Banchaabouchi, M., Campillos, M., Jensen, L. J., Gross, C., Gavin, A. C., & Bork, P. (**2013**). Systematic identification of proteins that elicit drug side effects. Molecular systems biology, 9(1), 663.
2. Galletti, C., Bota, P. M., Oliva, B., & Fernandez-Fuentes, N. (**2021**). Mining drug–target and drug–adverse drug reaction databases to identify target–adverse drug reaction relationships. Database, 2021, baab068.

3. Ioannidis, V. N., Song, X., Manchanda, S., Li, M., Pan, X., Zheng, D., ... & Karypis, G. (**2020**). DRKG - Drug Repurposing Knowledge Graph for Covid-19. https://github.com/gnn4dr/DRKG
4. Bonner, S., Barrett, I. P., Ye, C., Swiers, R., Engkvist, O., Hoyt, C. T., & Hamilton, W. L. (**2022**). Understanding the performance of knowledge graph embeddings in drug discovery. Artificial Intelligence in the Life Sciences, 2, 100036.
5. Bonner, S., Kirik, U., Engkvist, O., Tang, J., & Barrett, I. P. (**2022**). Implications of topological imbalance for representation learning on biomedical knowledge graphs. Briefings in bioinformatics, 23(5), bbac279.
6. Sparck Jones, K. (**1972**). A statistical interpretation of term specificity and its application in retrieval. Journal of documentation, 28(1), 11-21.
7. Himmelstein, D. S., & Baranzini, S. E. (**2015**). Heterogeneous network edge prediction: a data integration approach to prioritize disease-associated genes. PLoS computational biology, 11(7), e1004259.
8. Himmelstein, D. S., Lizee, A., Hessler, C., Brueggeman, L., Chen, S. L., Hadley, D., ... & Baranzini, S. E. (**2017**). Systematic integration of biomedical knowledge prioritizes drugs for repurposing. Elife, 6, e26726.
9. Binder, J., Ursu, O., Bologa, C., Jiang, S., Maphis, N., Dadras, S., ... & Oprea, T. I. (**2022**). Machine learning prediction and tau-based screening identifies potential Alzheimer's disease genes relevant to immunity. Communications Biology, 5(1), 125.

A05: Refined ADME Profiles for ATC Drug Classes

Luca Menestrina[1], Raquel Parrondo-Pizarro[1,2], Ismael Gómez[1], Ricard García-Serna[1], Scott Boyer[1], Jordi Mestres[1,2]

[1] *Chemotargets, S.L., Barcelona, Catalonia, Spain*

[2] *Institut de Química Computacional i Catàlisi, Facultat de Ciències, Universitat de Girona, Catalonia, Spain*

Drug discovery and development is a lengthy, costly, and risky process [1], making it essential to better understand and anticipate the pharmacokinetic properties of a drug, that is, its absorption, distribution, metabolism and excretion (ADME) profile. With rising amounts of data available, the advent of Artificial Intelligence (AI) methods is having a significant impact in drug discovery [2]. In generative AI drugs discovery programs, accurate Machine Learning (ML) models of ADME properties are essential for efficiently optimizing large generated molecular libraries and reducing the need for costly, time-consuming preclinical experiments [3].

The main objective of the study was to investigate the physicochemical and ADME property profiles and define the likely acceptable value distributions across indication-specific drug classes. This work introduces our efforts to build an automatic pipeline for the construction and evaluation of ML models to predict physicochemical and ADME properties of small molecules [4].

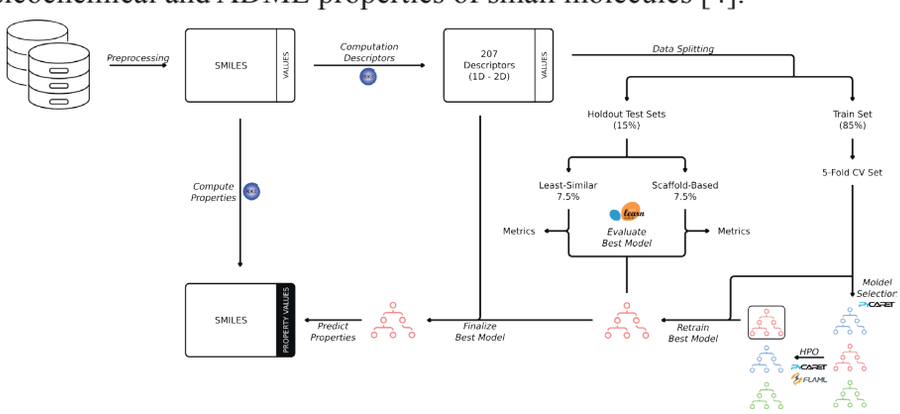


***Figure 1**: The PreCogs workflow for building machine learning models*

Although assessing the properties of drug candidates is an essential step in the drug discovery process, our analysis revealed that the availability of ADME properties for marketed drugs is limited, with coverage ranging from 3.9% to 53.2%. Overall, the generally low and uneven levels of experimental data completeness observed across ATC drug classes raises the need for constructing property models that provide full coverage through predicted data.

Those properties exhibiting broader experimental data coverage revealed less variability between experimental and predicted data distributions, underscoring the importance of having access to larger, wider, and high-quality experimental datasets to ensure that predictions derived from computational models align with the expected properties within drug classes. In addition, the comprehensive analysis of experimental and predicted data for 15 physicochemical and ADME properties across 14 different ATC drug classes reveals several trends that highlight the specificities of each class. Improving our ability to predict properties for drug classes can have an impact in multiple aspects of current generative chemistry trends, from narrowing the properties for certain routes of administration to directing the property profiles towards the intended pharmacological effects and therapeutic indications.

References

1. Wouters, O.J., et al., Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. JAMA, **2020**, 323, 9, 844-853, doi:10.1001/jama.2020.1166.
2. Gangwal, A., et al., Unleashing the Power of Generative AI in Drug Discovery. Drug Discov. Today, **2024**, 29, 6, 103992, doi:10.1016/j.drudis.2024.103992.

3. Danishuddin, et al., A Decade of Machine Learning-Based Predictive Models for Human Pharmacokinetics: Advances and Challenges. Drug Discov. Today, **2022**, 27, 2, 529-537, doi:10.1016/j.drudis.2021.09.013.
4. Menestrina, L., et al., Refined ADME Profiles for ATC Drug Classes. Pharmaceutics, accepted for publication.

A06: ADMET modelling with a quantum chemically pretrained Graphormer. Beyond benchmarking results

Alessio Fallani[1, 3], Ramil Nugmanov[1], Jose Arjona-Medina[1], Jörg Kurt Wegner[2], Alexandre Tkatchenko[3], and Kostiantyn Chernichenko[1]

[1] *Drug Discovery Data Sciences, Janssen Pharmaceutica NV, Beerse, Belgium*

[2] *Johnson & Johnson Innovative Medicine, Cambridge, USA*

[3] *Department of Physics and Materials Science, University of Luxembourg, Luxembourg City, Luxembourg*

We evaluate the impact of pretraining Graph Transformer1 architectures on atom-level quantum-mechanical features for the modeling of absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties of drug-like compounds.2 We compare this pretraining strategy with two others: one based on molecular quantum properties (specifically the HOMO-LUMO gap) and one using a self-supervised atom masking technique. After fine-tuning on a TDC ADMET datasets and a larger proprietary human liver microsome clearance dataset, the accuracy improvement between different models positioned atomic quantum mechanical properties as superior for pretraining.

We additionally analyzed the learned representations from differently pretrained models using several methods and found that the respective metrics varied significantly for the models that were similarly performant in terms of accuracy on a TDC dataset. For example, models pretrained on atomic quantum mechanical properties capture more low-frequency Laplacian eigenmodes of the input graph via the attention weights and produce better representations of atomic environments within the molecule (Figure 1). Learned representation analysis, including a newly introduced analysis of Graphormer's Attention Rollout Matrix, can guide pretraining strategy selection, as corroborated by a performance evaluation on a larger internal dataset.

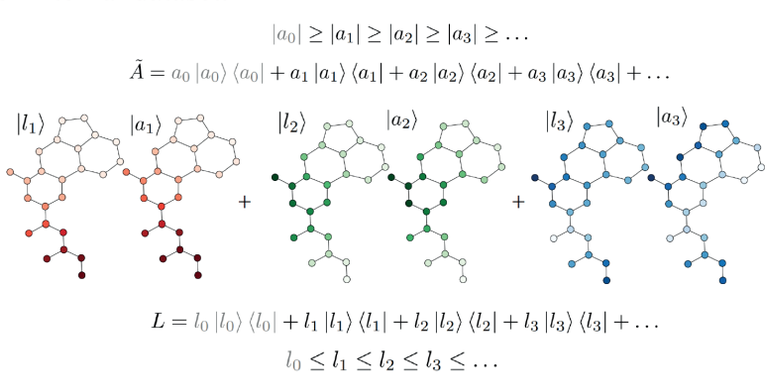


Figure 1*: Comparison between the most relevant eigenvectors of the Attention Rollout matrix from a model pretrained on atom-level QM properties and the low-frequency eigenvectors of the graph Laplacian associated to the molecular structure*

References

1. Nugmanov, R., et al., Bidirectional Graphormer for Reactivity Understanding: Neural Network Trained to Reaction Atom-to-Atom Mapping Task. J. Chem. Inf. Model. **2022**, 62, 14, 3307–3315, DOI: 10.1021/acs.jcim.2c00344
2. Fallani, A., et al., Pretraining Graph Transformers with Atom-in-a-Molecule Quantum Properties for Improved ADMET Modeling. Journal of Cheminformatics, *accepted manuscript,* DOI: 10.1186/s13321-025-00970-0

A07: Towards experiment-aware bioactivity model(er)s

Linde Schoenmaker[1], Enzo Sastrokarijo, Joost Beltman[1], Laura H. Heitman[1], Gerard J.P. van Westen[1], and Willem Jespers[1]

Leiden Academic Centre of Drug Research, Leiden University, The Netherlands

Protein-ligand interaction prediction with proteochemometric (PCM) models can provide valuable insights during early drug discovery and chemical safety assessment. These models have benefitted from the large amount of data available in large bioactivity databases. However, an issue that is often overlooked when using this data, is the large diversity in the biological assays present. The effect of perturbing a protein can be measured in various ways and this can influence the outcome.[1,2] Yet, currently there is a lack of standardized, specific assay metadata. Whilst this could help increase understanding of biological underpinnings, improve data curation and lead to better models.

To make use of the existing information on the biological context, this study set out to create and validate multiple assay descriptors and test their use in protein-ligand binding interaction models. Dimensionality reduction of embedded free text assay descriptions from ChEMBL showed that the BioBERT embeddings capture relevant features.[3] Additionally, clustering of these embedded descriptions groups the assays in a way that enriches purity, matches manually categorized assays and yields sensible topic keywords. From ligand-protein combinations measured in multiple assays, it can be concluded that the deviation between different assays in general is higher than the deviation within assay categories, with a mean absolute error of 0.83 and 0.66, respectively. Incorporating this biological context into the PCM models in the form of embeddings improved performance. Conversely, using simpler methods such as bag-of-words no improvement was seen. In addition, this novel method for assay categorization facilitates data curation and provides a useful overview of the biological context of studied targets.

Bibliography

1. Landrum, G. A. & Riniker, S. Combining IC50 or Ki Values from Different Sources Is a Source of Significant Noise. *J. Chem. Inf. Model.* **64**, 1560–1567 (2024).
2. Kenakin, T. & Christopoulos, A. Signalling bias in new drug discovery: detection, quantification and therapeutic impact. *Nat. Rev. Drug Discov.* **12**, 205–216 (2013).
3. Lee, J. *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).

A08: OpenMMDL – Simplifying the Complex: Building, Simulating, and Analyzing Protein–Ligand Systems in OpenMM

V Talagayev[1], C Yu[1], N.P. Doering[1], L Obendorf[2], K. Denzinger[1], K. Puls[1], K. Lam[1], S. Liu[1], C.A. Wolf[1], T. Noonan[1], M. Breznik[1], P. Knaus[2], G. Wolber[1]

[1] *Molecular Design Group, Institute of Pharmacy, Freie Universität Berlin, Germany*

[2] *Signal Transduction Group, Institute of Biochemistry, Freie Universität Berlin, Germany*

Molecular dynamics (MD) simulations are an important tool in scientific research. Their ability to provide insight into the behavior of protein-ligand complexes facilitates the rational design of novel drugs. OpenMM[1] is a powerful software package that allows the dynamics of molecular systems to be simulated with exceptional efficiency. The open-source nature of OpenMM, together with its support for multiple platforms, has made it an important tool in the field of computational drug design. One of the main difficulties in using OpenMM for protein-ligand complexes is the difficulty of setting up the simulation, which requires expertise in configuring force fields, selecting appropriate parameters and managing simulation protocols. In this work, we present a workflow called OpenMMDL[2] made for easy setup for OpenMM MD simulations of protein-ligand complexes and analysis of interactions during the simulation.

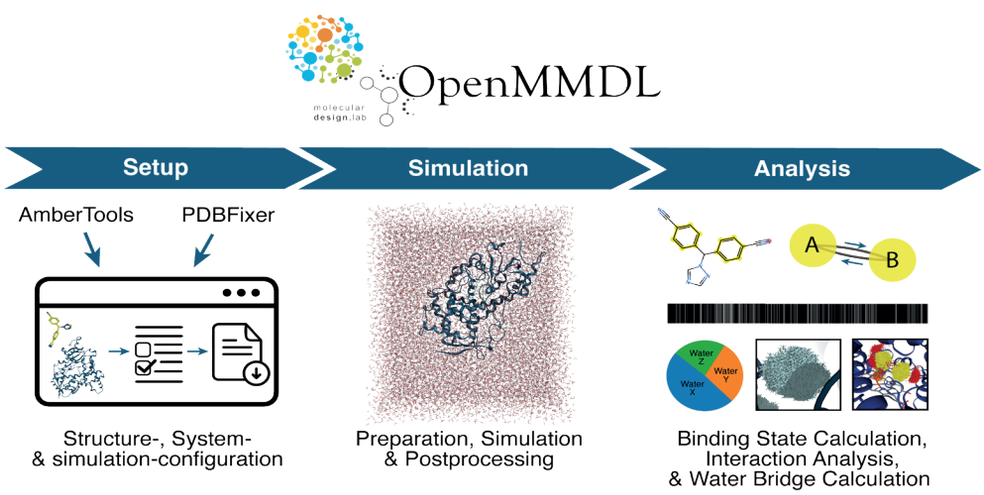


***Figure 1**: The OpenMMDL workflow consisting of three parts, first OpenMMDL Setup performing the preparation of the simulation script, OpenMMDL Simulation responsible for performing the simulation and OpenMMDL Analysis performing the analysis.*

The first part of the workflow is OpenMMDL Setup, an easy-to-use interface that allows the user to use their files as input for the simulation. The interface allows the user to modify all the necessary parameters of the simulation, thus allowing the generation of a simulation script that is suited to the user. The second part is called OpenMMDL Simulation and performs the MD simulation and post-processing, providing the user with the output of the MD simulation, which can be used directly for further analysis. The final part is OpenMMDL Analysis and allows the user to view the protein-ligand interactions during the simulation, including the most common binding modes present in the simulation, as well as the transitions of the binding modes.

References

1. Eastman, P., et al., OpenMM 8: Molecular Dynamics Simulation with Machine Learning Potentials. JPC B., **2023**, 128, 1, 109–116, 10.1021/acs.jpcb.3c06662
2. Talagayev, V., et al., OpenMMDL - Simplifying the Complex: Building, Simulating, and Analyzing Protein–Ligand Systems in OpenMM. JCIM., **2025**, 65, 4, 1967–1978, 10.1021/acs.jcim.4c02158

A09: High-accuracy QM in life sciences: From drug properties to binding modes

Christoph Riplinger[1], Michael Edmund Beck[2,3], Frank Neese[4]

[1] *FAccTs GmbH, Cologne, Germany*

[2] *Bayer AG, Research & Development, Crop Science, Monheim Germany*

[3] *Technische Universität Dortmund, Fakultät für Chemie und Chemische Biologie, Germany*

[4] *Department of Molecular Theory and Spectroscopy, Max-Planck Institut für Kohlenforschung, Mülheim an der Ruhr, Germany*

Quantum chemical (QM) calculations for life science applications have traditionally been constrained to small model systems, or computationally inexpensive but lower-accuracy methods, due to computational limitations. Most biomolecular modeling approaches rely on empirical parametrization and thus, at least indirectly, on some sort of training, which introduces biases and restricts their applicability. In particular, crucial drug targets such as RNA and metalloproteins often fall outside the scope of conventional methods, requiring extensive parametrization efforts.

Recent advances in quantum chemistry, alongside improvements in computing hardware, have enabled the application of high-accuracy QM methods – such as hybrid density functional theory (DFT) and coupled cluster (CCSD(T)) – to biomolecular systems.[1, 2, 3] These developments now allow for the accurate simulation of ligands in solution as well as in complex biological environments, including their chemical stability, potential degradation pathways, and photochemical reactivity. Moreover, the integration of QM calculations with machine learning models further enhances their predictive power, opening new avenues for drug discovery.

In this talk, we illustrate how state-of-the-art QM approaches are applied across various stages of drug development, from predicting photochemical degradation to elucidating binding mechanisms in metal-containing proteins.

References

1. Riplinger, C., et al., Sparse maps – A systematic infrastructure for reduced-scaling electronic structure methods. II. Linear scaling domain based pair natural orbital coupled cluster theory. J. Chem. Phys., **2016**, 144, 2, 024109, DOI: 10.1063/1.4939030
2. Beck, M.E., et al., Unraveling individual host–guest interactions in molecular recognition from first principles quantum mechanics: Insights into the nature of nicotinic acetylcholine receptor agonist binding, J. Comp. Chem., **2021**, 42, 293-302, DOI: 10.1002/jcc.26454
3. Kaltschnee, L., et al., Parahydrogen-enhanced magnetic resonance identification of intermediates in [Fe]-hydrogenase catalysis. Nature Catalysis, **2024**, 7, 1417–1429, DOI: 10.1038/s41929-024-01262-w

A10: Quantifying the Unknown: A Comparative Study of Deep Learning-Based Uncertainty Quantification Methods for Bioactivity Assessment

Bola Khalil[1,2], Kajetan Schweighofer[3], Natalia Dyubankova[1],

Günter Klambauer[3], Gerard van Westen [2], Herman van Vlijmen[1,2]

[1] *In silico discovery (ISD), Johnson & Johnson, Beerse, Belgium*

[2] *Division of Medicinal Chemistry, Leiden University, Netherlands*

[3] *ELLIS Unit Linz and LIT AI Lab, Institute for Machine Learning, Johannes Kepler University, Linz, Austria*

Reliable uncertainty quantification (UQ) is essential for bioactivity modeling, as it enables better risk assessment and trust in predictions. We benchmark six UQ approaches on large-scale xC_{50} and K_x bioactivity data from the Papyrus++ dataset. To ensure data quality, we applied molecular and assay-level filtering, using Ankh protein embeddings and ECFP for compound representations. Models were evaluated across stratified and scaffold-cluster splits for robustness. We compared Probabilistic Neural Networks (PNN), Deep Ensembles, Monte Carlo Dropout, Evidential Regression, and two novel hybrids: an ensemble of evidential models (EOE_{10}) and evidential Monte Carlo Dropout (EMC_{10}). EOE_{10} consistently outperformed all others in accuracy, even surpassing $Ensemble_{100}$, while excelling in interval-based metrics (CRPS, coverage). NLL remained elevated for EOE_{10}, possibly due to evidential distribution assumptions. RMSE-rejection analysis highlighted dataset-specific strengths: excelled on xC_{50}, while EOE_{10} was superior on K_x. Our results show UQ performance is task-dependent, and that hybrid evidential ensembles offer promising trade-offs between performance and efficiency.

Keywords: Uncertainty Quantification, Deep Learning, Drug Discovery, Bioactivity Assessment, Ensembles, Evidential Deep Learning, Monte Carlo Dropout.

References:

1. Mervin et al. Drug Discov. Today 2021, 26(2), 474–489. https://doi.org/10.1016/j.drudis.2020.11.027.
2. Béquignon et al. J. Cheminform. 2023, 15(1), 3. https://doi.org/10.1186/s13321-022-00672-x.

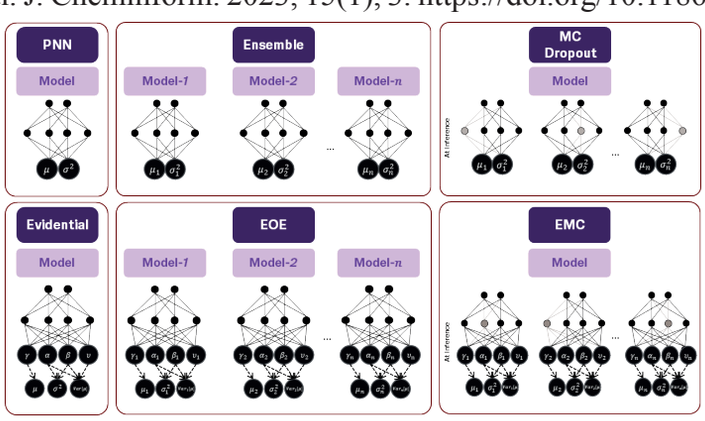


Figure: Uncertainty Quantification Approaches: *Ensemble models aggregate outputs from multiple PNN models. MC Dropout estimates uncertainty through stochastic forward passes. Evidential models output distribution parameters (α, β, γ, υ) to derive mean, aleatoric, and epistemic uncertainty. Hybrid approaches (EOE and EMC) integrate ensemble diversity or mc dropout stochasticity with evidential uncertainty estimation.*

A11: Exploration of Synthesis Space by Application of Evolutionary Strategies

Emma Armstrong[1], Michael Bieler[2], Val Gillet[1]

[1] *Information School, University of Sheffield, UK*

[2] *Boehringer Ingelheim, Biberach an der Riß, Germany*

De novo drug design aims to generate novel chemical compounds computationally which satisfy given constraints. Navigating the large chemical space associated with drug-like compounds proves one of the most difficult aspects of de novo design. Advances in AI and ML techniques have aided in searching more quickly and effectively[1], however, these generative models do not necessarily account for synthetic feasibility and can produce solutions with hard to discern reasoning.

Reaction-based models apply known chemical reactions sequentially to building blocks resulting in novel molecule designs[2,3]. By exploring chemical space through reactions, synthetic feasibility becomes intrinsic to the model and the final solutions present synthesis pathways alongside the molecular designs. Our proposed algorithm is a forward synthesis method that takes building blocks, known reactions and optimization criteria to produce a set of compounds and their synthesis routes. We combine evolutionary methods with reaction-based de novo design to evolve molecules that fit the design constraints while ensuring synthetic accessibility.

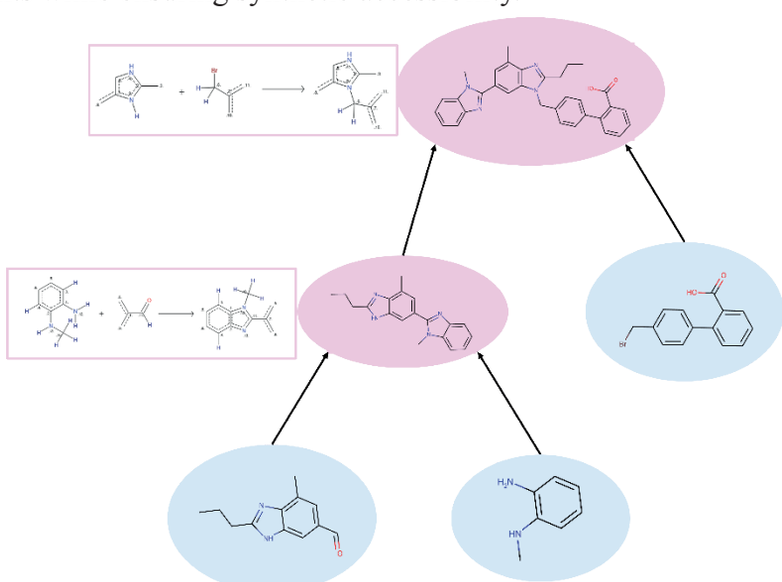


***Figure 1**: Synthesis tree for Telmisartan.*

Synthesis pathways are represented using tree data structure[4,5]. Input molecules (starting materials) are input to in-silico reactions, the output of which can then be input to further reactions, thereby building tree structures of compounds and reactions. Structural information of the building blocks and intermediate products, as well as reaction information, are stored within the nodes of the tree. Therefore, by traversing a tree and evaluating the result, a single product molecule is generated that can be scored using the desired fitness function.

An evolutionary strategy has been implemented to find optimal compounds from an initial population of candidate solutions. Bespoke genetic operators have been developed including crossover, mutation and grow operators to direct the population to higher scoring, synthetically feasible solutions. Compatibility between reactants and reactions within population trees limits the application of the operators to ensure all solutions present valid synthesis routes. A local optimiser has been implemented to aid the exploration of chemical space, individually optimising leaf nodes for a given sequence of reactions found within the population.

Our method has been applied to rediscovery tasks of known drug structures and synthesis pathways. In addition to the expected solution, other high scoring alternatives are also presented.

References

1. Tang, Y., et al., Recent Advances in Automated Structure-Based De Novo Drug Design. J. Chem. Inf. Model., **2024**, 64, 6, 1794-1805, 10.1021/acs.jcim.4c00247
2. Ding, Y., et al., Exploring Chemical Reaction Space with Machine Learning Models: Representation and Feature Perspective. J. Chem. Inf. Model., **2024**, 64, 8, 2955-2970, 10.1021/acs.jcim.4c00004
3. Patel, H., et al., Knowledge-Based Approach to de Novo Design Using Reaction Vectors. J. Chem. Inf. Model., **2009**, 49, 5, 1163-1184, 10.1021/ci800413m
4. Daeyaert, F. and Deem, M.W., A Pareto Algorithm for Efficient De Novo Design of Multi-Functional Molecules. Mol. Inf., **2017**, 36. 10.1002/minf.201600044
5. Bradshaw, J., et al., Barking up the right tree: An approach to search over molecule synthesis DAGs. NeurIPS., **2020**, 575, 6852-6866, 10.48550/arXiv.2012.11522

A12: Visualization and Clustering of Ultra-Large Chemical Space

Johannes Kaminski, Oliver Koch

Koch Group, Institute of Pharmaceutical and Medicinal Chemistry, University of Münster, Germany

The investigation of the relationships between high-dimensional objects is a recurring task in the field of data analysis. This issue also pertains to domain of chemistry and drug discovery, where compounds are often represented as vectors with numerous dimensions and where the intricate relationships between these representations are referred to as the '*chemical space*'. Particularly in drug discovery the need for such information arises from the similar property principle, stating that chemically similar compounds should exhibit similar biological effects, and its investigation is often employed for compound library analysis and design.

A multitude of algorithms have been devised to attain this objective. One such is the Self-Organizing Map (SOM) 1 algorithm, which has been extensively utilized in drug discovery. However, there has been a decline in interest in this approach in recent years, due to the rise of other dimensionality and clustering techniques such as t-SNE 2 and UMAP 3. In our view, this is an unjustified neglect, as the approach is both very intuitive and can potentially lead to a more faithful representation of the data space, due to its regularized grid 4. Especially the Emergent-SOM variant (ESOM) 5, which utilizes up to 3 neurons per data point, offers great advantages in terms of interpretability but is computationally demanding, as the number of necessary neurons scales linearly with the number of data points.

To meet the specific needs of chemical space analysis and adopt modern computing capabilities not utilized by the previous implementations, we developed a novel tool called *SOMba*. By leveraging parallel computing, including GPU acceleration, it enables the visualization of increasingly expansive chemical libraries comprising millions of compounds. In addition to the universally applicable algorithm, SOMba also includes both a graphical user interface and an interface to the PyTorch 6 dataloader, which facilitates intuitive use with chemical information such as visual inspection of individual data points/compounds.

We demonstrate the applicability of our software for the faithful representation of compound relationships in the *ChEMBL* 7 database, comprising 1.8 million compounds projected onto a map of approx. 3.8 million neurons, the largest ESOM to our knowledge. Furthermore, we utilize the SOM algorithm to generate combined, diverse subsets of multiple ultra-large compounds with tens of millions of compounds, to showcase its use in library design applications.

References

1. Kohonen, T., Self-organized formation of topologically correct feature maps. Biological Cybernetics, **1982**, 43, 1, 59-69, 10.1007/BF00337288
2. Van der Maaten L., Hinton G., Visualizing Datat using t-SNE. Journal of Machine Learning Research, **2008**, 9, 11, 2579-2605
3. McInnes L., et al., UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. ArXiv eprints, **2018**, 10.48550/arXiv.1802.03426
4. Orlov A., et al., From High Dimensions to Human Insight: Exploring Dimensionality Reduction for Chemical Space Visualization. Molecular Informatics, **2024**, 44, 1, 10.1002/minf.202400265
5. Ultsch A., Data Mining and Knowledge Discovery with Emergent Self-Organizing Feature Maps for Multivariate Time Series. Kohonen Maps (Elservier Science), **1999**, 33-45, 10.1016/B978-044450270-4/50003-6
6. Paszke A., et al., PyTorch: an imperative style, high-performance deep learning library. Proceedings of the 33rd International Conference on Neural Information Processing Systems, **2019**, 10.48550/arXiv.1912.01703

7. Zdrazil B., et al., The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. Nucleic Acids Research, **2023**, 51, D1, D1180 - D1192, 10.1093/nar/gkad1004

A13: CACHE Challenge #1: Searching for Hit Molecules in Ultra-Large Chemical Libraries Guided By De Novo Design

P. Polishchuk, G. Minibaeva, A. Ivanova, A. Kutlushina

Institute of Molecular and Translational Medicine, Faculty of Medicine and Dentistry, Palacky University, Olomouc, Czech Republic

CACHE challenge is a scientific community effort to provide assessment of computational hit-finding experiments. Each challenge is focused on a specific protein target of biological or pharmaceutical relevance. The hits predicted by participants are experimentally validated that creates a solid ground of comparison of different approaches and pipelines. The first CACHE challenge was focused on finding binders of WD40 repeat domain of LRRK2 kinase involved in Parkinson's disease development and there were no previously known binders. The only available information was X-ray structure of the domain in its apo-form. The search space was Enamine accessible chemical space, including Enamine REAL containing 23B molecules at that time.

To efficiently explore such a large space, we suggested a two-stage pipeline. On the first stage several representative protein conformations were derived from molecular dynamic simulations molecules and potential binders were designed using CReM. On the second stage the designed molecules were used as queries to retrieve similar molecules from Enamine REAL space. Molecules on each stage were subjected to the comprehensive virtual screening protocol including consensus docking by multiple methods to several protein conformations and final re-scoring of the most promising hits with MM-GBSA method. As a result, among 82 synthesized compounds eight compounds demonstrated the binding ability with K_d = 25-117 μM. The suggested pipeline proved its ability to identify promising hits. However, the contribution of individual virtual steps could not be evaluated. Therefore, we initiated further studies to more systematically evaluate the pipeline and its individual steps.

Output of all top performing teams were recently published in a joint publication [1] and we will briefly overview some of them to compare their computational efficacy and outputs.

References

Li, F.; Ackloo, S.; Arrowsmith, C. H.; et al. CACHE Challenge #1: Targeting the WDR Domain of LRRK2, A Parkinson's Disease Associated Protein. J. Chem. Inf. Model. **2024**, 64, 22, 8521-8536, https://doi.org/10.1021/acs.jcim.4c01267.

A14: Discovery of Novel CYP19A1 Inhibitors Using Machine Learning-Driven Virtual Screening and Structure-Based Approaches

Sijie Liu [1,2], Jie Wu [3], Ya Chen [2], Clemens Alexander Wolf [1], Matthias Bureik [3], Mario Marchisio [3], Johannes Kirchmair [2], Gerhard Wolber [1]

1. Institute of Pharmacy, Freie Universität Berlin, Germany

2. Department of Pharmaceutical Sciences, Division of Pharmaceutical Chemistry, Faculty of Life Sciences, University of Vienna, Austria

3. School of Pharmaceutical Science and Technology, Faculty of Medicine, Tianjin University, China

Cytochrome P450 19A1 (CYP19A1, aromatase) is a heme-containing enzyme responsible for converting androgens to estrogens during 17β-estradiol biosynthesis. Its critical role in hormone-related diseases, including breast cancer, endometriosis, and infertility, makes CYP19A1 a key therapeutic target. However, the discovery of novel CYP19A1 inhibitors is hindered by significant challenges posed by the enzyme's conformational flexibility and complex coordination chemistry, which complicate structure-based methods such as molecular docking. These factors can lead to inaccurate binding pose predictions and reduced success in conventional virtual screening campaigns.

To overcome these challenges, we developed a machine learning (ML)-driven virtual screening workflow to predict CYP19A1 inhibition and explore a broader chemical space. ML models were trained on bioactivity data from the ChEMBL and PubChem BioAssay databases using Morgan fingerprints and RDKit-2D descriptors. By optimizing the models for early enrichment using the Boltzmann-enhanced discrimination of receiver operating characteristic (BEDROC) metric, we achieved improved performance for virtual screening compared to traditional classifiers. Our workflow combined ML predictions with molecular docking, structural clustering, and visual inspection, reducing a library of 4.6 million commercially available compounds to 1,500 candidates.

Ten compounds were selected for experimental validation based on predicted activity and structural diversity. Enzymatic assays using heterologous expression of human CYP19A1 in yeast identified seven active inhibitors with previously unknown scaffolds. Among these, compound **9**, a non-covalent inhibitor containing coumarin and imidazole substructures, demonstrated the highest potency, with an IC_{50} value of 271 ± 51 nM. Molecular dynamics simulations provided insights into its binding interactions, confirming its stable and favorable binding conformation. This study highlights the power of integrating ML and structure-based approaches to address the limitations of conventional docking, offering a robust strategy for discovering novel CYP19A1 inhibitors.

A15: SpectruMS: A Multi-modal Foundation Model for Better Generalizability on Tandem MS2 Data

Aya Abdelbaky[1], Daniel Crusius[1], Tornike Onoprishvili[2], Jui-Hung Yuan[1], Vijay Ingalalli[1], Lila Khederlarian[3], Gustavo Bremo[1], Kamen Petrov[1], Niklas Leuchtenmuller[4], Sona Chandra[1], Aurelien Duarte[2], Andreas Bender[1], Yoann Gloaguen[1]

[1] *Pangea Botanica Germany GmbH, Berlin, Germany*

[2] *Independent consultant*

[3] *Pangea Botanica Ltd, London United Kingdom*

[4] *Wilde Ventures GmbH, Düsseldorf, Germany*

Mass spectrometry (MS) is able to effectively determine the metabolite fingerprint of a biological sample, such as a plant, and hence to provide information about its constituent compounds. Although MS is considered a primary method for structure elucidation, its power is restricted by the scarcity of available annotated tandem MS data. Recent machine- and deep-learning approaches, including Graph Neural Networks, Transformers, and Graph Transformers, aimed to bridge this gap by predicting structures from spectra, and *vice versa*. Despite the competitive performance of these models, their debatable generalizability on unseen chemical spaces hampers their reliability in real-world applications.

We present a series of large transformer models and tokenizers incorporating semi-supervised learning to mitigate data scarcity, enabling better generalization across unseen molecular structures and spectra. Our state-of-the-art multi-modal foundation model along with its tokenizer are pre-trained on tens of millions of chemical compounds and unlabeled MS2 spectra data independently, leveraging robust translation between these learnt languages. Moreover, extra attention is given to chemical space exploration across the models' predictions to ensure the generalizability of our models in downstream applications such as spectra-to-structure and structure-to-spectra translation.

In order to evaluate our models' performances, we first filtered out any chemical compounds from the training data that belong to the benchmark dataset of interest. This filtering step ensures that the model is able to predict new molecular structures. Our primary models are able to generate over 90% valid molecular structures with up to 5% of the time ranking the correct prediction as the top 1 candidate, 14% and 20% in the top 10 and 50 candidates, respectively. In conclusion, our work shows that utilising the large available unlabeled data for pre-training is essential for transformer models to learn MS language, and consequently, better generalize to new chemical spaces.

A16: Leveraging institutional data to improve Large Language Model (LLM) performance

Valery Tkachenko[1], Antony Williams[2], Greg Janesch[3] and Erik "Tyler" Carr[3]

[1] *Science Data Experts, USA*

[2] *U.S. Environmental Protection Agency, Office of Research and Development, Center for Computational Toxicology and Exposure, USA*

[3] *ORAU Student Services Contractor to Center for Computational Toxicology and Exposure, Office of Research and Development, U.S. Environmental Protection Agency, USA*

Large language models (LLMs) have shown great promise in chemistry, but several challenges persist. Chemical texts contain specialized terminology, complex molecular representations (e.g., SMILES, InChI), and reaction mechanisms that are not always well-represented in the general training data. This can lead to misunderstandings or inaccuracies in the generated output including „hallucinations" by producing chemically implausible or incorrect information. One of the primary reasons is that high-quality, annotated chemical data is less abundant compared to general language data. This scarcity makes it challenging to fine-tune models effectively for chemistry-specific tasks, leading to gaps in the model's understanding. Also, the field of chemistry is constantly advancing and LLMs might become outdated quickly if not regularly updated with the latest research findings, patent data, or experimental results. Overcoming these challenges requires domain-specific training, robust validation protocols, and often hybrid approaches that combine LLMs with specialized chemical databases or simulation tools.

We will present on our latest efforts of leveraging open source LLM models by retraining and augmenting them with both publicly available and institution-specific data.

This abstract does not necessarily represent the views or policies of the U.S. Environmental Protection Agency.

Session B: New Modalities and Large Chemical Data Sets

B01: A Workflow Pairing Rational and Computational PROTACs Design Yields a Novel BRD4 Degrader

Olga Tarkhanova[1], Mykola Protopopov[1,4], Olha Semenenko[1], Anna Kapeliukha[1,4], Serhii Hlotov[1,4], Diana Alieksieieva[2], Anna Beshtynarska[2], Diana Khotinets[2], Varvara Melnyk[2], Ilona Saraieva[3], Oleh Shyshlyk[2,5], Vladyslav Stadnichenko[2], Volodymyr Brovarets[5], Petro Borysko[2], Yurii Moroz[2,4]

[1] Chemspace LLC, Kyiv, Ukraine

[2] Enamine Ltd., Kyiv, Ukraine

[3] Enamine PL sp. z o.o., Wroclaw, Poland

[4] Taras Shevchenko National University of Kyiv, Ukraine

[5] V. P. Kukhar Institute of Bioorganic Chemistry and Petrochemistry, Kyiv, Ukraine

The effectiveness of PROTACs depends on multiple interdependent factors, including target binding, linker flexibility, and ternary complex stability, making their rational design a demanding task that benefits from a structured, data-driven approach. The most promising strategy to address these challenges lies in combining computational methods with rational design, enabling efficient screening, optimization, and predictive modeling to guide PROTAC development.

This project presents a PROTACs development pipeline that was successfully applied to BRD4. The process began with fragment screening and hit expansion, identifying 15 binders, of which six were crystallized. The crystallized ligands shared two scaffolds, that were selected for further PROTACs development. Based on these scaffolds, two potential linker attachment points were identified, that were facing outside of the pocket. To ensure diverse linker attachment chemistry, 24 building blocks were synthesized to represent the BRD4 ligands. These compounds were linked to three known CRBN ligands using Enamine in-stock ligand-linker conjugates, generating a virtual library of 1,742 compounds.

To prioritize candidates for synthesis, we employed a three-layer computational strategy: an ML ensemble trained on PROTAC-DB, MolScreen paired with QSAR modeling, and PROTAC docking using ICM-Pro Software. Docking was performed only on top-ranked candidates due to its computational cost. This approach led to the selection of 182 prioritized compounds, of which 135 were successfully synthesized.

To confirm the engagement of these compounds in CRBN:PROTAC:BRD4 complex formation and its potential to facilitate proteasomal degradation in vitro, we utilized the following assays: TR-FRET to assess binary complex formation with CRBN, SPR to evaluate binary complex formation with BRD4, TR-FRET for ternary complex formation, and a cell-based TR-FRET assay to measure BRD4 degradation. Through this 4-stage assessment, we identified three PROTACs that demonstrated 23-80% degradation at 30 μM, while one showed a DC50 of 774.7 nM.

This workflow was accomplished within 6 months and demonstrates how integrating computational predictions with rational design can accelerate PROTAC discovery while reducing synthesis costs and experimental workload.

B02: Benchmarking Searching in Combinatorial Spaces with the Approved Drug Space

Modest v. Korff[1], Thomas Liphardt[2], Thomas Sander[1]

[1] *Alipheron AG, Switzerland*

[2] *Idorsia Pharmaceuticals Ltd., Switzerland*

In this study, we introduce an ultra-large combinatorial space for benchmarking search algorithms. To ensure a meaningful and representative combinatorial library, this space was constructed using approved drug molecules.

Combinatorial chemistry has revolutionized drug discovery since its inception in the 1990s. By the 2010s, combinatorial spaces had expanded significantly, becoming ultra-large combinatorial libraries (ULCLs) that increasingly influenced the drug discovery process. The growing number of molecules in these libraries has pushed the limits of enumeration, making efficient search algorithms essential.

While multiple software tools exist for similarity and substructure searches in enumerated libraries, the situation is different for ULCLs. Several companies provide similarity search tools, as summarized by Warr et al. However, searching for exact substructures—including query features such as atom lists, wildcards, and substitution patterns—has only recently been addressed by Liphardt and Sander.

A missing component in this field is a publicly available combinatorial library space for benchmarking search algorithms. To address this gap, we introduce the Approved Drug Space. This space was derived from approved drug molecules, which were fragmented using the RECAP algorithm. RECAP cleaves molecules at synthetically accessible points, generating molecular fragments that were further processed into clipped fragments as described by Wahl & Sander. The maximum number of cleavage points was limited to two, resulting in clipped fragments with one or two reaction sites.

Using this methodology, we processed 3,200 approved drug molecules extracted from the ChEMBL database (Version 35). This resulted in 259 reaction sets containing approximately 20,000 unique clipped fragments. The combinatorial assembly of these fragments generated a space of 1 billion molecules, forming an ultra-large combinatorial library derived from approved drugs.

To evaluate the utility of the Approved Drug Space, we conducted a benchmarking test using the Hyperspace algorithm to assess the recovery rate of input molecules. Hyperspace enables exhaustive substructure searches within seconds. Our findings revealed a recovery rate of 50%, with non-recovered molecules primarily consisting of condensed ring systems, sugars, and macrocycles, which lack RECAP cleavage sites. The 1,700 recovered drug molecules can serve as benchmark molecules, many of which have publicly available crystal structures, making them well-suited for structure-based search method evaluations.

Despite containing 1 billion molecules, the Approved Drug Space remains enumerable, allowing direct comparisons between combinatorial space-based search methods and conventional enumerated library search methods. The Approved Drug Space will be publicly available at the time of the conference and can already be requested from the authors.

We will present benchmarking results from both the Hyperspace algorithm and the newly developed Pharos-3D algorithm. Pharos-3D is an advanced technology for screening ultra-large combinatorial spaces, identifying molecules with similar shapes and pharmacophore profiles to a given query structure. Through these two case studies—topological search (Hyperspace) and pharmacophore-based search (Pharos-3D)—we demonstrate that the Approved Drug Space is a valuable benchmarking tool for evaluating search algorithms in the rapidly expanding landscape of combinatorial libraries.

References

1. Warr, W. A., et al. (**2022**): Exploration of ultralarge compound collections for drug discovery. Journal of Chemical Information and Modeling, 62.9: 2021-2034.

2. Liphardt, T., & Sander, T. (**2023**): Fast Substructure Search in Combinatorial Library Spaces. Journal of Chemical Information and Modeling, 63.16 5133-5141.
3. Wahl, J, & Thomas S. (**2022**): Fully automated creation of virtual chemical fragment spaces using the open-source library OpenChemLib. Journal of Chemical Information and Modeling, 62.9, 2202-2211.

B03: COCONUT 2.0: A Comprehensive Improved Open Database for Natural Products Research

Venkata Chandrasekhar, Kohulan Rajan, Sagar Kanakam, Nisha Sharma, Viktor Weißenborn, Jonas Schaub, and Christoph Steinbeck*

Institute for Inorganic and Analytical Chemistry, Friedrich Schiller University Jena, Germany

Natural products (NPs) are small molecules made by living organisms, offering notable benefits to pharmaceutical research and various industries due to their biological attributes. Despite multiple databases, no unified digital collection has gathered NP structures in a single portal. The COlleCtion of Open Natural prodUcTs (COCONUT) presents a comprehensive, openly accessible repository, unifying data from numerous public sources. Initially introduced in 2021, COCONUT has evolved significantly to version 2.0, refining its underlying framework, enhancing data accuracy, and delivering an improved user interface.

COCONUT 2.0 [1] holds more than 695,000 unique NP structures compiled from 63 public sources, offering core details such as chemical structures, synonyms, organism data, worldwide distribution, and literature references. It also includes calculated molecular properties and chemical classifications. Key version 2.0 improvements feature a complete rebuild of the web and backend systems, more substantial curation and standardization processes, new tools for community-based editing and data entry, upgraded structure, substructure, and similarity searches, a REST API, multi-format data downloads, and comprehensive provenance tracking plus detailed logging.

The platform's architecture employs microservices and Docker containers to ensure scalability and maintainability. A PostgreSQL database integrated with RDKit enables efficient chemical queries. By functioning as a central, community-curated repository, COCONUT advances studies in drug discovery, cheminformatics, and a wide range of domains that rely on the diverse chemistry of natural products. Thus, it paves the way for innovative investigative approaches.

COCONUT 2.0 can be freely accessed at https://coconut.naturalproducts.net, with its open-source code and datasets hosted on GitHub and Zenodo. These resources support transparency, reuse, and alignment with FAIR data principles.

References

1. Chandrasekhar V, Rajan K, Kanakam SRS, Sharma N, Weißenborn V, Schaub J, Steinbeck C: COCONUT 2.0: a comprehensive overhaul and curation of the collection of open natural products database. Nucleic Acids Res **2025**, 53:D634–D643.

B04: StrAcTable – Combining Structural and Bioactivity Data with Atomic Precision for Protein-Ligand Complex Datasets

Torben Gutermuth[1],Emanuel Ehmki[1], Florian Flachsenberg[1], Patrick Penner[1], Tobias Harren[1], Sophia Hönig[1,2], Matthias Rarey[1]

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

[2] BioSolveIT GmbH, Sankt Augustin, Germany

Predicting of protein-ligand binding affinity is known to be challenging but crucial for every drug discovery project. Irrespective of the method employed, the importance of suitable datasets, both in size and quality is uncontested. They are the bedrock of method development and accurate performance estimation. Existing datasets that combine affinity and structural data are often built through a partially manual process[1,2,3], leading to difficulties in keeping up with the ever-growing literature data, human errors, and unclear decision criteria. As an alternative, the recently developed BioChemGraph[4] automatically crosslinks activity data between ChEMBL and PDB using Uniprot Identifiers and InChI Keys, but this approach underestimates the amount of data that could be connected.

To address this aim, we developed ActivityFinder, which implements a novel workflow to crosslink bioactivity data to structural data, such as ChEMBL and PDB. It utilizes the ATOM and SEQRES sequences and the SEQADV records to align the sequences in a given PDB file to all sequences in ChEMBL. This approach precisely describes how well sequences in the respective databases match, records structural differences in the binding site of the investigated small molecule, and even matches identically mutated bioactivity and structural data. Similarly, multiple molecule matching levels are utilized to crosslink found small molecules in ChEMBL and PDB, connecting perfect matches and, for example, matches with unspecified stereocenters. While developed for PDB and ChEMBL, ActivityFinder can query structures outside of PDB – for example, in proprietary databases – for data in ChEMBL. Furthermore, the workflow can be adjusted to other bioactivity databases.

Utilizing this work, we present StrAcTable, an automatically constructed dataset of over 18000 unique PDB-ligand combinations with at least one recorded activity. It goes beyond identifier matching and allows finding sequentially similar targets (e.g., different organisms) or bioactivity data of racemic mixtures while granularly annotating all differences. Any mutations within the protein binding site are annotated, as well as detailed data on the structure quality, the ligands' embedding in respective electron densities, potential problems or anomalies with specified ligands, such as incompletely modeled or covalently bound ligands, and molecular descriptors. Besides ActivityFinder, StrAcTable is based on two other software components, LigandFinder[5] and StructureProfiler[6]. These methods enable users to process all available data automatically and periodically. Due to its general approach, it builds on the diversity found in both PDB and ChEMBL, which can, for example, be seen in the target diversity (Figure 1).

StrAcTable stands out as a dataset that is void of any manual process. We expect StrAcTable to grow sustainably with the expanding literature data and can be regularly updated. It is designed as the foundation for new developments in computational molecular design, machine learning, and dataset generation.

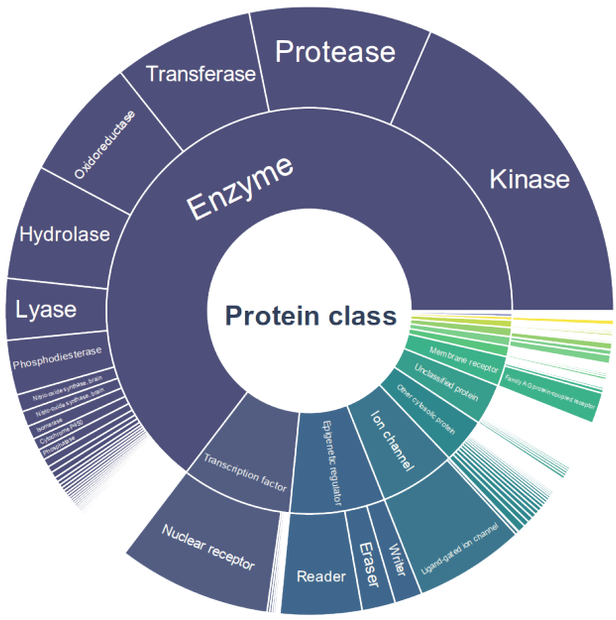


Figure 1: *Target Categorization for StrAcTable, highlighting both the reproduction of existing highly elaborated targets and the diversity of the dataset.*

References

2. Liu,Zhihai, et al., Forging the basis for developing protein-ligand interaction scoring functions, Accounts of chemical research, **2017**, 50, 2, 302-209, 10.1021/acs.accounts.6b00491
3. Wagle, Swapnil., et al., Sunsetting Binding MOAD with its last data update and the addition of 3D-ligand polypharmacology tools. Scientific Reports, **2023**, 13, 1, 3008, 10.1038/s41598-023-29996-w
4. Gilson, Michael K, et al., BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic Acids research, **2016**, 44, D1, D1045-D1053, doi.org/10.1093/nar/gkv1072
5. BioChemGraph Announcement, https://www.ebi.ac.uk/about/news/updates-from-data-resources/biochemgraph-data/, Accessed : 05.02.2025
6. Flachsenberg, Florian, et al., Redocking the PDB. JCIM, **2023**, 64, 1, 219-237, 10.1021/acs.jcim.3c01573.
7. Meyder, Agnes., et al., StructureProfiler: an all-in-one tool for 3D protein structure profiling. Bioinformatics, **2019**, 35, 5, 874-876, 10.1093/bioinformatics/bty692

Session C: Advanced Cheminformatics Techniques

C01: Transformers for Molecular Property Prediction: Pre-training is limited, and domain adaptation is a "significant" remedy

Afnan Sultan[1], Max Rausch-Dupont[2], Xi Yu[2], Dietrich Klakow[2], Andrea Volkamer[1]

[1] Data Driven Drug Design, Center for Bioinformatics, Saarland University, Saarbrücken, Germany

[2]Spoken Language Systems, Saarland Informatics Campus, Saarland University, Saarbrücken, Germany[2]

Over the past six years, molecular transformer models have become integral to computational drug discovery pipelines. These models are typically pre-trained on massive unlabeled datasets—ranging from millions to billions of molecules—sourced from public chemical libraries like ZINC or ChEMBL [1, 2]. However, the actual benefits of such large-scale pre-training for molecular property prediction (MPP) remain uncertain [2].

In this study, we investigate the performance bottlenecks of transformer models in MPP and evaluate strategies to overcome them via domain adaptation (DA). Specifically, we examine the impact of pre-training dataset size and introduce a chemically informed DA step using multi-task regression (MTR) objectives. Our findings [3] indicate that expanding pre-training beyond 400K–800K molecules does not yield statistically significant performance gains across seven datasets representing five ADME endpoints.

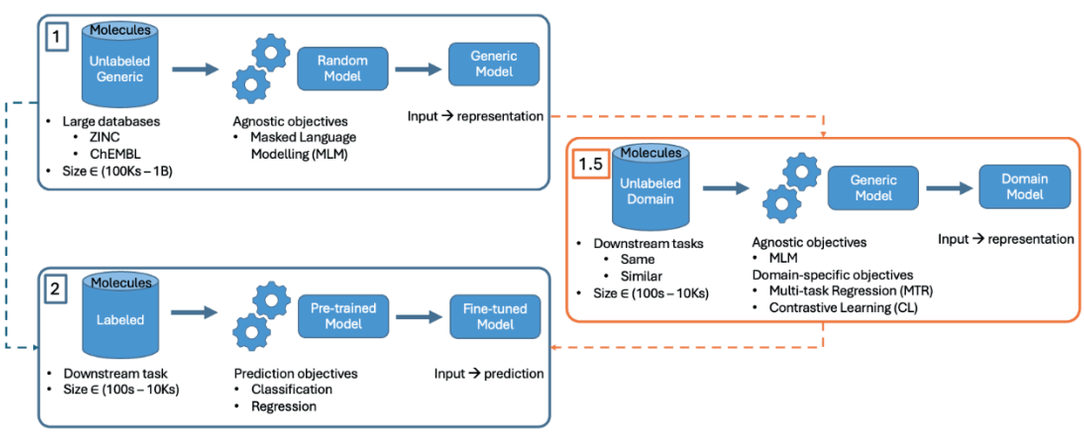


Figure 1: *Transformer models train in two steps. 1) Pre-training and 2) Fine-tuning. We propose an intermediate step (orange box), 1.5) domain adaptation.*

In contrast, domain adaptation on a small set of domain-relevant molecules (≤ 4K) leads to consistent and significant improvements ($P < 0.001$) [3]. This domain relevant dataset is the molecules used in each dataset. Furthermore, we explore retrieving chemically similar molecules from large chemical libraries using Tanimoto similarity and influence functions [4]. Preliminary results suggest that this strategy—leveraging similar molecules—offers further performance gains beyond the endpoint molecules alone.

Models pre-trained on 400K molecules and adapted using domain-specific data outperformed larger-scale models such as MolFormer and achieved performance on par with state-of-the-art models like MolBERT. Furthermore, comparison to baselines using physicochemical descriptors and Morgan fingerprints, chemically informed objectives and feature representations consistently led to superior performance—regardless of the model architecture [3].

Overall, this study highlights the crucial role of domain relevance and chemical similarity in the training and adaptation of transformer models for MPP. Efficient, domain-aware strategies like DA not only reduce resource requirements but also deliver performance gains, offering a practical and scalable approach to predictive modeling in cheminformatics.

References

[1] Wu, Z., et al., Chem. Sci., 2018, 9, 513–530.
[2] Sultan, A., et al., J. Chem. Inf. Model., 2024, 64, 6259–6280.
[3] Sultan, A., et al., arXiv:2503.03360 (2025).
[4] Chalkidis, I., et al., Findings EMNLP, 2020, 2898–2904.

C02: Navigating Synthon Space: Property-Driven Molecular Optimization for Pharmacokinetics

Rafał A. Bachorz, Michael S. Lawless, David W. Miller, Vladimir Chupakhin, Jeremy O. Jones, Robert Frączkiewicz

Simulations Plus, Inc., USA

The exploration of chemical space is a challenge due to its ultra-large scale, encompassing an immense number of possible molecular structures. Enumerated databases, which explicitly capture billions of compounds, are computationally expensive to process, making their efficient navigation a critical issue in drug discovery and materials science. Synthons, as an alternative representation, offer a more tractable approach to exploring chemical space. However, their effective utilization requires specialized algorithmic frameworks to ensure meaningful and efficient search strategies.

To address these challenges, this study leverages the Chemspace Freedom 3.0 database[1], which provides a structured repository of synthons, which are the combination of building blocks and associated linkage chemistries, for combinatorial molecular assembly. To guide the exploration toward chemically relevant and pharmacologically promising regions, ADMET Predictor® properties were employed, enabling a property-driven selection of candidate molecules. Furthermore, a Multicriteria Decision Analysis (MCDA) technique was applied to optimize molecular selections based on multiple physicochemical and pharmacokinetic criteria. This approach ensures that the most promising compounds are identified within the vast combinatorial landscape, balancing synthetic feasibility with desirable drug-like properties. In this study, we present selected examples of property-driven molecular optimization within the synthon space, utilizing the HTPK (High-Throughput Pharmacokinetics) module. The predicted in vivo objectives guide the chemical space exploration toward compounds with the desired pharmacokinetic (PK) profile.

By integrating a synthons-based representation, predictive ADMET modeling, and MCDA-driven selection, this work provides an efficient strategy for navigating ultra-large chemical spaces. The proposed framework enhances the prioritization of high-value molecular candidates, facilitating more targeted and computationally efficient chemical space exploration.

References

1. Protopopov MV, et al., The freedom space – a new set of commercially available molecules for hit discovery. Molecular Informatics., **2024**, 43, 12, 1-15, 10.1002/minf.202400114

C03: Scaffold Hopping with Generative Reinforcement Learning

L. Rossen [1,2], F. Sirockin [2], N. Schneider*[2], F. Grisoni [1,3*]

[1] *Eindhoven University of Technology, Dept. Biomedical Engineering, Institute for Complex Molecular Systems (ICMS), Eindhoven AI Systems Institute (EAISI), Eindhoven, The Netherlands*

[2] *Novartis BioMedical Research, Basel, Switzerland*

[3] *Centre for Living Technologies (CLT), Alliance TU/e, WUR, UU, UMC Utrecht, The Netherlands*

Scaffold hopping – the design of novel scaffolds for existing lead candidates – is a multi-faceted and non-trivial task, for medicinal chemists and computational approaches alike [1]. Generative reinforcement learning can iteratively optimize desirable properties of de novo designs, thereby offering opportunities to accelerate scaffold hopping [2, 3]. Current approaches confine the generation to a pre-defined molecular substructure (e.g., a linker or scaffold) for scaffold hopping [4, 5]. This confined generation may limit the exploration of the chemical space and require intricate molecule (dis) assembly rules [4,5,6]. In this work, we aim to advance reinforcement learning for scaffold hopping, by allowing 'unconstrained', full-molecule generation. This is achieved via the **RuSH** (**R**einforcement Learning for **U**nconstrained **S**caffold **H**opping) approach. **RuSH** steers the generation towards the design of full molecules having a high three-dimensional and pharmacophore similarity to a reference molecule, but low scaffold similarity. In this first study, we show the flexibility and effectiveness of RuSH in exploring analogs of known scaffold-hops and in designing scaffold-hopping candidates that match known binding mechanisms. Finally, a comparison between RuSH and two established methods (DeLinker [4] and LinkInvent [5]) highlights the benefit of its unconstrained molecule generation to systematically achieve scaffold diversity while preserving optimal three-dimensional properties.

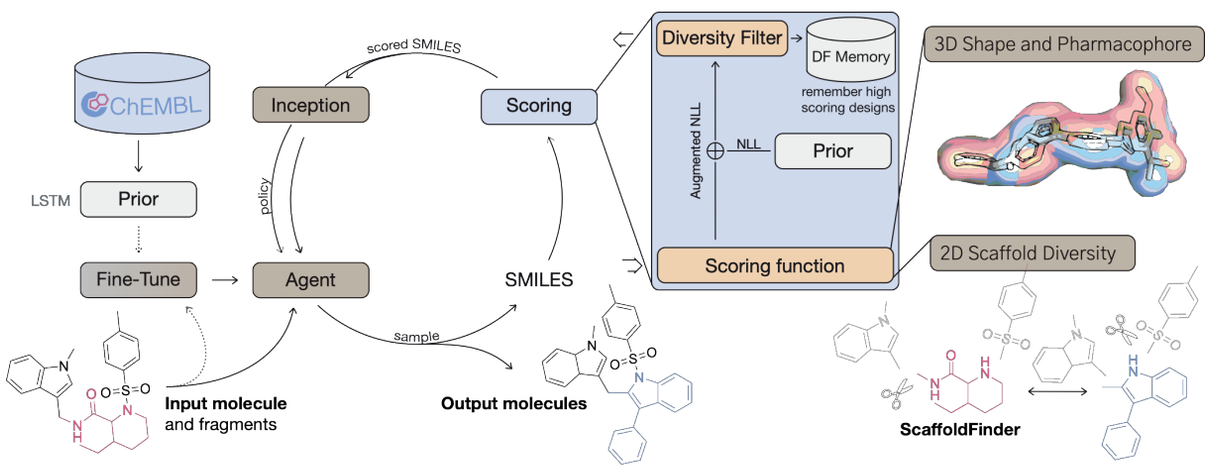


***Figure 1: RuSH (Reinforcement Learning for Unconstrained Scaffold Hopping**. A generator (long short-term memory network, LSTM) is trained on ChEMBL [7, 8] (Prior), and subsequently fine-tuned by transfer learning on a single reference molecule. The Prior will act as the reinforcement learning agent, and at each epoch, SMILES strings are sampled from the model [9]. These strings are scored by the scoring function, which rewards for (a) high three-dimensional and pharmacophore similarity (via ROCS [10]) and low scaffold similarity [11] to the same reference molecule. To detect scaffolds in generated designs, we developed an algorithm, ScaffoldFinder. During a reinforcement learning run, a diversity filter (DF) 'memory' is used to remember all high-scoring designs. An augmented Negative Log-Likelihood (NLL) is computed to update the policy of the agent across cycles. Inception [9] is used to speed up the reinforcement learning process by periodically exposing the agent to high scoring designs of earlier epochs.*

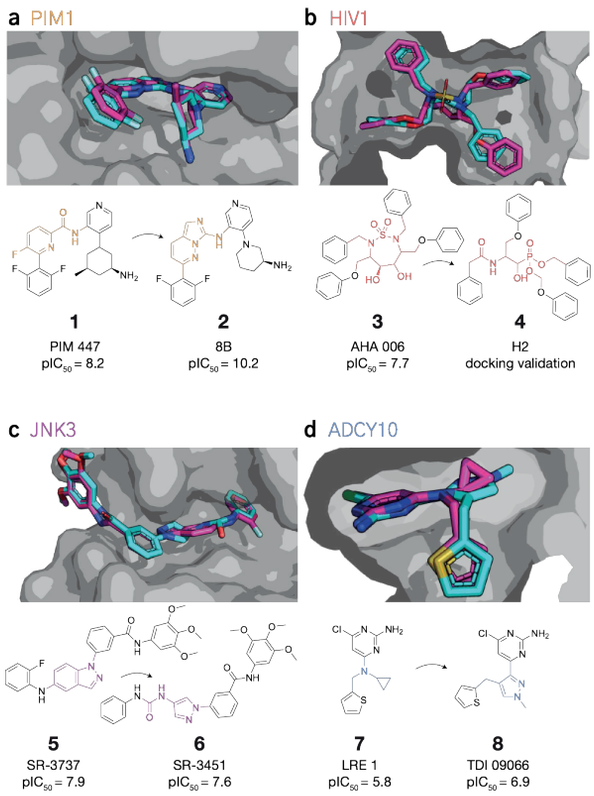


Figure 2: Selected scaffold hopping cases. *(**a**) PIM1. Reference molecule **1** (PDB-ID: 5DWR) and scaffold-hop **2** (PDB-ID: 5KZI) [12]. (**b**) HIV1, **3** (PDB-ID:1AJV) and scaffold-hop **4** (computational validation [13]). Superimposed instead is AHA-001 (PDB-ID: 1AJX) in the HIV1 active site. (**c**) JNK3, **5** (PDB-ID: 3FI3) and hop **6** (PDB- ID: 3FI2) [14], (**d**) ADCY10, **7** (PDB-ID: 5IV4) and scaffold-hop **8** (PDB-ID: 8CO7) [15].*

RuSH was able to generate designs that retain a high three-dimensional similarity to a bioactive reference, while at the same time possessing novel scaffolds. Moreover, unconstrained generation allows for scaffold hopping with an arbitrary number of decorations, and to tune the desired structural variability in molecular decorations. Similarly, the ability to perform transfer learning proved particularly useful to increase the speed and efficiency of chemical space exploration, and to promote the matching of shape and pharmacophore properties. In principle, RuSH's pipeline also allows for multiple references to be used, by combining the individual 2D and 3D scores via an aggregation function. Finally, the ScaffoldFinder algorithm can be adapted to virtually any type of substructure matching, e.g., generate designs with warheads, or coupling groups for improved synthetic accessibility.

Considering recent developments in reinforcement learning frameworks [16], RuSH's scoring function could be further combined with secondary objectives for multi-objective optimization and curriculum learning, e.g., to jointly improve the inclusion of decorations, considering several reference molecules simultaneously, and optimize for molecular properties like solubility or drug-likeness.

References

1. Schneider, G.; Schneider, P.; Renner, S. QSAR & Combinatorial Science **2006**, 25, 1162–1171.
2. Thomas, M.; Boardman, A.; Garcia-Ortegon, M.; Yang, H.; de Graaf, C.; Bender, A. Artificial Intelligence in Drug Design **2022**, 1–59.
3. Korshunova, M.; Huang, N.; Capuzzi, S.; Radchenko, D. S.; Savych, O.; Moroz, Y. S.; Wells, C. I.; Will- son, T. M.; Tropsha, A.; Isayev, O. Communications Chemistry **2022**, 5, 129
4. Imrie, F.; Bradley, A. R.; van der Schaar, M.; Deane, C. M. Journal of Chemical Information and Modeling **2020**, 60, 1983–1995

5. Guo, J.; Knuth, F.; Margreitter, C.; Janet, J. P.; Papadopoulos, K.; Engkvist, O.; Patronov, A. Digital Discovery **2023**, 2, 392–408.
6. Igashov, I.; St¨ark, H.; Vignac, C.; Satorras, V. G.; Frossard, P.; Welling, M.; Bronstein, M.; Correia, B. arXiv preprint arXiv:2210.05274 2022.
7. Brown, N.; Fiscato, M.; Segler, M. H.; Vaucher, A. C. Journal of Chemical Information and Modeling **2019**, 59, 1096–1108.
8. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B., et al. Nucleic acids research **2012**, 40, D1100–D1107.
9. Blaschke, T.; Arus-Pous, J.; Chen, H.; Margreitter, C.; Tyrchan, C.; Engkvist, O.; Papadopoulos, K.; Patronov, A. Journal of chemical information and modeling **2020**, 60, 5918–5922.
10. Hawkins, P. C.; Skillman, A. G.; Nicholls, A. Journal of medicinal chemistry **2007**, 50, 74–82.
11. Jaccard, P. New Phytologist **1912**, 11, 37–50.
12. Wurz, R. P.; Sastri, C.; D'Amico, D. C.; Herberich, B.; Jackson, C. L.; Pettus, L. H.; Tasker, A. S.; Wu, B.; Guerrero, N.; Lipford, J. R., et al. Bioorganic & medicinal chemistry letters **2016**, 26, 5580–5590.
13. Bergmann, R.; Linusson, A.; Zamora, I. Journal of medicinal chemistry **2007**, 50, 2708–2717.
14. Kamenecka, T.; Habel, J.; Duckett, D.; Chen, W.; Ling, Y. Y.; Frackowiak, B.; Jiang, R.; Shin, Y.; Song, X.; LoGrasso, P. Journal of Biological Chemistry **2009**, 284, 12853–12861.
15. Sun, S.; Fushimi, M.; Rossetti, T.; Kaur, N.; Ferreira, J.; Miller, M.; Quast, J.; van den Heuvel, J.; Steegborn, C.; Levin, L. R., et al. Journal of Chemical Information and Modeling **2023**, 63, 2828–2841.
16. Loeffler, H. H.; He, J.; Tibo, A.; Janet, J. P.; Voronov, A.; Mervin, L. H.; Engkvist, O. Journal of Cheminformatics **2024**, 16, 20.

C04: Honey, I shrunk the database: Making multi-billion compound libraries as small as possible

J. Mayfield, R. Sayle

NextMove Software Ltd, Cambridge, UK

Chemical databases grow larger every year, current computer hardware allows scaling of classical systems to hundreds of billions of compounds (with sufficient investment). However, if computer resources are limited, working with databases of this size can be challenging. The increased popularity and use of virtual and DNA encoded libraries requires ever more efficient techniques to store, distribute, and search these ultra-large datasets.

We refer to the *footprint* of a molecule as how much space it takes to store either in memory or on disk. The *footprint* depends on the number of atoms and bonds in the molecule; however, estimates can be given for a "*typical"* small molecule:

Method	Bytes/Molecule
Chemistry Toolkit (e.g. RDKit)	10,000 - 20,000
SDfile/MOLfile	1000 - 4000
SMILES	40 - 100
Compressed SMILES (e.g. GZip)	4 - 10
This work	1

In this work we aim to get the size of a molecule record down to a single byte. To put this in perspective ten billion molecules would require ten gigabytes to store.

Virtual libraries are made from a set of known reagents/building blocks/synthons that can be combined in different ways. In this work we turn this around and determine a set of virtual synthons from a set of known compounds. We ignore (don't need) how the molecules are actually made so the number of virtual synthons may be fewer and better suited for compression or searching.

Our primary motivation was for virtual libraries that have been filtered for desirable properties, however in practice any chemical library can be converted to a set of virtual synthons. This means existing algorithms that explore and search virtual spaces could (in theory) work with any large chemical library making ultra-large datasets more accessible than ever before.

C05: Docking-based geometric graph models for kinase-ligand affinity prediction

Andrius Bernatavicius[1,2], Chen Ji Rong Jiang[1,4], Martin Šícho[1,3], Gerard J.P. van Westen[1]

[1] *Leiden Academic Centre for Drug Research, Leiden University, The Netherlands*

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

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

[4] *Utrecht University, The Netherlands*

Kinases are critical drug targets due to their involvement in numerous cellular processes and diseases, particularly cancer and inflammation. Accurate in silico estimation of kinase-ligand binding affinity is crucial for efficient drug discovery. In our previous work [1] we have shown that the quality of affinity predictions can be improved using 3D data derived from molecular docking in the form of drug-target interaction fingerprints. Now we demonstrate that representing kinase pocket and ligand pairs as joint 3D graphs, augmented by additional edges that represent their physical interactions, and modelling them using graph neural networks (GNNs) further increases model predictive capabilities. We focus on highlighting the computational advantages of GNNs in terms of accuracy, compute efficiency, and explainability over traditional drug-target affinity model counterparts that use 2D or interaction fingerprints. An extensive ablation study also underlines the importance of feature selection for molecular graph representations and tuning of model architectures.

Keywords: Graph Neural Networks, Molecular Docking, Explainability

References

1. Jordy Schifferstein, Andrius Bernatavicius, and Antonius PA Janssen. “Docking-Informed Machine Learning for Kinome-wide Affinity Prediction”. In: Jour-nal of Chemical Information and Modeling 64.24 (**2024**), pp. 9196–9204.

C06: Validating the prediction of lowest-energy tautomers and conformers against experimental techniques

Bernardo de Souza, Jeroen Koopman, Anneke Dittmer, Christoph Riplinger

FACCTs GmbH, Cologne, Germany

Accurate determination of the lowest-energy conformers and tautomers is of crucial importance in drug design as well as chemical synthesis. To find the correct structures, most of the chemistry community rely on simple or purely stochastic approaches, but as molecules get larger and the conformational space gets more complex, with easily more than millions conformers even for medium-sized molecules, these are guaranteed to become too inefficient. Another problem from the field is how to validate these results, since experimental data in solution is usually the result of a whole ensemble of conformers instead of a measurement of a single molecular entity.

In order to address both issues, the new Global Optimization Algorithm (GOAT), a novel conformational search algorithm already implemented in ORCA 6, will be presented in this talk.[1, 2] We will show how GOAT can be used to determine the lowest-energy tautomer-/conformer-ensemble of molecular structures. These ensembles can be later used to predict various molecular properties, as e.g. molecular rotational resonance spectroscopy or collisional cross section data from ion-mobility mass spectrometry. The predictions in gas phase can be directly compared to the experiments, thus - in a direct way - assessing the accuracy of the predictions against reliable and robust experimental data. We will show how GOAT can outperform popular and widely used algorithms for generating conformers such as the ETKDG (from RDKIT) and even the current state-of-the-art CREST.

References

1. De Souza, B., GOAT: A Global Optimization Algorithm for Molecules and Atomic Clusters. *Angewadte Chemie*, **2025**, https://doi.org/10.1002/anie.202500393
2. Neese, F., et al., The ORCA quantum chemistry program package, J. *Chem. Phys.*, **2020**, 152, 224108, https://doi.org/10.1063/5.0004608

C07: Modernising the Reaction Vector Framework: From Legacy Code to Validated Synthesis

James Webster[1], Rebecca Craik[1], Dinakaran Murugesan[1], Gary Tarver[1], Valerie J. Gillet[2], & Michael J. Bodkin[1]

[1]Drug Discovery Unit, School of Life Sciences, University of Dundee, Dow Street, UK
[2]Information School, University of Sheffield, UK

Generative design using reaction-based methods focuses on creating novel chemical entities using *in-silico* predicted reactions. A significant advantage of reaction based *de novo* design over atom and fragment-based approaches is that the generated molecules will exhibit higher propensity for being synthetically tractable.

The **"reaction vector" (RV)** is an established methodology for encoding reaction transformations[1] and has been explored extensively as an *in-silico* reaction-based molecule construction approach[2–7]. However, like many academic software projects the RV framework to date has become a legacy cheminformatics system with significant complexity arising over a 16-year development timeline.

I will present how we reimplemented the RV framework from a legacy Java implementation to a modernised **Python/C++ library**. I will describe our use of **open-source large language models (LLMs)**[8] as tools for automating the translation and documentation of legacy cheminformatics software and the advantages and challenges faced.

I will then discuss how we have integrated the RV toolkit into a **modern synthesis stack** coupled with novel synthetic methodologies for plate and flow chemistry. Several case studies will be highlighted where we have utilised the RV approach to design compounds *de novo* utilising reaction ELN data in an automated manner with a **75-95% synthesis success rate** across a diverse range of reaction classes.

References

1. Broughton, H., Hunt, P. & MacKey, M. Methods for classifying and searching chemical reactions. (**2003**).
2. Patel, H. Knowledge-Based De Novo Design using Reaction Vectors. (University of Sheffield, **2009**).
3. Hristozov, D., Bodkin, M., Chen, B., Patel, H. & Gillet, V. J. Validation of reaction vectors for de novo design. in *ACS Symposium Series* vol. 1076 29–43 (**2011**).
4. Gillet, V. J., Bodkin, M. J. & Hristozov, D. Multiobjective De Novo Design of Synthetically Accessible Compounds. in *De novo Molecular Design* 267–285 (Wiley-VCH Verlag GmbH & Co. KGaA, **2013**). doi:10.1002/9783527677016.ch11.
5. Wallace, J. Structure generation and de novo design using reaction networks. (University of Sheffield, **2016**).
6. Ghiandoni, G. M. Enhancing Reaction-based de novo Design using Machine Learning. (University of Sheffield, **2019**).
7. Webster, J. Development of a novel tree search algorithm for reaction vector-based de novo design. (University of Sheffield, **2021**).
8. Jiang, J., Wang, F., Shen, J., Kim, S. & Kim, S. A Survey on Large Language Models for Code Generation. Preprint at https://doi.org/10.48550/arXiv.2406.00515 (**2024**).

Session D: Integrative Structure-Based Drug Design

D01: DockM8: All-in-One Open-Source Platform for Consensus Virtual Drug Screening

A. M. L. Lacour[1,2,3], H. Ibrahim[2], M. Bähr[2], A. H. K. Hirsch[1,3], A. Volkamer[1,2]

[1] *Helmholtz Institute for Pharmaceutical Research Saarland (HIPS) – Helmholtz Centre for Infection Research (HZI), Saarbrücken, Germany*

[2] *Data Driven Drug Design, Center for Bioinformatics, Saarland University, Saarbrücken, Germany*

[3] *Department of Pharmacy, Saarland University, Saarbrücken, Germany*

Structure-based virtual screening (SBVS) is a crucial tool in modern drug discovery. It expedites the identification of potential therapeutic compounds by attempting to predict a molecule's biological activity. This reduces the need for time-consuming experimental screening. SBVS also allows for exploration of vast chemical spaces, uncovering promising candidates that traditional approaches might miss.

DockM8[1] is an open-source SBVS workflow which seamlessly integrates diverse open-source packages to perform consensus docking and scoring to prioritize candidate molecules for further evaluation. The software offers a comprehensive suite of capabilities, including protein and library preparation, docking (5 methods), pose selection (over 30 methods), rescoring, (16 scoring functions) and consensus ranking (10 methods).

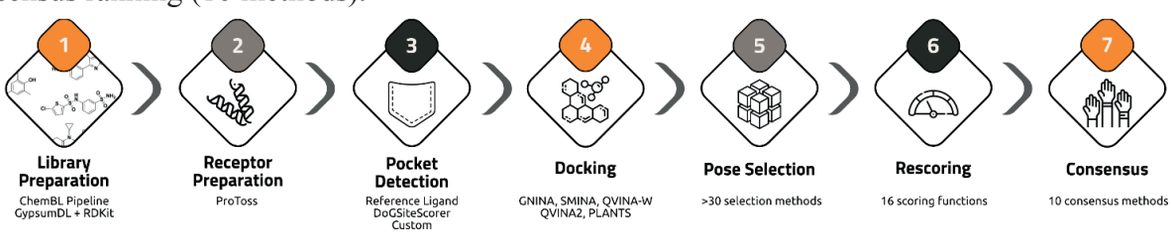


***Figure 1**: Representation of all the steps handled by the DockM8 workflow, along with some of the options available.*

DockM8 is also able to largely automate the optimization of the screening procedure through automatic decoy generation and determination of the optimal screening conditions. DockM8's performance has been rigorously assessed on the DUD-E[2], DEKOIS[3] and Lit-PCBA[4] benchmarks. We demonstrate that DockM8 achieves state-of-the-art enrichment performance when compared to other SBVS methods. We also present an extension for DockM8 called LearnM8 designed to streamline active-learning in the context of consensus docking.

Notably, DockM8's design invites active modification and updating by the user community, fostering an open-source ecosystem that can integrate novel docking and scoring protocols as they emerge. This inherent flexibility and adaptability set DockM8 apart as a future-proof tool in the rapidly evolving field of computer-aided drug design. DockM8 is available at: https://github.com/DrugBud-Suite/DockM8.

References

1. Lacour, A.; Ibrahim, H.; Volkamer, A.; Hirsch, A. K. H. DockM8: An All-in-One Open-Source Platform for Consensus Virtual Screening in Drug Design. ChemRxiv July 23, **2024**. https://doi.org/10.26434/chemrxiv-2024-17k46.
2. Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **2012**, *55* (14), 6582–6594. https://doi.org/10.1021/JM300687E/SUPPL_FILE/JM300687E_SI_004.TXT.
3. Bauer, M. R.; Ibrahim, T. M.; Vogel, S. M.; Boeckler, F. M. Evaluation and Optimization of Virtual Screening Workflows with DEKOIS 2.0 - A Public Library of Challenging Dock-

ing Benchmark Sets. *J. Chem. Inf. Model.* **2013**, *53* (6), 1447–1462. https://doi.org/10.1021/CI400115B/SUPPL_FILE/CI400115B_SI_001.PDF.

4. Tran-Nguyen, V. K.; Jacquemard, C.; Rognan, D. LIT-PCBA: An Unbiased Data Set for Machine Learning and Virtual Screening. *J. Chem. Inf. Model.* **2020**, *60* (9), 4263–4273. https://doi.org/10.1021/ACS.JCIM.0C00155.

D02: Computational Challenges in Modeling Metal-Binding Sites in Proteins: A Multiscale Approach to Copper-Ligand Interactions in the Plant Receptor ETR1

Lisa Sophie Kersten[1], Michele Bonus[1], Stephan Schott-Verdugo[2], Holger Gohlke[1,2]

[1] *Institute for Pharmaceutical and Medicinal Chemistry, Heinrich Heine University Düsseldorf, Germany*

[2] *Institute of Bio- and Geosciences (IBG-4: Bioinformatics), Forschungszentrum Jülich GmbH, Germany*

Metal-binding sites in proteins are essential for enzymatic catalysis, structural stability, and ligand recognition, making them key targets for drug discovery. However, accurate modeling of these sites is challenging due to the complex electronic structure of metal ions, their diverse coordination geometries, and their dynamic interactions with the protein environment [1].

In this work, we address the challenge of elucidating the structural dynamics of the plant receptor ETR1, which is involved in ethylene signaling and associated with post-harvest spoilage, making it a key target for securing food security. ETR1 has a copper-binding site where both agonists and antagonists compete for the same binding site at the copper cofactor [2]. To explore the ligand binding models, we used molecular docking and MD simulations with a non-bonded ion model to describe the copper binding site. Density functional theory (DFT) calculations, combined with geometry, bonding, charge distribution, electrostatic potential, and orbital decomposition analyses provided deeper insight into the electronic structure and coordination preferences of the copper-binding site. Our data revealed specific properties of the copper-binding site depending on whether an agonist or antagonist was present. The optimized QM complexes guided the parameterization of a copper-specific force field for MD simulations using a bonded model, aiming to capture ligand-induced side-chain fluctuations and structural dynamics. Finally, QM/MM simulations will refine the MD results to improve the accuracy of the copper-ligand interaction model in ETR1.

The combination of the methods indicates promising results in the case of ETR1 and provides valuable insights into its copper-ligand interactions. Our approach highlights best practices for describing metal-ligand interactions and presents a multiscale strategy that may apply to other metal-containing systems in computational drug discovery.

References

1. Patil, V. M. et al., Experimental and computational models to understand protein-ligand, metal-ligand and metal-DNA interactions pertinent to targeted cancer and other therapies. European Journal of Medicinal Chemistry Reports., **2024**, 10, https://doi.org:10.1016/j.ejmcr.2024.100133
2. Rodriguez, F. I. et al. A copper cofactor for the ethylene receptor ETR1 from Arabidopsis. Science., **1999**, 283, 996-998, https://doi.org:10.1126/science.283.5404.996

D03: AI and MD-aided computational docking pipeline to elucidate TGF-beta type I receptor and downstream signaling mediator interaction

Leon Obendorf[1], Amelie Bohrmann[1], Jasmine Okita[1], Robert Westphahl[1], Daniel Gehrke[1], Gerhard Wolber[1,2], Petra Knaus[1]

[1] *Biochemistry, AG Knaus, Freie Universität Berlin, Germany*

[2] *Pharmacy, AG Wolber, Freie Universität Berlin, Germany*

By designing a pipeline using molecular dynamics simulations, AlphaFold2 multimer [1], and Rosetta Commons Ensemble docking [2], we mapped crucial interfaces in the binding of the intracellular signaling complex of the transforming-growth-factor-beta family (TGF-beta). Within the interface of the receptor kinase domain, and its signaling mediator (called SMADs) binding domain, two loops are known to play a crucial role [3]. Therefore, we developed a semi-flexible pipeline considering multiple loop conformations prominent in MD-simulations for Rosetta docking.

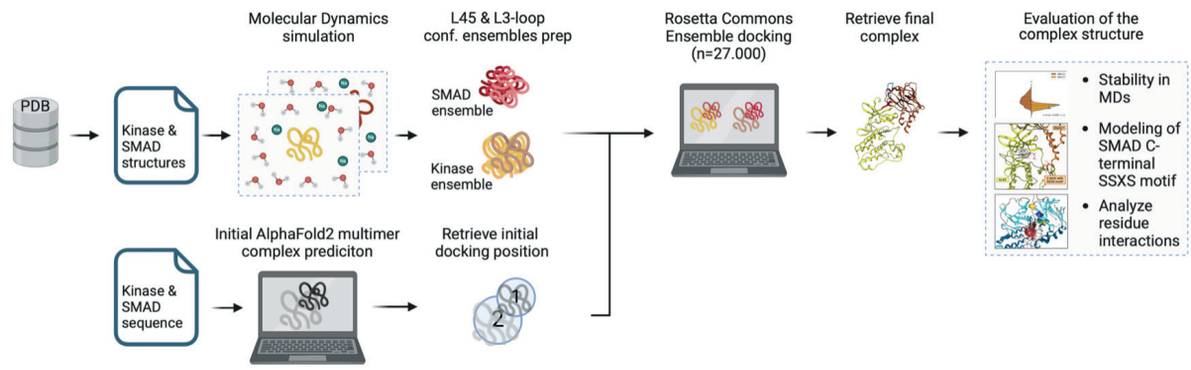


***Figure 1**: Schematic depiction of the computational pipeline using MD-simulations which clustered [4] on the flexible regions within the protein-protein interface. These ensembles of loop conformations were used in combination with an initial AlphaFold2 multimer prediction to perform Rosetta Ensemble docking generating a final complex via a "semi-flexible" backbone docking pipeline.*

TGF-beta family signaling governs critical cellular functions such as inflammation, genome stability, metabolism, and cell differentiation, making precise therapeutic intervention a key goal [5][6]. Within the TGF-beta family there are two sub-pathways, the TGF-beta and the bone morphogenetic protein (BMP) pathway. These are tightly regulated, and their imbalance plays a role in various diseases [7] [8]. Three different receptors in combination with two different SMADs were investigated as examples of the TGF-beta and BMP-signaling pathways.

Our predicted complexes highlight key residues driving binding specificity and defining distinct binding positions. Based on these predictions, we designed point mutant constructs to evaluate our findings in the cell culture using phosphorylation, subcellular localization and downstream signaling assays. The findings could facilitate development of protein therapeutics or to find new allosteric binding sites for inhibitors directed to one sub-pathway within the TGF-beta family signaling.

References

1. Richard Evans et al. "Protein complex prediction with AlphaFold-Multimer". In: bioRxiv (**2021**). DOI: 10.1101/2021.10.04.463034.
2. Sidhartha Chaudhury and Jeffrey J. Gray. "Conformer selection and induced fit in flexible backbone protein-protein docking using computational and NMR ensembles". In: Journal of Molecular Biology 381.4 (Sept. 12, **2008**), pp. 1068–1087. ISSN: 1089-8638. DOI: 10.1016/j.jmb.2008.05.042.
3. Urban Persson et al. "The L45 loop in type I receptors for TGF-beta family members is a critical determinant in specifying Smad isoform activation". In: FEBS Letters 434.1 (**1998**), pp. 83–87. ISSN: 1873-3468. DOI: 10.1016/S0014-5793(98)00954-5.

4. Thibault Tubiana et al. “TTClust: A Versatile Molecular Simulation Trajectory Clustering Program with Graphical Summaries”. In: Journal of Chemical Information and Modeling 58.11 (Nov. 26, **2018**). Publisher: American Chemical Society, pp. 2178–2182. ISSN: 1549-9596. DOI: 10.1021/acs.jcim.8b00512.
5. Erine H. Budi, Dana Duan, and Rik Derynck. “Transforming Growth Factor-beta Receptors and Smads: Regulatory Complexity and Functional Versatility”. In: Trends in Cell Biology 27.9 (Sept. 1, **2017**). Publisher: Elsevier, pp. 658–672. ISSN: 0962-8924. DOI: 10.1016/j.tcb.2017.04.005.
6. Farshad Nassiri et al. “Endoglin (CD105): A Review of its Role in Angiogenesis and Tumor Diagnosis, Progression and Therapy”. In: Anticancer Research 31.6 (Jan. 6, **2011**), pp. 2283–2290. ISSN: 0250-7005, 1791-7530
7. Richard N. Wang et al. “Bone Morphogenetic Protein (BMP) signaling in development and human diseases”. In: Genes and Diseases 1.1 (**2014**), pp. 87–105. DOI: https://doi.org/10.1016/j.gendis.2014.07.005.
8. Darja Obradovic Wagner et al. “BMPs: From Bone to Body Morphogenetic Proteins”. In: Science Signaling 3.107 (Feb. 2, **2010**). Publisher: American Association for the Advancement of Science Section: Meeting Report, mr1–mr1. ISSN: 1945-0877, 1937-9145. DOI: 10.1126/scisignal.3107mr1.

D04: Advancing free-energy calculations by combining multiscale modeling and multistate enhanced sampling

Domen Pregeljc, Sereina Riniker

Department of Chemistry and Applied Biosciences, ETH Zürich, Switzerland

Free energy calculations are one of the central themes in computational chemistry. Arguably the most rigorous approaches are methods based on molecular dynamics (MD) simulations. These are often termed “computational microscope” due to the detailed insight into the studied system they provide. This comes at a cost, though, and the attainable accuracy is limited by the available computational resources. In practice, a trade-off between two main sources of error in MD needs to be struck. These are model approximations and insufficient phase-space sampling. The former can be addressed through the use of multiscale modeling, such as QM/MM, enabling a more precise description of a part of the system in comparison to the widely used classical force fields, potentially capturing important effects such as polarization. Exploration of the relevant phase-space, on the other hand, is facilitated by the use of enhanced sampling techniques, among which multistate methods, such as replica-exchange enveloping distribution sampling (RE-EDS), are most efficient. We introduce a combined multiscale-multistate free energy method by integrating the established QM/MM scheme with RE-EDS, discuss the details of its implementation, and validate it on hydration free energies. Furthermore, we highlight the importance of QM-MM model compatibility and quantify the effects of QM method and classical water model selection.

D05: How useful are protein folding tools for drug design?

Henriëtte Willems

ALBORADA Drug Discovery Institute, University of Cambridge, UK

There has been a lot of excitement recently regarding protein folding tools, such as Boltz-1 or AlphaFold3 (AF3)[1], that appear to promise a much better way of doing drug design than more traditional methods, such as homology modelling combined with docking. But not much is known publicly yet about how useful these folding algorithms are for drug design. The reported RMSD to known protein structures is often very good, but that does not necessarily mean that key details of the binding site or pose are correct. This is important, because inaccuracies in the binding site of a model can have a large impact on drug design.

Here we report the results of a case study on protein folding of PI5P4K lipid kinases. Three crystal structures[2,3,4] that were not in the training set for the Boltz-1 and AF3 algorithms are compared with the models produced by AF3, Boltz-1, Chai and RosettaFold-AllAtom (RFAA). One structure has an allosteric pocket, one shows a novel loop location, and the third an unusual ligand pose, so all three were interesting test cases for protein folding.

The results show that some methods have difficulties with getting ligand conformations correct. Allosteric pockets may not be found, unless information of interacting residues is provided. Also, flexible loops near the binding site are not reliably positioned correctly. Nevertheless, most models outperformed docking to apo structures.

References

1. Abramson, J. et al *Nature* **2024**, **630**, 493-500. DOI: 10.1038/s41586-024-07487-w
2. Boffey, H. K., et al., Development of Selective Phosphatidylinositol 5-Phosphate 4-Kinase γ Inhibitors with a Non-ATP-competitive, Allosteric Binding Mode. *J.Med. Chem.* **2022**, 65(4), 3359–3370. DOI: 10.1021/acs.jmedchem.1c01819.
3. Rooney, T. et al., The Identification of Potent, Selective, and Brain Penetrant PI5P4Kγ Inhibitors as In Vivo-Ready Tool Molecules. *J. Med. Chem,* **2023**, 66 (1), 804-821. DOI: 10.1021/acs.jmedchem.2c01693.
4. Willems, H.M.G. et al., Identification of ARUK2002821 as an isoform-selective PI5P4Kα inhibitor. *RSC Med. Chem.,* **2023**, 14(5), 934-946. DOI: /10.1039/D3MD00039G

D06: Quantum Mechanics for Ligand Design, Binding Affinities, Toxicity Risk and Process Chemistry: Successes and Obstacles

Andreas H. Göller

Bayer AG, Pharmceuticals R&D, Structural Biology & Computational Design, Wuppertal, Germany

Quantum Mechanics calculations provide insights into weak non-classical interactions of molecules with itself, the condensed phase or with protein or DNA targets, that are not properly handled by force-field methods, like for instance weak dispersive interactions, strain energies or long-range electrostatic interactions. QM furthermore allows for understanding and optimizing chemical reactions in process chemistry, and can provide reactivity descriptors for machine learning.

I will provides examples for application of QM in pharmaceuticals R&D, as well as limitations of QM one has to deal with due to still suboptimal description of the condensed phase.

Nitrosamine impurities in formulations can pose a risk on carcinogenicity via DNA adducts of the degradation products. From calculations of complete activation pathways of small nitrosamines[1] we deduced criteria for risk assessment of Bayer market products. Our assessment based on QM calculations are now confirmed by long-term in-vivo safety studies.

Calculations on the free energy landscape of the 16 diastereomers of proposed late stage synthesis intermediate for Vilaprisan explained, why some could not be synthesized.[2] The results were part of the chemistry documentation package provided to authorities.

What if ligands bind to a target in exactly the same orientation (as confirmed by x-ray structures) but have huge differences in binding affinity that can not be explaiend by FF-based FEP? QM can provide the answer solely from the small molecules' steric and electronic properties.

Finally, I will report on QM calculations on protein-ligand complexes[3] with up to 5500 atoms using Google's tensor processing units chips and standard DFT with double-zeta quality basis sets, that challenge current computational limits and show the directions for the future of QM.

References

1. Göller, A.H., et al., Quantum Chemical Calculations of Nitrosamine Activation and Deactivation Pathways for Carcinogenicity Risk Assessment. *Frontiers in Pharmacology*., **2024**, 15, 10.3389/fphar.2024.1415266
2. Plöger, T.A., et alCombined Experimental and Quantum Mechanical Elucidation of the Synthetically Accessible Stereoisomers of Hydroxyestradienone(HED), the Starting Material for Vilaprisan Synthesis. JCAMD., **2021**, 35, 505-516, 10.1007/s10822-020-00353
3. Göller, A.H., et al., Towards Full-Protein Quantum Mechanics with Tensor Processing Units, in preparation.

D07: MDPath: Unraveling Allosteric Communication Pathways through Molecular Dynamics Simulations

Niklas P. Doering[1], Marvin Taterra[2], Marcel Bermúdez[2], Gerhard Wolber[1]

[1] *FU Berlin, Institute of Pharmacy, Germany*

[2] *Universität Münster, Institute of Pharmaceutical and Medicinal Chemistry, Germany*

Despite significant advances in the prediction of protein-ligand binding and the analysis of protein-ligand interactions, the identification of correlated motions leading to conformational shifts remains a key challenge in computational drug discovery. To address this, we present MDPath, a comprehensive Python-based CLI tool that uses normalized mutual information (NMI) analysis to elucidate allosteric pathways from molecular dynamics simulations. MDPath builds on the discovery by McClendon et al. that protein motions can be connected between distant regions, by mutual information calculations (**McClendon et al. 2009**). Their theoretical framework is now easily accessible via MDPath's CLI tool, which incorporates normalized mutual information (NMI) with graph-based pathfinding - an approach previously validated in G protein-coupled receptor studies (**Bhattacharya & Vaidehi 2014; Dutta & Shukla 2023**). Consequently, MDPath facilitates the systematic mapping and visualization of allosteric communication networks.

MDPath identifies allosteric pathways by analyzing the NMI of residue backbone dihedral angle motions. A graph is constructed where nodes represent residues, and edges connect proximal residues weighted by their NMI. The shortest high NMI paths between distant residues are then computed, clustered by spatial overlap, and can then be visualized in Jupyter notebooks using NGLView, PyMOL or Blender. Through detailed case studies of three pharmaceutically relevant targets - the β_2-adrenoceptor, the adenosine A_2A receptor and the μ-opioid receptor - we are able to demonstrate MDPath's ability to uncover both, established and previously unidentified signaling mechanisms. In addition, MDPath is able to reveal ligand-specific allosteric networks by tracing pathways from given protein-ligand interactions studied in the β_2-adrenoceptor and μ-opioid receptor.

Ultimately, the workflow enables the systematic mapping and analysis of allosteric communication pathways, providing atomistic explanations for conformational shifts. MDPath is not only usable for the tested GPCR systems, but its general implementation suggests usability across all protein domains and is freely available from our GitHub repository (https://github.com/wolberlab/mdpath) and PyPI (https://pypi.org/project/mdpath/).

References

4. McClendon, C. L., Friedland G., Mobley D. L., Amirkhani H., Jacobson M. P. (**2009**): Quantifying Correlations Between Allosteric Sites in Thermodynamic Ensembles. Journal of Chemical Theory and Computation, 5, 2486-2502; https://doi.org/10.1021/ct9001812
5. Bhattacharya, S. & Vaidehi, N. (**2014**): Differences in allosteric communication pipelines in the inactive and active states of a GPCR. Biophysical Journal, 107, 422–434, https://doi.org/10.1016/j.bpj.2014.06.015
6. Dutta, S. & Shukla, D. (**2023**): Distinct activation mechanisms regulate subtype selectivity of cannabinoid receptors. Communications Biology, 6, 485, https://doi.org/10.1038/s42003-023-04868-1

D08: How unsociable is the fragment space, and can we do better?

P. Janssen[1], F. Becker[1,2], T. Matviyuk[3,4], I. Kondratov[3,5,6], D. Kümmel[2], M. S. Weiss[7], O. Koch[1]

[1] *Institute of Pharmaceutical and Medicinal Chemistry, University of Münster, Germany*

[2] *Institute of Biochemistry, University of Münster, Germany*

[3] *Enamine Ltd., Kyiv, Ukraine*

[4] *HSPH, Harvard University, MA, USA*

[5] *Enamine Germany GmbH, Frankfurt am Main., Germany*

[6] *V. P. Kukhar Institute of Bioorganic Chemistry and Petrochemistry, National Academy of Sciences of Ukraine, Kyiv, Ukraine*

[7] *Macromolecular Crystallography, Helmholtz-Zentrum Berlin, Germany*

Fragment-based drug discovery has become an essential technique for developing small molecular probes or drugs. [1] Performing fragment screening with available fragment libraries is often tedious and time-consuming when selecting fragments for further extension. This is frequently due to the required direction being difficult or impossible to access synthetically. St. Denis et al. referred to these as unsociable fragments that lack the necessary growth vector for fragment extensions. [1] We faced this issue in our fragment campaigns, necessitating laborious scaffold replacement and the development of synthetic routes to allow for the required growth. This issue of synthetic tractability is seemingly widespread and commonly hampers fragment-to-lead projects. [2]

The first part of this ongoing project involves analysing the commercial fragment space. We have collected nearly one hundred commercially available libraries containing over half a million unique fragments and more than 4.5 million potential growth vectors. Using the recently developed ultra-large virtual chemical spaces as a framework, we are currently examining which growth vectors are readily accessible. [3] Unfortunately, this presents a grim picture but may provide a pathway to enhance commercial libraries.

To alleviate the issue of intractable fragments, we also designed an entirely sociable fragment library suitable for crystallographic screening. By using Enamine's REALSpace [4] we assembled a small library where every growth vector can be elaborated with a multitude of different substituents. As we are working towards novel and urgently needed antituberculotic drugs, we screened this library crystallographically on our established target, the thioredoxin reductase. [5] This has yielded multiple binding events, for which follow-up derivatives could directly be ordered and tested.

References

7. St. Denis. et al., Fragment-based drug discovery: opportunities for organic synthesis, RSC Med. Chem., **2021**, 12, 3, 321-329, 10.1039/D0MD00375A
8. Erlanson, D. Practical Fragments: Poll results: synthetic challenges are pervasive in FBLD. Practical Fragments. practicalfragments.blogspot.com/2021/07/poll-results-synthetic-challenges-are.html
9. Warr et al., Exploration of Ultralarge Compound Collections for Drug Discovery, J. Chem. Inf. Model. **2022**, 62 (9), 2021–2034, 10.1021/acs.jcim.2c00224
10. Grygorenko et al., Generating Multibillion Chemical Space of Readily Accessible Screening Compounds, iScience, **2020**, 23 (11), 101681, 10.1016/j.isci.2020.101681
11. Koch et al., Identification of M. tuberculosis thioredoxin reductase inhibitors based on high-throughput docking using constraints, J. Med. Chem. **2013**, 56 (12), 4849 – 4859, 10.1021/jm3015734

Poster session RED

Poster Session RED: Advanced Cheminformatics Techniques

P01: Using ChEMBL to improve molecule generation

Eloy Félix, Noel M. O'Boyle

Chemical Biology Services, EMBL's European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, UK

De novo molecule generation has become increasingly popular in recent years due to the application of AI methods such as Large Language Models and Graph Neural Networks to chemical structures [1]. After training on a large dataset of drug-like molecules (such as ChEMBL), methods such as reinforcement learning and fine-tuning can be used to guide generation towards structures that fulfill particular design goals (e.g. high docking score, low LogP). However, despite the use of approximate synthesisability measures such as RAScore, generated molecules may not be easily synthesisable or even adhere to basic chemical reasonableness.

We show how the ChEMBL database of synthesised drug-like molecules can be used to perform basic structural sanity checks that can be used as filters or components in a multi-objective optimisation as part of a generative model. These methods have been developed by ourselves [2] and others in the community [3,4] to assess the reasonableness of ring systems and scaffolds, as well as atom and bond environments.

Another common molecule generation task is to generate molecules similar to a reference via R group and linker enumeration. This can be used for idea generation, for example, or to build focussed virtual libraries. A naive enumeration procedure will suggest many groups that have already been considered (and either made or discounted), as well as groups which are not appropriate replacements in terms of properties or the synthetic space available. We describe how ChEMBL assay data can be used to infer the typical order in which R group (and linker) replacements occur, and provide the resulting dataset to support use in enumeration. A webapp (ReplaceR) has been developed that shows suggested replacements [5].

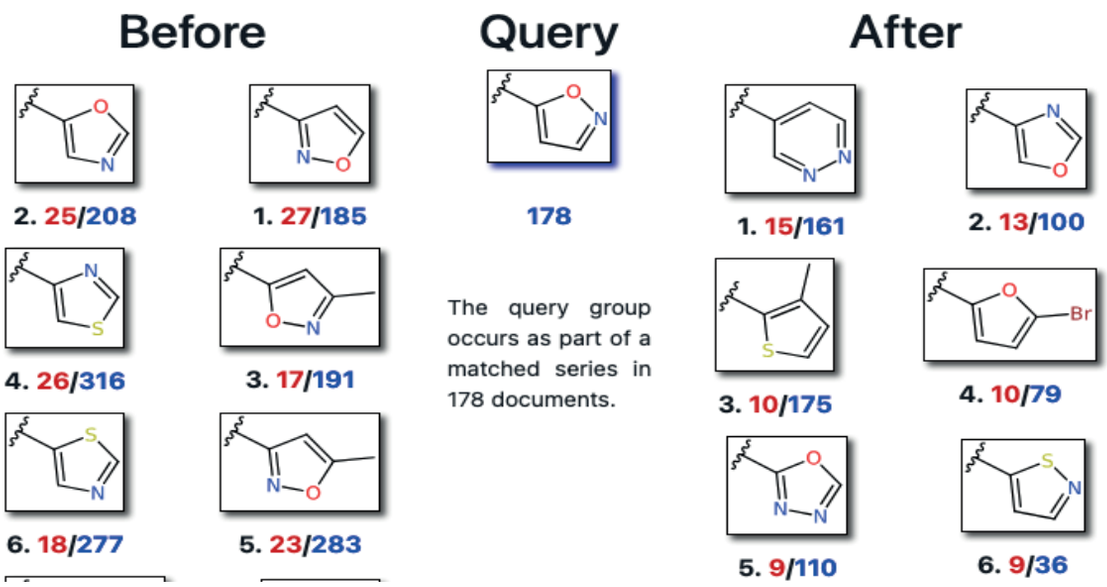


***Figure 1**: The ReplaceR webapp*

References

1. Grisoni, F. Chemical language models for de novo drug design: Challenges and opportunities. Curr. Opin. Struct. Biol. **2023**, 79, 102527
2. Félix, E. chembl_gen_check. https://github.com/eloyfelix/chembl_gen_check
3. Dehaen, W. LACAN. https://github.com/dehaenw/lacan/
4. Walters, W.P. Silly walks. https://github.com/PatWalters/silly_walks
5. O'Boyle, N.M. ReplaceR. https://baoilleach.github.io/replacer/

P03: Chemical probes in the scientific literature: Illuminating novel target-disease associations and strengthening the existing evidence

Melissa F. Adasme[1], David Ochoa[2], Irene Lopez[2], Hoang-My-Anh Do[1], Ellie MacDonagh[2], Andrew Leach[1], Noel O'Boyle[1] and Barbara Zdrazil[1]

[1] Chemical Biology Services, EMBL's European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, UK

[2] Open Targets, Wellcome Genome Campus, Hinxton Cambridgeshire, UK

Chemical probes, potent and selective small molecules, have emerged as indispensable tools in drug discovery. By specifically and selectively binding to target proteins, they enable researchers to investigate the functional roles of these proteins in cellular processes and disease states. This information is crucial for prioritizing targets for drug development. The scientific literature provides valuable insights into the use of chemical probes, revealing novel target-disease connections and strengthening previously identified links. Furthermore, when independent studies employ different, well-characterized probes to perturb the same target and consistently observe similar effects, it significantly bolsters the evidence for its role in disease. Surprisingly, despite the clear value of these published findings, systematic, large-scale investigations into the collective impact of chemical probe data remain scarce.

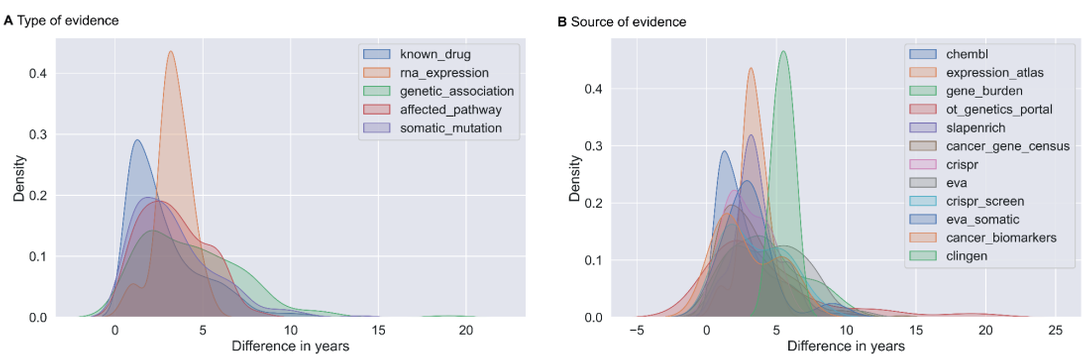


Figure 1: Chemical Probe based Evidence Precedes non-literature Open Targets evidence by decades. *This figure shows the significant time lag (years) between chemical probe publication and first non-literature evidence for target-disease associations in Open Targets, highlighting the predictive power of probe studies. (A) Grouped by evidence type. (B) Grouped by evidence data source.*

Through a natural language processing (NLP) analysis of scientific literature, we have quantified the growing impact of chemical probes in establishing novel target-disease links, particularly in contexts with limited prior evidence. We further explore how probe-based studies enhance the weight of evidence for these associations, increasing confidence in target prioritization. A time-stamp analysis of chemical probes in the literature reveals a significant time gap between the first chemical-probe-based evidence and other types of non-literature evidence in Open Targets. This highlights the potential of chemical probes to predict novel target-disease associations, especially in preclinical phases. By illuminating the intricate relationship between proteins and disease phenotypes, chemical probes are accelerating the identification of promising therapeutic targets and paving the way for more efficient and effective drug discovery.

References

1. Arrowsmith, Cheryl H., et al. "The promise and peril of chemical probes." Nature chemical biology 11.8 (2015): 536-541. DOI:10.1038/nchembio.1867
2. Antolin AA, Workman P, Al-Lazikani B. Public resources for chemical probes: the journey so far and the road ahead. Future Med Chem. 2021;13(8):731-747. DOI:10.4155/fmc-2019-0231

P05: Federated representation learning using knowledge distillation

Thierry Hanser, Jeffrey Plante, Rob Thomas Stephane Werner

Lhasa Limited, Molecular Informatics and AI team, Leeds, UK

As AI continues to transform scientific research, the need for high-quality data remains crucial. However, valuable insights are often locked within private corporate data silos, limiting collaboration and progress. FLuID (Federated Learning using Information Distillation) [1] addresses this challenge by introducing a novel approach that combines federated learning [2], information distillation, and multitask learning to enable knowledge sharing while preserving data privacy.

FLuID shifts from the conventional model-driven federated learning paradigm to a data-driven approach, utilizing surrogate non-sensitive data and a Teacher-Student [3] knowledge distillation mechanism as the primary anonymization engine. This framework allows valuable knowledge to be extracted from private datasets and shared to improve QSAR model performance and applicability domain. We demonstrate how FLuID can improve models by combining private data with federated data. Additionally, we investigate the use of Multitask Learning (MTL) to consolidate federated labels and construct a robust federated representation.

This work demonstrates that data-driven federated learning, combined with multitask learning, offers a lightweight and scalable alternative to traditional infrastructure-heavy solutions. By enabling organizations to share insights without exposing sensitive data, FLuID facilitates the development of more accurate and widely applicable predictive models.

Keywords: Federated Learning, Knowledge Sharing, Knowledge Transfer, Knowledge Distillation, Drug Discovery. Representation Learning.

References

1. Hanser T et al *Data-driven federated learning in drug discovery with knowledge distillation*. Nature Mach. Intell.(**2025**), https://doi.org/10.1038/s42256-025-00991-2 (provisory DOI)
2. Hanser T. *Federated learning for molecular discovery*. Curr. Opin. Struct Biol. (**2023**)
3. Papernot, N., Abadi, M., Erlingsson, Ú., Goodfellow, I. & Talwar, K. *Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data*. ArXiv161005755 Cs Stat (**2016**).

P07: Taming Tautomers: Extending ECFP for Tautomer Invariance

R.M.E. Pirie, J.W. Mayfield, R.A. Sayle

NextMove Software, Cambridge, UK

The problem of handling tautomers (isomers of a molecule that can readily interconvert through migration of a hydrogen) has long been considered a challenge for database registration and searching in Chemoinformatics.[1] The extent of this challenge is exemplified by considering the DNA base Guanine (**Figure 1**), which, excluding keto-enol forms, yields fifteen different possible tautomers.[2]

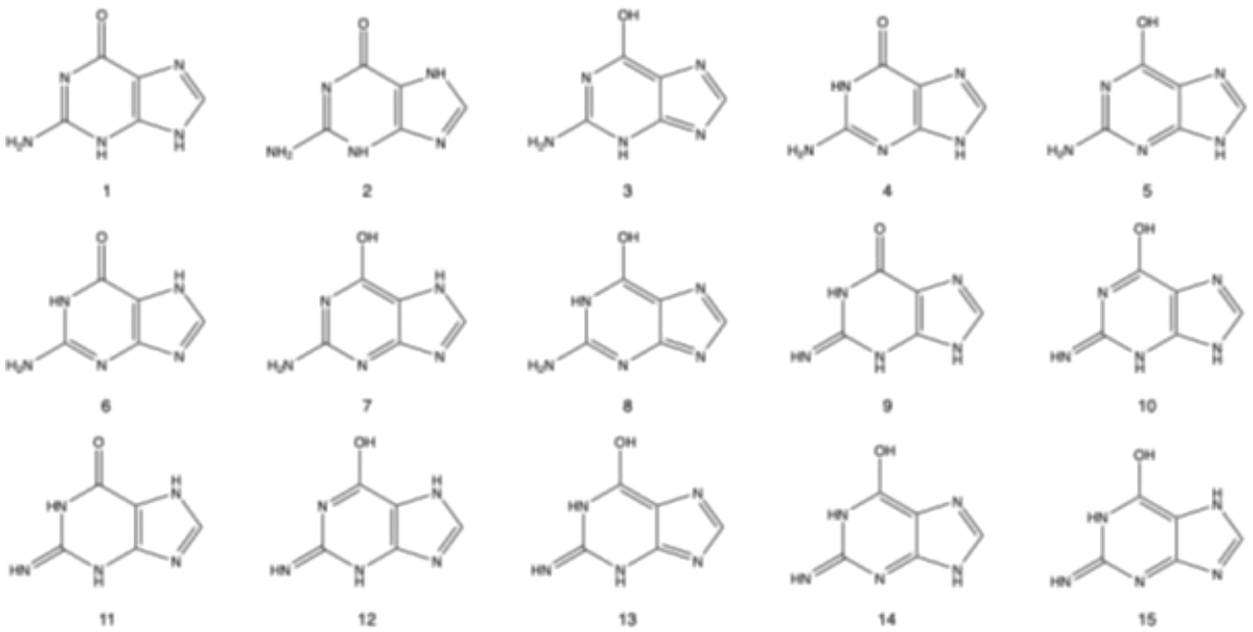


***Figure 1**: Tautomers of Guanine*

NextMove Software's ***Arthor*** provides a platform for real time search of ultra-large chemical libraries by similarity and substructure. The similarity search functionality scores molecules represented by extended-connectivity fingerprints (ECFP) based on their Tanimoto similarity. Our current implementation of the ECFP algorithm treats tautomers as distinct entities, and as such would return them as unique hits in a search, where in theory they should be treated as equivalent.

This poster will discuss the possible modifications to the atom and bond hashing used to generate the ECFP to allow for tautomer invariance. The possibilities and limitations will be discussed considering the Tautobase[3] and VEHICLe[4] databases, and a benchmarking study presented evidencing retention of speed and accuracy of the Arthor search with these modifications.

References

1. Sayle, R. A., So You Think You Understand Tautomerism?, *J Comput Aided Mol Des*, **2010**, 24, 485-496, 10.1007/s10822-010-9329-5
2. Sayle, R. A & Delaney, J. Canonicalization and Enumeration of Tautomers, *EuroMUG99*, **1999**
3. Wahl, O. & Sander, T., Tautobase: An Open Tautomer Database, *J Chem Inf Model*, **2020**, 60, 1085-1089, 10.1021/acs.jcim.0c00035
4. Pitt, W. R., Parry, D. M., Perry, G. B. & Groom, C. R, Heteroaromatic Rings of the Future, *J Med Chem*, **2008**, 52, 2952-2963, 10.1021/jm801513z

P09: Practical Molecular Property Prediction with MolPipeline – A Scientific Industry Perspective

Conrad Stork, Christian W. Feldmann, Jennifer Hemmerich, Jochen Sieg, Frederik Sandfort, Philipp Eiden, Miriam Mathea

BASF SE, Ludwigshafen, Germany

Molecular Property Prediction (MPP) is crucial for BASF in crop protection and environmental sciences. Accurate forecasting of molecular properties aids in prioritizing compounds for testing and strategic decision-making. Various computational methods, from basic physical and chemical assessments to advanced deep learning techniques, have evolved over time. [1-2]

This poster outlines the development of machine learning models using the open-source MolPipeline package. [3] MolPipeline extends scikit-learn's capabilities [4] and is designed for tasks involving small molecules as input. By utilizing RDKit, [5] MolPipeline offers an intuitive Python package for creating comprehensive pipelines that handle large datasets efficiently and manage errors and serialization, simplifying model building and enhancing reproducibility.

We will share insights from developing and implementing MolPipeline in industrial and scientific settings. The first analysis compares molecular representations like neural fingerprints in terms of confidence estimation and interpretability. The second analysis shows that MolPipeline now features explainable AI (XAI) methods to enhance user understanding and model refinement. This section details generating explanations using Random Forest models and Morgan fingerprints with SHAP's Shapley Values. The third analysis examines the benefits of integrating chemical and biological descriptors [6] for challenging in vivo endpoints using transformer models. This poster aims to highlight MolPipeline's potential for molecular property prediction in both industry and academia.

References:

1. Muratov, Eugene N., et al. "QSAR without borders." Chemical Society Reviews 49.11 (**2020**): 3525-3564.
2. Yang, Kevin, et al. "Analyzing learned molecular representations for property prediction." Journal of chemical information and modeling 59.8 (**2019**): 3370-3388.
3. Sieg, Jochen, et al. "MolPipeline: A python package for processing molecules with RDKit in scikit-learn." (**2024**).
4. Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." the Journal of machine Learning research 12 (**2011**): 2825-2830.
5. "RDKit: Open-source cheminformatics. https://www.rdkit.org"
6. Garcia de Lomana, Marina, et al. "ChemBioSim: enhancing conformal prediction of in vivo toxicity by use of predicted bioactivities." Journal of Chemical Information and Modeling 61.7 (**2021**): 3255-3272.

P11: Integrating *in Silico* Enrichment Prediction with *de Novo* Building Block Selection for Efficient Combinatorial Library Design

Remco L. van den Broek[1], A. Nandkeolyar[2], E. A. van der Nol[1], S. M. McKenna[1], M. Šícho[1,3], S. Pomplun[1], D. L. Mobley[2], G. J. P. van Westen[1], and W. Jespers[1]

[1] *Division of Medicinal Chemistry, Leiden Academic Centre for Drug Research (LACDR), Leiden University, Leiden, The Netherlands*

[2] *Department of Pharmaceutical Sciences, University of California Irvine, Irvine, California, United States of America*

[3] *Laboratory of Informatics and Chemistry, University of Chemistry and Technology, Prague, Czech Republic*

A significant challenge facing de novo molecular design is the translation of generated molecules into real-life molecules. While computer-aided synthesis planning (CASP) has accelerated the small-scale synthesis of generated molecules, it is not yet viable for obtaining large libraries. The critical limitation is the tremendous effort required by chemists to set up and optimize each reaction within an often multi-step synthesis route. To mitigate this limitation, chemists have developed a combinatorial synthesis method that allows the synthesis of thousands to millions of molecules while requiring only a few synthesis steps and a couple hundred building blocks (BBs). Through affinity selection, these libraries are screened en masse resulting in the identification of enriched BBs for binding to a specific target[1]. Here, we are developing 1) an in silico approach to predict enrichment of BBs based on experimental results and 2) a de novo platform aimed at generating high-enrichment BBs. By utilizing Thompson sampling[2] in conjunction with interaction fingerprint (IFP) docking, we determine which BBs contribute most significantly to molecule-target binding. In-house obtained affinity selection results are used for obtaining experimental validation of this approach. By coupling this approach to the DrugEx molecular generator[3], we aim to generate novel BBs that demonstrate high enrichment for target binding and map them to their nearest vendor-purchasable equivalent. Together, this platform supplements traditional combinatorial drug design with de novo-guided building block selection to achieve increased effectiveness in identifying novel high-affinity binding molecules without increasing the chemist's effort required to synthesize them.

References

1. Mata, J. M.; van der Nol, E.; Pomplun, S. J. Advances in Ultrahigh Throughput Hit Discovery with Tandem Mass Spectrometry Encoded Libraries. *J. Am. Chem. Soc.* **2023**, *145* (34), 19129–19139. https://doi.org/10.1021/jacs.3c04899.
2. Klarich, K.; Goldman, B.; Kramer, T.; Riley, P.; Walters, W. P. Thompson Sampling─An Efficient Method for Searching Ultralarge Synthesis on Demand Databases. *J. Chem. Inf. Model.* **2024**, *64* (4), 1158–1171. https://doi.org/10.1021/acs.jcim.3c01790.
3. Šícho, M.; Luukkonen, S.; van den Maagdenberg, H. W.; Schoenmaker, L.; Béquignon, O. J. M.; van Westen, G. J. P. DrugEx: Deep Learning Models and Tools for Exploration of Drug-Like Chemical Space. *J. Chem. Inf. Model.* **2023**, *63* (12), 3629–3636. https://doi.org/10.1021/acs.jcim.3c00434.

P13: MolMeDB – Molecules on Membranes Database

Storchmannová K[1], Juračka J[1,2], Martinát D[1], Bazgier V[1], Galgonek J[3], Türková A[4], Zdrazil B[5], Berka K[1]

[1] *Department of Physical Chemistry, Palacky University in Olomouc, Czech Republic*

[2] *Department of Computer Science, Palacky University in Olomouc, Czech Republic*

[3] *Institute of Organic Chemistry and Biochemistry of the Czech Academy of Sciences, Prague, Czech Republic*

[4] *Al|ffinity s.r.o., Brno-Medlánky, Czech Republic*

[5] *European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton CB10 1SA, United Kingdom*

Biological membranes serve as essential barriers that protect cells, playing a crucial role in cellular function and influencing the pharmacokinetics of drug-like small molecules. A small molecule can pass through membranes in two ways: via passive diffusion or actively via membrane transport proteins. A vast amount of data is available reporting interactions of small molecules and membranes and also interactions between small molecules and transporters.

MolMeDB (molmedb.upol.cz) is a free, comprehensive, and interactive database of interactions of small molecules with membranes.[1] From the start, we have collected data about partitioning and penetration of the small molecules crossing biological membranes. Recently, we have expanded our area of interest to include interactions of small molecules with transporters and ion channels. Nowadays, more than 930,000 interactions for almost 500,000 molecules are available in MolMeDB.

The data within the MolMeDB is collected from scientific papers, our in-house calculations (COSMOmic/COSMOperm[2]), and obtained by data mining from several databases (e.g. ChEMBL[3], PubChem[4], or IUPHAR/BPS Guide to PHARMACOLOGY[5]). Data in the MolMeDB are fully searchable and browsable by name, SMILES, membrane, method, transporter, or dataset. Also, the content of the database is available via REST API and the RDF model of MolMeDB (docs.molmedb.upol.cz).

References

1. Juračka, J., et al., MolMeDB: Molecules on Membranes Database. Database, **2019**, 2019, baz078, database/baz078.
2. Schwöbel, J. A. H., COSMO Perm: Mechanistic Prediction of Passive Membrane Permeability for Neutral Compounds and Ions and Its pH Dependence. J. Phys. Chem. B, **2020**, 124, 16, 3343–3354, 10.1021/acs.jpcb.9b11728.
3. Zdrazil, B., The ChEMBL Database in 2023: A Drug Discovery Platform Spanning Multiple Bioactivity Data Types and Time Periods. Nucleic Acids Res., **2024**, 52, D1, D1180–D1192, 10.1093/nar/gkad10044. Kim, S., PubChem 2023 Update. Nucleic Acids Res., 2023, 51, D1, D1373–D1380, 10.1093/nar/gkac956.
4. Harding, S. D., The IUPHAR/BPS Guide to PHARMACOLOGY in 2024. Nucleic Acids Res., **2024**, 52, D1, D1438–D1449, 10.1093/nar/gkad944.

P15: The next generation of the IUPAC International Chemical Identifier (InChI)

Gerd Blanke[1], Andrey Yerin[2], Clare Tovee[3], Ian Bruno[3], Jonathan Goodman[4], Richard Hartshorn[5], Ulrich Schatzschneider[6], Djordje Baljozovic[7], Felix Bänsch[8], Frank Lange[7], Jan Brammer[7], Nauman Ullah Khan[7], Sonja Herres-Pawlis[7]

[1] StructurePendium GmbH, Essen, Germany

[2] ACD/Labs, Porto, Portugal

[3] Cambridge Crystallographic Data Centre, Cambridge, UK

[4] Yusuf Hamied Department of Chemistry, University of Cambridge, Lensfield Road, UK

[5] School of Physical and Chemical Sciences, University of Canterbury, Christchurch, New Zealand

[6] Institut für Anorganische Chemie, Julius-Maximilians-Universität Würzburg, Germany

[7] Institut für Anorganische Chemie, RWTH Aachen, Germany

[8] Beilstein-Institut zur Förderung der Chemischen Wissenschaften, Frankfurt am Main, Germany

The InChI (IUPAC International Chemical Identifier) is a standardized and machine-readable alpha-numeric code that serves as a unique identifier for chemical compounds [1,2] It encodes structural information, allowing for precise identification and communication of chemical structures across databases and research platforms, promoting interoperability in the field of chemistry. It helps to make chemical data FAIR [3] and open. In the last 20 years, the InChI has been intensively used in organic chemistry (e.g. in databases with more than 100 million compounds such as PubChem) and is now heading towards applications in machine learning, restricted similarity searches and clustering analyses that aids the identification of potential drug candidates and support the understanding of chemical relationships within and between datasets.

InChI applications like the Mixture InChI (MInChI), Nano InChI (NInChI), and Reaction InChI (RInChI) extend the InChI usage into the identification of mixtures and formulations, nano materials and reactions,

Due to cheminformatic constraints, metal bonds have been disconnected right from the early days during the InChI processing which renders InChIs for metal complexes and organometallic compounds less meaningful or even meaningless caused by the information loss. Since molecular inorganic compounds gain more and more importance as catalysts, their analysis is of large interest for databases and machine learning application. Hence, we are working on the next InChI generation providing identifiers based on a correct topological description for these substance classes that has to include the complex stereochemistry of inorganics and organometallics as well while InChIs for organic compounds must not change.[4]

The base of the new functionality is the InChI Trust move away from the traditional programming to open-source development on GitHub allowing in kind contributions beside the traditional financial trust model under the scientific supervision of the IUPAC.

References

1. Heller, S., et al, InChI – the worldwide chemical structure identifier standard. Cheminform 5 7 (**2013**). DOI: https://www.doi.org/10.1186/1758-2946-5-7
2. Goodman, J. M., et. al., InChI version 1.06: now more than 99.99% reliable., J Cheminform 13 40 (**2021**). DOI: https://www.doi.org/10.1186/s13321-021-00517-z
3. Wilkinson, M.D. et al., The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data, 3, 160018 (**2016**). DOI: https://doi.org/10.1038/sdata.2016.18
4. Blanke, G., et. al., Making the InChI FAIR and sustainable while moving to inorganics; Faraday Discussions, **2025**, 256, 503, DOI: https://www.doi.org/10.1039/d4fd00145a

P17: ProQSAR: An End-to-End Framework for the Automated Construction of Predictive Models

Tuyet-Minh Phan[1], Tieu-Long Phan*[,2,3], Tuyen Ngoc Truong[1]

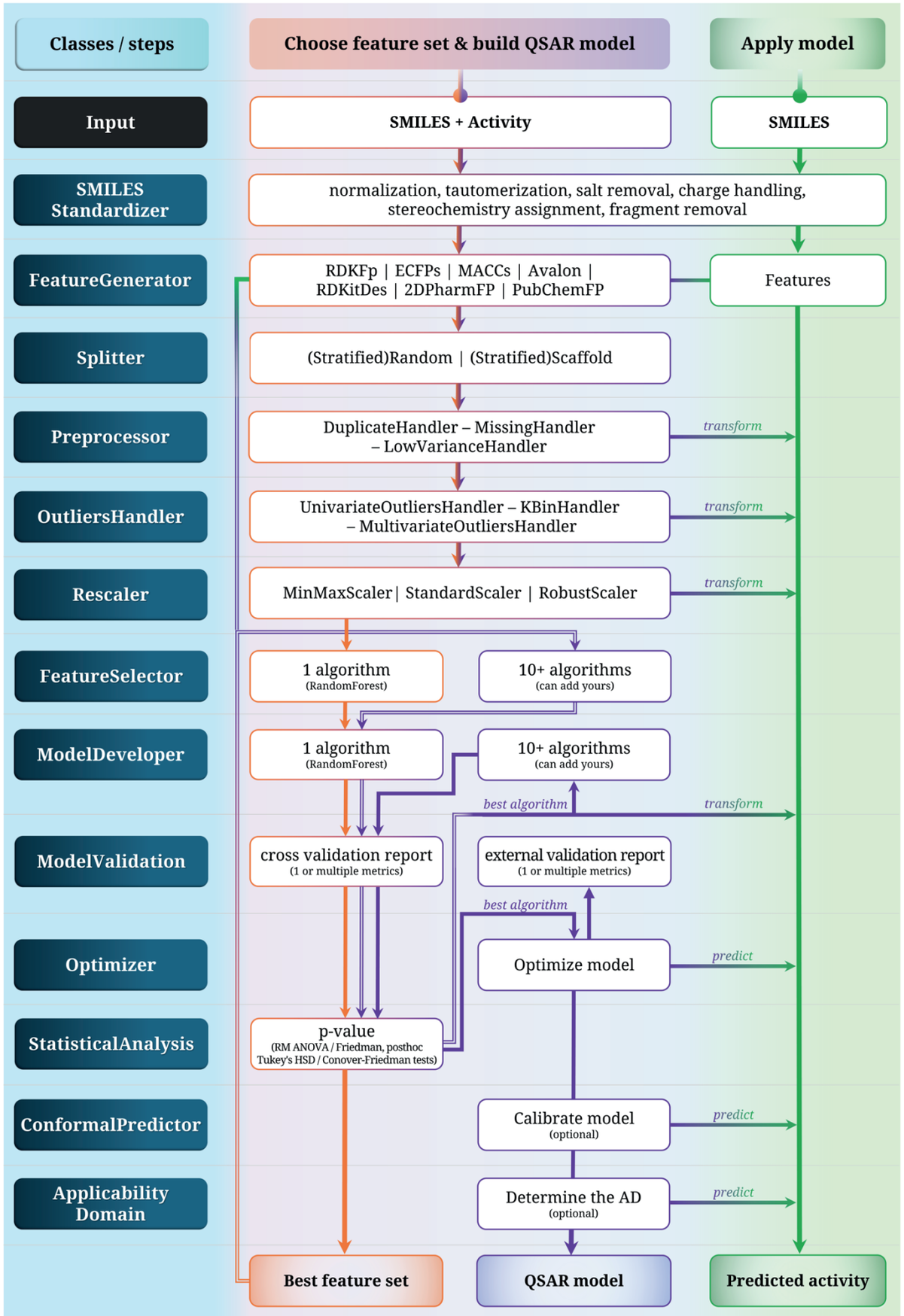
[1] *Falcuty of Pharmacy, University of Medicine and Pharmacy at Ho Chi Minh City, Ho Chi Minh City, Vietnam*

[2] *Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, Universität Leipzig, Germany*

[3] *Department of Mathematics and Computer Science, University of Southern Denmark, Odense M, Denmark*

[*]*Corresponding author. Email: tieu@bioinf.uni-leipzig.de*

The Quantitative Structure-Activity Relationship (QSAR) modeling is an essential tool in computer-aided drug design, using existing chemical compounds to predict the properties of new molecules. It's applied in various areas, such as drug discovery and environmental toxicology. Integrating deep learning has improved QSAR models' accuracy, with tools like *AutoQSAR*, *ChemProp*, *DeepTox*, *Uni-QSAR*, and *QSARtuna* broadening the scope of biological targets they can predict. However, these models often face challenges due to the limited availability of validated experimental data, which can lead to overfitting and limited generalizability. Innovative approaches like pre-training and self-supervised learning have been developed to address these issues, although the accessibility of pre-trained models and the specialized knowledge required to use them effectively remain limited. Furthermore, decisions about molecular fingerprint selection, choosing machine learning algorithms, and setting other key training parameters often lack thorough statistical validation. To overcome these challenges, we introduce *ProQSAR*, a comprehensive framework that streamlines the QSAR modeling process, requiring minimal user expertise. *ProQSAR* includes a full pipeline from choosing molecular representations and reducing features to selecting algorithms, tuning hyperparameters, and performing statistical tests. This framework ensures stringent statistical validation throughout the development process and uses conformal prediction techniques to enhance the reliability of the models. *ProQSAR* also sets a standard applicability domain for traditional QSAR models, making the tool both robust and user-friendly. *ProQSAR* is implemented in Python, utilizing major libraries like *Scikit-learn*, *RDKit*, and *Optuna*, offering an automated and efficient approach to developing QSAR models with minimal adjustments needed for high accuracy.


Classes / steps
Choose feature set & build QSAR model
Apply model
Input
SMILES + Activity
SMILES
SMILES Standardizer
normalization, tautomerization, salt removal, charge handling, stereochemistry assignment, fragment removal
FeatureGenerator
RDKFp | ECFPs | MACCs | Avalon | RDKitDes | 2DPharmFP | PubChemFP
Features
Splitter
(Stratified)Random | (Stratified)Scaffold
Preprocessor
DuplicateHandler – MissingHandler – LowVarianceHandler
transform
OutliersHandler
UnivariateOutliersHandler – KBinHandler – MultivariateOutliersHandler
transform
Rescaler
MinMaxScaler| StandardScaler | RobustScaler
transform
FeatureSelector
1 algorithm (RandomForest)
10+ algorithms (can add yours)
ModelDeveloper
1 algorithm (RandomForest)
10+ algorithms (can add yours)
best algorithm
transform
ModelValidation
cross validation report (1 or multiple metrics)
external validation report (1 or multiple metrics)
best algorithm
Optimizer
Optimize model
predict
StatisticalAnalysis
p-value (RM ANOVA / Friedman, posthoc Tukey's HSD / Conover-Friedman tests)
ConformalPredictor
Calibrate model (optional)
predict
Applicability Domain
Determine the AD (optional)
predict
Best feature set
QSAR model
Predicted activity

P19: Prefix-based decision tree for faster generation of SMARTS-based fingerprints

Andrew Dalke

Andrew Dalke Scientific AB, Sweden

Fingerprints like the MACCS[1] and PubChem[2] substructure keys and the Klekota-Roth[3] fingerprints are defined with hundreds or even thousands of substructure tests. Many implementations use one SMARTS search per test, which can be quite slow – an RDKit implementation of the 4,860 Klekota-Roth patterns processes about 72 molecules per second.

Sayle[4] described a high-performance alternative which merges multiple substructure queries into an optimized search then generates a specialized C++ or Java implementation for one of several supported cheminformatics toolkits. The implementation, Patsy, is not available as free/open-source software, and as a practical matter, distributing precompiled Python extensions for RDKit is more complicated than distributing a Python-only package which uses RDKit's Python API.

This poster presents an intermediate code-generation approach which analyzes the SMARTS patterns to generate a decision tree based on atom type screens, shared SMARTS prefixes, and the knowledge that most matches are rare. For example, the checks for "BrC(C)(Br)Br", "BrCC(C)O", and 5 other Klekota-Roth patterns can be skipped if the common prefix "BrCC" does not match. Prefix identification is improved by re-rooting the SMARTS to start with the least common element at the end of a long chain.

The decision tree is used to generate a Python module with a set of nested-if statements and RDKit SMARTS searches. The resulting decision tree fingerprinter for the Klekota-Roth patters processes about 900 molecules per second, which is over an order of magnitude faster than the original. The success is likely because a tree structure is a good match to the method used to create the Klekotha-Roth patterns. By comparison, the decision tree fingerprinter for the 755 SMARTS-defined bits in the 881-bit PubChem fingerprint is only about twice as fast as testing each of the 755 substructure patterns sequentially.

The source is available from https://hg.sr.ht/~dalke/talus under an open source license.

References

1. MACCS Structural Keys, Molecular Design Ltd., San Leandro, California, USA
2. PubChem Substructure Fingerprint V1.3, https://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.txt
3. Klekota, J. and Roth, F. P., Chemical substructures that enrich for biological activity. *Bioinformatics*, **2008,** *24*(21), 2518-2525, doi:10.1093/bioinformatics/btn479
4. Sayle, R., Efficient matching of multiple chemical subgraphs. 9th ICCS, Noordwijkerhout, The Netherlands, 9th June 2011. https://nextmovesoftware.com/products/MultipleSMARTS.pdf

P21: A Comprehensive Mapping of Chemical and Pharmacological Spaces for Drug Discovery and Repurposing: Insights from Marketed Drugs and Drug Candidates

Candida Manelfi[1], Valerio Tazzari[1], Carmen Cerchia[2], Pieter F. W. Stouten[1,3], Andrea Rosario Beccari[1]

[1] *EXSCALATE, Dompé Farmaceutici SpA, Napoli, Italy*

[2] *Department of Pharmacy, University of Naples "Federico II", Napoli, Italy*

[3] *Stouten Pharma Consultancy BV, Sint-Katelijne-Waver, Belgium*

In the era of big data, navigating the pharmacological landscape has become increasingly challenging, owing to the growing number of biological activities associated with molecules[1]. Consequently, various strategies have been developed to map the boundaries of chemical, target, and disease spaces. Among all molecular entities, drugs have received the most thorough pharmacological evaluation, as they must undergo rigorous preclinical and clinical testing. Furthermore, approved drugs are often investigated for repurposing, which broadens their biological profiles[2]. In this context, large, annotated databases play a pivotal role in enabling comprehensive analyses of chemical and pharmacological spaces, thereby facilitating drug discovery and repurposing efforts. Here, we offer an overview of the current landscape of drug discovery and development by examining the chemical and pharmacological space of both marketed drugs and investigational compounds, including those in clinical and preclinical stages. To accomplish this, we aggregated data from commercial and public sources, classifying compounds according to their highest phase of development (Figure 1).

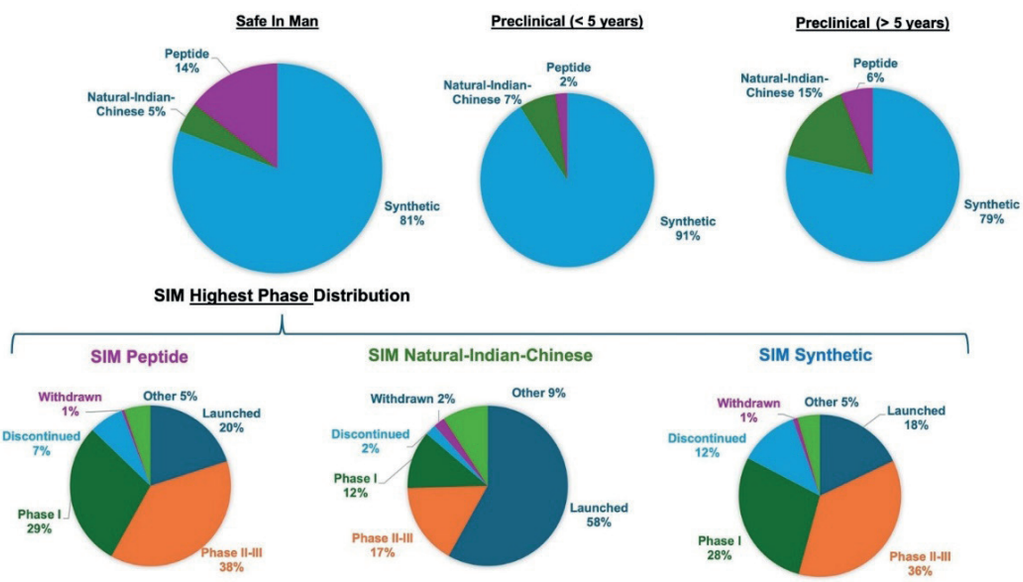


Figure 1. Pie charts showcasing the distribution of drugs according to their chemical classification (Safe in Man includes marketed drugs, compounds in clinical phases, withdrawn and discontinued).

We also performed focused analyses of specific compound classes, such as peptides and PROTACs (Figure 2).

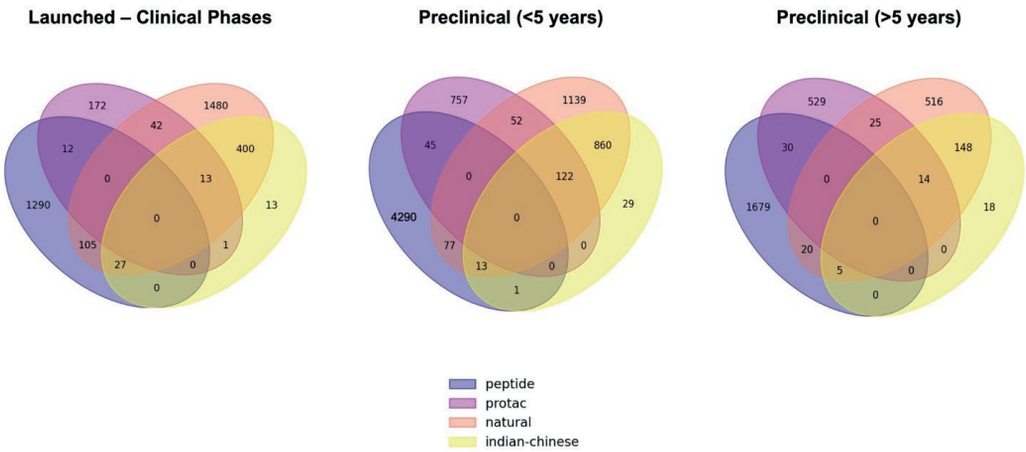


Figure 2. Venn diagrams showing unique and overlapping structures between peptides (slate), PROTACs (violet), natural products (orange), including Indian and Chinese natural products (yellow).

In addition, we cross-referenced molecular target information with protein families and associated diseases to investigate pharmacological space, revealing key trends, gaps, and strategic opportunities. Ultimately, we aim to illuminate current drug discovery trends and provide valuable guidance for future research and development strategies.

References

1. Korn, M.; Ehrt, C.; Ruggiu, F.; Gastreich, M.; Rarey, M. Navigating Large Chemical Spaces in Early-Phase Drug Discovery. *Curr Opin Struct Biol* **2023**, *80*, 102578.
2. Pushpakom, S.; Iorio, F.; Eyers, P. A.; Escott, K. J.; Hopper, S.; Wells, A.; Doig, A.; Guilliams, T.; Latimer, J.; McNamee, C. Drug Repurposing: Progress, Challenges and Recommendations. *Nat Rev Drug Discov* **2019**, *18*, 41–58.

P23: Tangible Chemical Space: Easy Access to a Fully Annotated Database of Compounds by Enumeration of Two- or Three-Component Reactions

Valerio Tazzari[1], Candida Manelfi[1], Carmen Cerchia[2] , Anna Fava[1,] and Andrea Rosario Beccari[1]

[1] *EXSCALATE, Dompé Farmaceutici SpA, Naples, Italy*

[2] *Department of Pharmacy, University of Naples "Federico II", Naples, Italy*

Recently, drug discovery projects have been dealing with increasingly large compound collections, containing millions or even billions of compounds. Currently, the chemical space ranges in size and properties, from "small" libraries of few thousands' bioactive molecules with associated data, such as ChEMBL, to larger ones, up to 10^8 compounds, such as Molport and PubChem, and even huge virtual collections, not fully enumerable, covering beyond 10^{10} compounds. The continuous growth of virtual libraries highlights the need of appropriate resources to accomplish key tasks such as virtual screening or analogue search on these libraries. We herein describe the Tangible Chemical Space (TCS), containing trillions of compounds (10^{12}-10^{15}), built starting from a database of commercially available reagents combined by robust, one-step synthetic reactions. The TCS can be employed for virtual screening applications, analogue search, and scaffold hopping, and it is seamlessly integrated with Exscalate's proprietary tools such as DompéKeys [3], a set of substructure-based descriptors, Molecular Anatomy[1], which describes a molecule on different abstraction levels allowing to rapidly analyze large compound libraries, and LiGen [2], which is able to perform docking of billions of compounds in few hours. We extensively compared TCS with large, publicly available enumerated libraries. Moreover, TCS has been thoroughly characterized for chemical novelty and pharmaceutically relevant properties, showing overlap with marketed drugs or bioactive compounds annotated in DrugBank and ChEMBL, as well as with other larger spaces.

References

[1]. Manelfi, C.; Gemei, M.; Talarico, C.; Cerchia, C.; Fava, A.; Lunghini, F.; Beccari, A.R. "Molecular Anatomy": a new multi-dimensional hierarchical scaffold analysis tool. J. Cheminform. 2021, 13, 54.

[2]. Beccari, A.R.; Cavazzoni, C.; Beato, C.; Costantino, G. LiGen: A High Performance Workflow for Chemistry Driven de Novo Design. J. Chem. Inf. Model. 2013, 53, 1518–1527.

[3].] "DompéKeys": a set of novel substructure-based descriptors for efficient chemical space mapping, development and structural interpretation of machine learning models, and indexing of large databases. Manelfi, C. et al., J. Cheminform. 2024, 16, 21.

Poster Session RED: Artificial Intelligence, Machine Learning, and QSAR

P25: The Compound Mapper: bridging practitioners and bioactivity data with automated quality control

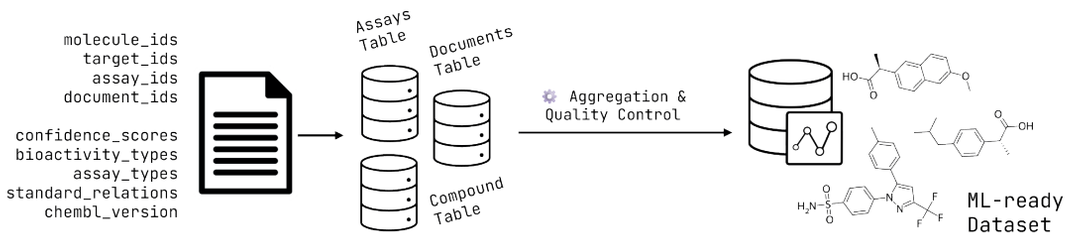
David Alencar Araripe[1,2], Linde Schoenmaker[1], Olivier Béquignon[1,3,4], Gerard J.P. van Westen[1]

[1] *Division of Medicinal Chemistry, Leiden Academic Centre for Drug Research, Leiden University, The Netherlands*

[2] *Department of Human Genetics, Leiden University Medical Centre (LUMC), The Netherlands*

[3] *Amsterdam UMC location Vrije Universiteit Amsterdam, Neurosurgery, The Netherlands*

[4] *Cancer Center Amsterdam, Cancer Biology and Immunology, Amsterdam, The Netherlands*



Advances in computing power and data availability have accelerated QSAR modeling development. However, QSAR datasets often lack consistent releases incorporating the latest ChEMBL version, limiting access to current experimental data. Fetching and aggregating ChEMBL data has been made more accessible and reproducible. (1) standardizes dataset generation through shareable SQL queries. Similarly, (2) provides a reproducible protocol for curating benchmarking compound-target activity pairs for each ChEMBL release. However, when curating bioactivities for specific targets, QSAR practitioners need to make clear decisions about data selection and aggregation methods. Research by (3) demonstrated that quality control is crucial when combining data from similar or different assay types, leading to max-curated dataset standards. While IC50 values show higher variability than Ki values, combining these can be valuable (4) when properly accounted for in analyses. Further, careful data merging is critical - a recent study by (5) highlighted how overlooking undefined stereochemistry in a dataset distorted model results. To address these challenges, we developed CompoundMapper, a Python package with a robust and flexible framework for ChEMBL dataset aggregation with comprehensive quality control measures. Essential quality checks such as data duplication and annotation error detections are applied by default. Same-compound readouts are captured through molecular fingerprints, with the option of adding further metadata to readout aggregation, such as assay description or identifiers, as the max-curated standard. All curation parameters are saved, enabling researchers to share reproducible data collection and aggregation protocols. Both Python API and command-line interface are provided, supporting data collection through any combination of molecule, target, assay, or document IDs.

References

1. Hoyt CT. chembl-downloader: Write reproducible code for getting and processing ChEMBL [Internet]. Github; [cited 2025 Feb 6]. https://github.com/cthoyt/chembl-downloader

2. Heinzke AL, Zdrazil B, Leeson PD, Young RJ, Pahl A, Waldmann H, et al. A compound-target pairs dataset: differences between drugs, clinical candidates and other bioactive compounds. Sci Data., 2024 Oct 21;11(1):1160. https://www.nature.com/articles/s41597-024-03582-9

3. Landrum GA, Riniker S. Combining IC50 or Ki values from different sources is a source of significant noise. J Chem Inf Model., 2024 Mar 11;64(5):1560–7. https://pubs.acs.org/doi/10.1021/acs.jcim.4c00049

4. Kalliokoski T, Kramer C, Vulpetti A, Gedeck P. Comparability of mixed IC□□ data - a statistical analysis. PLoS One., 2013 Apr 16;8(4):e61007. https://pmc.ncbi.nlm.nih.gov/articles/PMC3628986/

5. Tetko IV, van Deursen R, Godin G. Be aware of overfitting by hyperparameter optimization! J Cheminform., 2024 Dec 9;16(1):139. https://jcheminf.biomedcentral.com/articles/10.1186/s13321-024-00934-w

P27: Discovering novel beta-lactamase inhibitors with an AI-based virtual pipeline

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

[1]Medicinal Chemistry, Leiden Academic Centre for Drug Research, The Netherlands

[2] Systems Pharmacology and Pharmacy, Leiden Academic Centre for Drug Research, Leiden University, The Netherlands

[3] Certara, University Road, Canterbury Innovation Centre, Unit 43, Kent, UK

Over the past few decades, the discovery of new antibiotics has stagnated [1]. Consequently, the rise of antibiotic resistance is a major global health crisis with severe health and economic consequences [2]. Artificial intelligence (AI) has the potential to accelerate antibiotic discovery through the effective use of available data [3], [4]. Generative *de novo* drug design, in particular, provides a promising approach to efficiently explore the vast chemical space for novel antibiotics by learning from compounds with known antibiotic activity. Beta-lactam antibiotics are one of the most widely used classes of antibiotics [5]. Their primary resistance mechanism is through the production of beta-lactamase enzymes, making beta-lactamase inhibitors crucial for maintaining the effectiveness of beta-lactam antibiotics [6]. Therefore, this study aims to use a generative AI pipeline to discover new beta-lactamase inhibitors.

We generated 10,000 novel beta-lactamase inhibitors using DrugEx, a reinforcement learning *de novo* generation framework [7], [8]. This framework balances the exploration of novel chemical space with the exploitation of known beta-lactamase inhibitor structures. The DrugEx graph-based neural network was optimized using feedback from a quantitative structure-activity relationship (QSAR) model for beta-lactamase inhibition. This QSAR model was trained on open-source bioactivity data retrieved from Papyrus [9], a large-scale curated bioactivity dataset. The filtered dataset contained data for approximately 1,100 compounds targeting beta-lactamases from various bacterial strains. From the generated compounds, a selection of 54 purchasable compounds was made that we are planning to *in vitro* validate on a panel of gram-negative bacterial strains.

References

1. Lewis, K., The Science of Antibiotic Discovery. Cell, **2020**, vol. 181, no. 1, pp. 29–45, doi: 10.1016/j.cell.2020.02.056.
2. O'Neill, J., Antimicrobial Resistance: Tackling a Crisis for the Health and Wealth of Nations. Review on Antimicrobial Resistance, **2014**.
3. Stokes, J. M., et al., A Deep Learning Approach to Antibiotic Discovery. Cell, **2020**, vol. 180, no. 4, pp. 688-702.e13, doi: 10.1016/j.cell.2020.01.021.
4. Melo, M. C. R., et al., Accelerating antibiotic discovery through artificial intelligence. Commun Biol, **2021**, vol. 4, no. 1, pp. 1–13, doi: 10.1038/s42003-021-02586-0.
5. World Health Organization, WHO report on surveillance of antibiotic consumption: 2016-2018 early implementation. **2018**. Accessed: Feb. 24, 2025. [Online]. Available: https://www.who.int/publications/i/item/who-report-on-surveillance-of-antibiotic-consumption
6. Tooke, C.L., et al., β-Lactamases and β-Lactamase Inhibitors in the 21st Century. Journal of Molecular Biology, **2019**, vol. 431, no. 18, pp. 3472–3500, doi: 10.1016/j.jmb.2019.04.002.
7. Liu, X., et al., An exploration strategy improves the diversity of de novo ligands using deep reinforcement learning: a case for the adenosine A2A receptor. Journal of Cheminformatics, **2019**, vol. 11, no. 1, p. 35, doi: 10.1186/s13321-019-0355-6.
8. Šícho, M., et al., DrugEx: Deep Learning Models and Tools for Exploration of Drug-Like Chemical Space, J. Chem. Inf. Model., **2023**. vol. 63, no. 12, pp. 3629–3636, doi: 10.1021/acs.jcim.3c00434.

9. Béquignon, O. J. M., et al., Papyrus: a large-scale curated dataset aimed at bioactivity predictions. Journal of Cheminformatics, **2023**, vol. 15, no. 1, p. 3, doi: 10.1186/s13321-022-00672-x.

P29: Balancing Complexity and Efficiency: Scalable Machine Learning Approaches for Reaction Yield Prediction

Idil Ismail & Sereina Riniker

Department of Chemistry and Applied Biosciences, ETH Zürich,Switzerland

Chemical reaction yields serve as a key metric for assessing the efficiency of synthetic routes. Traditionally, experimentalists have relied on heuristics and chemical intuition to optimize reaction conditions. More recently, machine learning (ML) has emerged as a powerful tool for yield prediction, often employing expensive Density Functional Theory (DFT)-derived descriptors or deep learning architectures. While effective, these approaches introduce significant computational overhead and require large datasets, thereby limiting their applicability for high-throughput screening and reaction discovery. In this work, we show that classical ML models, when coupled with inexpensive topological fingerprints, achieve predictive performance comparable to more complex methods. Topological descriptors, which encode molecular connectivity and structure, provide a computationally efficient alternative to DFT-derived features without compromising on accuracy. We demonstrate this approach on two widely studied catalytic reactions, namely, the Buchwald-Hartwig amination and Suzuki coupling, showing that our model effectively captures key factors influencing reaction yield. Additionally, we address a critical but often overlooked challenge: dataset bias. Commonly used datasets tend to overrepresent either low-yielding reactions (high-throughput experimental data) or unreliable high-yielding examples (reaction databases such as USPTO). These biases can distort model performance and limit generalizability to real-world reaction conditions. To mitigate this, we construct more balanced training sets, improving model robustness and applicability. By balancing complexity and interpretability, we establish a practical and scalable methodology for reaction yield prediction. Overall, our findings suggest that highly engineered, resource-intensive models are not strictly necessary for reliable predictions. Instead, simpler ML frameworks provide an effective and accessible baseline, particularly valuable for applications requiring rapid reaction evaluation, such as combinatorial synthesis and automated reaction optimization.

References

1. Rinehart, N. I., et al., A machine-learning tool to predict substrate-adaptive conditions for Pd-catalyzed C–N couplings. Science, **2023**, 381, 6661, 965–972, DOI: 10.1126/science.adg2114
2. Ahneman, D. T., et al., Predicting reaction performance in C–N cross-coupling using machine learning. Science, **2018**, 360, 6385, 186–190, https://doi.org/10.1126/science.aar5169.

P31: Bio-Isostere Guided Molecular Property Prediction

Anatol Ehrlich[1], Nils M. Kriege[1], Christoph Flamm[2]

[1] *Faculty of Computer Science, University of Vienna, Austria*

[2] *Department of Theoretical Chemistry, University of Vienna, Austria*

Accurate prediction of molecular properties is crucial for drug discovery and materials science. While self-supervised learning leverages abundant unlabeled data, it frequently fails to incorporate crucial chemical information limiting its effectiveness. We developed a novel set of rules for generating biologically similar molecules based on bio-isosteres[1], which are used to learn a pre-trained model. A triplet mining strategy[2] as shown in Figure 1, allows a graph neural network (GNN) to learn meaningful embeddings by placing two similar molecules close to another and a dissimilar molecule farther apart.

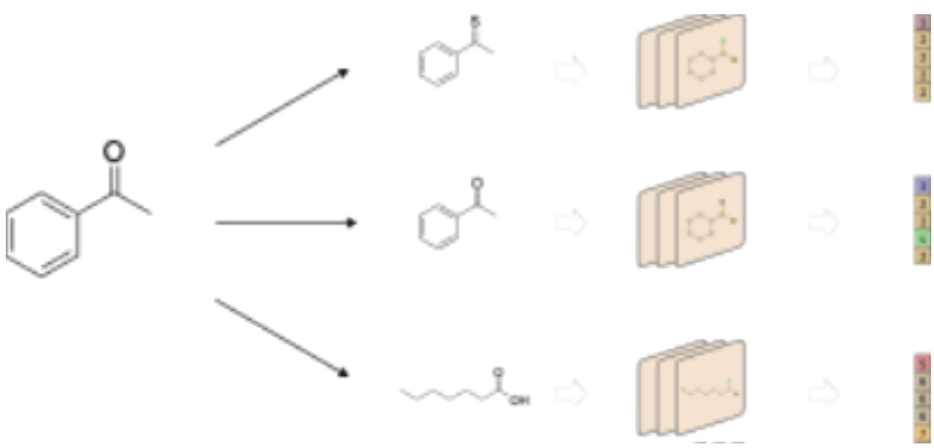


Figure 1*: Generation of a triplet consisting of two similar, and one dissimilar molecule, and their embedding values after being passed through a graph neural network.*

The quality of the learned embeddings was evaluated during fine-tuning by comparing their predictive performance against a randomly initialized pre-trained model and Extended Connectivity Fingerprints[3] (ECFPs). Our results show that the pre-training procedure outperforms the randomly initialized model in all 4 critical classification tasks related to drug activity and toxicity and surpasses ECFPs in one task. This demonstrates the potential of our pre-training strategy to generate highly informative molecular embeddings.

References

1. Mannhold et al., *Methods and Principles in Medicinal Chemistry*. Wiley-VCH, 2012, 10.1002/SERIES6138
2. Schroff et al., 'FaceNet: A Unified Embedding for Face Recognition and Clustering'. In 2015 IEEE *Conference on Computer Vision and Pattern Recognition* (CVPR), 815–23. Boston, MA, USA: IEEE, 2015. https://doi.org/10.1109/CVPR.2015.7298682.
3. Rogers, David, and Mathew Hahn. 'Extended-Connectivity Fingerprints'. *Journal of Chemical Information and Modeling* 50, no. 5 (24 May 2010): 742–54. https://doi.org/10.1021/ci100050t.

P33: Prediction of *in vivo* PK Profiles from Chemical Structures and *in vitro* ADME Experiments

Moritz Walter[1], Bettina Gerner[2], Hermann Rapp[2], Hannes Wendelin[3], Christofer Tautermann[1], Miha Skalic[1], Jens Markus Borghardt[2], Lina Humbeck[1]

[1 2] Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach, Germany

[1] Medicinal Chemistry [2] Drug Discovery Sciences

[3] Boehringer Ingelheim RCV GmbH & Co KG, Cancer Research DDS ADME, Wien, Austria

A successful drug needs to combine several properties including high potency and a good pharmacokinetic (PK) profile to sustain efficacious plasma concentration over time. To estimate required doses for preclinical animal efficacy models or for the clinics, *in vivo* PK studies need to be conducted. Here, we present machine learning (ML) models to predict *in vivo* PK profiles for both i.v. (intravenous) and p.o. (*per os*, i.e. oral) application routes.

In a first step, compartmental models were fitted to *in vivo* profiles from in-house PK studies. Next, multi-task graph-based neural network were trained to predict the compartmental parameters of test compounds. As input features, both chemical structures and predictions of *in vitro* ADME data from a separate multi-task model[1] were used (see Figure 1). Improving on our study using preclinical in-house rat PK data (i.v.)[2], accurate predictions were obtained with a geometric mean fold error (GMFE) of <3fold for around 62% of compounds when comparing to fitted PK profiles. When compounds have been characterized in *in vitro* ADME assays, this information can be included to refine predictions. The framework allows the integration of multiple relevant preclinical PK species for i.v. and p.o. administration in a single ML model.

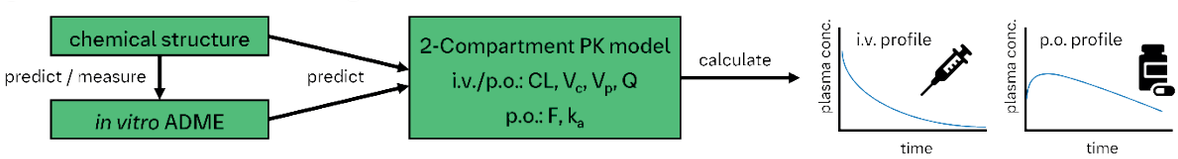


***Figure 1**: Visual summary of the PK profile prediction approach.*

When combined with on-target potency information, our presented models may accelerate drug discovery projects by enabling (i) the prioritization of new compound designs before synthesis as well as (ii) the selection among compounds characterized *in vitro* for *in vivo* studies.

References

1. Walter, M., et al., Multi-Task ADME/PK prediction at industrial scale: leveraging large and diverse experimental datasets. Molecular Informatics, **2024**, 43, 10, e202400079, https://doi.org/10.1002/minf.202400079
2. Walter, M., et al., In silico PK predictions in Drug Discovery: Benchmarking of Strategies to Integrate Machine Learning with Empiric and Mechanistic PK modelling. bioRxiv., **2024.** https://doi.org/10.1101/2024.07.30.605777

P35: Censored Loss: Inclusion of Censored Data in Molecular Affinity Modelling

Marc A. Boef*, R. L. van den Broek, G. J. P. van Westen

Division of Medicinal Chemistry, Leiden Academic Center for Drug Research, The Netherlands

Modelling of molecular affinities requires large amounts of high quality datapoints. A variety of datasets are available, such as Papyrus[1], that have compiled and standardized bioactivity datapoints from published sources. This dataset currently contains over of 60 million datapoints, roughly 2% of which are considered 'high quality', viable for training of regression models. Another 0.5% of data is censored data, representing a range of values in which the exact affinity lies. The censoring is typically due to an upper or lower limit sensitivity threshold of an assay. While censored data is not ideal for training regression models, it does contain valid information from which the model can learn. Properly incorporating censored data into training can allow for better, more representative modelling of affinities and physicochemical properties. In the case of the Papyrus dataset, it would mean expanding the total number of samples by 25%.

An approach to this, SurvLoss[2], describes the use of a mean squared error (MSE) loss in the application of survival modelling on censored data. Based on this work, we have reimplemented an MSE loss function that considers the relation of censored data, allowing for its use in regression modelling. Our implementation allows for the use of both right- and left-censored data simultaneously. We apply the censored loss function to the PyBoost algorithm. As a set for benchmarking, the kinase affinity modelling sets defined by Sohvi et al.[3] were used. These were then modified to include censored data with a separate table defining the relation of each datapoint. Impact of inclusion and exclusion of censored data on performance and applicability domain will be assessed.

References

1. Béquignon, O. J. M., et al. "Papyrus: A Large-Scale Curated Dataset Aimed at Bioactivity Predictions." Journal of Cheminformatics, **2023**, 15(1), pp 3, 10.1186/s13321-022-00672-x.
2. Rahat, Mahmoud, and Zahra Kharazian. "SurvLoss: A New Survival Loss Function for Neural Networks to Process Censored Data." PHM Society European Conference, **2024**, 8(1), pp 7, 10.36001/phme.2024.v8i1.4052.
3. Luukkonen, Sohvi, et al. "Large-Scale Modeling of Sparse Protein Kinase Activity Data." Journal of Chemical Information and Modeling, **2023**, 63 (12), pp 3688–3696, 10.1021/acs.jcim.3c00132.

P37: Evaluating Machine Learning Models for Molecular Property Prediction: Performance and Robustness on Out-of-Distribution Data

Hosein Fooladi[1,2,3], Thi Ngoc Lan Vu[1,2,3], and Johannes Kirchmair[1,2]

[1]*Christian Doppler Laboratory for Molecular Informatics in the Biosciences, Department for Pharmaceutical Sciences, University of Vienna, Austria*

[2]*Department of Pharmaceutical Sciences, Division of Pharmaceutical Chemistry, University of Vienna, Austria*

[3]*Vienna Doctoral School of Pharmaceutical, Nutritional and Sport Sciences (PhaNuSpo), University of Vienna, Austria*

Today, machine learning models are employed extensively to predict the physicochemical and biological properties of molecules.[1] Yet, their performance is typically evaluated on in-distribution (ID) data, i.e., data originating from the same distribution as the training data. However, the real-world applications of such models often involve molecules that are more distant from the training data, which necessitates assessing their performance on out-of-distribution (OOD) data.[2] In this work, we investigate and evaluate the performance of 12 machine learning models, including classical approaches like random forests, as well as graph neural network (GNN) methods, such as message-passing graph neural networks, across 8 data sets using 7 splitting strategies for OOD data creation.

First, we investigate what indeed constitutes OOD data in the molecular domain for ADMET and high-throughput screening prediction tasks. In contrast to the common point of view, we show that both classical ML and GNN models work well on scaffold splitting based on Bemis-Murcko scaffolds (not significantly different from random splitting). Splitting based on chemical similarity clustering (K-Means using ECFP4 fingerprints) poses the hardest challenge for both types of models.

Second, we investigate the extent to which ID and OOD data have a positive linear relationship. If a positive correlation holds, models with the best performance on the ID data can be selected with the promise of having the best performance on OOD data.[3] We show that the strength of this linear relationship is strongly related to how the OOD data is created, i.e., which splitting strategies are used for creating OOD data. While the correlation between ID and OOD for scaffold splitting is strong (Pearson r ~0.9), this correlation decreases significantly for cluster-based splitting (Pearson r ~0.4). Therefore, a strong positive correlation is not guaranteed for all OOD scenarios, and the relationship can be more nuanced. These findings suggest that OOD evaluation and model selection strategies should be carefully aligned with the intended application domain.

References

1. Askr, H., et al., Deep Learning in Drug Discovery: An Integrative Review and Future Challenges. Artificial Intelligence Review, **2023**, 56, 5975–6037, DOI: 10.1007/s10462-022-10306-1
2. Tossou, P., et al., Real-World Molecular Out-Of-Distribution: Specification and Investigation. Journal of Chemical Information and Modeling, **2024**, 64, 697–711, DOI: 10.1021/acs.jcim.3c01774
3. Miller, John P., et al., Accuracy on the Line: on the Strong Correlation Between Out-Of-Distribution and In-Distribution Generalization. International Conference on Machine Learning, **2021**, 7721–7735, DOI: 10.48550/arXiv.2107.04649

P39: Trialblazer: A Chemistry-Focused Predictor of Toxicity Risks in Late-Stage Drug Development

Huanni Zhang[1,2,3], Matthias Welsch[1,2,3], William Schueller[1], Johannes Kirchmair*[1,2,3]

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

[2]*Christian Doppler Laboratory for Molecular Informatics in the Biosciences, Department for Pharmaceutical Sciences, University of Vienna, Austria*

[3]*Vienna Doctoral School of Pharmaceutical, Nutritional and Sport Sciences (PhaNuSpo), University of Vienna, Austria*

Failures in drug discovery research remain a major concern, as they block substantial

resources for efforts that ultimately prove futile.[1] A recent review reports that unmanageable toxicity accounts for approximately 30% of clinical trial failures (second only to the primary reason, which is lack of clinical efficacy, accounting for 40–50% of failures).[2] However, methods to predict and evaluate such toxicity remain limited, especially when toxicity arises from off-target or target-unrelated mechanisms during the late stage. In this study, we compiled a novel dataset comprising 1610 benign compounds and 246 toxic compounds, specifically designed to address late-stage unexpected toxicity. Using Morgan2 fingerprints and similarity-based bioactivity fingerprints, we developed a Multi-layer Perceptron (MLP) classification model to predict late-stage unexpected toxicity. In validation experiments, the model achieved ROC-AUC and MCC values of up to 0.87 and 0.44, respectively. While these results demonstrate the model's ability to differentiate between toxic and benign compounds even without prior knowledge of the relevant target, the results also reveal limitations related to the training on small datasets, as altering training features led to only marginal improvements in performance metrics. This study provides a knowledge-based dataset and an auxiliary model for predicting late-stage unexpected toxicity, offering a valuable tool to enhance decision-making and efficiency in early drug discovery and development .

References

1. Hay, M.; Thomas, D. W.; Craighead, J. L.; Economides, C.; Rosenthal, J. Clinical Development Success Rates for Investigational Drugs. *Nat. Biotechnol.* **2014**, 32 (1), 40–51. https://doi.org/10.1038/nbt.2786.
2. Sun, D.; Gao, W.; Hu, H.; Zhou, S. Why 90% of Clinical Drug Development Fails and How to Improve It? *Acta Pharm. Sin. B* **2022**, *12* (7), 3049–3062. https://doi.org/10.1016/j.apsb.2022.02.002.

P41: Fast and Scalable 3D Pharmacophore Screening with PharmacoMatch

Daniel Rose[1,2,3], Oliver Wieder[1,2], Thomas Seidel[1,2], Thierry Lager[1,2]

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

[2] *Christian Doppler Laboratory for Molecular Informatics in the Biosciences, Department of Pharmaceutical Sciences, University of Vienna, Austria*

[3] *Vienna Doctoral School of Pharmaceutical, Nutritional and Sport Sciences, University of Vienna, Austria*

3D pharmacophore screening is a powerful virtual screening method, offering computational efficiency, scaffold hopping for retrieval of diverse hit lists, and an intuitive representation that facilitates communication with medicinal chemists[1]. However, its application to ultra-large molecular libraries is constrained by the computational cost of pharmacophore alignment, which underlies the matching of a query pharmacophore to database molecules. While feasible for millions of compounds, scaling to ultra-large libraries remains a challenge[2].

In this work, we introduce PharmacoMatch[3], a novel contrastive learning approach that leverages learned representations for virtual screening. By formulating pharmacophore screening as an approximate subgraph matching problem, our method enables fast and efficient pharmacophore matching within a structured embedding space that captures query-target relationships[4]. Our model is trained in a self-supervised fashion on unlabeled data, making it adaptable across datasets without explicit supervision.

We evaluate the zero-shot pre-screening performance of our model on several benchmark datasets, demonstrating comparable accuracy to existing methods while significantly reducing runtime. Our results highlight the potential of deep learning-driven pharmacophore screening for scaling virtual screening workflows to large-scale molecular libraries.

References

1. Gerhard Wolber et al., LigandScout: 3-d pharmacophores derived from protein-bound ligands and their use as virtual screening filters. J. Chem. Inf. Model., **2005**, 45, 1, 160-169, DOI: 10.1021/ci049885e.
2. Wendy A. Warr et al., Exploration of ultralarge compound collections for drug discovery. J. Chem. Inf. Model., **2022**, 62, 9, 2021-2034, DOI: 10.1021/acs.jcim.2c00224.
3. Daniel Rose et al., PharmacoMatch: Efficient 3D pharmacophore screening via neural subgraph matching, The Thirteenth International Conference on Learning Representations, **2025**, URL: https://openreview.net/forum?id=27Qk18IZum.
4. Rex Ying et al., Neural subgraph matching, arXiv preprint, **2020**, DOI: 10.48550/arXiv.2007.03092.

P43: Unlocking the Potential of C2-Carboxylated 1,3-Azoles: A Computational Design-Make-Test-Analyze (DMTA) Approach

Kerrin Janssen[1], Johannes Kirchmair[2,3], Jonny Proppe[1]

[1] *Institute of Physical and Theoretical Chemistry, TU Braunschweig, Germany*

[2] *Department of Pharmaceutical Sciences, University of Vienna, Austria*

[3] *Christian Doppler Laboratory for Molecular Informatics in the Biosciences, Department for Pharmaceutical Sciences, University of Vienna, Austria*

C2-carboxylated 1,3-azoles represent a structurally diverse yet underutilized class of compounds with significant potential in pharmaceuticals, cosmetics, and agrochemicals [1]. However, only a fraction of the millions of available 1,3-azoles are carboxylated at the C2 position, highlighting opportunities for expanding their synthetic accessibility and functional applications [1,2]. To systematically explore this chemical space, we employ a data-driven design-make-test-analyze (DMTA) cycle, integrating machine learning, virtual screening, and molecular docking.

To improve synthetic accessibility, we developed an explainable machine learning model to predict reaction yields for amide-coupled C2-carboxylated 1,3-azoles [2]. To further enhance molecular design, we implemented a heat-mapping algorithm, PIXIE (Predictive Insights and Xplainability for Informed chemical space Exploration), which visualizes substructural contributions to predicted yields, enabling rational molecule selection [2]. These computational tools integrate computer-aided synthesis planning into the DMTA cycle, streamlining the transition from design to experimental validation.

For biological exploration, we performed pharmacophore-based target prediction and fastROCS shape-based screening of over 3,000 commercially available C2-carboxylated 1,3-azoles, identifying Vanilloid Receptor 1 (TRPV1) as a novel target for this substance class. Expanding on this, we conducted structure-based docking and virtual screening to prioritize hits. These approaches are now being integrated into a collaborative experimental workflow, where selected compounds are evaluated in biological assays.

References

1. Janssen, K., et al., Relevance and Potential Applications of C2-Carboxylated 1,3-Azoles. ChemMedChem., **2024**, 19, 21, e202400307, 10.1002/cmdc.202400307
2. Janssen, K., et al., Predicting and Explaining Yields with Machine Learning for Carboxylated Azoles and Beyond. Journal of Chemical Information and Modeling., **2025**, ASAP, 10.1021/acs.jcim.4c02336

P45: Application of DFT to assess nitrosamine formation risks: *Tertiary amines aren't a risk, except when they are*

E. Pye, M. Kawamura, M. Burns, C. Barber

Lhasa Limited, Granary Wharf House, Leeds, UK

Importance: This work holds significant regulatory value and aims to enhance our software, Mirabilis, in line with the ICH M7 guidelines on assessing and controlling DNA-reactive (mutagenic) impurities in pharmaceuticals.

Background: Nitrosamines, ICH M7 and the importance of assessing tertiary amines

Nitrosamines as a Regulatory Concern

Nitrosamines are a class of indirect-acting mutagens that can form DNA-alkylating diazonium ions through metabolic activation. Nitrosamines are classified as probable human carcinogens by the WHO's International Agency for Research on Cancer (IARC) and belong to the cohort of concern in the ICH M7 guideline, which sets strict control measures for mutagenic impurities in pharmaceuticals.

The discovery of nitrosamines in several pharmaceuticals – starting with valsartan (2018) and ranitidine (2019) – prompted global regulatory action. These incidents have led to drug recalls and regulatory guidance that requires the re-evaluation of synthetic and formulation processes to assess for nitrosamine formation risks. Comprehensive risk assessments, analytical testing and stricter impurity limits to prevent nitrosamine contamination are now required from regulators, including the FDA and EMA.[1,2]

ICH M7 and Nitrosamine Risk Management

The ICH M7 guideline – which is designed to control mutagenic impurities arising from manufacturing processes and degradation pathways using the Threshold of Toxicological Concern (TTC) approach – requires (due to their high-potency and thus included in the *cohort of concern*) nitrosamines to have compound-specific acceptable intake (AI) limits, which can be significantly lower than the general 1.5 µg/day threshold for other mutagenic impurities.

The Importance of Studying Tertiary Amines

Nitrosamine generation necessitates three elements: nitrosatable substrate, nitrosating agent and specified reaction conditions. Tertiary amines (nitrosatable substrate) – in the presence of a nitrosating agent – are generally considered less prone to nitrosamine formation than secondary amines (primary amines are not considered a risk for nitrosamine formation), as they require an additional nitrosative dealkylation step to generate a nitrosatable secondary amine intermediate (Figure 1).

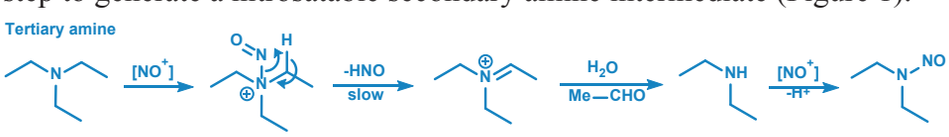


Figure 1: *Mechanistic scheme for the nitrosation of tertiary amines to form nitrosamines.*

However, increasingly, there is concern over tertiary amines, particularly those with activating structural features that could facilitate nitrosamine formation under certain conditions. For example, gramine is well documented to form N-Nitrosodimethylamine (NDMA) as a result of direct intramolecular nucleophilic attack on the nitrosammonium ion to form a cyclic structure and displace NDMA.3

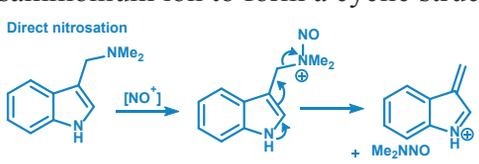


Figure 2: *Nitrosative dealkylation of gramine.*

Whilst secondary amines are more likely to undergo direct nitrosation, they are more likely to lead to higher CPCA (Cohort of Concern for Carcinogenicity Assessment) category compounds. Whereas, tertiary amines – despite being less reactive – account for the majority of potential CPCA category 1 and 2 compounds. This suggests that tertiary amines, especially those with specific structural predispositions, should not be overlooked in nitrosamine risk assessments. [4]

This poster presents the research that was conducted to develop a DFT-derived training set that is anticipated to be used in the development of a machine learning model in order to predict the propensity of tertiary amines to form nitrosamines (this model will subsequently be integrated into Mirabilis in order to significantly improve Mirabilis' ability to predict the formation of nitrosamines from tertiary amines). This study explores the potential relationship between the bond dissociation energy (BDE) of the C-N bond in the nitrosated tertiary amines and their propensity for nitrosamine formation (Figure 3).

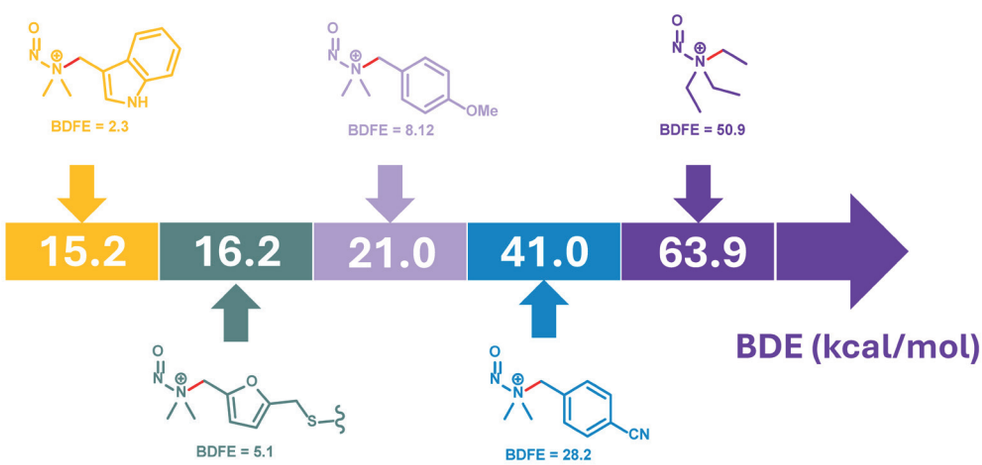


***Figure 3**: Trend in C-N BDE and Nitrosamine Formation Risk.*

References

1. Organic Process Research & Development., **2022**, 27, 10, 1693-1702;
2. Organic Process Research & Development., **2020**, 24, 12, 2915-2926
3. Organic Process Research & Development., **2023**, 27, 10, 1714-1718
4. Journal of Pharmaceutical Sciences., **2023**, 112, 12, 3005-3011.

P47: Closing the generative AI for SBDD loop: From GPCR structure to reinforcement learning guided de novo ligand design and back again

Chris de Graaf

Structure Therapeutics, USA / China

Generative chemical language models have demonstrated success in learning language-based molecular representations for de novo drug design. Here, we integrate structure-based drug design (SBDD) principles with chemical language models to present a modern hit-finding workflow to go from protein structure to novel small-molecule ligands, without a priori knowledge of ligand chemistry. Using Augmented Hill-Climb we successfully optimised multiple objectives within a practical timeframe, including protein-ligand complementarity.

Generated de novo molecules contained both known and promising adenosine A2A receptor ligand chemistry that is not available in commercial vendor libraries, accessing commercially novel areas of chemical space. Experimental validation identified three nanomolar ligands with confirmed functional activity, two of which contain novel chemotypes. Overall, demonstrating a binding hit rate of 88% with 50% of the binders demonstrating confirmed functional activity emphasising the complex relationships in translating binding to downstream pharmacology. The two strongest binders were co-crystallised with the A2A receptor revealing their binding mechanisms that can be used to inform future iterations of structure-guided de novo design, closing the AI SBDD loop.

Additional generative AI for SBDD approaches will be presented for peptide Class B GPCRs, addressing challenges in targeting larger, flexible, ligand-induced binding sites.

P49: Conformal calibration of QSAR classifiers

Sébastien Guesné, Stéphane Werner and Thierry Hanser

Lhasa Limited, Granary Wharf House, Leeds, United Kingdom

Class probabilities estimated by a Quantitative Structure Activity Relationship (QSAR) classifier are used to quantify the uncertainty of predictions made. This quantification is of great importance when assessing the potential liability of a single chemical to cause toxic/adverse effect(s) and promoting the use of predictions in the context of risk assessment. This quantification must be calibrated so that the QSAR classifier user is confident in using the model.

Imbalanced training sets with respect to the distribution between classes and the type of the QSAR classifier algorithm generate non-calibrated probability estimates. Given that for many endpoints the datasets investigated are imbalanced, it is therefore extremely important to calibrate the probabilistic predictions of the QSAR classifier.

This presentation will describe how a modification of the Mondrian conformal prediction algorithm for classification allows the modeller to calibrate a probabilistic *in silico* binary classifier. In this novel calibration approach the prediction set generated by the conformal prediction algorithm are replaced by calibrated class probabilities while avoiding information loss or distortion. E.g. balancing the training set.

P51: ANNalog – Generation of MedChem-similar molecules

Wei Dai[1], Jonathan D. Tyzack[2], Chris de Graaf[3], Arianna Fornili[1], Noel M.O'Boyle[4]

[1] *School of Physical and Chemical Science, Queen Mary University of London, United Kingdom*

[2] *Nxera Pharma, Steinmetz Building, Granta Park, Great Abington, Cambridge, United Kingdom*

[3] *Structural Therapeutics, South San Francisco, USA*

[4] *EMBL's European Bioinformatics Institute, WHinxton, Cambridgeshire, United Kingdom*

Generative models have been widely used in the pharmaceutical field in recent years, particularly for small molecule generation. However, most existing generative models either produce molecules randomly or with a specific focus on chemical or physical property optimization. In practical applications, medicinal chemists often lack clarity about the specific generation objectives, and the indiscriminate use of generative models has not proven effective in helping them design the desired molecules. To address this challenge, we developed ANNalog, a generative model based on a sequence-to-sequence architecture designed to generate potential drug molecules that medicinal chemists can intuitively recognize as relevant for given input molecules.

The dataset used to train ANNalog was derived from the similarity benchmark proposed by O'Boyle et al.[1], which assumes that, under specific conditions, two molecules within the same ChEMBL assay can be perceived as similar by medicinal chemists. These analogous molecular pairs were employed to train ANNalog. For a molecule within an assay in the test set, on average more than half (52%) of 100 generated molecules exhibited high similarity to other molecules within the same assay.

References

1. N. M. O'Boyle and R. A. Sayle, *Journal of cheminformatics*, **2016**, 8, 36.

P53: Improving the affinity of EpCAM-binding peptides using AlphaFold2 for *in vivo* tumor imaging

Nada Badr[1*], Nicky Janssen[3], Mark Fonteyne[1], Jochem R. van der Burgt[1], Taryn March[1], Lysanne D.A.N. de Muynck[1], Shadhvi Bhairosingh[1], Robbert Kim[2], Alexander L. Vahrmeijer[1], Gerard J.P. van Westen[3], Peter J. K. Kuppen[1]

[1] *Department of Surgery, Leiden University Medical Center, Leiden, The Netherlands*

[2] *Department of Cell and Chemical Biology, Leiden University Medical Center, Leiden, The Netherlands*

[3] *Medicinal Chemistry, Leiden Academic Centre for Drug Research, Leiden, The Netherlands*

[*] *Presenting author*

INTRODUCTION. Surgical treatment of cancer is often hindered by poor tumor visibility. To address this, we aim to develop peptide-based near-infrared fluorescent (NIRF) probes. Recent advances in deep learning, such as AlphaFold2 (AF2), have improved protein-peptide interaction predictions, enabling probe development. This study targets the epithelial cell adhesion molecule (EpCAM), a transmembrane protein overexpressed in carcinomas. We evaluated AF2's ability to predict *in vitro* binding and to enhance the affinity of an EpCAM-binding peptide, EP1.

METHODS. Protein-peptide multimer predictions were conducted using AF2 (v 2.3.2) with the extracellular sequence of EpCAM (EpEX) and EP1. Using an alanine scan, important amino acids (AA) with the lowest (interface) predicted template modeling scores (iPTM+PTM) were identified and optimized by substituting them with all possible AAs. EP0 (lowest score), EP2 (highest score), and the reference peptide EP1 were synthesized and conjugated to FITC for further analysis. The affinity of the probes was determined via fluorescence polarization assays (FPA). Additionally, EpCAM-positive HT-29 and EpCAM-negative HT-1080 tumor cells were incubated with the probes for 1 hour at 37 °C, followed by fluorescence microscopy imaging.

RESULTS & DISCUSSION. Two peptides were selected for further evaluation after analyzing peptide-protein interactions between EP1 derivatives and EpEX. EP0 ranked lowest (Figure 1A), had a relative iPTM+PTM score of 0.47, compared to the baseline score of 1.00 for the peptide EP1 (Figure 1B). The iPTM+PTM score of EP2 in interaction with EpEX was 1.10 (Figure 1C). No changes in polarized fluorescence were observed for EP0-FITC (Figure 2A), indicating no detectable binding to EpEX. EP1-FITC showed binding to EpEX with an affinity K_D of 724 nM, as determined by a nonlinear fit of polarization intensity (Figure 2B). EP2-FITC demonstrated a higher binding affinity with a K_D of 14 nM (Figure 2A). These findings were further supported in tumor cell lines (Figure 2C), where EP0-FITC did not stain the cell membrane of EpCAM-positive cells. Both EP1-FITC and EP2-FITC demonstrated specific binding to EpCAM, with EP2-FITC showing an intense stain of the cell membrane, indicating enhanced binding.

CONCLUSION. This study demonstrates that AF2 can accurately predict *in vitro* binding of peptides to proteins. Using AF2, we significantly improved the affinity of a peptide targeting EpCAM specifically. Our optimized peptide, EP2, might be suitable for use as an NIRF probe in the surgical treatment of carcinomas.

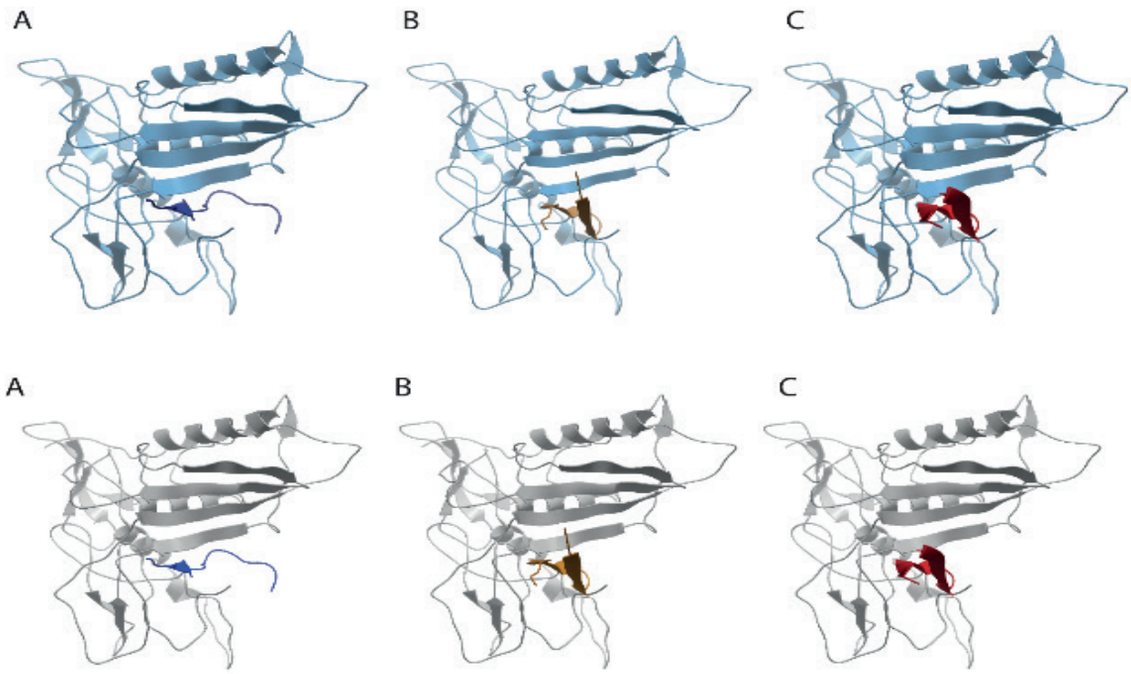


Figure 1. *Visualization of the predicted protein-peptide complex formation using AF2. Interaction between EpEx (extracellular domain of EpCAM), shown in* **grey** *and* **A.** *EP0 shown in* **blue**, **B.** *the reference peptide, EP1 shown in* **brown**, *and* **C.** *improved peptide, EP2 shown in* **red**.

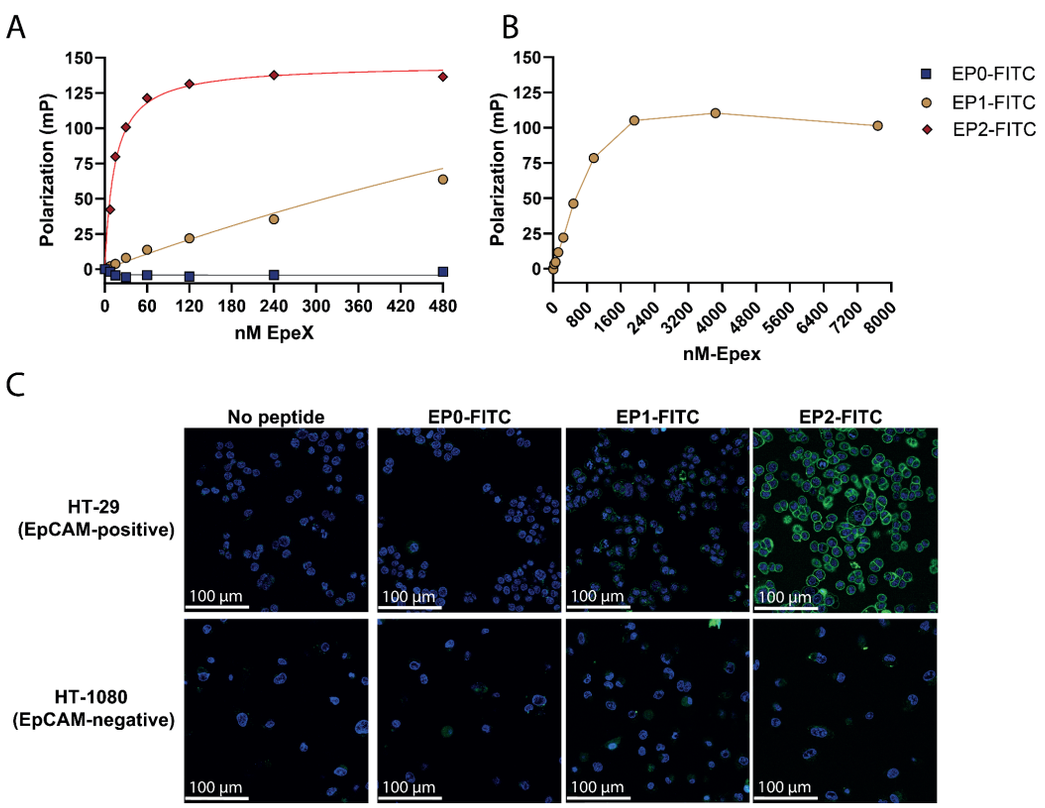


Figure 2. *Experimental evaluation of EpCAM binding by EP0-FITC, EP1-FITC, and EP2-FITC. A. Fluorescence polarization assay (FPA) measuring fluorescence polarization (mP) of purified EpEX protein up to 480 nM, demonstrating the binding of the selected peptides. B. FPA of purified EpEX protein concentrations up to 8000 nM, showing the binding of EP1-FITC. C. Imaging of HT-29 and HT-1080 cells incubated with 1 µM of EP0-FITC, EP1-FITC, or EP2-FITC. The cell nuclei were stained with Hoechst 3342, and imaging was performed using a Leica SP8 microscope.*

P55: Integration of stereochemistry within DrugEx for better sample efficiency

Chiel Jespers[1], Martin Sicho[1,3], Mike Preuss[2], Gerard van Westen[1]

[1] *Systems Pharmacology and Pharmacy, LACDR, Leiden University, The Netherlands*

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

[3] *National Infrastructure for Chemical Biology, Department of Informatics and Chemistry, Faculty of Chemical Technology, University of Chemistry and Technology Prague, Czech Republic*

Structure-based methods have in recent times been integrated with de novo drug generators, offering a complementary methodology to the better explored ligand-based scoring functions that can even be used with little to no prior ligand information available [1]. However, the introduction of such methods have brought with them comparatively long computational time, thus raising the need to optimize the learning taken from structure-based methods to achieve the training of the language model in a reasonable timeframe. Recent development with regards to this sample efficiency have seen the introduction of better reinforcement strategies [2], as well as ways to make more use of the learning taken from a single molecule [3].

In this study stereochemistry is integrated into the vocabulary of the SMILES-based de novo drug generator DrugEx [4]. By generating molecules with at least partially defined stereochemistry, only the stereoisomers that agree with the definition have to be scored by the respective structure-based method, potentially limiting the number of oracle calls needed to perform reinforcement learning. The efficacy of this methodology is tested with known benchmarks such as Guacamol [5] and Practical Molecular Optimization[6], as well as by a new benchmark derived from PoseBusters [7] of known hard-to-dock targets in which multiple 3D methods including shape-matching and docking are tested out. The methodology can potentially be used in tandem with previously introduced improvements to sample efficiency, thus further bringing down the computational cost of training a de novo generator with structure-based methods.

References

1. Thomas, M., Smith, R.T., O'Boyle, N.M. et al. Comparison of structure- and ligand-based scoring functions for deep generative models: a GPCR case study. J Cheminform 13, 39 (**2021**). doi: 10.1186/s13321-021-00516-0
2. Thomas M, O'Boyle NM, Bender A, de Graaf C. Augmented Hill-Climb increases reinforcement learning efficiency for language-based de novo molecule generation. J Cheminform 14, 68 (**2022**). doi: 10.26434/chemrxiv-2022-prz2r.
3. Bjerrum, E. J., Margreitter, C., Blaschke, T. & de Castro, R. L.-R. Faster and more diverse de novo molecular optimization with double-loop reinforcement learning using augmented SMILES. doi: arXiv:2210.12458
4. Šícho M, et al. DrugEx: Deep Learning Models and Tools for Exploration of Drug-Like Chemical Space. J Chem Inf Model. doi: 10.1021/acs.jcim.3c00434
5. Brown, N., Fiscato, M., Segler, M. H. S. & Vaucher, A. C. GuacaMol: Benchmarking Models for de Novo Molecular Design. J Chem Inf Model 59, 1096–1108 (**2019**). doi: 10.1021/acs.jcim.8b00839
6. Gao, W., Fu, T., Sun, J. & Coley, C. W. Sample Efficiency Matters: A Benchmark for Practical Molecular Optimization. ArXiv (**2022**). doi:10.48550/arXiv.2206.12411
7. M. Buttenschoen, G. M. Morris and C. M. Deane, PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences, Chem. Sci., **2024**, 15, 3130–3139. doi: 10.1039/D3SC04185A

P57: Multi-task is what you need! Multi-task machine learning models for molecular property prediction

Eelke Bart Lenselink[1], Giovanni A. Tricarico[1], Marie-Pierre Dréanic[2], Johan Hofmans[1], Kenneth Goosens[1], Stephane de Cesco[1]

1. Galapagos NV, Mechelen, Belgium

2. Galapagos SASU, Romainville, France

The prediction of molecular properties- such as physicochemical properties or biological activities - based on chemical structures is a well-established field. In traditional modelling approaches (i.e., Quantitative Structure Activity Relationship/QSAR) one single property is modelled at a time (single-task). By now sufficient evidence has emerged that multi-task models, where multiple properties are modelled simultaneously can outperform single-task models.[1–5]

In this talk we share the experience that we gained using multi-task Directed-Message Passing Neural Networks (D-MPNNs).[6] This is illustrated by a number of internal and external use-cases; first we introduce a novel way to split the multi-task datasets by chemistry,[7] using this approach as a benchmark for 1) Kinome wide datasets.[8] 2) internal ADME data 3) SAMPL7 challenge and 5) finally our approach in the Polaris open ADME challenge. Moreover, we show some "tricks of the trade" – such as adding calculated properties to the predictive models. We demonstrate that multi-task machine learning models consistently perform better than, or at least as good as, single-task machine learning models.

References:

1. Ramsundar, B.; Liu, B.; Wu, Z.; Verras, A.; Tudor, M.; Sheridan, R. P.; Pande, V. Is Multitask Deep Learning Practical for Pharma? *J. Chem. Inf. Model.* **2017**, *57* (8), 2068–2076.
2. Feinberg, E. N.; Sur, D.; Wu, Z.; Husic, B. E.; Mai, H.; Li, Y.; Sun, S.; Yang, J.; Ramsundar, B.; Pande, V. S. PotentialNet for Molecular Property Prediction. *ACS Cent. Sci.* **2018**, *4* (11), 1520–1530.
3. Lenselink, E. B.; Ten Dijke, N.; Bongers, B.; Papadatos, G.; Van Vlijmen, H. W.; Kowalczyk, W.; IJzerman, A. P.; Van Westen, G. J. Beyond the Hype: Deep Neural Networks Outperform Established Methods Using a ChEMBL Bioactivity Benchmark Set. *J. Cheminformatics* **2017**, *9* (1), 1–14.
4. Kumar, K.; Chupakhin, V.; Vos, A.; Morrison, D.; Rassokhin, D.; Dellwo, M. J.; McCormick, K.; Paternoster, E.; Ceulemans, H.; DesJarlais, R. L. Development and Implementation of an Enterprise-Wide Predictive Model for Early Absorption, Distribution, Metabolism and Excretion Properties. *Future Med. Chem.* **2021**, *13* (19), 1639–1654. https://doi.org/10.4155/fmc-2021-0138.
5. Walter, M.; Borghardt, J. M.; Humbeck, L.; Skalic, M. Multi-Task ADME/PK Prediction at Industrial Scale: Leveraging Large and Diverse Experimental Datasets**. *Mol. Inform.* **2024**, *43* (10), e202400079. https://doi.org/10.1002/minf.202400079.
6. Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59* (8), 3370–3388.
7. Tricarico, G. A.; Hofmans, J.; Lenselink, E. B.; Ramos, M. L.; Dréanic, M.-P.; Stouten, P. F. Construction of Balanced, Chemically Dissimilar Training, Validation and Test Sets for Machine Learning on Molecular Datasets. **2022**.
8. Luukkonen, S.; Meijer, E.; Tricarico, G. A.; Hofmans, J.; Stouten, P. F. W.; Van Westen, G. J. P.; Lenselink, E. B. Large-Scale Modeling of Sparse Protein Kinase Activity Data. *J. Chem. Inf. Model.* **2023**, *63* (12), 3688–3696. https://doi.org/10.1021/acs.jcim.3c00132.

P59: Fantastic SMILES augmentation methods and where to find them

H. Brinkmann[1], A. Argante[1], H. ter Steege[1], F. Grisoni[1,2]

[1] *Institute for Complex Molecular Systems (ICMS), Eindhoven AI Systems Institute (EAISI), Department of Biomedical Engineering, Eindhoven University of Technology, The Netherlands*

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

Chemical language models (CLMs) are generative deep learning algorithms that have shown incredible promise for de novo drug design [1-3]. CLMs learn directly from molecular string representations, such as the Simplified Input Line Entry System (SMILES) strings [4], which capture bonds, branches, and stereochemistry of a molecule. A key characteristic of using molecular strings is the possibility to perform 'data augmentation', i.e., to artificially inflate the number of data available for training. This is achieved by using multiple SMILES strings referring to the same molecule, obtained by traversing the molecular graph differently [5]. Such SMILES enumeration (Fig. 1a) enhances the quality of molecular generative models, especially in low-data regimes [6-7].

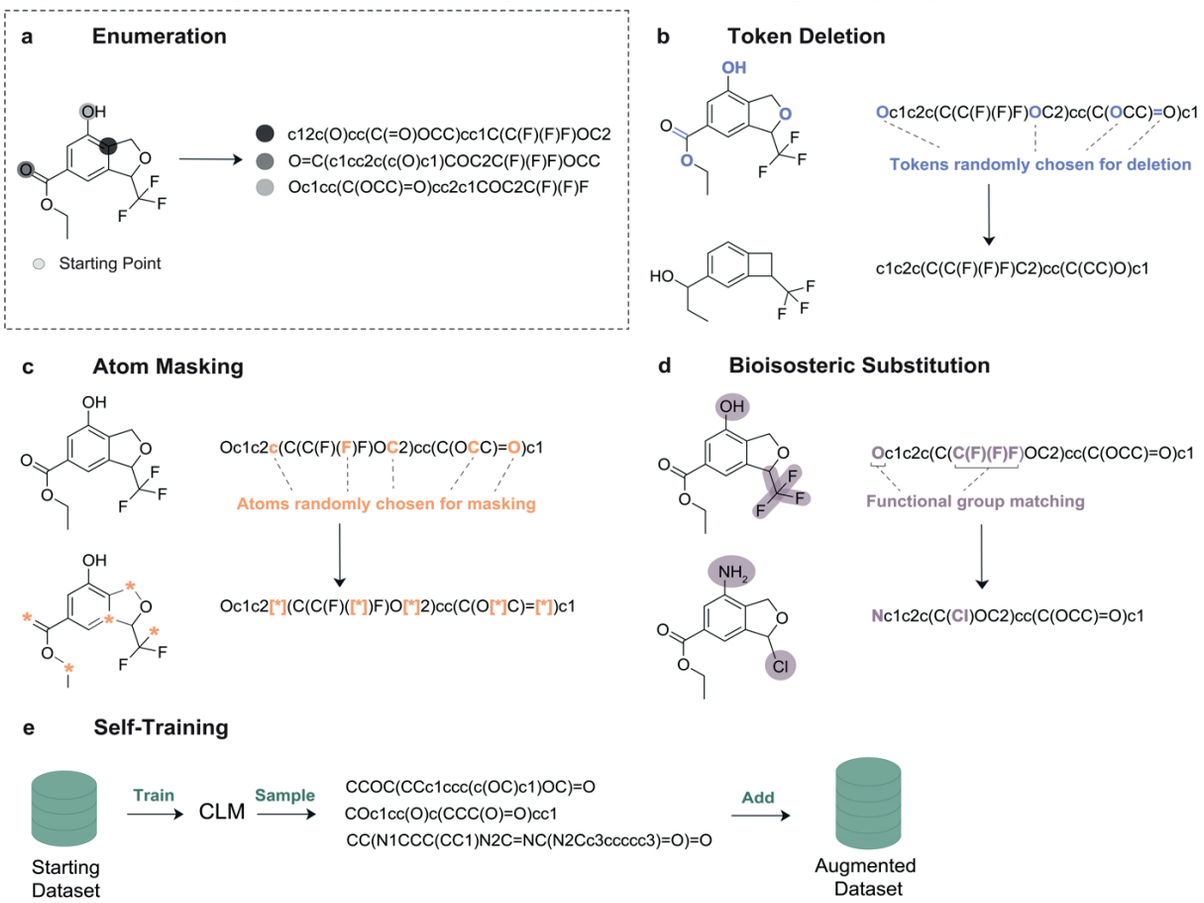


Figure 1. *Overview of available and used SMILES augmentation methods. (a) SMILES enumeration [3] is used as a baseline in this work. We compare them with following methods: (b) Token deletion, (c) atom masking, (b) bioisosteric substitution, and (e) self-training.*

Despite the great success of SMILES enumeration, no additional strategies for augmentation have been explored for CLMs yet. Stemming from this gap, in this work [8], we introduce four data augmentation approaches for de novo design and investigate their potential for chemical space exploration. The four approaches are based on random (a) deletion of tokens from the original string (Fig. 1b), (b) masking of atoms with a dummy atom '[*]' (Fig. 1c), (c) substitution of molecular substructures with reported bioisosters (Fig. 1d), and (d), self-training, where the model is used to augment its own training set (Fig. 1e). For each strategy, we investigate its effect on the quality of molecular design, in comparison with the well-established SMILES enumeration, in different low- to high-data scenarios, and in combination with transfer learning. Every strategy showed distinct advantages and could be

used for different tasks, for example in matching the distribution of the training properties and / or scaffolds in high or low similarity datasets. This new repertoire of SMILES augmentation strategies expands the available toolkit to design molecules with bespoke properties in low-data scenarios.

References

1. Grisoni, F., Chemical language models for de novo drug design. Current Opinion in Structural Biology, **2023**, 79, 102527, https://doi.org/10.1016/j.sbi.2023.102527
2. Segler, M. H. et al., Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks, ACS Cent. Sci., **2018**, 4, 1, 120–131, https://doi.org/10.1021/acscentsci.7b00512
3. Moret, M. et al., Generative molecular design in low data regimes. Nat. Mach. Intell., **2020**, 2, 171–180, https://doi.org/10.1038/s42256-020-0160-y
4. Weininger, D., SMILES, a chemical language and information system, J. Chem. Inf. Comput. Sci., **1988**, 28, 1, 31-36, https://doi.org/10.1021/ci00057a005
5. Bjerrum, E.J., SMILES Enumeration as Data Augmentation for Neural Network Modeling of Molecules, arXiv, **2017**, https://doi.org/10.48550/arXiv.1703.07076
6. Skinnider, M.A. et al., Chemical language models enable navigation in sparsely populated chemical space, Nat. Mach. Intell., **2021**, 3, 9, 759-770, https://doi.org/10.1038/s42256-021-00368-1.
7. Arús-Pous, J. et al., Randomized SMILES strings improve the quality of molecular generative models. J. Cheminform., **2019**, 11, 1, https://doi.org/10.1186/s13321-019-0393-0.
8. Brinkmann, H. et al, Going beyond SMILES enumeration for generative deep learning in low data regimes, ChemRxiv, **2025**, https://doi.org/10.26434/chemrxiv-2025-fdnnq

P61: Constrained generation of molecules using Diffusion Models

Cristian Pop[1], James Longden[2], Aniket Ausekar[2], Andreas Bender[1]

[1] *Iuliu Hatieganu University of Medicine and Pharmacy, Faculty of Pharmacy, Cluj-Napoca, Romania*

[2] *Evolvus Inc., Frankfurt am Main, Germany*

Diffusion models have emerged as powerful generative frameworks, achieving state-of-the-art performance in various tasks, such as image synthesis and molecular design. We explored the application of diffusion models in the generation of novel chemical compounds, and the potential improvements in the model architecture toward directed generation of novel compounds based on predefined criteria. The model training data is compiled from both public (Chembl: 2 million compounds, PubChem: 100 million compounds) and proprietary resources (Evolvus liceptor: 10 million compounds). The validation of the generated molecules was done though *in silico* methods such as compound rediscovery, diversity and uniqueness.

P63: A Novel Statistical Machine Learning Framework for Enhanced Drug Safety Prediction in Zebrafish Assays

Filippo Lunghini[1], Christian Cortes Campos[2], Vincenzo Pisapia[3], Gentzane Sánchez Elexpuru[2], Sylvia Dyballa[2], Francesco Sacco[3], Daniela Iaconis[1], Vincenzo Di Donato[3], Andrea Beccari[1]

[1] EXSCALATE, Dompé Farmaceutici SpA, Italy

[2] ZeClinics SL, Barcelona

[3] Professional Service Department, SAS Institute, Milan, Italy

The zebrafish model is valuable in predictive toxicology due to its genetic similarity to humans and transparent embryos, enabling real-time observation of organ-level processes. Assessing drug-induced organ toxicity in zebrafish embryos provides rapid and relevant predictions for human drug safety. Existing quantitative structure-activity relationship (QSAR) models are often constrained by a limited applicability domain, hindering their utility in the pharmaceutical sector. Our research presents a novel machine learning framework using an extensive library of 800 compounds tested in zebrafish assays to evaluate cardiotoxicity.

To this aim, we imaged 120 hpf zebrafish larvae after incubation for 24 hours with each compound. Videos of fluorescent hearts were recorded for 30 seconds per larva and via in-house developed software we were able to measure 8 different readouts. Measuring beat frequency (in atrium and Ventricle), QTc interval, cardiac arrest and ejection fraction allow us to identify potential chronotropic drugs that may change the heart rate and rhythm Moreover, recording arrhythmias heart chamber areas enable us to detect potential inotropic drugs that can modulate the force of heart contractions.

The amount of data generated allowed the development of a two-step modeling framework in SAS-viya4.0: 1) builds local models on 8 zebrafish readouts; and 2) combines these into a global cardiotoxicity outcome. Local models showed strong predictive power, with an average Pearson correlation of 0.70, sensitivity of 0.91, and specificity of 0.76. Testing on an external set of 92 drugs with known cardiotoxicity showed good predictive power, validating our framework for identifying potential adverse drug reactions in new compounds.

This study advances predictive toxicology with a scalable model adaptable to various compound libraries, enhancing drug discovery and safety evaluation. By prioritizing early toxicity detection, our research aids in creating safer medications and supports the EU's 3Rs principle to minimize animal experiments.

References

1. Dyballa S, Miñana R, Rubio-Brotons M, Cornet C, Pederzani T, Escaramis G, Garcia-Serna R, Mestres J, Terriente J. Comparison of Zebrafish Larvae and hiPSC Cardiomyocytes for Predicting Drug-Induced Cardiotoxicity in Humans. Toxicol Sci. **2019** Oct 1;171(2):283-295. doi: 10.1093/toxsci/kfz165. PMID: 31359052; PMCID: PMC6760275.

P65: Predicting the Dissipation Kinetics of Agrochemicals in Soil

Vincent-Alexander Scholz,[1,2] Richard Marchese-Robinson,[3] Sevil Payvandi,[3] Timothy J. C. O'Riordan[3] and Johannes Kirchmair[1,4]

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

[2] *Vienna Doctoral School of Pharmaceutical, Nutritional and Sport Sciences (PhaNuSpo), University of Vienna, Austria*

[3] *Syngenta UK, Jealott's Hill International Research Centre, Bracknell, Berkshire, United Kingdom*

[4] *Christian Doppler Laboratory for Molecular Informatics in the Biosciences, Department for Pharmaceutical Sciences, University of Vienna, Austria*

Safe and efficacious plant protection products (PPPs) are essential to agriculture.[1] Insights into the dissipation kinetics of candidate compounds provide valuable support during the development of novel PPP active ingredients. Computational approaches can help to optimize the dissipation profile of small organic compounds to produce chemicals with an optimal holistic product profile. Classical machine learning models have been deployed for half-life prediction of PPP active ingredients in soil.[2,3] Contrary to this approach, pharmacokinetic studies indicate that the prediction of the complete time course, utilizing time as a feature, or different models for different points in time, is advantageous compared to the direct prediction of endpoints.[4,5]

In order to predict the dissipation time course of agrochemicals, we employ graph neural networks to estimate the parameters of single first-order (SFO), double first-order in parallel (DFOP) and single first-order reversible binding (SFORB) kinetics. Thus, the model is conditioned on well-established assumptions about dissipation processes. The incorporation of expert-knowledge within a Deep-Learning framework makes our models' predictions more insightful and consistent regarding the time course. Models that separately predict the recovery at each time point could not guarantee this prediction consistency.

References

1. Tudi et al., Agriculture development, pesticide application and its impact on the environment. *Int. J. Environ. Res. Public Health*, **2020**, 18, 3, 1112, 10.3390/ijerph18031112.
2. Latino et al., Eawag-Soil in enviPath: a new resource for exploring regulatory pesticide soil biodegradation pathways and half-life data. *Environ. Sci.: Processes Impacts*, **2017**, 19, 3, 449-464, 10.1039/C6EM00697C.
3. Lunghini et al. Publicly available QSPR models for environmental media persistence. *SAR QSAR Environ. Res.*, **2020**, 31, 9, 655-675, 10.1080/1062936X.2020.1776387.
4. Handa et al., Prediction of compound plasma concentration–time profiles in mice using random forest. *Mol. Pharmaceutics*, **2023**, 20, 6, 3060-3072, 10.1021/acs.molpharmaceut.3c00071.
5. Obrezanova et al., Prediction of in vivo pharmacokinetic parameters and time–exposure curves in rats using machine learning from the chemical structure. Mol. Pharmaceuticals, **2022**, 19, 5, 1488-1504, 10.1021/acs.molpharmaceut.2c00027.

P67: NLP-inspired Operators for De novo Design Augmentation

Hanz Tantiangco[1], James Webster[2], Beining Chen[3], Val Gillet[1]

[1] *Information School, The University of Sheffield, U.K.*

[2] *Drug Discovery Unit, The University of Dundee, U.K.*

[3] *Department of Chemistry, The University of Sheffield, U.K.*

De novo drug design is a method that generates molecules from scratch. A recent trend in *de novo* design is the use of generative deep learning models, particularly chemical language models (CLMs) such as recurrent neural networks (RNNs). CLMs use concepts from natural language processing (NLP) to learn the rules that govern molecules in the form of string representations, such as SMILES (Grisoni, 2023). However, despite the promises of CLMs, they can be outperformed by traditional methods such as the Graph GA on benchmark goal-directed generation tasks (Brown *et al.*, 2019).

A potential approach to improve the performance of CLMs is to use SMILES augmentation. SMILES augmentation expands the training data of CLMs by incorporating different SMILES representation of the same molecule, known as randomised or enumerated SMILES. Results indicate that the addition SMILES augmentation in RNNs improved goal-directed generation performance (*Bjerrum et al.*, 2023; Guo and Schwaller, 2024).

In NLP, a common approach in data augmentation is the use of text editing operators to introduce new samples and improve model performance (Wei and Zou, 2019). Here, we propose the use of NLP-inspired operators as an additional method for SMILES augmentation in the context of goal-directed generation using RNNs. We show that using the NLP-inspired operators improves the goal-directed generation performance of RNNs.

References

1. Bjerrum, E.J. *et al.* (**2023**) 'Faster and more diverse de novo molecular optimization with double-loop reinforcement learning using augmented SMILES', *Journal of Computer-Aided Molecular Design* [Preprint]. Available at: https://doi.org/10.1007/s10822-023-00512-6.
2. Brown, N. *et al.* (**2019**) 'GuacaMol: Benchmarking Models for de Novo Molecular Design', *Journal of Chemical Information and Modeling*, 59(3), pp. 1096–1108. Available at: https://doi.org/10.1021/acs.jcim.8b00839.
3. Grisoni, F. (**2023**) 'Chemical language models for de novo drug design: Challenges and opportunities', *Current Opinion in Structural Biology*. Elsevier Ltd. Available at: https://doi.org/10.1016/j.sbi.2023.102527.
4. Guo, J. and Schwaller, P. (**2024**) 'Augmented Memory: Sample-Efficient Generative Molecular Design with Reinforcement Learning', *JACS Au* [Preprint]. Available at: https://doi.org/10.1021/jacsau.4c00066.
5. Wei, J. and Zou, K. (**2019**) 'EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks', in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 6381–6387. Available at: https://doi.org/10.18653/v1/D19-1670.

Poster Session RED:
Integrative Structure-Based Drug Design

P69: Extending Enveloping Distribution Sampling Towards NE-EDS: A Non-Equilibrium Approach to Free-Energy Estimation

Shu-Yu Chen[1,†], Enrico Ruijsenaars[1,†], Philippe H. Hünenberger[1], Sereina Riniker[1]

[1] Department of Chemistry and Applied Biosciences, ETH Zurich, Switzerland

[†] S.-Y.C. and E.R. contributed equally to this work

Enveloping distribution sampling (EDS) is an efficient method for free-energy estimation of closely related states, but its effectiveness hinges on finding suitable parameters – a non-trivial challenge that impacts its practical implementation. To navigate these challenges, we introduce non-equilibrium enveloping distribution sampling (NE-EDS) that combines EDS with non-equilibrium work measurements to dynamically perturb the reference potential and drive transitions between physical states. By leveraging a time-dependent reference potential, NE-EDS facilitates controlled state transitions, improving sampling efficiency and robustness in free-energy calculations. NE-EDS is first validated within the controlled framework of harmonic potential systems, where it demonstrates improved transition success rates and reduced statistical uncertainties relative to conventional equilibrium EDS. The results highlight a fundamental trade-off between protocol-based transitions, which ensure transitions at arbitrarily low temperatures but induce higher irreversible work, and diffusion-based transitions, which are thermally driven but require sufficient system diffusion. Subsequently, we apply NE-EDS to the computation of relative hydration free energies for 193 small organic molecules to demonstrate the feasibility in practical free-energy calculations. The results are compared against multistate Bennett acceptance ratio (MBAR) and standard non-equilibrium switching methods, identifying key advantages and trade-offs in computational accuracy and efficiency. By extending the utility of EDS to non-equilibrium transformations, NE-EDS provides a robust computational tool for accelerating free-energy calculations in complex molecular systems.

References

1. Christ, C. D., et al., Enveloping distribution sampling: A method to calculate free energy differences from a single simulation. The Journal of Chemical Physics, **2007**, 126, (18), 184110, https://doi.org/10.1063/1.2730508.
2. Shirts, M. R., et al., Equilibrium Free Energies from Nonequilibrium Measurements Using Maximum-Likelihood Methods. Physical Review Letters, **2003**, 91, (14), 140601, https://doi.org/10.1103/physrevlett.91.140601.
3. Feng, E. H., Length of time's arrow. Physical Review Letters, **2008**, 101, (9), 090602, https://doi.org/10.1103/physrevlett.101.090602.

P71: IMERGE-FEP 2.0: Generating intermediate R-groups for challenging free energy perturbations

Daan A. Jiskoot[1,2], Linde Schoenmaker[2], Jeroen L.A. Pennings[3], Willie J. G. M. Peijnenburg[1,4], David L. Mobley[5], Pim N.H. Wassenaar[1], Gerard J.P. van Westen[2], Willem Jespers[2,6]

[1] Centre for Safety of Substances and Products, National Institute for Public Health and the Environment (RIVM), Bilthoven, the Netherlands

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

[3] Centre for Health Protection, National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands

[4] Institute of Environmental Sciences, Leiden University, The Netherlands

[5] Department of Pharmaceutical Sciences, University of California, Irvine, United States

[6] Department of Medicinal Chemistry, Photopharmacology and Imaging, Groningen Research Institute of Pharmacy, Groningen, The Netherlands

Optimizing a ligand's binding affinity is a key objective in early-stage drug discovery. Free energy perturbation (FEP) has proven itself as a reliable method for computing binding free energy differences of closely related ligands. However, FEP becomes inaccurate when the structural changes between ligands are large, due to insufficient phase space overlap. One approach to address this limitation is to introduce an intermediate molecule between two distinct molecular endpoints. This effectively breaks down a single large perturbation into two more manageable perturbations through the intermediate molecule. Schoenmaker et al. proposed IMERGE-FEP, where the R-groups of the original endpoint molecules are recombined, yielding a structural hybrid of both parent molecules[1]. However, this approach is unable to create intermediate molecules where a single R-group position is the bottleneck. In recent research, Furui et al. demonstrated that adding generated intermediate molecules could increase the accuracy of binding free energy predictions for single topology FEPs[2]. Here, we propose IMERGE-FEP 2.0, an R-group intermediate generator that creates suitable intermediate molecules for challenging FEP calculations.

Our workflow, as depicted in Figure 1, begins by identifying the maximum common substructure (MCS) of the parent molecules and decomposing each molecule's R-groups. By utilizing the STONED-SELFIES algorithm on these R-groups, the relevant chemical space between the endpoints is efficiently explored and generated R-groups are extracted[3]. After reattaching these R-groups to the MCS, the resulting molecules are filtered and scored using metrics such as Tanimoto similarity, LOMAP scoring, and ROCS scoring to identify the most promising candidate intermediates for FEP.

Incorporating these intermediates into perturbation maps is expected to improve convergence of challenging edges. Even though addition of extra edges might increase the total computational cost, the increased molecular similarity will decrease convergence time, resulting in little additional cost for the added data point. Combining the original intermediate generator[1] with our updated version may further enhance FEPs applicability domain. As such a powerful strategy is proposed for bridging even larger structural gaps, by first recombining the parent R-groups and then introducing additional intermediates for especially challenging R-groups.

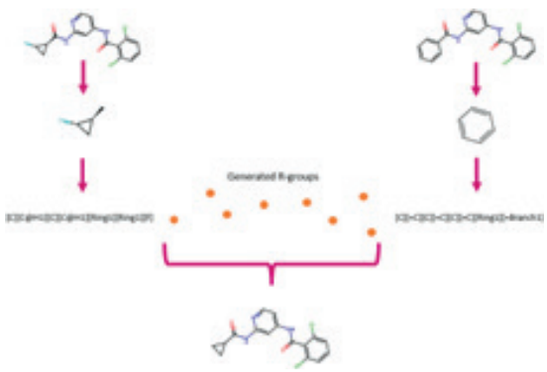


Figure 1. *Overview of the R-group generator.*

References

1. L. Schoenmaker et al., "IMERGE-FEP: Improving Relative Free Energy Calculation Convergence with Chemical Intermediates," J. Phys. Chem. B, Feb. **2025**, doi: 10.1021/acs.jpcb.4c07156.
2. K. Furui, T. Shimizu, Y. Akiyama, S. R. Kimura, Y. Terada, and M. Ohue, "PairMap: An Intermediate Insertion Approach for Improving the Accuracy of Relative Free Energy Perturbation Calculations for Distant Compound Transformations," J. Chem. Inf. Model., vol. 65, no. 2, pp. 705–721, Jan. **2025**, doi: 10.1021/acs.jcim.4c01634.
3. A. Nigam, R. Pollice, M. Krenn, G. dos P. Gomes, and A. Aspuru-Guzik, "Beyond generative models: superfast traversal, optimization, novelty, exploration and discovery (STONED) algorithm for molecules using SELFIES," Chem. Sci., vol. 12, no. 20, pp. 7079–7090, May **2021**, doi: 10.1039/D1SC00231G.

P73: Assessing the role of machine learning-based pose sampling in virtual screening

Thi Ngoc Lan Vu[1,2,3], Hosein Fooladi[1,2,3], Johannes Kirchmair[1,2]

[1] *Christian Doppler Laboratory for Molecular Informatics in the Biosciences, Department for Pharmaceutical Sciences, University of Vienna, Austria*

[2] *Department of Pharmaceutical Sciences, Division of Pharmaceutical Chemistry, University of Vienna, Austria*

[3] *Vienna Doctoral School of Pharmaceutical, Nutritional and Sport Sciences (PhaNuSpo), University of Vienna, Austria*

Ligand docking has been a fundamental component of structure-based virtual screening (VS) for decades. While classical docking approaches have traditionally dominated the field, machine learning (ML)-based pose sampling methods offer a promising complementary alternative. Although these methods have demonstrated competitive performance in pose prediction compared to classical approaches,[1,2] their effectiveness in virtual screening (VS) remains an open question. In this study, we integrate DiffDock-L[3], one of the most promising ML-based pose sampling methods, into VS workflows by combining it with established scoring functions and systematically evaluating its performance using a standard VS benchmarking dataset. We assessed its effectiveness, complementarity, and pose quality. Our findings show that DiffDock-L performs comparably to the classical AutoDock Vina's pose sampling approach[4], suggesting the potential of ML-based pose sampling methods for future VS applications.

References

4. CORSO, G., et al., DiffDock: Diffusion steps, twists, and turns for molecular docking. ArXiv, **2022**, DOI: arXiv:2210.01776.
5. BUTTENSCHOEN, M., et al., PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. Chem. Sci., **2024**, 15, 9, 3130-3139, DOI: 10.1039/d3sc04185a.
6. CORSO, G., et al., Deep confident steps to new pockets: Strategies for docking generalization. ArXiv, **2024**, DOI: arXiv:2402.18396v1.
7. TROTT, O., et al., AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J. Comp. Chem., **2010**, 31, 2, 455-461, DOI: 10.1002/jcc.21334.

P75: Ultralarge Tailored Library of Amino Acid Derivatives Designed to Target Peptide GPCRs

Pach S.[1], Carlsson J.[1]

[1] *Science for Life Laboratory, Department of Cell and Molecular Biology-BMC, Uppsala University, Uppsala, Sweden*

G protein-coupled receptors (GPCRs) represent a significant family of drug targets, with hundreds of approved therapeutics currently available. [1] Among these, peptide receptors play a crucial role in regulating a plethora of physiological functions. [2] Poor pharmacokinetic properties of peptides entail the development of small molecule ligands in the medicinal chemistry of peptide GPCRs. The large binding pockets and complex interaction networks between peptides and their receptors hinder the identification of small molecule ligands.

The neurotensin receptor (NTSR) is a promising target among peptide GPCRs for non-opioid analgesics. [3] Analysis of crystal structures identified the C-terminal leucine of neurotensin as a hallmark of NTSR ligands and a potential anchor for non-peptide small molecule ligands. To address the challenge of the NTSR ligand design, we created a tailored in-silico library of 6.3 million compounds through amide coupling of the amino group of leucine.

Subsequent docking studies and visual inspections of this library led to the synthesis of 17 compounds, six of which demonstrated low micromolar efficacy in functional assay. Further optimization efforts for the most potent scaffold led to the identification of a sub-micromolar agonist of NTSR. Additionally, X-ray crystallography confirmed the anchoring role of leucine and validated the predicted binding mode.

To advance the design of small molecule ligands for peptide GPCRs, we developed and shared a 4.7 billion compound library comprising 20 proteinogenic amino acids, coupled via amidation at their respective N- and C-termini with commercially available building blocks (ANCHOR library: Amino acid N/C-termini Hybrids Optimized for Receptors).

References

1. Hauser, A. S., et al., Trends in GPCR drug discovery: new agents, targets and indications. Nature Reviews Drug Discovery, **2017**, 16 (12), 829–842.
2. Davenport, A. P., et al., Advances in therapeutic peptides targeting G protein-coupled receptors. Nature Reviews Drug Discovery, **2020**, 19 (6), 389–413.
3. Kleczkowska, P., et al., Neurotensin and neurotensin receptors: Characteristic, structure–activity relationship and pain modulation—A review. European Journal of Pharmacology, **2013**, *716* (1–3), 54–60.

P77: Oxytocin-Signaling-Inspired Allosteric Modulator Design for the μ-Opioid Receptor

Marvin Taterra, Marcel Bermudez

Universität Münster, Institute of Pharmaceutical and Medicinal Chemistry, Germany

The primary pharmaceutical target for the treatment of severe pain is the μ-receptor (MOR). However, classical opioids are associated with severe side effects, including respiratory depression and dependance, which are significant contributors to North American opioid crisis. (**Vadivelu et al.**)

While allosteric modulators provide advantages regarding subtype selectivity and potentially better safety profiles. Their discovery is hampered by resource-intensive methodologies, such as high-throughput screening, or the challenges of identifying cryptic binding pockets.

To address this challenge, we employed, MDPath, a novel tool which identifies allosteric pathways in proteins by analyzing correlated dihedral angle motions in molecular dynamics simulations. (**Doering et al.**) Integration of MDPath with classical *in silico* methods provided mechanistic insights and unveiled the differential receptor modulation induced by the opioid agonists fentanyl and morphine.

It has previously been discovered that oxytocin acts as a positive allosteric modulator (PAM) at the MOR. (**Meguro et al.)** Building on this finding, our approach led to the identification of a novel oxytocin-binding pocket in the MOR, which is located outside the seven transmembrane (TM) domains between TM3, TM4 and TM5. We show how oxytocin enhances signaling efficiency without increasing binding affinity on a molecular level.

In addition, the molecular basis of probe dependence in the MOR system was elucidated, thereby demonstrating how morphine and fentanyl exhibit different potency gains when combined with oxytocin. These findings lay the foundation for the development of oxytocin-inspired drug candidates as PAMs for the MOR.

These novel compounds have the potential to enhance the safety profile of opioid-based pain treatments by minimizing adverse effects. The present findings propose a novel, rational, structure-based approach to the identification of allosteric modulators, thereby establishing new avenues for the development of safer opioid therapeutics.

References

1. Vadivelu, N. et al., The Opioid Crisis: a Comprehensive Overview. Current Pain and Headache Report., **2018**, 22, 16, https://doi.org/10.1007/s11916-018-0670-z
2. MDPath, **2024**, Doering, N. et. at., https://github.com/wolberlab/mdpath
3. Meguro, Y. et al., Neuropeptide oxytocin enhances μ opioid receptor signaling as a positive allosteric modulator. Journal of Pharmacological Sciences, **2018**, 137(1), 67–75. https://doi.org/10.1016/j.jphs.2018.04.002

P79: Improved Protein-Ligand Structure Modeling using Conserved Scaffold Placement

Jonathan Pletzer-Zelgert[1], Matthias Rarey[1], Bernd Kuhn[2]

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

[2] *Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd., Basel, Switzerland*

Despite recent improvements in structure prediction through deep learning methods, accurately predicting protein-ligand binding modes remains challenging and often unsuccessful, particularly when the structural conformation of the binding site significantly deviates from available experimental data. The lack of a reliable binding mode hypothesis greatly impedes efforts to rationally optimize molecules for these drug targets. Conversely, small molecule docking methods, despite their limitations in predicting affinity, are generally regarded as capable of identifying correct binding poses when the binding site is in the correct structural configuration.

Here we employ molecular docking in a reverse approach. Instead of using a protein structure to identify new binders, we utilize known binders to select structural models of the protein. For this purpose, we use small data sets of ligand series with measured affinities and a common molecular scaffold, which are typically available early in a project. We also leverage the observation that experimental structures for different molecules in a chemical series usually exhibit a highly conserved scaffold placement. If a molecule is indeed a binder, this scaffold show a consistent binding pose across the ligand series with high probability.

Starting from a set of structural models of any origin, we generate a large and diverse set of rotamer states by employing the backbone dependent Dunbrack rotamer library[1] for residues deemed relevant. We select diverse rotational states by clustering of the rotamers resulting in a few dozen to thousands of models. We subsequently dock the series compounds using JAMDA with softened potential [2] into all models and identify plausible scaffold poses by calculating a density cluster with the DBSCAN algorithm [3]. Each combination of scaffold placement and protein model is evaluated using a consensus score of the cluster coverage, docking scores, measured affinities and penalties for strained torsion states for the cluster members.

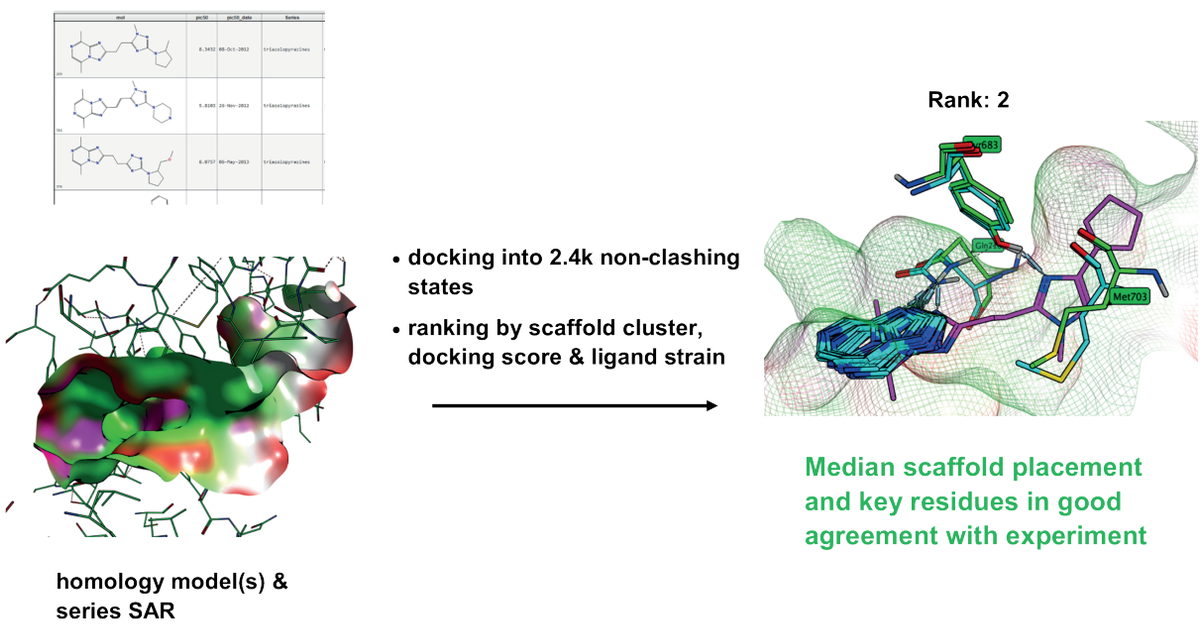


***Figure 1**: One PDE10A homology model showed correct backbone conformation but had a side chain blocking the binding pocket. By sampling side chain rotamer states we could recreate a near native conformation that resulted in consistent ligand placement.*

In initial proof of concept studies, we applied this methodology to two common scenarios involving the prediction of binding modes for pharmaceutically relevant targets. In the first case study, we addressed a situation where initial homology models of a protein of interest were available, but the binding mode for a chemical series was unclear. This involved using homology models of phosphodiesterase 10A (PDE10A) based on PDE5 templates to predict the correct binding poses, including significant induced fit effects, for a series of triazolopyrazines [4]. In the second case study, we aimed to predict a large conformational change upon binding of a series of triazolopyrimidinones to fatty acid-binding protein 4 (FABP4), where only FABP4 structures without this conformational change were known. In both scenarios, our method successfully identified models that closely matched the corresponding experimental structures, capturing all relevant interactions within the top 3 solutions.

References

1. Shapovalov, M., et al., A Smoothed Backbone-Dependent Rotamer Library for Proteins Derived from Adaptive Kernel Density Estimates and Regressions, Structure **2011,** 19, 844-858. doi: 10.1016/j.str.2011.03.019
2. Flachsenberg, F., et al., Redocking the PDB. J. Chem. Inf. Model. **2024**, 64, 1, 219–237, doi: 10.1021/acs.jcim.3c01573
3. Ester, M., et al., A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, KDD'96: 2. International Conference on Knowledge Discovery and Data Mining. **1996**, 226 – 231, doi: 10.5555/3001460.3001507
4. Tosstorf, A., et al., A high quality, industrial data set for binding affinity prediction: performance comparison in different early drug discovery scenarios. J Comput Aided Mol Des, **2022**, 36(10):753-765., doi: 10.1007/s10822-022-00478-x

P81: Structure-based Pharmacophore Screening of TREM2

Kevin Lam[1], Gerhard Wolber[1], Moustafa Gabr[2]

[1] *Molecular Design Group, Institute of Pharmacy, Department of Biology, Chemistry & Pharmacy, Freie Universität Berlin, Germany*

[2] *Moustafa Gabr Laboratory, Department of Radiology, Weill Cornell Medical College, New York, United States*

The Triggering Receptor of Myeloid Cells 2 (TREM2) is a receptor, which is predominantly expressed on microglia and is linked to a range of neurodegenerative diseases, including Alzheimer's disease[1]. Currently, no small molecules other than non-specific endogenous ligands are known to bind to TREM2. Therefore, a de novo structure-based pharmacophore screening workflow was constructed to discover small molecules targeting TREM2, by using PyRod[2] to reveal crucial intermolecular interactions by tracing water molecules in molecular dynamics (MD) simulation.

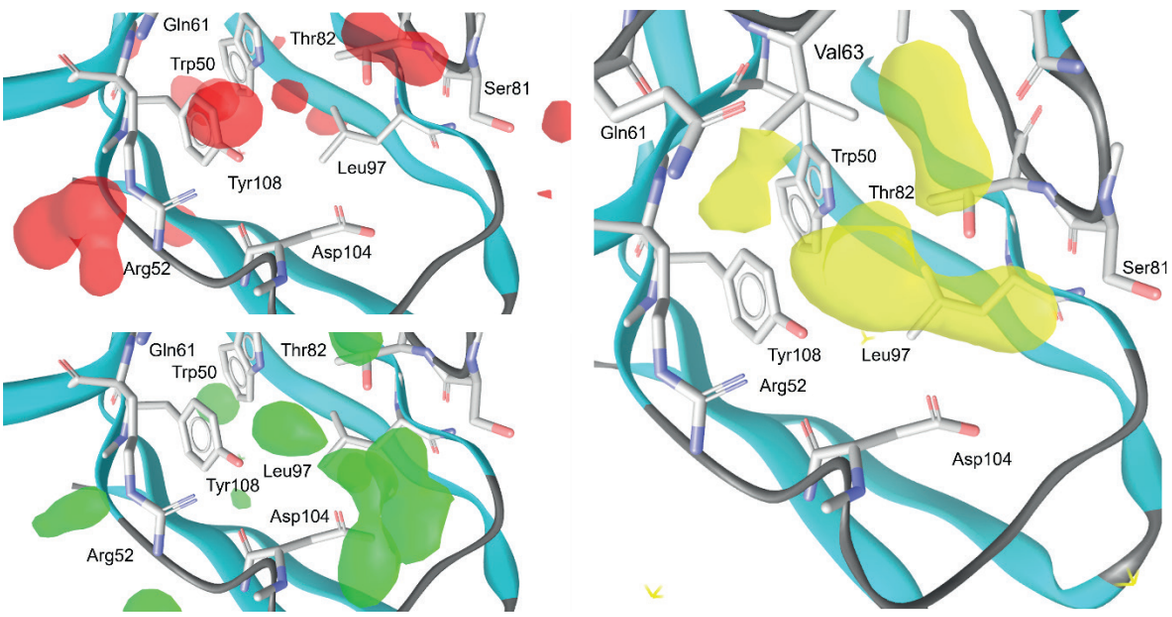


***Figure 1**: Dynamic molecular interaction fields of TREM2 generated with PyRod with hotspots for hydrogen acceptor (red) and donor (green) bonds and for hydrophobic contacts (yellow).*

A suitable binding cavity was identified by applying LigandScout's[3] pocket detection on an apo-structure of human TREM2 (PDB: 5UD7). To generate the 3D pharmacophore, PyRod was used to locate potential hotspots for small molecule intermolecular interactions with the protein, called dynamic molecular interaction fields (dMIFs), by tracing water molecule interactions with the binding site in MD simulation (**Fig. 1**). Pharmacophore features were then placed within the dMIFs to create a 3D pharmacophore query, which was refined in multiple screenings of the Enamine Screening Collection (ESC) (~4.4 mio. molecules) and visual inspection of the resulting hits, leading to the final screening pharmacophore. (**Fig. 2**)

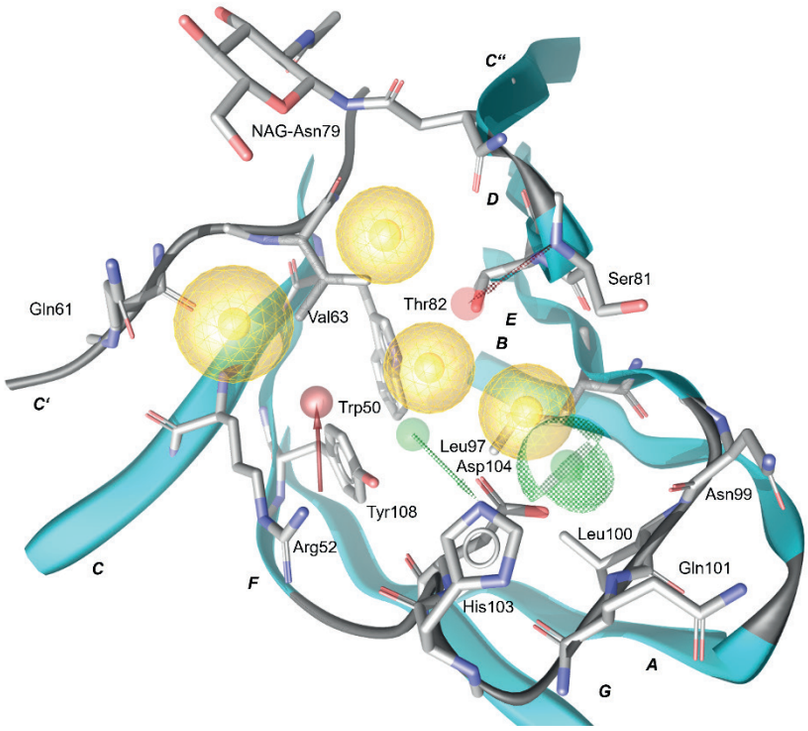


Figure 2: *3D Pharmacophore of TREM2 with yellow spheres depicting hydrophobic contacts, red arrows depicting hydrogen acceptor bonds and green arrow and sphere depicting hydrogen donor bonds. Compounds had to fulfill at least one of the dashed features.*

The ESC was screened against the final pharmacophore, resulting in 1312 hits. To validate these results, protonation states of all hits were generated and docked into the binding site in GOLD[4]. After discarding binding poses with less than three hydrophobic contacts, 9540 binding poses remained for visual inspection. Finally, twenty compounds were selected for experimental validation.

References

1. Colonna, M., The biology of TREM receptors. Nat. Rev. Immunol, **2023**, 23, 580–594, doi.org/10.1038/s41577-023-00837-1
2. Schaller, D. et al., PyRod: Tracing Water Molecules in Molecular Dynamics Simulations, J. Chem. Inf. Model. **2019**, 59, 6, 2818–2829, doi.org/10.1021/acs.jcim.9b00281
3. Wolber, G. et al., LigandScout: 3-D Pharmacophores Derived from Protein-Bound Ligands and Their Use as Virtual Screening FiltersJ. Chem. Inf. Model. **2005**, 45, 1, 160–169, doi.org/10.1021/ci049885e
4. Jones, G. et al, Development and validation of a genetic algorithm for flexible docking, **1997**, 267, 3, 727-748

P83: Identification, Synthesis and Biological Assessment of Allosteric ligands for the C-C Chemokine Receptor 5

Kian Noorman van der Dussen[1], Martin Šícho[1,2*‡], Khaled Essa[1,3‡], Yao Yao[1], Benthe Bleijs[1], Laura Heitman[1,3], Gerard van Westen[1,3*], Daan van der Es[1], Willem Jespers[1]

[1] Division of Medicinal Chemistry, Leiden Academic Centre for Drug Research (LACDR), Leiden University, 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, Czech Republic

[3] Oncode Institute, Leiden, The Netherlands

The C-C chemokine receptor 5 (CCR5) is a G-protein coupled receptor (GPCR) which has been associated with many diseases, including cancer and autoimmune disorders. It is therefore considered to be an important potential pharmacological target[1], though few drugs which act on CCR5 have been developed. If the extracellular orthosteric site of CCR5 is targeted, the ligand would have to compete with the endogenous chemokine ligands to bind to the receptor. This study however, has aimed to identify compounds that bind to the intracellular allosteric binding site, thereby avoiding this competition. Furthermore, this study focused on developing compounds which are selective to CCR5, which is not trivial as the allosteric binding site is similar to that of CCR2. The ligands that were developed in this study are based on a previously identified biaryl sulfonamide compound for CCR4 and CCR5[2]. In order to improve the potency and attain selectivity, a virtual screening strategy with a combinatorial library of biaryl sulfonamide compounds was employed based on docking. A smaller set of promising compounds was then assessed with molecular dynamics. There, protein-ligand interactions are predicted based on the structure of the compounds and CCR5. The most promising compound was identified as 4-methyl-N-(6-morpholino pyridin-2-yl)benzene sulfonamide and served as a novel scaffold. In total, nineteen compounds were synthesised with different substituents on the benzene ring and varying aliphatic rings. A functional CCR5 GTPγS assay showed moderate affinity of multiple compounds to CCR5. Radioligand binding assays showed a trend of selectivity towards CCR5 over CCR2. This scaffold could therefore be promising in the further development of selective and potent allosteric antagonists for CCR5.

References

1. Jiao, X., Nawab, O., Patel, T., Kossenkov, A. V., Halama, N., Jaeger, D., & Pestell, R. G. Recent Advances Targeting CCR5 for Cancer and Its Role in Immuno-Oncology, **2019**, *Cancer Research (Chicago, Ill.)*, 79(19), 4801–4807. DOI:10.1158/0008-5472.CAN-19-1167.
2. Andrews, G., Jones, C., & Wreggett, K. A., An Intracellular Allosteric Site for a Specific Class of Antagonists of the CC Chemokine G Protein-Coupled Receptors CCR4 and CCR5, **2008**, *Molecular Pharmacology*, *73*(3), 855–867. DOI:10.1124/mol.107.039321.

Poster Session RED: New Modalities and Large Chemical Data Sets

P85: Generation of Custom Synthetically Accessible Combinatorial Chemical Spaces using Machine Learning-Based Reagent Filtering – Design of Freedom Space 4.0

Anna Kapeliukha[1,3], Serhii Hlotov[1,3], Marina Vasylchuk[1], Mykola Protopopov[1,3], Olga Tarkhanova[1], Yurii Moroz[2,3]

[1] *Chemspace LLC, Kyiv, Ukraine*

[2] *Enamine Ltd., Kyiv, Ukraine*

[3] *Taras Shevchenko National University of Kyiv, Ukraine*

Ensuring synthetic accessibility in large combinatorial spaces remains a major challenge in chemistry. Enamine's REAL Space addresses this through experimental reagent validation, but the approach is resource-intensive and not transferable to external collections. Traditional ML-based methods predict reaction success at the final compound level, which is computationally impractical for billion-scale spaces, limiting the integration of proprietary building blocks while maintaining high synthesis success rates.

To overcome these challenges, we developed the Custom Space Generation workflow, which shifts the focus from final compounds to building blocks. Using custom ML-based filters trained on experimental reaction success data, this approach enables accurate reagent selection before enumeration, making the process scalable, computationally efficient, and adaptable to diverse chemical collections. Currently, the workflow supports 40 reaction protocols, with further expansion planned.

The workflow has been applied for generation of Freedom Space 4.0 utilizing commercially available BBs provided by different vendors. The space contains 142 billion molecules in synthon-based format, and is available as two enumerated subsets, offering great flexibility for exploration.

The success of this approach is demonstrated by Freedom Space 3.0, a 5-billion-molecule space built on commercially available BBs and 10 reaction protocols, 8 of which were ML-filtered. We experimentally confirmed the 80%+ synthesis success rate based on the synthesis of over 1000 compounds from the space.

Searching through such large collections of molecules is a challenge, so for Freedom Space 4.0 a custom synthon-based search platform is being developed. The platform will be freely available to researchers and will feature exact match and substructure searches inside the Freedom Space 4.0 of 142 billion molecules.

Poster Session RED: Open Science, Omics, and Natural Products

P87: Identifying off-target drug interactions mediated via DNA methylation

Delaney A. Smith[1], Russ B. Altman[2]

[1] *Department of Biochemistry, Stanford University Medical School, Stanford CA, USA*

[2] *Departments of Genetics, Bioengineering, and Bioinformatics, Stanford University Medical School, Stanford CA, USA*

Metformin is a common treatment for Diabetes Mellitus Type II. Metformin is known to inhibit hepatic gluconeogenesis and oppose the action of glucagon, however, interindividual efficacy of this treatment varies.[1] Some of the variance observed in individual response to metformin can be attributed to genetic variants.[1] However, recent studies have shown that epigenetic changes, such as DNA methylation, may also impact drug response.[2] Integrating DNA methylation information into genetic analysis of drug response can be useful as DNA methylation markers represent stable, partially heritable markers of lifetime exposures and can explain additional variance in interindividual drug response.[2] In this study, we implement a causal inference framework that estimates the effects of DNA methylation sites on metformin response (as measured by fasting glucose levels). Finding 38 CpG sites (methylation markers) with significant effects on fasting glucose levels, we noticed that 18.4% of these sites also had previously been associated with inflammatory bowel disease (IBD), Ulcerative Colitis (UC), or Crohn's Disease (CD).[3] Moreover, we found that for most of 5/7 (71.4 %) of the sites associated with IBD/UC/CD, a methylated status was associated with higher relative fasting glucose levels post-treatment with metformin (reduced clinical efficacy). Based on these findings, we found that within the Stanford Medical Center Electronic Health Record database, IBD/UC/CD patients taking metformin had higher fasting glucose levels compared with patients taking metformin without these diagnoses. These findings contribute to a growing body of work showing that metformin may play a role in regulating gut inflammation and gut microbiome.[4] Our work implicates DNA methylation markers as potential mediators of this interplay and demonstrates the need to incorporate omics data in future works aimed at understanding the full spectrum of a drug's clinical indications and counterindications.

References

1. Niu, N. et al. Metformin pharmacogenomics: a genome-wide association study to identify genetic and epigenetic biomarkers involved in metformin anticancer response using human lymphoblastoid cell lines. Hum Mol Genet 25, 4819–4834 (**2016**).
2. Smith, D. A., Sadler, M. C. & Altman, R. B. Promises and challenges in pharmacoepigenetics. Camb Prism Precis Med 1, e18 (**2023**).
3. The EWAS Catalog: a database of ... | Wellcome Open Research. https://wellcomeopenresearch.org/articles/7-41/v2.
4. Lee, H. & Ko, G. Effect of Metformin on Metabolic Improvement and Gut Microbiota. Appl Environ Microbiol 80, 5935–5943 (**2014**).

P89: VHP4Safety Compound Wiki: an open science approach to collect domain specific knowledge

Willighagen E[1], Zare Jeddi, M[2], Sinke L[3]

[1] *Dept of Translational Genomics, Maastricht University, The Netherlands*

[2] *Shell Global Solutions International BV, The Netherlands*

[3] *Leiden Academic Centre for Drug Research, University of Leiden, The Netherlands*

Collecting chemical structures is an old science. Datasets large and small are released routinely, with both open and closed licenses. Publishing these dedicated databases, however, remains a time-consuming task, often warranting a standalone research article. This work demonstrates a low-cost, open science, and scalable approach to collecting chemical structure data through a crowdsourced, continuously evolving model, providing sufficient FAIR-ness and enabling automation of much of the data curation and management. We previously presented how Wikidata stores information for around 1.5 million chemicals, and here we introduce the VHP4Safety Compound Wiki (compoundcloud.wikibase.cloud, see Figure 1), which complements Wikidata using the Wikibase software underlying Wikidata to support project-specific data.

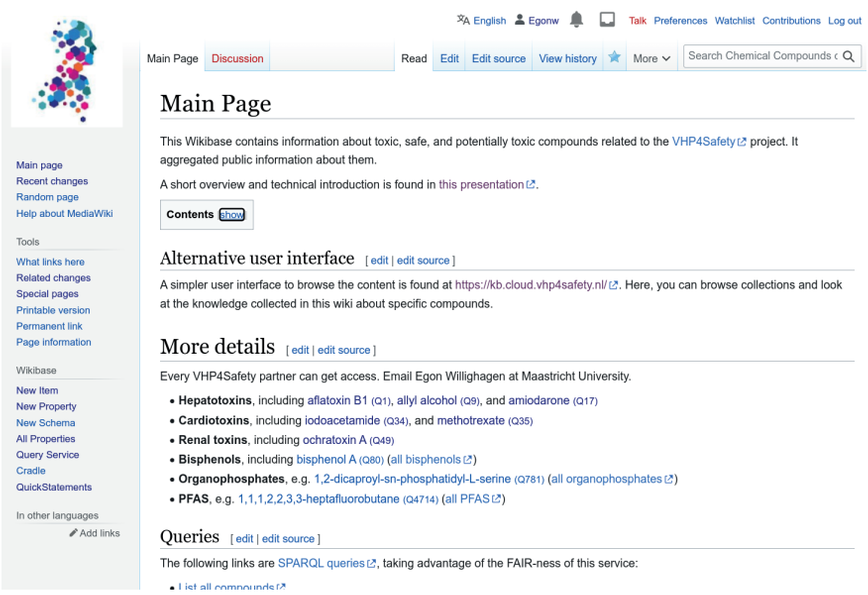


***Figure 1**: Screenshot of the VHP4Safety Compound Wiki landing page.*

Specifically, we add custom Wikibase properties relevant to the toxicology studied in VPH4Safety (grant number NWA.1292.19.272) including links to, for example, ToxBank Wiki, WikiPathways (for metabolism pathways), and AOP-Wiki (for stressors). As in our previous work, chemical structures are added to the knowledge base using Bioclipse scripts using the CDK 2.10, the OpenSMILES standard, and the InChI toolkit to ensure uniqueness identification and matching against other databases. To minimize redundancy, identifiers are automatically included for databases already supported by Wikidata. Furthermore, this wiki includes links to reports and datasets specifically describing the toxicology of included compounds. Finally, we showcase the wiki's SPARQL endpoint, which enables integration into other knowledge bases and computational tools in the VHP4Project, and supports federated SPARQL queries, allowing for extended functionality, such as the automated retrieval of physicochemical properties.

References

1. Willighagen, E. et al. Scholia Chemistry: access to chemistry in Wikidata, *ChemRxiv*, **2025**, https://doi.org/10.26434/chemrxiv-2025-53n0w"

P91: A mutator effect caused by two amino acid changes in the DNA binding region of M. smegmatis DnaE1: Novel insights into DNA polymerase fidelity using in silico and in vivo approaches

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

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

[2] *Department of Cell & Chemical Biology, Leiden University Medical Center (LUMC), The Netherlands*

Drug resistance in *Mycobacterium tuberculosis* presents a major challenge in tuberculosis treatment, highlighting the need to understand the underlying mechanisms[1]. DNA replication plays an important role in the acquisition of drug resistance and the expression of the DNA polymerase DnaE2 during adverse conditions has been associated with increased mutation rates [2].

Here we investigate the functional differences between the high-fidelity replicative DNA polymerase DnaE1 and the predicted error-prone DNA polymerase DnaE2, focusing on which amino acid changes affect polymerase fidelity. For this we identify potential fidelity-altering positions using a two-entropy sequence analysis [3]there are still many unresolved questions: Was the optimal subdivision of a protein family achieved? Do the definitions at different levels of the phylogenetic tree affect the prediction of specificity positions? Can the whole phylogenetic tree be used instead of only one level in it to predict specificity positions?Results: Here we present a novel method, TEA-O (Two-entropies analysis—Objective combined with experimental validation to test whether changes of these positions affect the mutation rates.

We find that a double mutation in the DNA binding region of DnaE1: D431S/R432D, increases mutation frequencies both *in vivo* and *in vitro*. The location of these two residues adjacent to the DNA backbone of the template strand (**Figure 1**) suggests that the amino acid change results in a loser grip on the DNA, allowing for the incorporation of incorrect nucleotides.

These insights improve our understanding of the mechanisms underlying drug resistance in *M. tuberculosis* and could help in the development of future strategies to combat it.

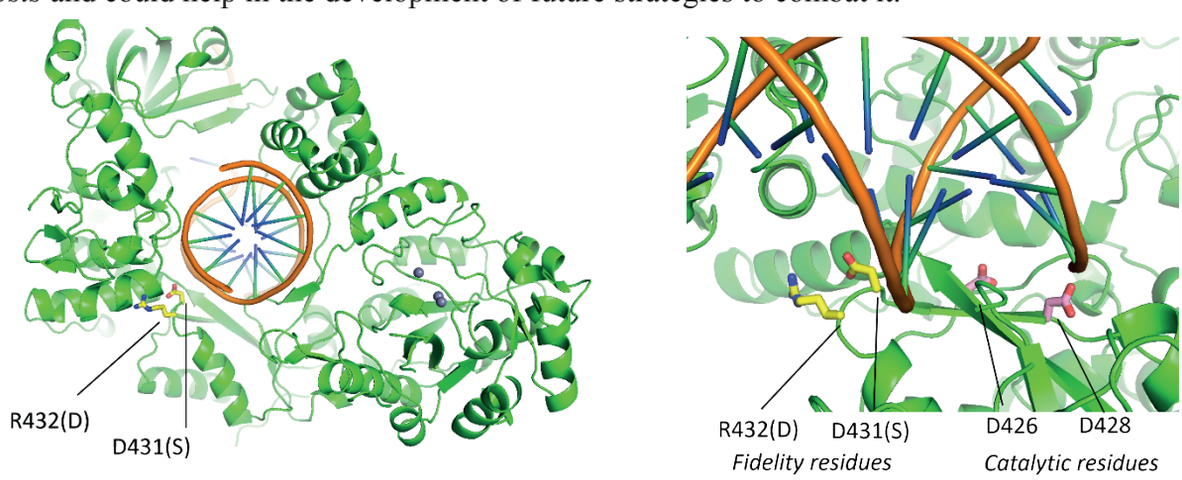


Figure 1: *Location of the fidelity-altering and catalytic residues in M. tuberculosis DnaE1. DNA is shown orange and blue sticks. The cryo-EM structure of M. tuberculosis DnaE1 bound to DNA was obtained from Chengalroyen et al.* [4]

References

1. Singh, R.; Dwivedi, S. p.; Gaharwar, U. s.; Meena, R.; Rajamani, P.; Prasad, T. Recent Updates on Drug Resistance in Mycobacterium Tuberculosis. *Journal of Applied Microbiology* **2020**, *128* (6), 1547–1567. https://doi.org/10.1111/jam.14478.

2. Boshoff, H. I. M.; Reed, M. B.; Barry, C. E.; Mizrahi, V. DnaE2 Polymerase Contributes to In Vivo Survival and the Emergence of Drug Resistance in Mycobacterium Tuberculosis. *Cell* **2003**, *113* (2), 183–193. https://doi.org/10.1016/S0092-8674(03)00270-8.
3. Ye, K.; Vriend, G.; IJzerman, A. P. Tracing Evolutionary Pressure. *Bioinformatics* **2008**, *24* (7), 908–915. https://doi.org/10.1093/bioinformatics/btn057.
4. Chengalroyen, M. D.; Mason, M. K.; Borsellini, A.; Tassoni, R.; Abrahams, G. L.; Lynch, S.; Ahn, Y.-M.; Ambler, J.; Young, K.; Crowley, B. M.; Olsen, D. B.; Warner, D. F.; Barry III, C. E.; Boshoff, H. I. M.; Lamers, M. H.; Mizrahi, V. DNA-Dependent Binding of Nargenicin to DnaE1 Inhibits Replication in Mycobacterium Tuberculosis. *ACS Infect. Dis.* **2022**. https://doi.org/10.1021/acsinfecdis.1c00643.

Poster session BLUE

Poster Session BLUE: Advanced Cheminformatics Techniques

P02: STELLAR: Developing and optimizing a novel advanced docking protocol for processing large peptides without AI assistance. A cancer context application

Alejandro Rodríguez-Martínez, Jochem Nelen, Miguel Carmena-Bargueño, Carlos Martínez-Cortés, Horacio Pérez-Sánchez

Structural Bioinformatics and High Performance Computing Research Group (BIO-HPC), UCAM Universidad Católica de Murcia (UCAM), Murcia, Spain

Docking methods have improved significantly with AI and parameter optimization, but docking large molecules like peptides remains challenging. Existing tools such as DockThor (1) and DLPepDock (2) struggle with accuracy and efficiency for peptides of certain lengths, and generative AI approaches still lack robust ranking metrics (3-4).

To address this, we developed **STELLAR** (Score-Tuning for Efficient ranking in Large Ligands libraries using an Accurate and Refined docking configuration), a workflow for docking peptides longer than 10 residues. Built on our MetaScreener platform (https://github.com/bio-hpc/metascreener), STELLAR includes optimized algorithms for handling large peptides. Validation was conducted using benchmark complexes (10–30 residues) from Propedia (5) and PDB database.

STELLAR achieved low RMSD values. For instance, 2.91 Å for a 12-residue peptide (2W10). It scales linearly in computing time (increasing with every three amino acids), runs on CPUs, and avoids reliance on AI approximations. Target docking takes 10–20 minutes; blind docking, 30–50 minutes.

STELLAR also supports virtual screening to identify bioactive peptides, successfully generating ranked hit lists for cancer-related targets. Future developments aim to extend STELLAR to other large biomolecules like nucleic acids, carbohydrates, and fatty acids.

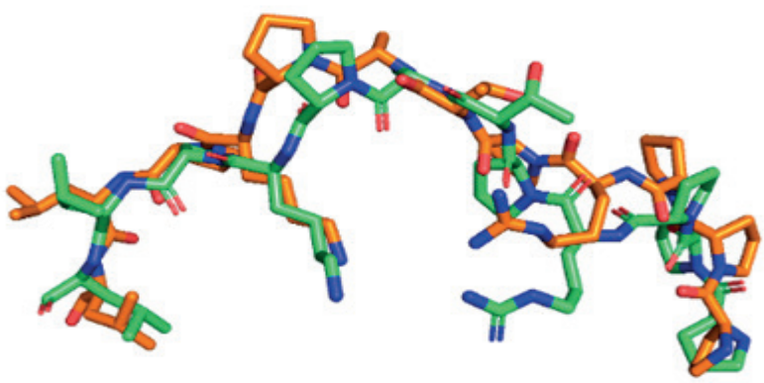


***Figure 1**: Comparison of the docking-predicted pose (orange) obtained using STELLAR with the crystallographic pose of the peptide (green) in the complex (PDB:2W10)*

References

1. Santos KB, Guedes IA, Karl ALM, Dardenne LE. Highly Flexible Ligand Docking: Benchmarking of the DockThor Program on the LEADS-PEP Protein–Peptide Data Set. J Chem Inf Model. **2020** Feb 24;60(2):667–83.
2. Sun L, Fu T, Zhao D, Fan H, Zhong S. Divide-and-link peptide docking: a fragment-based peptide docking protocol. Phys Chem Chem Phys. **2021**;23(39):22647–60.
3. Jin X, Chen Z, Yu D, Jiang Q, Chen Z, Yan B, et al. TPepPro: a deep learning model for predicting peptide–protein interactions. Bioinformatics. **2025** Jan 1;41(1):btae708.
4. Zhang Z, Verburgt J, Kagaya Y, Christoffer C, Kihara D. Improved Peptide Docking with Privileged Knowledge Distillation using Deep Learning. bioRxiv. **2023** Dec 4;2023.12.01.569671.
5. Martins PM, Santos LH, Mariano D, Queiroz FC, Bastos LL, Gomes I de S, et al. Propedia: a database for protein–peptide identification based on a hybrid clustering algorithm. BMC Bioinformatics. **2021** Jan 2;22(1):1.

P04: Towards More Reliable Distance Geometry-Based Conformer Generation

Niels Maeder, Gregory A. Landrum, Sereina Riniker

Department of Chemistry and Applied Biosciences, ETH Zürich, Switzerland

Conformational flexibility plays a crucial role in molecular recognition, directly impacting the accuracy of docking, virtual screening, and structure-based drug design. The quality of generated conformer ensembles -- defined by their likelihood to represent bio-active poses, their diversity, and their agreement with experimentally determined structures -- strongly influences the success of computational models (1). Various tools have been developed for estimating conformational ensembles. Typically, the validation of generated ensembles is done by comparing them to (experimental) reference data (2). The RDKit's ETKDG (3,4) conformer generator leverages crystallographic data to guide torsional sampling toward relevant regions of conformational space. Here, we present several improvements to the conformer generation workflow, ensuring physically realistic conformers, with a special focus on bond lengths and angles. With conformer evaluation workflows mainly focusing on correct torsional states, we found that these metrics often miss more fundamental flaws in conformers and propose further metrics to identify bad conformers. A detailed analysis of ETKDG-generated conformers, emphasizing physical validity, provides new insight into the performance and accuracy of the conformer generator that are often not considered in validation. This understanding enables precise optimization of the method, leading to an improved and more reliable conformer generator.

References

1. Hawkins, P. C. D., Conformation Generation: The State of the Art. J. Chem. Info. Model., **2017**, 57, 8, 1747–1756, 10.1021/acs.jcim.7b00221
2. Friedrich, N., et al., High-Quality Dataset of Protein-Bound Ligand Conformations and Its Application to Benchmarking Conformer Ensemble Generators. J. Chem. Info. Model., **2017**, 57, 3, 529-539, 10.1021/acs.jcim.6b00613
3. Riniker and Landrum, Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. J. Chem. Info. Model., **2015**, 55, 12, 2562–2574, 10.1021/acs.jcim.5b00654
4. Wang, S., et al., Improving Conformer Generation for Small Rings and Macrocycles Based on Distance Geometry and Experimental Torsional-Angle Preferences. J. Chem. Info. Model., **2020**, 60, 4, 2044–2058, 10.1021/acs.jcim.0c00025

P06: Investigating Structural Information in Graph-based Neural Fingerprints for Similarity Searches

S. Homberg[1], M. Modlich[2], B. Risse[2], O. Koch[1]

[1] *Institute of Pharmaceutical and Medicinal Chemistry, University of Münster, Germany*

[2] *Computer Vision and Machine Learning Systems, University of Münster, Germany*

Fingerprints are essential for virtual screening as they compactly represent chemical structures, allowing for quick similarity searches, machine learning, and bioactivity predictions. Extended-Connectivity Fingerprints (ECFPs) have been traditionally used, while neural fingerprints, particularly those from graph neural networks (GNNs), offer potential advantages in capturing molecular structure. However, ECFPs outperform neural fingerprints trained on the same molecular graphs, limiting the practical use of graph-based neural fingerprints in similarity searches [1].

To address this limitation, we investigated the information flow through various GNN architectures using explainable AI. In particular, we employ Myerson values, inspired by cooperative game theory, which can be directly applied to visualize the contribution of individual fingerprint positions on the molecular graph [2].

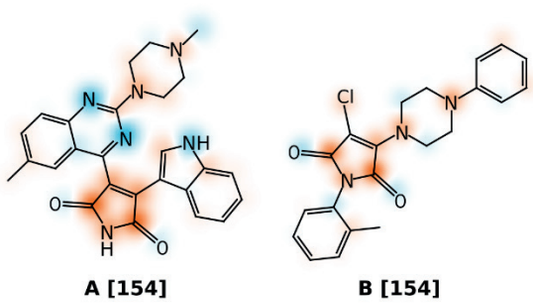


***Figure 1**: Visualizing contributions to an individual fingerprint position on the molecular graph. Piperazine and pyrroledione substructures have similar Myerson values.*

Unexpectedly, and contrary to their poor virtual screening performance, our method shows that GNN-based neural fingerprints still capture distinct structural features, as indicated by the Myerson values (Figure 1). We further investigate the role of over-smoothing, which we hypothesize may negatively impact virtual screening performance, despite high predictive accuracy and the use of GNN architectures designed to mitigate over-smoothing.

References

1. Menke, J., et al., Using Domain-Specific Fingerprints Generated Through Neural Networks to Enhance Ligand-Based Virtual Screening, *J. Chem. Inf. Model.* **2021**, *61*, 664, https://doi.org/10.1021/acs.jcim.0c01208.
2. Homberg, S., et al., Interpreting Graph Neural Networks with Myerson Values for Cheminformatics Approaches, *ChemRxiv Preprint*, **2024**, https://doi.org/10.26434/chemrxiv-2023-1hxxc-v2.

P08: Automatic Annotation of Sites of Metabolism from Substrate-Metabolite Pairs

R. A. Jacob[1,2,3], J. Kirchmair[1,2]

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

[2] *Christian Doppler Laboratory for Molecular Informatics in the Biosciences, Department of Pharmaceutical Sciences, University of Vienna, Austria*

[3] *Vienna Doctoral School of Pharmaceutical, Nutritional and Sport Sciences (PhaNuSpo), University of Vienna, Austria*

Computational metabolism prediction tools have emerged as valuable aids for assessing the metabolic fate of xenobiotics, enabling faster and more cost-effective evaluations[1]. Among them, site of metabolism (SOM) predictors identify the specific atoms where metabolic reactions occur and are used independently or alongside biotransformation rules to predict likely metabolite structures[2]. However, despite the availability of several SOM predictors, their predictive performance has stagnated — primarily due to the limited quantity and quality of training data.

Most of the currently available SOM predictors rely on just two datasets[3,4], encompassing approximately 680 and 2,000 biotransformations, respectively. Yet, vast amounts of potential training data exist in the form of substrate-metabolite pairs — data that remains untapped due to the absence of SOM annotations. Manual SOM annotation is labor-intensive and requires expert evaluation[4]. Automating this process would dramatically improve SOM prediction and, by extension, metabolite structure prediction[2].

Here, we introduce AutoSOM, the first open-source tool for extracting SOMs directly from substrate-metabolite pairs. AutoSOM identifies SOMs by mapping structural differences between substrate and metabolite molecules to a defined set of transformation rules. It is both fast, processing over 5,000 reactions in minutes, and highly accurate, achieving over 90% labeling accuracy across diverse metabolic reactions. Moreover, AutoSOM's annotation process is fully transparent and interpretable, increasing its reliability for medicinal chemists and regulatory agencies.

References

1. Kirchmair, J., et al. Predicting drug metabolism: experiment and/or computation? Nat. Rev. Drug Discov., **2015**, 14, 6, 387-404, 10.1038/nrd4581
2. de Bruyn Kops, C., et al. GLORYx: Prediction of the metabolites resulting from phase 1 and phase 2 biotransformations of xenobiotics. Chem. Res. Toxicol., **2020**, 34, 2, 286-299, acs.chemrestox.0c00224
3. Zaretzki, J., et al. RS-predictor: a new tool for predicting sites of cytochrome P450-mediated metabolism applied to CYP 3A4. J. Chem. Inf. Model., **2011**, 51, 7, 1667-1689, 10.1021/ci2000488
4. Pedretti, A., et al. MetaQSAR: an integrated database engine to manage and analyze metabolic data. J. Med. Chem., **2018**, 61, 3, 1019-1030, 10.1021/acs.jmedchem.7b01473

P10: Bayesian Illumination: Inference and Quality-Diversity Accelerate Generative Molecular Models

Jonas Verhellen[1,2]

[1] University of Copenhagen, Department of Drug Design and Pharmacology, Denmark

[2] University of Oslo, Centre for Integrative Neuroplasticity, Norway

Designing high-performing small molecules remains a central challenge in computational drug discovery. While deep learning has advanced generative molecular design, traditional methods like genetic algorithms (GAs) remain competitive [1]. Recent work has shown that integrating quality-diversity (QD) methods [2] into GAs can improve efficiency and avoid stagnation. Building on these ideas, we introduce **Bayesian Illumination**, a novel generative model that unifies Bayesian optimization, QD methods, and domain-specific molecular kernels [3].

Small Molecule Protein Binders				
Algorithm	**Minimum Docking (↓)**	**Mean Docking (↓)**	**Quality-Diversity Score (↓)**	**Archive Coverage (↑)**
Target Protein: Dopamine D3 Receptor (DRD3)				
GB-BI	**-12.05 ± 0.25**	**-10.77 ± 0.17**	**-638.76 ± 21.36**	**45.11 % ± 1.39 %**
GB-EPI	-11.10 ± 0.30	-9.98 ± 0.18	-471.89 ± 19.47	34.89 % ± 3.15 %
GB-GA	-10.81 ± 0.18	-9.64 ± 0.20	N/A	N/A
Target Protein: Tyrosine-Protein Kinase ABL (ABL1)				
GB-BI	**-11.99 ± 0.44**	**-10.97 ± 0.37**	**-652.82 ± 4.39**	**45.11 % ± 1.54 %**
GB-EPI	-11.10 ± 0.34	-9.93 ± 0.06	-443.82 ± 20.74	33.78 % ± 2.14 %
GB-GA	-10.72 ± 0.24	-9.53 ± 0.23	N/A	N/A
Target Protein: Epidermal Growth Factor Receptor (EGFR)				
GB-BI	**-12.22 ± 0.08**	**-11.17 ± 0.11**	**-674.63 ± 19.32**	**46.67 % ± 4.16 %**
GB-EPI	-11.06 ± 0.07	-10.01 ± 0.15	-461.80 ± 20.24	35.11 % ± 1.68 %
GB-GA	-10.85 ± 0.32	-9.69 ± 0.23	N/A	N/A

Table 1: *Optimisation results obtained in three independent docking tasks for Bayesian Illumination (GB-BI), Illumination (GB-EPI) and Genetic Algorithms (GB-GA), limited to 1000 fitness function calls and subject to structural filters from ChEMBL and Veber's rule of druglikeness.*

Bayesian Illumination consistently outperforms state-of-the-art techniques—including deep generative models, standard QD algorithms, and conventional GAs—by efficiently producing diverse, high-quality molecules. Its effectiveness was demonstrated across three benchmarking tasks: a) **Fingerprint-Based Rediscovery** using Guacamol benchmarks [4], the model successfully recovered known compounds. b) **Descriptor-Based Optimization**: We introduced a novel benchmark using USRCAT [5] and Zernike descriptors [6] to test optimization beyond traditional fingerprints. c) **Docking-Based Design**: Within the Therapeutics Data Commons [7], Bayesian Illumination generated synthetically feasible candidates using docking scores, applying filters from ChEMBL and Veber's rule to ensure druglikeness. The model's use of Gaussian processes within a QD framework enables a balance of exploration and exploitation—a critical aspect of efficient molecular search. Results show improved performance under tight evaluation budgets (e.g., 1,000 fitness calls), as summarized in Table 1.

Conclusion

Bayesian Illumination sets a new benchmark in molecular generation by blending GAs, QD methods, and Bayesian inference. Although it currently relies on predefined molecular representations, future work may incorporate data-driven embeddings to further enhance performance. This work opens new avenues in reaction optimization, data-driven drug design, and efficient chemical space exploration.

References

1. Jensen, Jan H. "A graph-based genetic algorithm and generative model/Monte Carlo tree search for the exploration of chemical space." *Chemical science* 10.12 (**2019**): 3567-3572.
2. Verhellen, Jonas, and Jeriek Van den Abeele. "Illuminating elite patches of chemical space." *Chemical science* 11.42 (**2020**): 11485-11491.

3. Griffiths, Ryan-Rhys, et al. "GAUCHE: a library for Gaussian processes in chemistry." *Advances in Neural Information Processing Systems* 36 (**2024**).
4. Brown, Nathan, et al. "GuacaMol: benchmarking models for de novo molecular design." *Journal of chemical information and modeling* 59.3 (**2019**): 1096-1108.
5. Schreyer, Adrian M., and Tom Blundell. "USRCAT: real-time ultrafast shape recognition with pharmacophoric constraints." *Journal of cheminformatics* 4 (**2012**): 1-12.
6. Venkatraman, Vishwesh, Padmasini Ramji Chakravarthy, and Daisuke Kihara. "Application of 3D Zernike descriptors to shape-based ligand similarity searching." *Journal of cheminformatics* 1 (**2009**): 1-19.
7. Huang, Kexin, et al. "Therapeutics Data Commons: Machine Learning Datasets and Tasks for Drug Discovery and Development." *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

P12: The Molecule Fragmentation Framework (MORTAR): a rich client application for algorithmic substructure extraction

F. Bänsch[1], C. Steinbeck[2], A. Zielesny[1], J. Schaub[2]

[1] *Institute for Bioinformatics and Chemoinformatics, Westphalian University of Applied Sciences, Recklinghausen, Germany*

[2] *Institute for Inorganic and Analytical Chemistry, Friedrich Schiller University Jena, Germany*

Developing computational algorithms for extracting specific substructures from molecular graphs (*in silico* molecule fragmentation) typically involves multiple cycles of implementing a rule set, applying it to relevant structure sets, inspecting the results, and adjusting the algorithm. The open MORTAR (MOlecule fRagmenTAtion fRamework) Java rich client application [1] supports this development process and makes new *in silico* molecule fragmentation algorithms readily available via its graphical user interface. Fragmentation results are visualised in various ways, and further analysis features are provided (Figure 1). In addition, fragmentation pipelines with any combination of the available fragmentation methods can be executed. These include the Ertl algorithm for functional group identification [2,3], the Sugar Removal Utility *in silico* deglycosylation algorithm [4], and the molecular scaffold functionalities of the Chemistry Development Kit-Scaffold module [5].

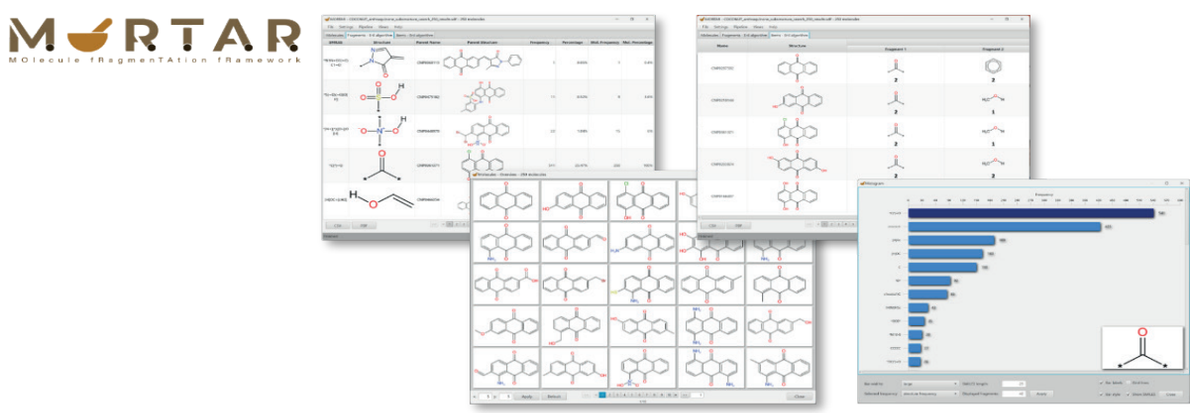


***Figure 1**: Overview of the MORTAR graphical user interface*

Additional features currently in development include an integration of the RECAP fragmentation scheme [6], a fragmentation algorithm for complex alkyl structures, and a clustering functionality that rapidly groups molecules based on fragment feature vectors with the adaptive resonance theory-based artificial neural network clustering method ART-2a [7].

References

1. Bänsch, F., et al., MORTAR: a rich client application for in silico molecule fragmentation. J Cheminform., **2023**, 15, 1, 1, https://doi.org/10.1186/s13321-022-00674-9
2. Ertl, P., An algorithm to identify functional groups in organic molecules. J Cheminform., **2017**, 9, 1, 36, https://doi.org/10.1186/s13321-017-0225-z
3. Fritsch, S., et al., ErtlFunctionalGroupsFinder: automated rule-based functional group detection with the Chemistry Development Kit (CDK). J Cheminform., **2019**, 11, 1, 37, https://doi.org/10.1186/s13321-019-0361-8
4. Schaub, J., et al., Too sweet: cheminformatics for deglycosylation in natural products. J Cheminform., **2020**, 12, 1, 67, https://doi.org/10.1186/s13321-020-00467-y
5. Schaub, J., et al., Scaffold Generator: a Java library implementing molecular scaffold functionalities in the Chemistry Development Kit (CDK). J Cheminform., **2022**, 14, 1, 79, https://doi.org/10.1186/s13321-022-00656-x

6. Lewell, X. Q., et al., RECAP-Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. J. Chem. Inf. Comput., **1998**, 38, 3, 511-522, https://doi.org/10.1021/ci970429i
7. Wienke, D., et al., An adaptive resonance theory based artificial neural network (ART-2a) for rapid identification of airborne particle shapes from their scanning electron microscopy images. Chemom Intell Lab Syst., **1994**, 25, 2, 367-387, https://doi.org/10.1016/0169-7439(94)85054-2

P14: Retrospective evaluation of molecule enumeration methods

Pierre-Yves Libouban, David Hahn, Natalia Dyubankova, Dries Van Rompaey, Gary Tresadern

In-Silico Discovery, Johnson & Johnson, Beerse, Belgium

With the advancements of in-silico methods in terms of speed and accuracy, we are able to triage large sets of novel compound ideas. Generating those compound ideas to feed into in-silico triaging methods is a critical step that significantly influences the success of the entire drug discovery process (Figure 1). Therefore, it is crucial to employ enumeration tools that ensure the generation of high-quality molecules while thoroughly exploring chemical space.

Various methodologies and tools have been developed for molecular enumeration, ranging from matched molecular pair and reaction-based methods to generative AI (GenAI).1-3 Evaluating and benchmarking these tools is challenging due to their distinct strengths and weaknesses and limited objective means to evaluate those methods.4,5 Nonetheless, their usefulness can be assessed by analyzing metrics such as the physicochemical properties, drug likeness,6 synthesizability, diversity, and similarity of the enumerated molecules to previously synthesized target compounds, alongside determining the percentage of compounds filtered out by standard in-silico filters.

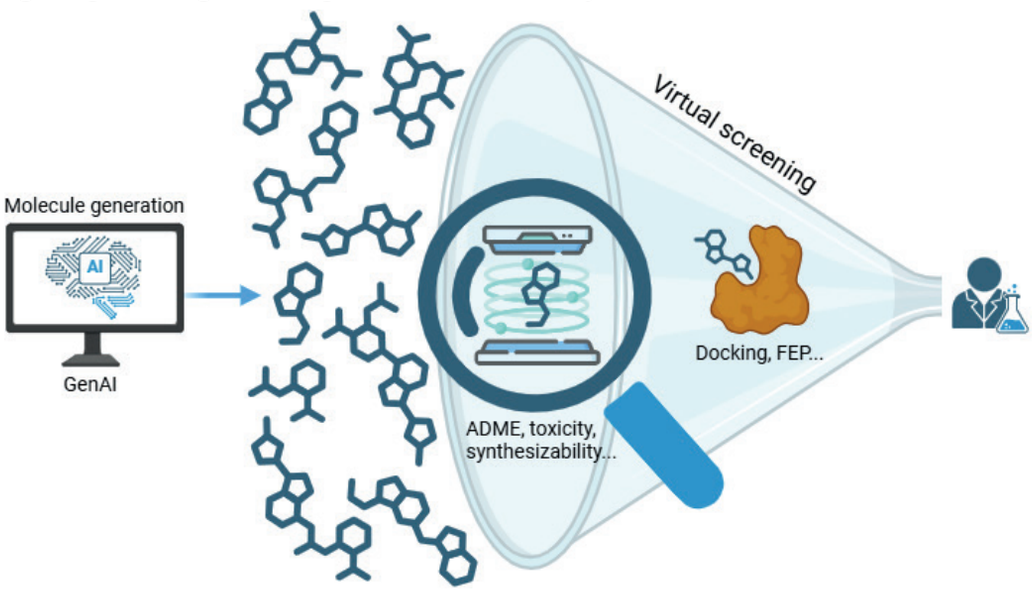


***Figure 1**: Molecule enumeration and triaging workflow*

We present a retrospective analysis of a diverse range of enumeration tools on industrial projects. Based on this analysis, we propose guidelines for selecting the most suitable tools to achieve specific objectives, such as evading intellectual property, optimizing molecular properties, performing scaffold hopping, or growing a specific vector.

Among other findings, we note that GenAI tools can adapt to diverse applications by leveraging reinforcement learning on ligand-based model predictions (QSAR, MPO) or structure-based, physics-aware approaches.

References

1. Dalke, A., et al., mmpdb: An Open-Source Matched Molecular Pair Platform for Large Multiproperty Data Sets. Journal of chemical information and modeling., **2018**, 58, 902-910, 10.1021/acs.jcim.8b00173
2. Loeffler, H. H., et al., Reinvent 4: Modern AI–driven generative molecule design. Journal of Cheminformatics., **2024**, 16, 20, 10.1186/s13321-024-00812-5

3. Bou, A., et al., ACEGEN: Reinforcement Learning of Generative Chemical Agents for Drug Discovery. Journal of chemical information and modeling., **2024**, 64, 5900-5911, 10.1021/acs.jcim.4c00895
4. Brown, N., et al., GuacaMol: Benchmarking Models for de Novo Molecular Design. Journal of chemical information and modeling., **2019**, 59, 1096-1108, 10.1021/acs.jcim.8b00839
5. Polykovskiy, D., et al., Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. Frontiers in Pharmacology., **2020**, 11, 10.3389/fphar.2020.565644
6. Choung, O.-H., et al., Extracting medicinal chemistry intuition via preference machine learning. Nature Communications., **2023**, 14, 6651, s41467-023-42242-1

P16: Deconvolution of "Origin-of-life" Reaction Networks

Nico Domschke[1], Richard Golnik[1], Chen Wang[2], Ales Charvat[2], Thomas Gatter[1], Bernd Abel[2], Peter F. Stadler[1]

[1] *Bioinformatics, Leipzig University, Germany*

[2] *Wilhelm-Ostwald-Institut für Physikalische und Theoretische Chemie, Leipzig University, Germany*

Modeling the origin of life is a well-established field of research encompassing many decades and various experiments such as the Miller-Urey experiment.[1,2] Here, the formation of amino acids in a simulated early earth atmosphere containing methane, ammonia, hydrogen, and water was demonstrated in cyclic conditions of warm and cool phases under electric discharges. To characterize such Origin-of-life experiments, appropriate analytical techniques need to be combined with a powerful formal modelling framework. We introduce a methodological pipeline tailored towards the analysis of MS/MS spectra associated with these kind complex reaction mixtures.

First MØD[3], a category theory-based cheminformatics tool for graph rewriting, is utilized to generate a reaction network starting from a few small organic compounds. It was initially designed for very fast chemical space exploration, employing double-pushout diagrams to specify reaction rules. This assures mathematical consistency, formal reaction reversibility, and associative composability of rules and reactions. In addition, application of the "AlChemy" rule set collection ensures adherence to chemistry-like behavior.[4] The framework iteratively computes all reachable products, mitigating a combinatorial explosion by different pruning strategies such that chemically unlikely products can be recognized and rejected.

Following the generation of potential products, an MS/MS spectra predictor such as CFM-ID is used to predict possible fragments.[5] These fragments, along with their corresponding base molecules and mass peaks, will be stored in a database. Subsequently, the database can be queried on a peak-by-peak basis to analyze experimental MS/MS spectra of complex reaction mixtures. The acquired information facilitates the indentification of the most probable candidate molecules and the most likely formation pathway (figure 1).

Our pipeline represents a first step to a more comprehensive simulation of origin of life models, such as the amino acids in the Miller-Urey experiment. Towards this goal, we aim to iteratively increase the set of reaction rules, while validating results based on experimental data in a closed loop. To demonstrate the feasibility of the approach, we start by investigating thermal decomposition of amino acids, simulated by the use of collision-induced dissociation (CID). Ultimately, we aim to provide a framework capable to form hypotheses for the source of spectra in real world settings. One putative example is the investigation of the oceanic compositions of the icy *Saturn* moon *Enceladus*. Through this pipeline, we offer a systematic and computationally efficient method for unraveling complexities inherent in reaction mixtures, paving the way for a better understanding of them.

Figure 1: *(left) Analysis of candidates in MS/MS Spectra of complex reaction mixture. Overlaps in Predicted fragments and predicted fragments are highlighted (right) Derived likeliest reaction path to found candidates.*

Bibliography

1. S L. Miller and H. C. Urey. Science, 130:1622–1623, **1959**.
2. Cooper GJT, Surman AJ, McIver J, et al., Angew. Chem. Int. Ed. Engl. **2017**;56(28):8079-8082.
3. J. L. Andersen; C. Flamm; D. Merkle; P. F. Stadler. J. Syst. Chem., 4 (**2013**) 4.
4. A. Wollos; R. Roszak; A. Żądlo-Dobrowolska; W. Beker; B. Mikulak-Klucznik; G. Spólni; M. Dygas; S. Szymkuć; B. A. Grzybowski Science, 369 (**2020**) 1-12.
5. F. Wang, D. Allen, S. Tian , E. Oler , V. Gautam, R. Greiner, TO Metz, DS Wishart. Nucleic Acids Research 50 (**2022**) W165-W174.

P18: Deciphering Molecular Embeddings with Centered Kernel Alignment

Matthias Welsch[1,2,3], Steffen Hirte[1,3], Johannes Kirchmair[*,1,2]

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

[2] *Christian Doppler Laboratory for Molecular Informatics in the Biosciences, Department for Pharmaceutical Sciences, University of Vienna, Austria*

[3] *Vienna Doctoral School of Pharmaceutical, Nutritional and Sport Sciences (PhaNuSpo), University of Vienna, Austria*

Analyzing machine learning models, particularly nonlinear ones, presents significant challenges. In this context, centered kernel alignment (CKA) has emerged as a valuable tool for assessing similarity between embeddings.[1] The effectiveness of CKA depends on selecting a kernel that accurately captures the relevant properties of the models being compared. Originally designed for neural networks (NNs), with their invariance to data rotation in mind, CKA has been successfully applied across various scientific fields. However, its adoption in cheminformatics has been limited, partly due to the widespread use of random forests (RF), which lack rotational invariance. In this work,[2] we adapt CKA by incorporating an RF-specific kernel to better align with RF properties. As part of our validation, we demonstrate that this adapted approach correlates well with RF model prediction similarity. Furthermore, we illustrate how CKA with the RF kernel can be used to analyze and interpret RF models trained on molecular and rooted fingerprints.

References

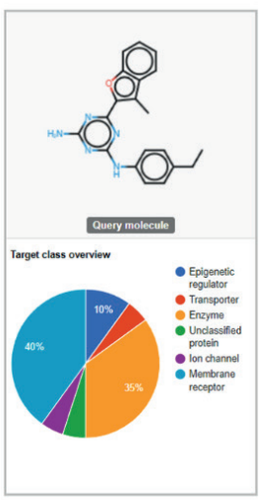
1. Kornblith, S. et al., Similarity of Neural Network Representations Revisited. Proceedings of the 36th International Conference on Machine Learning, in Proceedings of Machine Learning Research., **2019**, 97, 3519-3529 https://proceedings.mlr.press/v97/kornblith19a.html
2. Welsch, M. et al., Deciphering Molecular Embeddings with Centered Kernel Alignment. Journal of Chemical Information and Modeling., **2024**, 64, 19, 7303-7312 10.1021/acs.jcim.4c00837

P20: Predicting off-targets from ChEMBL data using the polypharmacology browser

Maedeh Darsaraee and Jean-Louis Reymond*

Department of Chemistry, Biochemistry and Pharmacy, University of Berne, Switzerland

The public archive ChEMBL, which collects bioactive compounds and their associated targets from the literature, has been used by many groups to build models predicting the possible targets of small molecules to guide the experimental search for off-targets. In our group we have developed the polypharmacology browsers (PPB and PPB2),[1,2] which assign possible targets to a query molecule based on molecular fingerprint similarities to ChEMBL molecules, and provided critical insights in several practical case studies such as the identification of LPAATβ as the actual target of a putative kinase inhibitor (Figure).[3,4]



Rank	ChEMBL ID	Common name	Nearest neighbours
1	CHEMBL4772	1-acylglycerol-3-phosphate O-acyltransferase beta	Show NN
2	CHEMBL253	Cannabinoid CB2 receptor	Show NN
3	CHEMBL1904	Glutamate [NMDA] receptor subunit epsilon 2	Show NN
4	CHEMBL1800	Corticotropin releasing factor receptor 1	Show NN
5	CHEMBL224	Serotonin 2a (5-HT2a) receptor	Show NN
6	CHEMBL228	Serotonin transporter	Show NN
7	CHEMBL2916	Telomerase reverse transcriptase	Show NN
8	CHEMBL1293222	Nucleotide-binding oligomerization domain-containing protein 1	Show NN
9	CHEMBL2971	Tyrosine-protein kinase JAK2	Show NN
10	CHEMBL3764	Urotensin II receptor	Show NN
11	CHEMBL3833	Trace amine-associated receptor 1	Show NN
12	CHEMBL204	Thrombin	Show NN
13	CHEMBL3018	Matriptase	Show NN
14	CHEMBL3286	Urokinase-type plasminogen activator	Show NN
15	CHEMBL244	Coagulation factor X	Show NN

However, our PPB and PPB2 models associated only a single target per ChEMBL molecule. To better integrate the existing polypharmacology information available in ChEMBL, we are updating our PPB to handle multi-target information for ChEMBL molecules, using various machine learning models into account, and exploiting the latest version of the database featuring a total of 1.6 million molecule-target associations.

Keywords: Computer-aided drug design, Polypharmacology, Target prediction, Web-based tool, Cheminformatics

References

1. M. Awale, J. L. Reymond, *J. Cheminf.* **2017**, *9*, 11, DOI: 10.1186/s13321-017-0199-x.
2. M. Awale, J.-L. Reymond, *J. Chem. Inf. Model.* **2019**, *59*, 10, DOI: 10.1021/acs.jcim.8b00524.
3. M. Poirier, M. Awale, M. A. Roelli, G. T. Giuffredi, L. Ruddigkeit, L. Evensen, A. Stooss, S. Calarco, J. B. Lorens, R.-P. Charles, J.-L. Reymond, *ChemMedChem* **2019**, *14*, 224, DOI: 10.1002/cmdc.201800554.
4. M. R. Cunha, R. Bhardwaj, A. L. Carrel, S. Lindinger, C. Romanin, R. Parise-Filho, M. A. Hediger, J.-L. Reymond, *RSC Med. Chem.* **2020**, *11*, 1032, DOI: 10.1039/D0MD00145G.

P22: COMPASS: COMputational Pocket Analysis and Scoring System

Akash Deep Biswas[1]*, Emanuela Sabato[2], Serena Vittorio[2], Parisa Aletayeb[2], Alessandro Pedretti[2], Angelica Mazzolari[2], Carmen Gratteri[3], Andrea R. Beccari[1], Giulio Vistoli[2], and Carmine Talarico[1]

[1] *Dompé Farmaceutici S.p.A., Napoli, Italy*

[2] *Dipartimento di Scienze Farmaceutiche, Università degli Studi di Milano, Italy*

[3] *LIGHT s.c.ar.l., Brescia, Italy*

We present a novel computational methodology for prioritizing protein binding sites through the integration of structural analysis, molecular docking, and molecular dynamics simulations. The approach introduces an innovative Pocket Frequency Score (PFS) algorithm that quantifies binding site relevance based on residue frequency patterns across multiple protein conformations. Our pipeline begins with comprehensive binding site detection across diverse protein structures, followed by systematic docking simulations with probe molecules. The PFS algorithm then evaluates each potential binding site by analyzing the distribution and conservation of key residues, generating a unique frequency-based metric that captures site significance across conformational states. This score is combined with geometric and interaction-based parameters to produce a Global Score, [1] enabling systematic ranking of binding sites. The methodology employs a multi-step validation process where top-ranked sites undergo molecular dynamics simulations and binding free energy calculations to assess their stability and druggability. The algorithm's ability to process multiple conformational states while considering residue frequency patterns represents a significant advance in binding site identification. This systematic approach enhances structure-based drug discovery by enabling rational selection of optimal protein conformations and binding sites, offering a robust framework for virtual screening campaigns that can be applied across diverse protein targets.

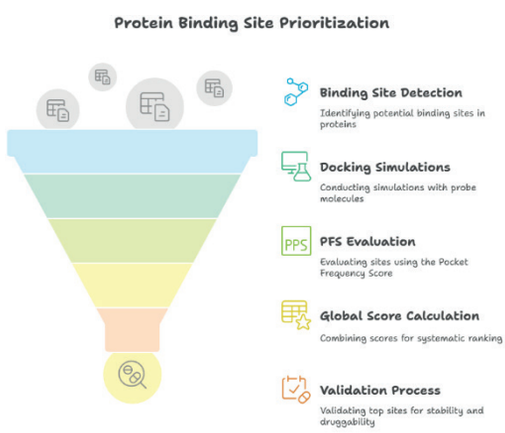


***Figure 1**: The workflow of COMPASS*

References

1. Bocchi, Giovanni, et al. "GENEOnet: A new machine learning paradigm based on Group Equivariant Non-Expansive Operators. An application to protein pocket detection." *arXiv preprint arXiv:2202.00451* (**2022**).

P24: LACAN: Leveraging adjacent co-occurrence of atomic neighborhoods

Wim Dehaen[1,2]

[1] *Department of Informatics and Chemistry, University of Chemistry and Technology Prague, Czech Republic;*

[2] *Department of Organic Chemistry, University of Chemistry and Technology Prague, Czech Republic*

Some molecular fragments commonly occur and co-occur in organic molecules, but when they co-occur, they tend to not be adjacent. For example, halides and amines are common motifs in druglike molecules, but their direct linkage, haloamines, is associated with toxicity and instability. Likewise, a druglike molecule will commonly contain several alkoxy radicals, but their direct linkage, peroxides, is underrepresented in databases of druglike molecules as this functionality is often explosive and toxic.

This adjacent co-occurrence phenomenon can be leveraged for local molecular filtering: for every bond in a molecule, two small atomic neighborhoods (each atom participating in the bond and their topologically connected neighbors, i.e. ECFP2-like) exist around the bond interface. For this combination of neighborhoods, the actual vs. expected co-occurrence in a large reference data set can then quantify how abnormal a given bond is.

We show this approach is able to localize chemical issues, including in cases where commonly used filters such as SAScore[1] and QED[2] fail. This data-driven filter can then be used for applying constraints on operations on the molecular graph such as molecular mutations or fragment recombinations, retaining only molecules with acceptable linkages. This results in chemically sensible mutation and crossover operations that are directly applied to a new molecular generator based on an Evolutionary Algorithm strategy.

References

1. Ertl, P., & Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of cheminformatics*, **2009**, *1*, 1, 11. DOI: 10.1186/1758-2946-1-8
2. Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S., & Hopkins, A. L. , Quantifying the chemical beauty of drugs. Nature chemistry, **2012**, 4, 2, 90-98. DOI:10.1038/nchem.1243

Poster Session BLUE: Artificial Intelligence, Machine Learning, and QSAR

P26: Enriching ChEMBL assay descriptions using Natural Language Processing

Ines A. Smit[1], Melissa F. Adasme[1], Emma Manners[1], Sybilla Corbett[1], Hoang-My-Anh Do[1], Noel O'Boyle[1], Andrew R. Leach[1], Barbara Zdrazil[1]

[1] *Chemical Biology Services, European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, United Kingdom*

Bioactivity data in ChEMBL links compounds to their biological targets via assays that measure biological activity in a given experimental set-up. Well-annotated assays are essential for users to combine bioactivity data measured in similar assay set ups for analysis and modelling. Here, we describe two applications of Natural Language Processing (NLP) to enrich ChEMBL assay descriptions (Fig. 1). First, we developed a novel broad assay categorization using a text classification model which provides more granular assay categories compared to the current binding (B) and functional (F) assay types. Second, we trained a Named Entity Recognition (NER) model to extract experimental method details and subsequently link these to the BioAssay Ontology (BAO)[1] using the text2term tool,[2] providing annotations for previously unmapped entities.

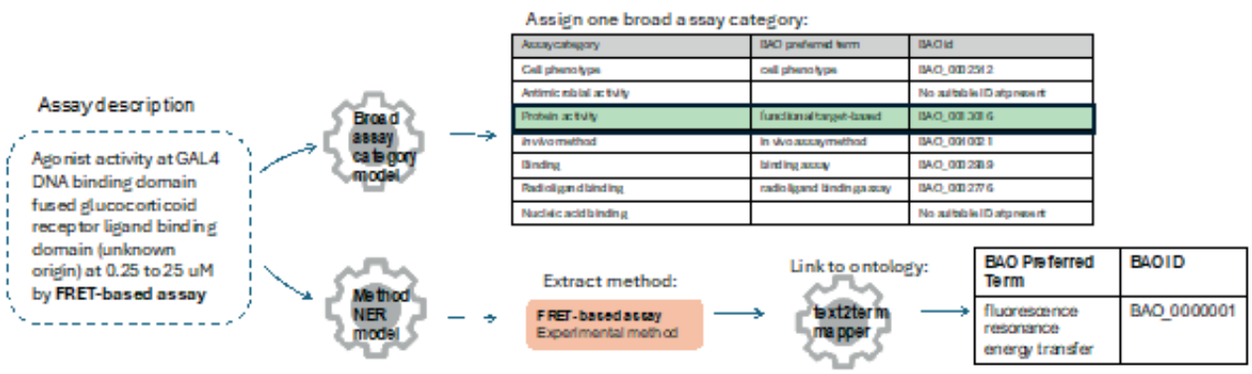


***Figure 1**: NLP pipeline to annotate ChEMBL assays with a broad assay category and experimental method annotations linked to the BioAssay Ontology (BAO)*

A total of 900 assays were manually annotated with broad assay categories and used to train a multi-class classification model predicting the following assay categories: protein activity, cell phenotype, antimicrobial activity, *in vivo* method, binding, radioligand binding and nucleic acid binding. The resulting models have average F-scores between 0.81-0.91 depending on the assay category during cross-validation. In ChEMBL 35, 89% of B and F assays can be reliably categorized with one of these categories. For the NER model, 800 manually annotated assays were used for training, yielding models with an average precision, recall, and F1-score of 0.93, 0.95, and 0.94, respectively during cross-validation. An experimental method could be identified in 57% of B and F assays in ChEMBL 35, while most of the remaining assay descriptions do not contain a method. Lastly, we created a gold standard of 600 assays with experimental methods mapped to BAO terms. We evaluated the text-2term[2] tool on this gold standard, finding it can link to correct BAO terms with a precision of ~0.8 and recall of ~0.6. Overall, the novel annotations can help users find and group assay data appropriately and support making ChEMBL data FAIRer and more ML-ready.

References

1. Abeyruwan, S. *et al.* Evolving BioAssay Ontology (BAO): modularization, integration and applications. J Biomed Semantics, **2014**, 5, S5, doi: 10.1186/2041-1480-5-S1-S5
2. Gonçalves, R. S. *et al.* The text2term tool to map free-text descriptions of biomedical terms to ontologies. Preprint, **2024**, at https://doi.org/10.48550/arXiv.2407.02626

P28: Multiretro – A Synthesis Planning Toolkit

Alan Kai Hassen[1,2], Jacquelyn L. Klug-McLeod[3], Roger M. Howard[3], Jason Mustakis[3], Antonius P. A. Janssen[4], Gerard J.P. van Westen[4], Mike Preuss[2], Djork-Arné Clevert[1]

[1] Machine Learning Research, Pfizer Research and Development, Berlin, Germany

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

[3] Pfizer Research and Development, Groton, CT, USA

[4] Leiden Academic Centre for Drug Research, Leiden University, The Netherlands

We introduce "Multiretro – a Synthesis Planning Toolkit" designed to deploy synthesis planning solutions at scale on high-performance computing (HPC) infrastructure. Our toolkit provides an end-to-end process, from executing synthesis planning on large sets of molecules to analyzing the resulting synthesis routes across multiple levels (reaction-level, synthesis route-level, molecule-level, and multi-target). By seamlessly integrating with AiZynthFinder [1] for synthesis planning and the Models Matter framework for state-of-the-art retrosynthesis models [2], we demonstrate the versatility of Multiretro through a range of synthesis planning applications.

References

1. Saigiridharan, L. et al. (**2024**) 'AiZynthFinder 4.0: Developments based on learnings from 3 years of industrial application', Journal of Cheminformatics, 16(1). doi:10.1186/s13321-024-00860-x.
2. Torren-Peraire, P. et al. (**2024**) 'Models matter: The impact of single-step retrosynthesis on Synthesis Planning', Digital Discovery, 3(3), pp. 558–572. doi:10.1039/d3dd00252g.

P30: Robust Prediction of the Pharmacophore Fit Scores with Active Learning

D. Goldmann, C. Grebner, G. Hessler

Synthetic Molecular Design, Integrated Drug Discovery, Sanofi, Frankfurt am Main, Germany

Virtual screening of large databases with ligand-based or structure-based pharmacophore models is a classical computational methodology applied in hit and lead optimization campaigns[1]. However, conformer generation and pharmacophore screening are still time-consuming steps for searching large chemical spaces for promising compounds. Therefore, we have evaluated and integrated machine learning into pharmacophore screening workflows to accelerate pharmacophore-based virtual screening.

One key ingredient is active learning which has been widely applied for building high performing machine learning models to predict molecular docking scores and binding free energies[2]. In this work we validated active learning for predicting pharmacophore scores for several diverse targets. We used classical (Random Forest) and deep learning (feedforward and graph convolutional neural networks) machine learning for generation of the most accurate models for the prediction of the pharmacophoric scores (Figure 1). This was combined with two active learning cycles based on the uncertainty quantification[3] of the virtual hits and further improved the model accuracy on the external test set. We observe a significant reduction of the execution time required for the generation of the machine learning based pharmacophore scores compared to traditional pharmacophore screening without sacrificing on the diversity of the hits.

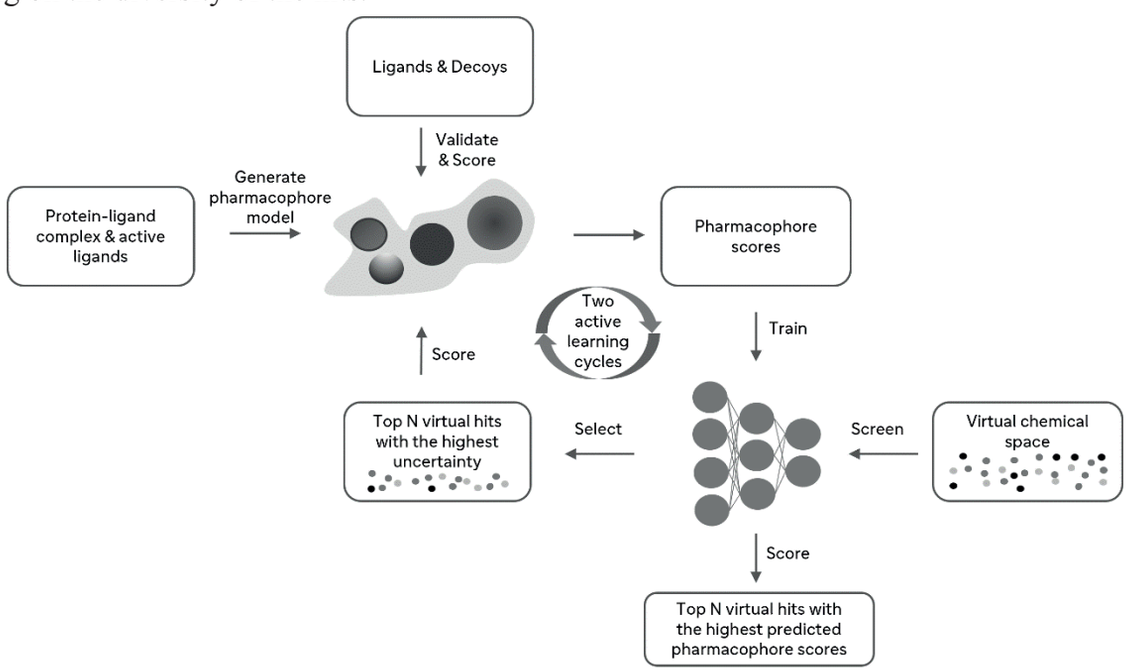


***Figure 1**: Active learning workflow for predicting pharmacophore scores*

References

1. WOLBER, G., et al., LigandScout: 3-D Pharmacophores Derived from Protein-Bound Ligands and Their Use as Virtual Screening Filters. J. Chem. Inf. Model., **2005**, 45, issue, 160-169, DOI:10.1021/CI049885E
2. GRAFF, D., et al., Accelerating high-throughput virtual screening through molecular pool-based active learning. Chem. Sci., **2021**, 12, 7866-7881, DOI:10.1039/DOSC6805E
3. BAILEY, M., et.al., Deep Batch Active Learning for Drug Discovery. BioRxiv, **2023**, DOI:10.1101/2023.07.26.550653.

P32: Using deep learning and machine learning-based docking to investigate metalloenzyme-substrate complexes

Daniil Lepikhov[1,2], Laura Sandner[1], Silke Leimkühler[2], Ariane Nunes-Alves[1]

[1] *Theoretical Structural Biology group, Technical University Berlin, Germany*

[2] *Molecular Enzymology group, University Potsdam, Germany*

Databases of experimentally determined structures such as the Protein Data Bank were used to develop AlphaFold2[1], which in turn was used to build the AlphaFold Protein Structure database of computed protein structures. This database regroups more than 200 millions entries, offering a high quality database of synthetic protein structures. Analogically, a great effort is ongoing to build such a database for protein-ligand complex structures. Here, we're interested in developing a pipeline to build a synthetic database of enzyme-substrate complex structures, which can be challenging due to the presence of metals and cofactors in enzymes. To evaluate the pipeline efficiency and accuracy in modeling enzyme-substrate complex structures, we built a non-redundant benchmark dataset of 1,325 complexes using Q-BioLiP[2].

One of the major limitations of canonical docking approaches[3,4] is that prior knowledge of the binding pocket is required. To address this drawback, AI tools were developed for simultaneous prediction of the binding site and the correct ligand conformation. In this work, we compared the performance of 7 AI tools[5-11](Equibind, TankBind, DiffDock, NeuralPlexer, DiffusionProteinLigand, DynamicBind and AlphaFold3) to the popular docking software AutoDock Vina in the prediction of substrate-enzyme complex structures for 3 different enzyme classes, A, B and C. A represents enzymes without metal cofactors, B enzymes with metals and C enzymes with metals in the binding pocket (figure 1). The distinction between these 3 classes was made in order to emphasize the ability of each tool to account for metal ions in the binding pocket, which is known to be challenging[12]. Our results show that no AI tools outperform Vina when binding site information is provided. There is no single solution to the problem of building a high quality synthetic database of enzyme-substrate complex structures, but rather a combination of AI tools for high-throughput binding pocket elucidation, followed by conformation fine-tuning, as suggested in previous work[6,7].

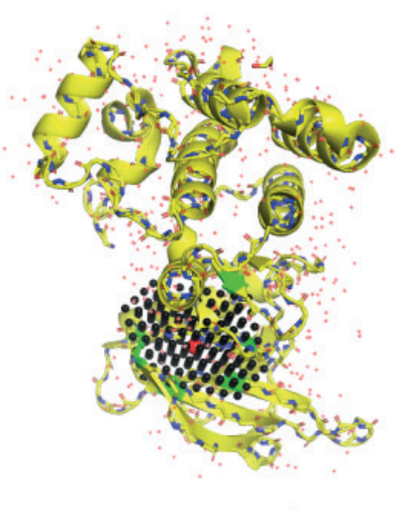


Figure 1*: Geometrically defined binding site (black spheres) with its center of mass (red sphere)*

To evaluate the implication of accurate models of enzyme-substrate complex structures in enzymology, we also created deep learning models to predict the constant of Michaelis (Km), which indicates the enzymatic activity for a given enzyme-substrate complex. Using neural networks, amino acid identity as features for the enzymes, and fingerprints and molecular descriptors as features for the substrates, we show that additional enzyme structural information improves generalization of the

model to unseen data. Accurately computing the Km *in silico* can accelerate the discovery of new substrates and broaden our understanding of biology.

References

1. Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (**2021**). https://doi.org/10.1038/s41586-021-03819-2
2. Wei et al, Q-BioLiP: A Comprehensive Resource for Quaternary Structure-based Protein–ligand Interactions, Genomics, Proteomics & Bioinformatics, 22(1), qzae001, **2024**.
3. Eberhardt, Jerome et al., AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. J. Chem. Inf. Model., **2021**, 10.1021/acs.jcim.1c00203
4. Verdonk ML et al., Improved protein-ligand docking using GOLD. Proteins. **2003**, doi: 10.1002/prot.10465
5. Li Y, Li L, Wang S, Tang X. EQUIBIND: A geometric deep learning-based protein-ligand binding prediction method. Drug Discov Ther. **2023,** doi: 10.5582/ddt.2023.01063
6. Wei Lu et al., TANKBind: Trigonometry-Aware Neural NetworKs for Drug-Protein Binding Structure Prediction. NeurIPS **2022**
7. Nakata, S., Mori, Y. & Tanaka, S. End-to-end protein–ligand complex structure generation with diffusion-based generative models. BMC Bioinformatics 24, 233, **2023**, 10.1186/s12859-023-05354-5
8. Qiao, Z., Nie, W., Vahdat, A. et al. State-specific protein–ligand complex structure prediction with a multiscale deep generative model. Nat Mach Intell 6, 195–208, **2024**, 10.1038/s42256-024-00792-z
9. Corso et al., DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking. ICLR **2023**, 10.48550/arXiv.2210.01776
10. Lu, W., Zhang, J., Huang, W. et al. DynamicBind: predicting ligand-specific protein-ligand complex structure with a deep equivariant generative model. Nat Commun 15, 1071 (**2024**). 10.1038/s41467-024-45461-2
11. Abramson, J., Adler, J., Dunger, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 630, 493–500, **2024**, 10.1038/s41586-024-07487-w
12. Matthijs L. A. Hakkennes, Francesco Buda, and Sylvestre Bonnet, MetalDock: An Open Access Docking Tool for Easy and Reproducible Docking of Metal Complexes. Journal of Chemical Information and Modeling 63 (24), 7816-7825. **2023**, 0.1021/acs.jcim.3c01582
13. Martin Buttenschoen, Garrett M. Morris, Charlotte M. Deane. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. **2023**, 10.48550/arXiv.2308.05777

P34: From Theory to Practice: Reaction Similarity Search in Real-World Applications at Astex

I.N. Derbenev, D. Branduardi, R.F. Ludlow

[1] *Computational Chemistry & Informatics Department, Astex Pharmaceuticals, Cambridge, United Kingdom*

Astex has been a pioneer in fragment-based drug discovery for over two decades, accumulating a database of approximately 50,000 in-house reactions. This collection continues to grow with contributions from contract research organizations (CROs) (~20,000 reactions) and published reactions (~20 million reactions). The challenge lies in transforming this vast dataset into actionable knowledge.

To tackle this, we have adopted the reaction similarity model developed by Philippe Schwaller et al. at IBM, Switzerland.[1] Based on bidirectional encoder representations from transformers (BERT), the model encodes reaction SMIRKS into embeddings, effectively creating reaction fingerprints. Implemented as an API within our internal custom electronic lab notebook (ELN), the system is backed by a PostgreSQL database, enabling efficient vector-based search and retrieval. This is of paramount importance to ensure the performance required for an acceptable user experience.

This presentation demonstrates how the model, initially designed in a controlled research setting, has been successfully adapted for real-world cheminformatics applications (see Figure 1). A case study will show how reaction similarity search aids in selecting optimal reaction conditions, ultimately improving synthetic efficiency. Our experience highlights the efforts involved in the transition from innovative research to a practical, impactful tool used every day in drug discovery.

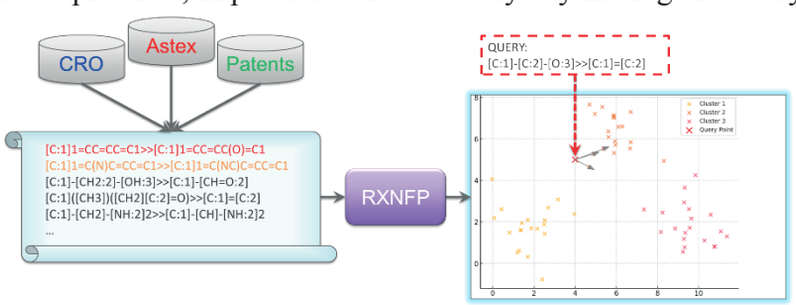


Figure 1. *Schematic depiction of reaction similarity search*

References

1. Schwaller, P., et al., Mapping the space of chemical reactions using attention-based neural networks. Nature machine intelligence, 2021, 3, 2, 144-152, https://doi.org/10.1038/s42256-020-00284-w

P36: Prediction of Pharmacokinetics Profile as Time Series

Uday Abu Shehab[1,3], Gerhard F. Ecker[1], Lina Humbeck[2], Miha Skalic[2], Moritz Walter[2], Andreas Bergner[3]

[1] *University of Vienna, Department of Pharmaceutical Chemistry, Austria*

[2] *Boehringer Ingelheim Pharma GmbH & Co KG, Medicinal Chemistry Department, Biberach an der Riss, Germany*

[3] *Boehringer Ingelheim RCV GmbH & Co KG, Drug Discovery Sciences, Vienna, Austria*

Early evaluation of a drug candidate's pharmacokinetics (PK) profile is critical for the success of any drug discovery project. Over the past two decades, improvements in PK profiling have significantly reduced drug attrition rates associated with poor PK profiles from approximately 40% to 10%.[1] Nonetheless, further refinements in PK profiling can continue to reduce attrition rates.

This project explored machine learning approaches – such as deep autoregression (DeepAR), temporal fusion transformer (TFT) algorithms, and physics-informed neural networks (PINNs) modeling – to predict the pharmacokinetics profile as a time-series. Both models used static molecular descriptors to predict the plasma drug concentration over time. While TFT outperformed DeepAR in predictive power, it demonstrated comparable performance to three in-house models at Boehringer Ingelheim[2]in vivo PK studies need to be conducted. While the prediction of ADME properties of compounds using Machine Learning (ML that employ different strategies and training sets.

Further analysis revealed that each model (TFT and three in-house models) outperformed all other models in predicting a unique set of compounds, leading to developing a consensus model. Although the consensus model did not improve the overall accuracy of the model when evaluating the entire external test set, the accuracy improvement was significant when evaluating only those predictions with the least uncertainty.

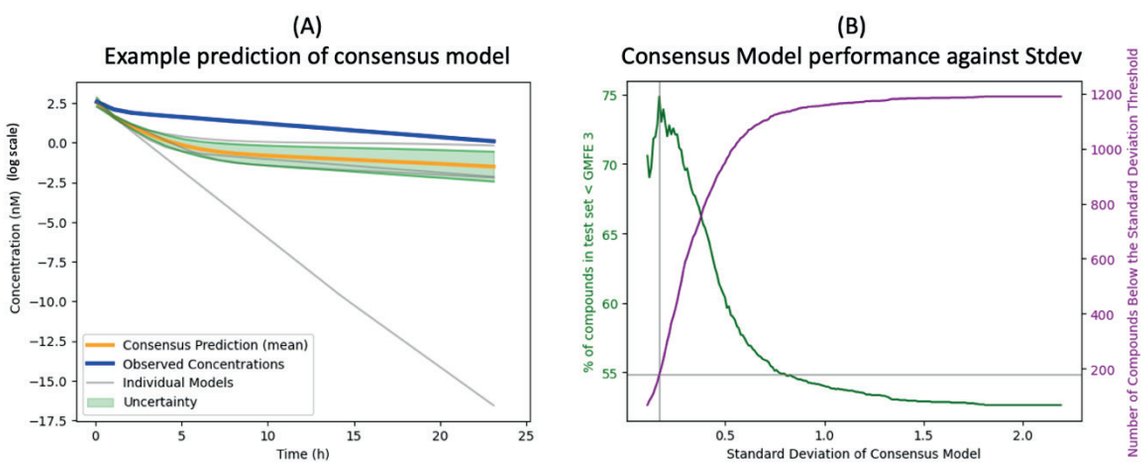


***Figure 1**: (A) Example prediction of the consensus model after outlier removal. (B) Model performance with Geometric Mean Fold Error (GMFE) < 3 against standard deviation in the consensus model. Higher standard deviation leads to lower prediction accuracy. Outliers are defined as curves outside the 25th to 75th percentile.*

While further investigation is needed to understand better the relevance of these findings to modeling PK data, our results serve as a proof of concept for using time series machine learning algorithms. We also demonstrate the potential advantages of developing a consensus model with uncertainty quantification for real-world screening scenarios. In addition, we highlight the power of using PINNs and their impact on modeling life science data, including PK data.

References

1. Kola, I.; Landis, J. Can the Pharmaceutical Industry Reduce Attrition Rates? *Nat. Rev. Drug Discov.* **2004**, *3* (8), 711–716. https://doi.org/10.1038/nrd1470.
2. Walter, M.; Aljayyoussi, G.; Gerner, B.; Rapp, H.; Tautermann, C. S.; Balazki, P.; Skalic, M.; Borghardt, J. M.; Humbeck, L. *In Silico* PK Predictions in Drug Discovery: Benchmarking of Strategies to Integrate Machine Learning with Empiric and Mechanistic PK Modelling. July 30, **2024**. https://doi.org/10.1101/2024.07.30.605777.

P38: Chemistry-aware foundation model for Small Molecule ADMET and Polypharmacology Property Estimation

Pietro Morerio[1], Filippo Lunghini[2], Alessio Del Bue[1], Andrea Beccari[2]

[1] *Istituto Italiano di Tecnologia, Italy*

[2] *EXSCALATE, Dompé Farmaceutici SpA, Naples, Italy*

Foundation models are transforming drug discovery by significantly enhancing the accuracy of molecular property predictions. Leveraging vast datasets and self-supervised learning techniques, these advanced models outperform traditional methods, uncovering novel patterns and relationships.

We propose an innovative architecture that leverages graph neural networks and a set of hand-crafted chemistry-meaningful fingerprint descriptors, "DompeKeys (DKs)" [1]. These DKs enrich the molecular graph (Figure 1) with an additional hierarchical layer encoding from specific functional groups and structural patterns to simpler pharmacophoric points.

The model is pre-trained in a self-supervised fashion on a set of meaningful pretext tasks, involving both standard atom-level and novel DK-related prediction tasks.

Finetuned for downstream tasks, our model achieved state of the art performances on a comprehensive set of 22 ADMET endpoints [2] and predicting polypharmacological drug target-interactions on >60 kinases [3].

Thanks to the incorporation of an attention layer, our model significantly enhances explainability in AI. This enables the model to identify the most significant functional groups responsible the activity, providing medicinal chemists with actionable insights to replace these groups during the lead optimization phase. Our framework thus offers critical support in drug design, leading to a more efficient and targeted therapeutic development.

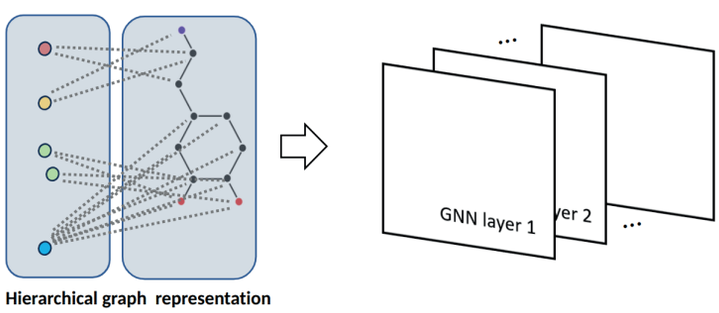


***Figure 1**: DK-based hierarchical molecule representation*

References

1. https://jcheminf.biomedcentral.com/articles/10.1186/s13321-024-00813-4
2. https://www.nature.com/articles/s41589-022-01131-2
3. https://jcheminf.biomedcentral.com/articles/10.1186/s13321-023-00728-6

P40: Exploiting SARkush and Free-Wilson Analysis to Accelerate an Antiviral Drug Discovery Project

Jess Stacey[1], Lauren Reid[1], Al Dossetter[1], Ed Griffen[1], Andrew Leach[1], Phillip de Sousa[1], Bashy Khan[1], Dan James[1], and David Cousins[1]

[1] MedChemica Ltd, Motorworks, Macclesfield, United Kingdom

ASAP Discovery[1] is an open-science, antiviral drug discovery consortium contributed to by scientists from all over the world. ASAP's aim is to deliver antiviral drug compounds in preparation for the next pandemic. It focuses on accelerating the identification and optimisation of compounds through structure-based and ligand-based approaches by collecting data for numerous biological assays. There are multiple active programs that are being run simultaneously. It is therefore important to be able to quickly and routinely extract SAR in each program to understand and develop the structure-activity relationship (SAR). The lead optimisation phase is led by MedChemica and exploit their tools to guide the design process. SARkush®[2] and Free-Wilson3 analysis are used together to extract and understand SAR knowledge.

Free-Wilson analysis is an established quantitative structure-activity relationship (QSAR) technique that can relate the presence or absence or specific structural fragments to a contribution factor of a property of interest by using regression.

Although first published concurrently with Hansch and Fujita's[4] work on numerical analysis that developed into QSAR, Free-Wilson analysis has been underused in comparison, with less than half the citations of Hansch and Fujita's work. The critical, practical issue in applying Free-Wilson analysis is in the data preparation stage of accurately assigning substituents into R group tables. Although trivial in small sets of compounds, "real world" applications where there may be partial or full symmetry between R groups positions (for example in the ortho and meta positions of aromatic rings) or the substitution of heterocycles for aryl rings makes this assignment difficult. SARkush® is a MedChemica tool that automatically clusters compounds into Markush like structures to explain the SAR around those structures. Importantly, it heuristically assigns R groups to cores making it an ideal solution to the data preparation stage of Free-Wilson analysis.

We show how rapid and routine Free-Wilson analysis can be used to develop time-series analysis of the MERS and SARS project, focus on gaps in the compounds made at a particular point and quickly spot where particular substituents are of benefit. This was used to guide decision-making in this lead optimisation project.

References

1. https://asapdiscovery.org/
2. https://www.medchemica.com/a-cheminformatics-journey-to-develop-sarkushr/
3. Free S.M., Wilson J.W. A Mathematical Contribution to Structure-Activity Studies. J Med Chem. **1964**, 7(4), 395-399.
4. Hansch, C., Fukita, T. *p*-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure J. Am. Chem. Soc. **1964**, 86, 8, 1616–1626.

P42: Efficient compound selection strategies in lead optimization: insights from retrospective analysis

Mas Pablo[1,2], Filoche-Rommé Bruno[2], Vuilleumier Rodolphe[1], Bianciotto Marc[2]

[1] PASTEUR Lab, École Normale Supérieure – PSL, Paris, France

[2] Molecular Design Sciences, Integrated Drug Discovery, Sanofi, Vitry-sur-Seine, France

In the Drug Discovery pipeline, the lead optimization stage is devoted to the multi-parametric optimization of a small set of lead compounds identified at the previous stage of the process. Within the Design-Make-Test-Analyze (DMTA) paradigm, the lead optimization takes place through an iterative process where new molecules are proposed and selected for synthesis at the "Design" stage of each DMTA round. This iterative selection process is akin to Active Learning (AL), a subfield of Machine Learning (ML) in which an algorithm selectively queries unlabeled samples to optimize model training.

AL algorithms have been applied to the iterative selection of molecules during virtual screening[1] or hit finding[2], but their application at the lead optimization stage presents significant challenges. One of them is that the pool of molecular hypotheses to choose from at each round of selection is dynamic, contrary to the lead identification stage where the chemical space of hypotheses to be selected is static. Additionally, lead optimization entails a multi-objective optimization, where potency must be balanced with selectivity and ADMET properties, further complicating the application of AL in this context.

We developed a multi-objective compound prioritization toolkit that integrates a set of selection strategies inspired by AL algorithms, multicriteria decision methods, and classical medicinal chemistry approaches. On top of this toolkit, we developed an analytical framework that quantitatively and qualitatively characterizes the exploitation and exploration capabilities of these strategies. We used this framework to evaluate retrospectively the performance of these different selection strategies on a dozen of legacy industrial lead optimization projects.

In this contribution, we present our analysis of these retrospective simulations to uncover the distinct behaviors of different selection strategies. Our findings highlight that while some strategies excel at rapidly identifying the best compounds, they do so at the expense of a less thorough exploration of the chemical space. Furthermore, we demonstrate that commonly used evaluation criteria for selection strategies can be misleading.

References

1. Graff, D.E., et al., Accelerating high-throughput virtual screening through molecular pool-based active learning. Chemical Science., 2021, 10.1039/d0sc06805e
2. Tilborg, D., et al., Traversing chemical space with active deep learning for low-data drug discovery. Nature Computational Science., 2024, 10.1038/s43588-024-00697-2

Pablo Mas, Bruno-Filoche-Rommé and Marc Bianciotto are Sanofi employees and may hold shares and/or stock options in the company. Rodolphe Vuilleumier has nothing to disclose.

P44: regAL: Python Package for Active Learning of Regression Problems

Elizaveta Surzhikova and Jonny Proppe

Institute of Physical and Theoretical Chemistry, Technische Universität Braunschweig, Germany

Machine learning models have been successfully developed for various applications, including different matters in chemistry. [1] While in many cases the efficiency of these models makes them more favorable than classical computational chemistry methods, their usage has a major constraint: the need for sufficient training data. This poses a serious limitation in contexts where labeled data is scarce or expensive to obtain. To overcome this issue, machine learning models can be trained with active participation. An active learning algorithm actively decides which training instance needs to be labeled next for the machine learning model to obtain the highest knowledge gain. [2] This process accelerates the convergence of model performance and therefore allows the use of less training data. A key performance factor of active learning is the criterion by which the next training instance to be labeled is chosen. Numerous different selection strategies are known. Selecting an active learning algorithm with superior convergence of model performance for a given dataset requires time for research, analysis and implementation. To bypass that obstacle, we present **regAL** (**A**ctive **L**earning for **Reg**ression) [3], a Python package enabling active learning in various flavors and in a black-box fashion for arbitrary datasets. In addition, we are currently developing regAL further by allowing users to start the active learning procedure with no prior knowledge of the dataset.

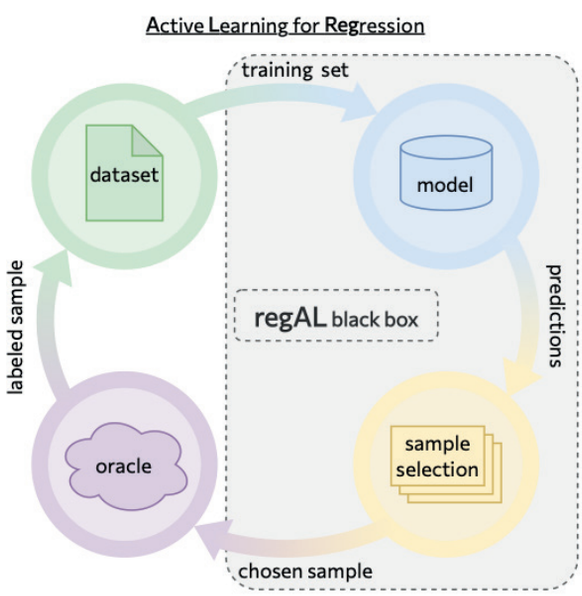


***Figure 1**: The workflow cycle of the active learning method. The model training and sample selection parts are covered by **regAL**.*

References

1. Friederich, P., et al., Machine-learned potentials for next-generation matter simulations. Nature Materials., **2021**, 20, 750–761, 10.1038/s41563-020-0777-6
2. Settles, B., Active Learning. Morgan & Claypool., **2012**, 1-20, 10.5555/3019233
3. Surzhikova, E., et al., regAL: Python Package for Active Learning of Regression Problems. arXiv preprint., **2024**, 10.48550/arXiv.2410.17917

P46: Exploration of Data from the Pharmaceutical Industry for Site-of-Metabolism Prediction

Ya Chen[1,2]

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

[2] *Medicinal Chemistry, Research and Early Development, Cardiovascular, Renal and Metabolism (CVRM), BioPharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden*

Metabolism presents significant challenges in drug discovery due to its crucial role in drug safety and efficacy. Experimental studies of drug metabolism are time-consuming and resource-intensive. Sites of metabolism (SoMs) are the atom positions where metabolic transformations occur. Several mature in silico SoM predictors, many based on machine learning, can be used to aid in optimizing metabolic properties for compounds. FAME 3 is the only publicly available predictor covering both phase I and II metabolic reactions with good performance.[1] However, the main barrier to advancing SoM prediction is the need for high-quality experimental data to capture novel atom environments and rare or complex biotransformations.[2]

This study aimed to explore AstraZeneca's proprietary metabolite identification data with publicly available parent compounds for predicting SoMs. After annotating, the newly curated dataset was compared with existing public SoM data to evaluate its coverage of chemical space and atomic environments. In silico SoM prediction models, built using FAME technology, were retrained and assessed against both the newly annotated and existing datasets. We are making the SoM annotations of parent compounds publicly available to support further research. While this is a significant step in SoM data sharing, more diverse and high-quality data are needed to improve prediction models.

References

1. Šícho, M., Stork, C., Mazzolari, A., de Bruyn Kops, C., Pedretti, A., Testa, B., Vistoli, G., Svozil, D., Kirchmair, J. FAME 3: Predicting the Sites of Metabolism in Synthetic Compounds and Natural Products for Phase 1 and Phase 2 Metabolic Enzymes. J. Chem. Inf. Model. **2019**, 59 (8), 3400–3412. DOI: 10.1021/acs.jcim.9b00376.
2. Chen, Y., Seidel, T., Jacob, R. A., Hirte, S., Mazzolari, A., Pedretti, A., Vistoli, G., Langer, T., Miljković, F., Kirchmair, J. Active Learning Approach for Guiding Site-of-Metabolism Measurement and Annotation. J. Chem. Inf. Model. **2024**, 64 (2), 348-358. DOI: 10.1021/acs.jcim.3c01588.

P48: PySSA: end-user protein structure prediction and visual analysis with ColabFold and PyMOL

H. Kullik[1], M. Urban[1], J. Schaub[2], A. Loidl-Stahlhofen[3], A. Zielesny[1]

[1] *Institute for Bioinformatics and Chemoinformatics, Westphalian University of Applied Sciences, Recklinghausen, Germany*

[2] *Institute for Inorganic and Analytical Chemistry, Friedrich Schiller University Jena, Germany*

[3] *Laboratory of Protein Chemistry, Westphalian University of Applied Sciences, Recklinghausen, Germany*

Recent advances in accurate protein folding prediction, such as AlphaFold or ColabFold, have yet to be widely adopted by bench scientists in relevant fields, who are often not familiar with software tools like scripting notebooks that require programming skills.

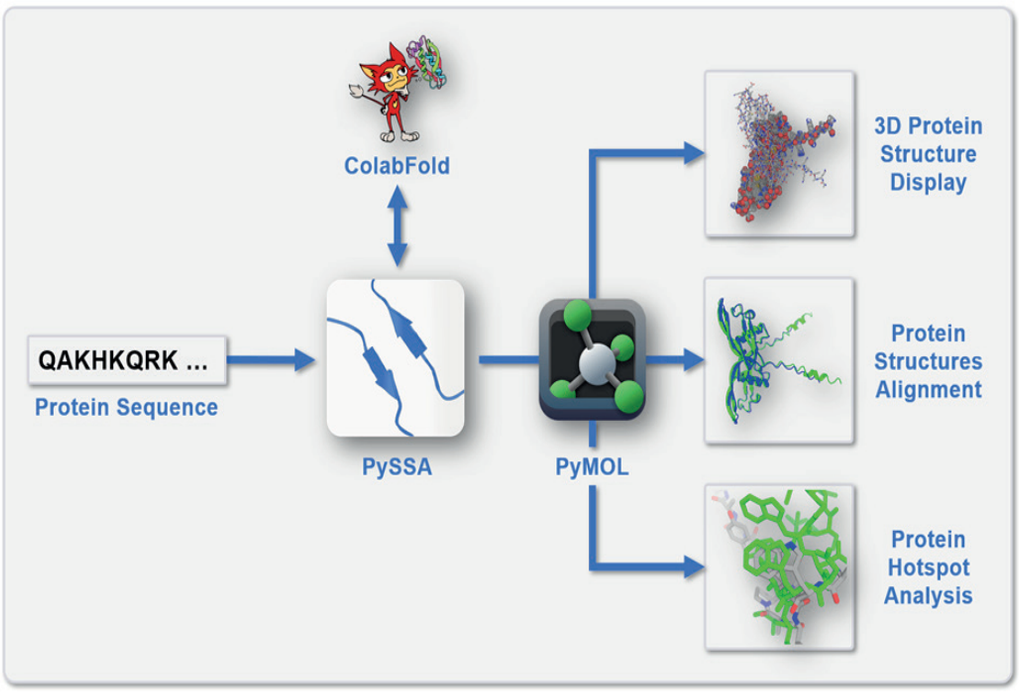


***Figure 1**: PySSA overview*

PySSA (Python rich client for visual protein Sequence to Structure Analysis) [1] combines ColabFold for protein structure prediction with the visualisation and analysis functionalities of PyMOL in a Python-based rich client application (Figure 1). It provides a convenient graphical user interface, especially for end-users without advanced computing expertise. The software project is openly available on GitHub, along with a "single-click" graphical installer executable for the Windows operating system: https://github.com/urban233/PySSA

References

1. [1] H. Kullik, M. Urban, J. Schaub, A. Loidl-Stahlhofen, A. Zielesny, PySSA: end-user protein structure prediction and visual analysis with ColabFold and PyMOL, ChemRxiv, **2024**, https://doi.org/10.26434/chemrxiv-2024-srx5d

P50: REST2-AMP/MM: Integrating Enhanced Sampling with Machine Learning Potentials for Molecular Conformational Sampling

Riccardo Solazzo[1], Igor Gordiy[1], Sereina Riniker[1]

[1]*Department of Chemistry and Applied Biosciences, ETH Zürich, Switzerland*

Numerical potentials from machine learning (ML) have been proposed to replace computationally expensive quantum-mechanical (QM) calculations. Anisotropic message passing (AMP) is a equivariant graph neural network potential that incorporates directional and long-range interactions, making it particularly suitable for QM/MM hybrid schemes [1,2]. While AMP is computationally much cheaper than DFT (at similar accuracy), ML/MM molecular dynamics (MD) simulations can still encounter sampling difficulties. In classical MD, replica exchange with solute scaling (REST2) is a popular sampling-enhancement method, which works by selectively scaling solute interactions while maintaining a fixed bath temperature [3]. This reduces the number of replicas required compared to standard temperature replica exchange and thus increases the computational efficiency. Here, we present the combination of REST2 with AMP, which is made possible because the AMP Hamiltonian – in contrast to QM Hamiltonians – can be decomposed into solute-solute, solvent-solvent, and solute-solvent interaction contributions. Due to this feature, solute interactions in AMP can be selectively scaled by REST2. We validate the REST2-AMP approach on the popular test system alanine dipeptide, achieving good sampling coverage and capturing dihedral transitions within a short simulation time. These results suggest that REST2-AMP enables efficient conformational sampling while providing a more accurate description of biomolecular systems, capturing effects that may be omitted by classical force fields.

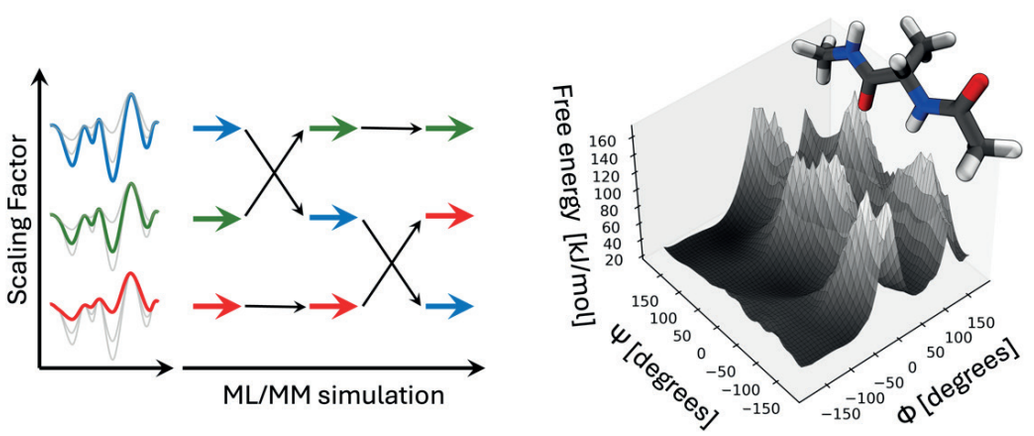


***Figure 1**: Schematic view of REST2-AMP/MM and predicted free energy surface of alanine dipeptide.*

References

1. Thürlemann, M., et al., Anisotropic Message Passing: Graph Neural Networks with Directional and Long-Range Interactions. Proceedings of the International Conference on Learning Representations (ICLR), **2023**.
2. Pultar, F., et al., Neural Network Potential with Multiresolution Approach Enables Accurate Prediction of Reaction Free Energies in Solution. Journal of the American Chemical Society, **2025**, Article ASAP. DOI: 10.1021/jacs.4c17015
3. Wang, L., et al., Replica Exchange with Solute Scaling: A More Efficient Version of Replica Exchange with Solute Tempering (REST2). The Journal of Physical Chemistry B, **2011,** 115 (30), 9431-9438, DOI: 10.1021/jp204407d

P52: Cooperative Free energy: Induced PPI, Solvation, and Conformation in Protein-Ligand-Protein Ternary Complexation

Shu-Yu Chen[1], Riccardo Solazzo[1], Marianne Fouche[2], Hans-Joerg Roth[2], Birger Dittrich[2], Sereina Riniker[1]

[1] *Department of Chemistry and Applied Biosciences, ETH Zurich, Switzerland*

[2] *Novartis Biomedical Research, Basel, Switzerland*

Cooperativity in the protein-ligand-protein ternary complex regulates the complexation efficiency and selectivity. In terms of free energy, cooperativity can be decomposed into three three-body contributions. The gas-phase entropy part accounts for the correlation between the protein-ligand interactions, the solvation-free energy part accounts for the solvent-shielding effect at the three-body interface, and the conformational free energy part accounts for the induced protein-protein interactions and the conformational penalty to bring the ligand, ligand-protein dimers, and protein-ligand-protein trimer to the shared conformational space. Under the theoretical framework, molecular dynamics simulations can be implemented to predict the cooperativity of the system of interest. The study discusses the thermodynamics of the cooperative process and is anticipated to facilitate the design of cooperative molecular glues.

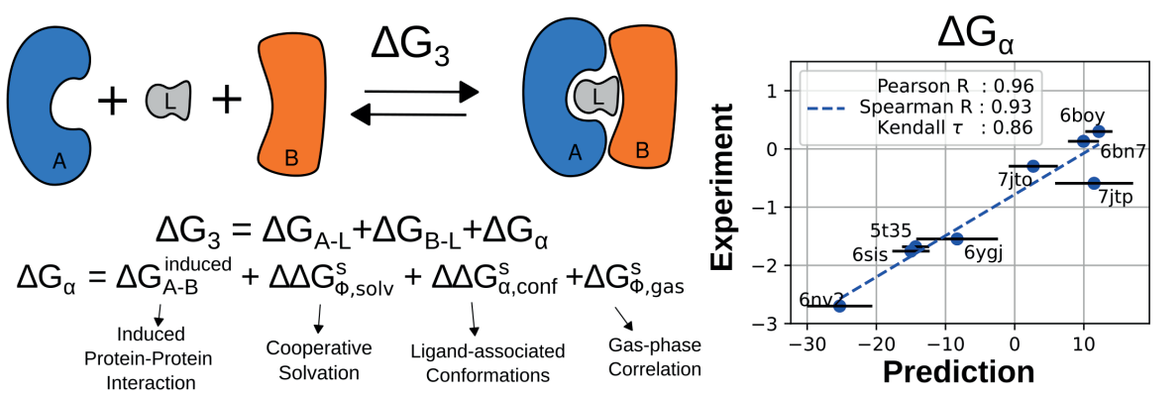


***Figure 1**: Components of Cooperative free energy ΔG_{α} and the correlation between computed and experimentally measured cooperativity. ΔG_{3} ,$\Delta G_{A\text{-}L}$,and $\Delta G_{B\text{-}L}$ are the effective three-body association free energy and ligand-proteins (A and B) binding free energy, respectively.*

P54: A Feature-Engineered Delta-ML Approach for Molecular Structure Refinement: Bridging Exploration and Exploitation in Computational Chemistry

Federico Lazzari

Scuola Superiore Meridionale, Napoli, Italy

In computational chemistry, achieving a balance between efficiency and accuracy is crucial—especially in high-resolution spectroscopy, which is highly sensitive to molecular geometry. While linear-scaling methods exist for energy and property calculations, accurate geometries remain a bottleneck. Yet, reliable energy and property evaluations require precise structures. Thermochemistry benchmarks energetic data, but high-resolution microwave spectroscopy offers the most accurate geometrical information, as rotational constants depend solely on geometry. Achieving bond length errors within 0.1 pm and angle errors within 0.1° is necessary for spectroscopic accuracy (~0.1% on rotational constants, ~0.3% on vibrational frequencies)[1,2], a level that current workflows often fail to reach. Structure prediction typically involves: (1) exploration—using DFT or semi-empirical methods to sample PESs[1] and identify conformers efficiently—and (2) exploitation—for high-accuracy refinement and experimental validation. While ML aids exploration, exploitation remains difficult since ML cannot surpass the accuracy of its training data. DL models, in particular, rely on large but often approximate datasets, limiting generalization. Feature-engineered ML with chemically meaningful descriptors and Δ-ML[3] strategies offers a promising alternative, improving interpretability and reducing data needs.

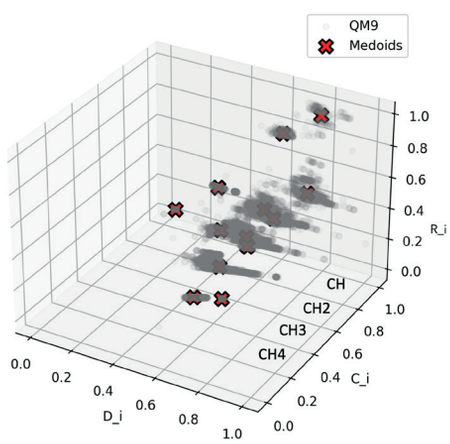


***Figure 1**: The QM9[4] feature space of carbon atoms in hydrocarbons and their 13 k-medoids clusters associated with atom types.*

In this work, we introduce an atomic feature space based on three heuristic and continuous descriptors of the local environment of an atom in a molecule—delocalization, coordination, and rigidity—which naturally cluster atoms into categories like the discrete atom types employed in the force fields underlying molecular mechanics simulations[5,6], as can be seen in Fig.1. This representation is applied to refine DFT equilibrium molecular structures for improved rotational spectroscopy parameters, leveraging a templating strategy where bonds of initial structures are matched to a set of 141 high-accuracy reference geometries, optimized by a state-of-the-art quantum chemical method (the Pisa composite scheme (PCS2) [1]), using a Tanimoto[7] similarity metric. Then, a Δ correction is applied, systematically improving bond lengths and approaching spectroscopic accuracy. The proposed templating approach bridges multi-scale modeling and enhances molecular simulations without the high cost of ab initio methods. Beyond spectroscopy, it applies to structural predictions in fields like materials science and drug discovery, where hybrid physics-informed ML models are increasingly relevant. By embedding chemical insight into the feature space, this work improves quantum mechanical predictions beyond baseline accuracy and provides a roadmap for integrating machine learning while preserving physical consistency in computational sciences.

References

1. Barone, V., et al., *WIREs Comput. Mol. Sci.*, **2025**, *15* (1), e70000.
2. Lazzari, F., et al., *J. Phys. Chem. A*, **2025**, 129 (2), 503–517.
3. Nandi, A., et al., *J. Chem. Phys.,* **2021**, 154 (5), 051102.
4. Ramakrishnan, R., et al., *Sci. Data*, **2014**, *1* (1), 140022.
5. Lazzari, F., et al., *J. Phys. Chem. A*, **2024**, 128 (7), 1385–1395.
6. Lazzari, F., et al., *J. Chem. Inf. Model.*, **2020**, 60 (6), 2668–2672.
7. Bajusz, D., et al., *J. Cheminform.*, **2015**, 7 (1), 20.

P56: Machine Learning Predictions of the Protein-Ligand Binding Affinity with Fingerprints, Shape and Electrostatics

K. Stanciakova[1], M. Krier[1], L. Eberlein[1], G. Stahl[1], J. Chen[2], S. Mandal[2], S. Nath[2], M. Geballe[2]

[1] *OpenEye, Cadence Molecular Sciences, Cologne Germany*

[2] *OpenEye, Cadence Molecular Sciences, Santa Fe, United States*

Fast and accurate prediction of binding affinities is one of the most sought capabilities in drug discovery. Here, we report two different machine learning (ML) based approaches for building Quantitative Structure-Activity Relationship models (QSAR models). The first approach builds regression models for affinity prediction using neural networks (NN) along with grid search for hyperparameters and 2D fingerprint-based feature vector. The second approach - 3D-QSAR methodology –leverages information about 3D structure of molecules and predicts binding affinity using shape and electrostatics (ROCS and EON) based similarities as descriptors in combination with two distinct ML approaches (Figure 1). Both methods were validated against industrially relevant PDE10A dataset [1] with a number of training-test splits that reflect different lead optimization stages.

Our results demonstrate that both 2D based NN models and 3D-QSAR perform comparably to or even surpass the performance of other methods from literature, including 2D- and 3D-ML methods as well as empirical scoring functions for predicting binding affinities. Compared to 2D based methods, 3D QSAR shows a stable performance across different splits, ensuring thus reliability of predictions.

Moreover, both 2D based NN models and 3D-QSAR provide built-in interpretable model which offers a visual insight into the predicted results at atomic or fragment level. Such interpretations can be used in generating new ideas, and they allow the model to be used as a potential generative design tool in the drug discovery cycle.

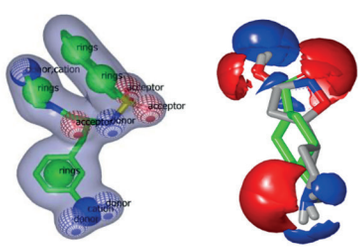


Figure 1. ROCS (left) and EON (right).

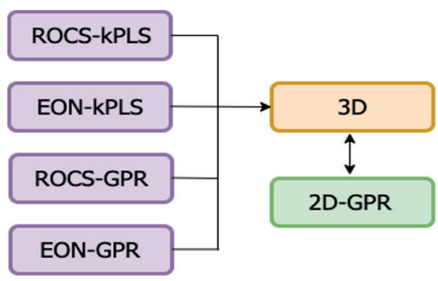


Figure 2. Illustration of models built in this work.

Figure 1*: The 3D-QSAR model is built as a composition of multiple models that combines orthogonal sets of 3D similarity descriptors (left) and machine learning techniques (right).*

References

1. TOSSTORFF, A., et al., A high quality, industrial data set for binding affinity prediction: performance comparison in different early drug discovery scenarios. J. Comput. Aided Mol. Des., **2022**, 36, 10, 753-765, DOI: 10.1007/s10822-022-00478-x.

P58: Balancing Data Quantity and Quality: Evaluating Curation Strategies for Bioactivity Prediction Models

Carl C.G. Schiebroek, Gregory A. Landrum, and Sereina Riniker

Department of Chemistry and Applied Biosciences, ETH Zurich, Switzerland

Building good models to predict the bioactivity of novel chemical series remains a challenging task. Accurate models require a training set with a large number of diverse samples and a low level of noise. When extracting data from public databases such as ChEMBL, different levels of curation rigor may be applied, resulting in training sets of varying size, diversity, and, presumably, noise levels. It is not possible to know *a priori* whether increasing the size of the dataset at the cost of adding more noise improves model generalization. To assess this trade-off, we compare three data curation and modeling approaches: (1) models trained on data for a single target, (2) models trained on target-specific data further restricted to a single set of assay conditions, and (3) multitask models where each assay condition is treated as a separate task. We evaluate these models using leave-assay-out, a surrogate for leave-chemical-series-out, with graph neural networks (GNNs) and random forests (RFs). We find no meaningful differences between these curation and modeling strategies for either model type, suggesting that adding more data at the expense of increased noise does not improve generalizability. Surprisingly, the GNN models show more variance between different random seeds than between different modeling approaches, highlighting the variability of these models and the importance of evaluating them using multiple seeds.

References

1. Landrum, G. A., Riniker, S., Combining IC50 or K i Values from Different Sources Is a Source of Significant Noise. Journal of Chemical Information and Modeling., **2024**, 64, 5, 1560–1567, 10.1021/acs.jcim.4c00049
2. Zdrazil, M., *et al.*, The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. Nucleic acids research., **2024**, 52, D1, D1180-1192, 10.1093/nar/gkad1004

P60: PROTAC-Splitter: An AI-Based System to Automatically Identify PROTAC Ligands

Stefano Ribes[1], Anders Källberg[1], Ranxuan Zhang[1], Eva Nittinger[2], Christian Tyrchan[2], and Rocío Mercado[1]

[1] Department of Computer Science and Engineering, Section for Data Science and AI, Chalmers University of Technology and University of Gothenburg, Sweden

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

PROTACs (PROteolysis TArgeting Chimeras) are a novel therapeutic modality that leverages the cellular degradation machinery by linking together a selected E3 ligase and a target protein, forming a ternary complex. To achieve that, PROTACs are designed with three main parts: an E3 ligase binder, a warhead that binds to the target protein, and a linker that connects the two. More detailed analysis and the impact of changes of E3 binder, linkers, or the POI warhead currently demand manual curation and definition of substructure matches, which quickly becomes unfeasible when applying to cross project PROTACs data with many alterations.

In this work, we developed the PROTAC-Splitter, a novel Transformer-based generative AI model, to automatically predict the substructures of a PROTAC given its SMILES representation. Due to the limited amount of publicly available PROTAC data (around 5K PROTAC structures) [1, 2], we release a synthetic data set of 1M new PROTACs generated by recombining different E3 ligands, linkers, and warheads of known open-source PROTACs. In doing so, we mimic the open source data distribution of functional groups at the substructure attachment points. The synthetic data set is then used to train the PROTAC-Splitter.

We evaluated our implementation on public and internal AstraZeneca data: the model can generate SMILES substructures that are perfectly matching the test labels in 94.28% of the open source samples. Additionally, we show that for 99.38% of the open data, the PROTAC-Splitter returns fragments that reassemble into the input PROTAC. The label-matching and reassembly scores increase to 99.52% and 99.93%, respectively, when considering the top-5 predictions. Our results show that the PROTAC-Splitter model can accurately predict the substructures of a PROTAC, providing a valuable tool for PROTAC design and development.

References

1. London, N., et al., “PROTACpedia”. https://protacpedia.weizmann.ac.il/, Accessed: **2025/01/21**.
2. Weng, G., et al. “PROTAC-DB 2.0: an updated database of PROTACs.” Nucleic acids research vol. 51,D1 (**2023**): D1367-D1372., doi:10.1093/nar/gkac946

P62: The predicTeam's All-Inclusive Strategy for Leveraging the Full Potential of ADMET Predictions in Drug Discovery

Lara Kuhnke, Uschi Dolfus

Bayer AG, Berlin, Germany

The predicTeam is a self-organized, interdisciplinary team that covers all steps in machine learning from endpoint selection, data collection and preparation, model building, retraining and deployment to user interaction and applicability assessment. Providing expert knowledge in all these fields within one team is key for the successful application of the provided prediction models in a plethora of use cases.

The team transforms experimental data into predictive models to create insights that feed back into experiments and thus supports drug discovery projects with state-of-the-art machine learning models for relevant pharmacological and ~40 ADMET (absorption, distribution, metabolism, excretion, toxicology) endpoints. Thanks to the close interaction and partnership with the experimentalists, optimal model application and trust within the scientific community are facilitated for the identification of new endpoints as well as the application of the existing prediction models. The seamless implementation of the prediction models into the internal drug design platform as Compute Tools ensures their widely applied utilization in various phases of drug discovery projects and de-novo design applications. A robust and standardized infrastructure for retrieving, preparing, and providing reliable data for (re-)training forms the basis for prediction models that serve the drug discovery projects with models covering their evolving chemical spaces. By fostering method development and scientific collaborations, the team also contributes to innovation in the field of machine learning and QSAR/QSPR (quantitative structure-activity/property relationships).

The predicTeam delivers the all-inclusive package for the prediction of ADMET properties of compounds in drug design. This holistic approach within one team ensures high quality in endpoint selection, data preparation, predictive modelling, model provision and user interaction. All these aspects combined enable the full leverage of the potential of machine learning for drug discovery.

References

1. Göller, Andreas & Kuhnke, Lara & Montanari, Floriane & Bonin, Anne & Schneckener, Sebastian & ter Laak, Antonius & Wichard, Joerg & Lobell, Mario & Hillisch, Alexander. (**2020**). Bayer's in silico ADMET platform: a journey of machine learning over the past two decades. Drug Discovery Today. 25. 10.1016/j.drudis.2020.07.001.

P64: Molecular deep learning at the edge of chemical space

Derek van Tilborg[1,2], Luke Rossen[1], and Francesca Grisoni[1,2*]

[1] *Institute for Complex Molecular Systems (ICMS), Department of Biomedical Engineering, Eindhoven University of Technology, The Netherlands*

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

Molecular deep learning models[1–5] often struggle to generalize beyond the chemical space of their training data, limiting their reliability when confronted with structurally novel molecules[6–8] (Fig. 1a). To address this challenge, we propose a joint modeling approach (Fig. 1b) that integrates molecular property prediction with molecular reconstruction. Our approach breaks with the well-established application of self-supervised learning for generative chemistry[9] or performance improvement[10–14], by using reconstruction capabilities as a direct proxy for out-of-distribution (OOD) estimation. Specifically, we hypothesize that poorly reconstructed molecules are less familiar to the model, indicating that they fall outside the distribution learned from the training data[15–18]. Building on this hypothesis, we introduce a new metric, termed *unfamiliarity* ($\mathbb{U}$), which is derived from a model's reconstruction ability and aims to quantifying how far a molecule deviates from the training distribution.

We conducted a systematic study using 33 experimentally labeled datasets[19–21] that were split into in-distribution and out-of-distribution molecules (Fig. 1c). After confirming the presence of molecular distribution shifts (Fig. 2a-d), we demonstrate that unfamiliarity robustly quantifies these distribution shifts (Fig. 2e-h). Additionally, unfamiliarity correlates strongly with model performance (Table 1), even though it is independent from the prediction uncertainty metric. Finally, we show that unfamiliarity performs comparably to well-established methods of prediction reliability estimation in a virtual screening study (Fig. 3).

Ultimately, the introduced concept of molecular *unfamiliarity* provides a principled approach to estimating model generalizability, even in the presence of molecular distribution shifts. Our approach offers a different perspective on estimating prediction reliability, complementing established concepts such as the applicability domain[22,23] and uncertainty estimation[24,25] – ultimately guiding the discovery of structurally novel molecules in a more precise and informed manner.

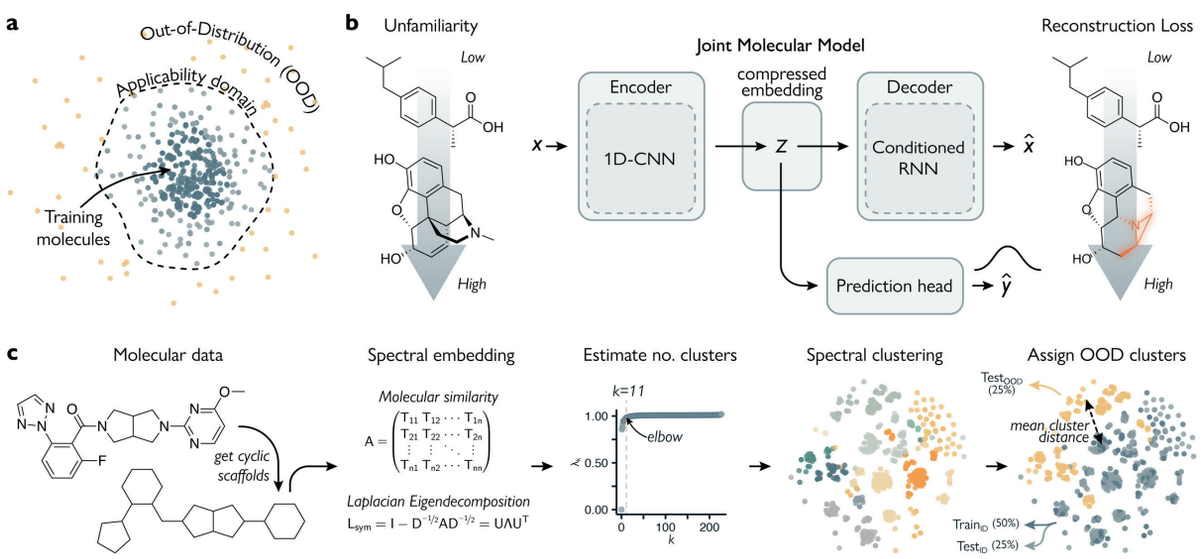


Figure 1 | Estimating unfamiliarity of molecular data using joint modelling. a. *Conceptual representation of the applicability domain. Molecules close to the training data in chemical space are within a models' applicability domain. Molecules outside of this boundary are considered out-of-distribution (OOD).* ***b.*** *The architecture of the Joint Molecular Model (JMM) estimates how 'unfamiliar' a molecule is to the model through its reconstruction loss.* ***c.*** *Inducing molecular distribution shifts by separating molecular data into in-distribution and out-of-distribution groups through spectral clustering. Results for the OX2R dataset are shown.*

	Rank-based correlation to bin order		
Binning metric	**Balanced accuracy**	**Hit rate**	**Precision**
Scaffold sim	0.42±0.06	0.51±0.04	0.50±0.06
Mol core overlap	0.28±0.07	0.22±0.09	0.25±0.07
Pharmacophore sim	0.19±0.07	0.37±0.09	0.43±0.08
Embedding distance	0.36±0.06	0.24±0.09	0.29±0.08
Uncertainty	0.51±0.08	**0.62±0.06**	**0.72±0.04**
Unfamiliarity	**0.58±0.04**	0.52±0.07	0.52±0.05

Table 1 | Correlation of reliability metrics to model performance

Kendall correlation between several bin-wise performance metrics and the bin order. Molecules are binned into ten bins by: mean pharmacophore similarity to the training set (cosine distance of CATS descriptors), mean scaffold (Tanimoto) similarity to the training set, mean molecular core overlap (MCS fraction) to the training set, Mahalanobis distance of embeddings (z-vectors) to the training set, prediction uncertainty, and unfamiliarity. Mean and SEM for all datasets are reported. A correlation of 1.0 indicates perfect model calibration. For every metric, bins are ordered to reflect low to high confidence. Highest correlations are reported in bold.

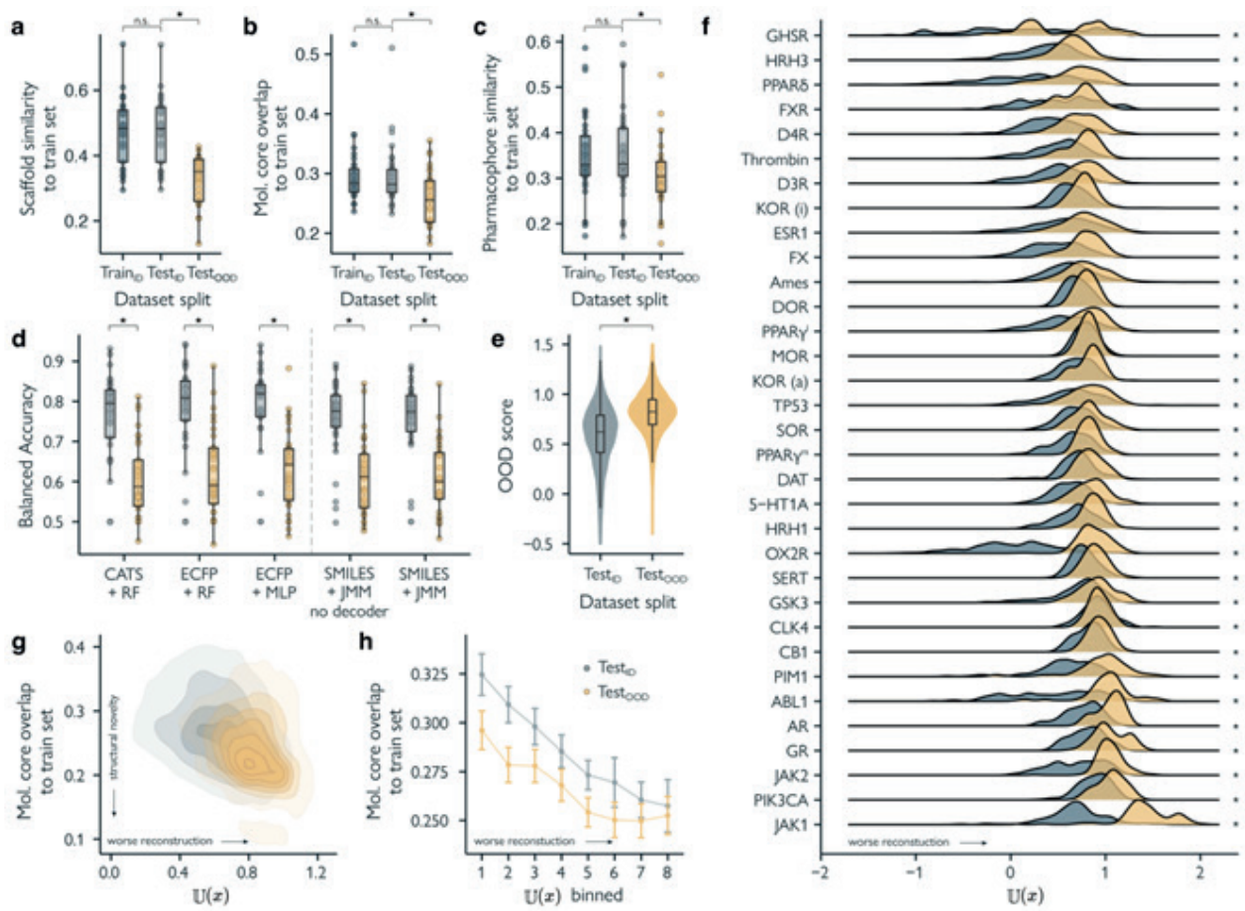


Figure 2 | Detecting induced molecular distribution shifts with the unfamiliarity score. a. *Mean scaffold similarity of data splits in the labelled data sets to their train set. Similarity is calculated as the Tanimoto coefficient between ECFPs of Bemis-Murcko scaffolds. Every point in the box plot represents a dataset.* ***b.*** *Mean Maximal Common Substructure Fraction (MCSF) between data splits in the labelled sets to their train set.* ***c.*** *Mean pharmacophore similarity (CATS12) between data splits in the labelled sets to their train set.* ***d.*** *Predictive performance on the finetuning sets. From left to right, random forest models using CATS descriptors, random forest using ECFPs, Multi-layer Perceptron (MLP) using ECFPs, MLP using a SMILES string encoder, Joint Molecular Model (JMM) using the same SMILES string encoder.* ***e.*** *Distribution of the JMM's out-of-distribution score for all molecules in testID and testOOD in all labelled datasets.* ***f.*** *Distributions of the JMM's*

*out-of-distribution score in testID and testOOD per datasets. Statistically significant differences ($p < 0.05$) in **a-f** are denoted as *, determined by paired, two-sided, Wilcoxon signed-rank tests (**a-d**) and two-sided Kolmogorov-Smirnov tests (**e, f**). **g.** Relationship between the JMM's out-of-distribution score and the mean MCSF similarity of all molecules in the labelled datasets to their respective train set. **h.** Relationship between binned unfamiliarity and MCSF similarities to the respective train set. Points represent the mean over all datasets, error bars represent the standard error.*

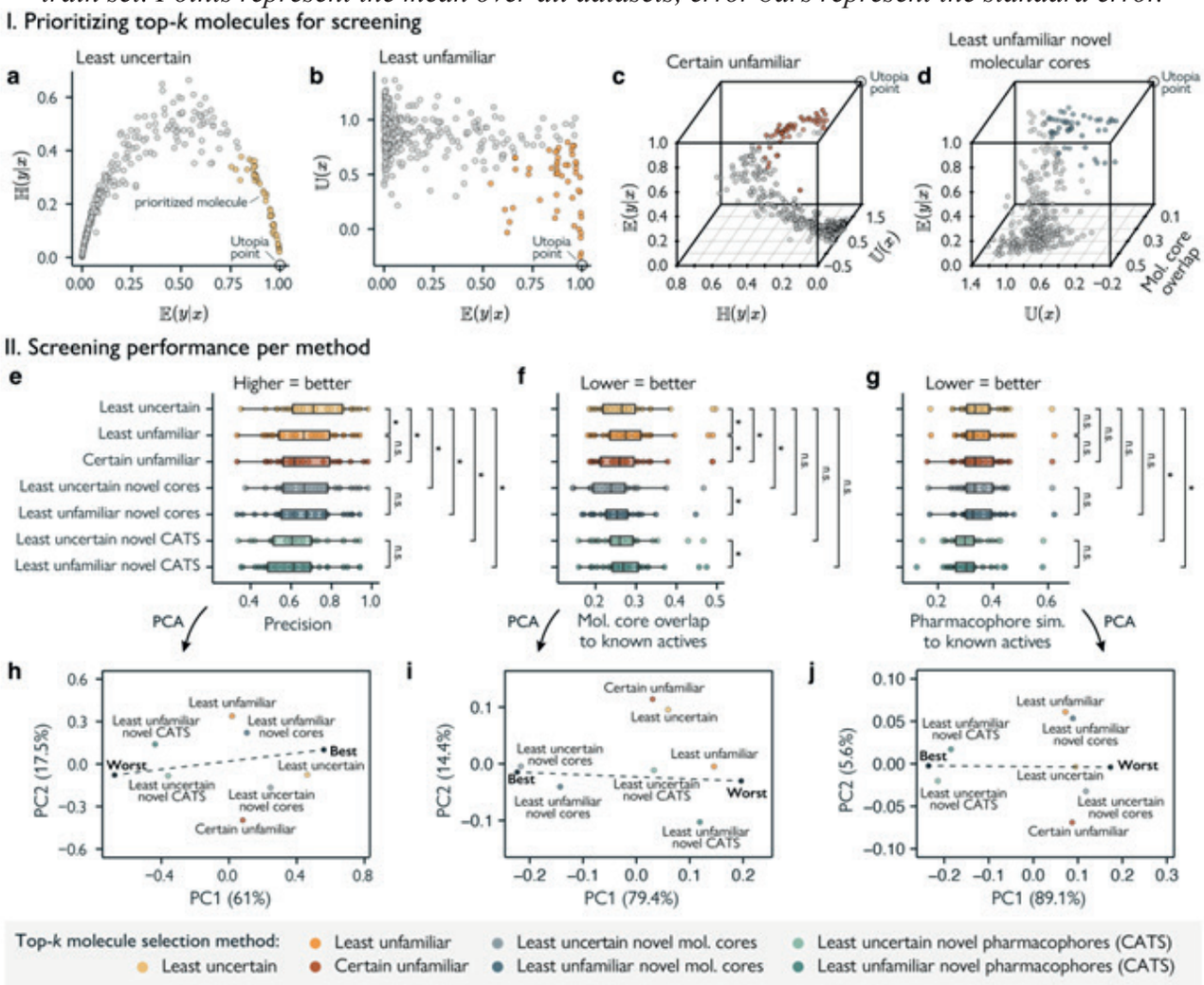


Figure 3 | Using reliability metrics to perform virtual hit screening. *Top-k (k=50) molecules from 33 out-of-distribution test set are prioritized for screening/labelling using different selection methods. **a.** Selecting molecules that have high bioactivity predictions and low prediction uncertainty scores. **b.** Selecting molecules that have high bioactivity predictions and low unfamiliarity. **c.** Selecting molecules that have high bioactivity predictions, low prediction uncertainty, and high unfamiliarity scores. **d.** Selecting molecules that have high bioactivity prediction, low unfamiliarity, and low molecular core overlap with the training data (i.e., high structural novelty). **e.** Precision on the top-k molecules. **f.** Mean molecular core overlap of discovered hits in the top-k molecules with known bioactive molecules in the training set. **g.** Pharmacophore similarity of discovered hits in the top-k molecules with known bioactive molecules in the training set. Statistically significant differences ($p < 0.05$) in **e-g** are denoted as *, determined by paired, two-sided, Wilcoxon signed-rank tests. **h-j**. Global ranking of all methods in **e-g**, respectively, using anchored PCA, scaled between best and worst performance across all data sets. First two principal components, PC1 and PC2, are shown with their explained variance (%). Distance orthogonal to the best-worst line resembles high variance of a specific method.*

References

1. Gawehn, E.; Hiss, J. A.; Schneider, G. Deep Learning in Drug Discovery. *Molecular Informatics* **2016**, *35* (1), 3–14. https://doi.org/10.1002/minf.201501008.
2. Jiménez-Luna, J.; Grisoni, F.; Weskamp, N.; Schneider, G. Artificial Intelligence in Drug Discovery: Recent Advances and Future Perspectives. *Expert Opinion on Drug Discovery* **2021**, *16* (9), 949–959. https://doi.org/10.1080/17460441.2021.1909567.
3. Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; Zhao, S. Applications of Machine Learning in Drug Discovery and Development. *Nat Rev Drug Discov* **2019**, *18* (6), 463–477. https://doi.org/10.1038/s41573-019-0024-5.
4. Stokes, J. M.; Yang, K.; Swanson, K.; Jin, W.; Cubillos-Ruiz, A.; Donghia, N. M.; MacNair, C. R.; French, S.; Carfrae, L. A.; Bloom-Ackermann, Z.; Tran, V. M.; Chiappino-Pepe, A.; Badran, A. H.; Andrews, I. W.; Chory, E. J.; Church, G. M.; Brown, E. D.; Jaakkola, T. S.; Barzilay, R.; Collins, J. J. A Deep Learning Approach to Antibiotic Discovery. *Cell* **2020**, *180* (4), 688-702.e13. https://doi.org/10.1016/j.cell.2020.01.021.
5. Liu, G.; Catacutan, D. B.; Rathod, K.; Swanson, K.; Jin, W.; Mohammed, J. C.; Chiappino-Pepe, A.; Syed, S. A.; Fragis, M.; Rachwalski, K.; Magolan, J.; Surette, M. G.; Coombes, B. K.; Jaakkola, T.; Barzilay, R.; Collins, J. J.; Stokes, J. M. Deep Learning-Guided Discovery of an Antibiotic Targeting Acinetobacter Baumannii. *Nat Chem Biol* **2023**. https://doi.org/10.1038/s41589-023-01349-8.
6. Ji, Y.; Zhang, L.; Wu, J.; Wu, B.; Huang, L.-K.; Xu, T.; Rong, Y.; Li, L.; Ren, J.; Xue, D.; Lai, H.; Xu, S.; Feng, J.; Liu, W.; Luo, P.; Zhou, S.; Huang, J.; Zhao, P.; Bian, Y. DrugOOD: Out-of-Distribution (OOD) Dataset Curator and Benchmark for AI-Aided Drug Discovery -- A Focus on Affinity Prediction Problems with Noise Annotations. arXiv 2022. https://doi.org/10.48550/ARXIV.2201.09637.
7. Dias, A. L.; Bustillo, L.; Rodrigues, T. Limitations of Representation Learning in Small Molecule Property Prediction. *Nat Commun* **2023**, *14* (1), 6394. https://doi.org/10.1038/s41467-023-41967-3.
8. Tossou, P.; Wognum, C.; Craig, M.; Mary, H.; Noutahi, E. Real-World Molecular Out-Of-Distribution: Specification and Investigation. *J. Chem. Inf. Model.* **2024**, *64* (3), 697–711. https://doi.org/10.1021/acs.jcim.3c01774.
9. Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4* (2), 268–276. https://doi.org/10.1021/acscentsci.7b00572.
10. Doersch, C.; Zisserman, A. Multi-Task Self-Supervised Visual Learning. In *2017 IEEE International Conference on Computer Vision (ICCV)*; IEEE: Venice, **2017**; pp 2070–2079. https://doi.org/10.1109/ICCV.2017.226.
11. Assran, M.; Duval, Q.; Misra, I.; Bojanowski, P.; Vincent, P.; Rabbat, M.; LeCun, Y.; Ballas, N. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Vancouver, BC, Canada, **2023**; pp 15619–15629. https://doi.org/10.1109/CVPR52729.2023.01499.
12. Kim, D.; Yoo, Y.; Park, S.; Kim, J.; Lee, J. SelfReg: Self-Supervised Contrastive Regularization for Domain Generalization. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: Montreal, QC, Canada, **2021**; pp 9599–9608. https://doi.org/10.1109/ICCV48922.2021.00948.

13. Sun, Y.; Wang, X.; Liu, Z.; Miller, J.; Efros, A.; Hardt, M. Test-Time Training with Self-Supervision for Generalization under Distribution Shifts. In *Proceedings of the 37th International Conference on Machine Learning*; PMLR, **2020**; pp 9229–9248.
14. Albuquerque, I.; Naik, N.; Li, J.; Keskar, N.; Socher, R. Improving Out-of-Distribution Generalization via Multi-Task Self-Supervised Pretraining. arXiv **2020**. https://doi.org/10.48550/ARXIV.2003.13525.
15. (15) Pimentel, M. A. F.; Clifton, D. A.; Clifton, L.; Tarassenko, L. A Review of Novelty Detection. *Signal Processing* **2014**, *99*, 215–249. https://doi.org/10.1016/j.sigpro.2013.12.026.
16. Zhou, Y. Rethinking Reconstruction Autoencoder-Based Out-of-Distribution Detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New Orleans, LA, USA, **2022**; pp 7369–7377. https://doi.org/10.1109/CVPR52688.2022.00723.
17. Chalapathy, R.; Chawla, S. Deep Learning for Anomaly Detection: A Survey. arXiv **2019**. https://doi.org/10.48550/ARXIV.1901.03407.
18. Ruff, L.; Vandermeulen, R. A.; Görnitz, N.; Binder, A.; Müller, E.; Müller, K.-R.; Kloft, M. Deep Semi-Supervised Anomaly Detection. arXiv **2019**. https://doi.org/10.48550/ARXIV.1906.02694.
19. Van Tilborg, D.; Alenicheva, A.; Grisoni, F. Exposing the Limitations of Molecular Machine Learning with Activity Cliffs. *J. Chem. Inf. Model.* **2022**, *62* (23), 5938–5951. https://doi.org/10.1021/acs.jcim.2c01073.
20. Tran-Nguyen, V.-K.; Jacquemard, C.; Rognan, D. LIT-PCBA: An Unbiased Data Set for Machine Learning and Virtual Screening. *J. Chem. Inf. Model.* **2020**, *60* (9), 4263–4273. https://doi.org/10.1021/acs.jcim.0c00155.
21. Hansen, K.; Mika, S.; Schroeter, T.; Sutter, A.; Ter Laak, A.; Steger-Hartmann, T.; Heinrich, N.; Müller, K.-R. Benchmark Data Set for in Silico Prediction of Ames Mutagenicity. *J. Chem. Inf. Model.* **2009**, *49* (9), 2077–2081. https://doi.org/10.1021/ci900161g.
22. Hanser, T.; Barber, C.; Marchaland, J. F.; Werner, S. Applicability Domain: Towards a More Formal Definition. *SAR and QSAR in Environmental Research* **2016**, *27* (11), 865–881. https://doi.org/10.1080/1062936X.2016.1250229.
23. Mathea, M.; Klingspohn, W.; Baumann, K. Chemoinformatic Classification Methods and Their Applicability Domain. *Molecular Informatics* **2016**, *35* (5), 160–180. https://doi.org/10.1002/minf.201501019.
24. Wang, D.; Wu, Z.; Shen, C.; Bao, L.; Luo, H.; Wang, Z.; Yao, H.; Kong, D.-X.; Luo, C.; Hou, T. Learning with Uncertainty to Accelerate the Discovery of Histone Lysine-Specific Demethylase 1A (KDM1A/LSD1) Inhibitors. *Briefings in Bioinformatics* **2023**, *24* (1), bbac592. https://doi.org/10.1093/bib/bbac592.
25. Soleimany, A. P.; Amini, A.; Goldman, S.; Rus, D.; Bhatia, S. N.; Coley, C. W. Evidential Deep Learning for Guided Molecular Property Prediction and Discovery. ACS Cent. Sci. **2021**, 7 (8), 1356–1367. https://doi.org/10.1021/acscentsci.1c00546.

P66: Integrating Structural and Morphological Fingerprints: Understanding Information for Pattern Identification and Better Toxicity Prediction

Floriane Odje[1], Elena von Coburg[2], Christopher Wolff[3], Jens Peter von Kries[3], Sebastian Dunst[2], Andrea Volkamer[1]

[1] *Data Driven Drug Design, Universität des Saarlandes, Saarbrücken, Germany*

[2] *German Federal Institute for Risk Assessment (BfR), Berlin, Germany*

[3] *Institute for Molecular Pharmacology, Berlin, Germany*

Molecular property prediction is essential for identifying bioactive molecules and deprioritizing harmful ones. Traditional models rely on structural fingerprints (SFPs), assuming that similar structures yield similar biological responses. However, this approach may overlook downstream phenotypic effects. Morphological fingerprints (MFPs), derived from Cell Painting assays, complement SFPs by capturing cellular phenotypes [1]. In the BMBF-funded MORPHEUS project (MORPHology-based Endocrine Disruptor Screening), we integrate SFPs and MFPs to predict endocrine activity and understand hormone-like compound patterns. Existing integration strategies—either combining input features or merging prediction outputs—enhance performance but lack insight into modality complementarity [2,3].

We introduce a two-level clustering approach: first by structural similarity, then by morphological profiles. This reveals the relationship between chemical structure and phenotypic response (Figure 1). Applied to publicly available Cell Image Library data [4] and in-house MCF-7 profiles, our method uncovers key mechanistic insights. Some structurally similar compounds cluster morphologically, while others diverge, indicating differing cellular responses despite similar scaffolds. Using maximum common substructures (MCS) and maximum common feature profiles (MCP), we dissect whether effects are driven by core structure or specific functional groups.

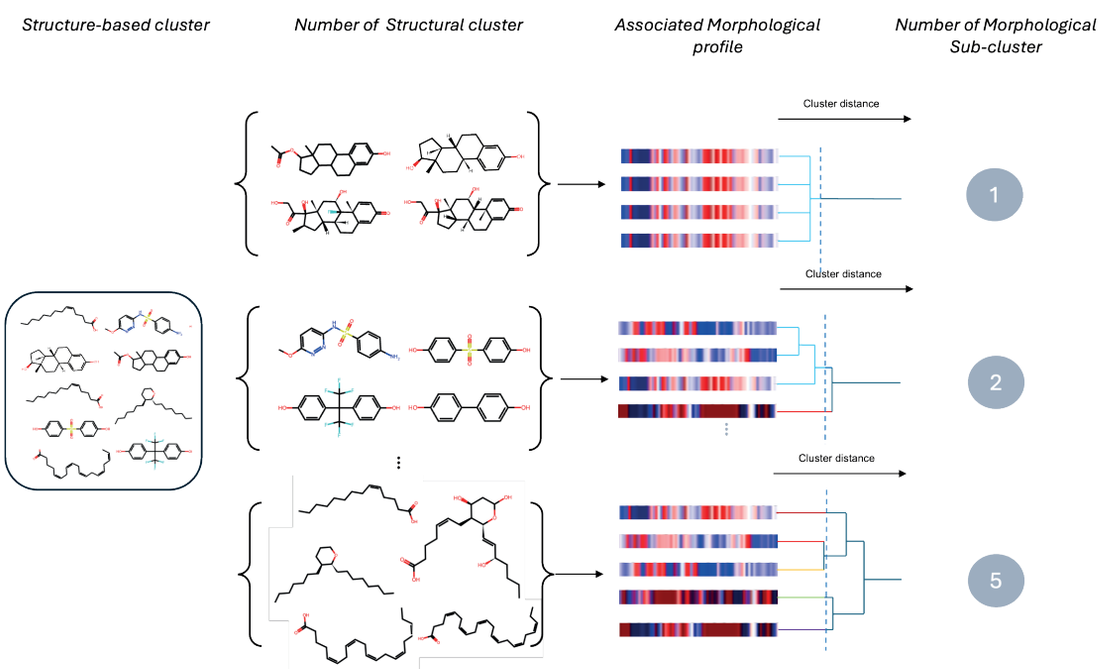


***Figure 1**: Clustering of molecules based on structure and second step clustering based on associated morphological profile*

Our results show that structural similarity does not always align with phenotypic similarity, reinforcing the need for integrated analyses. This approach deepens mechanistic understanding and improves toxicity prediction.

References

1. Cimini, B.A., et al., Optimizing the Cell Painting assay for image-based profiling. Nat. Protoc., **2023**, 18(7), 1981-2013, DOI: 10.1038/s41596-023-00840-9.
2. Chandrasekaran, S. N., et al., JUMP Cell Painting dataset: morphological impact of 136,000 chemical and genetic perturbations. **2023**, doi:10.1101/2023.03.23.534023.
3. Odje, F., et al., Unleashing the potential of cell painting assays for compound activities and hazards prediction. Front. Toxicol., **2024**, 6, Article 1401036, DOI: 10.3389/ftox.2024.1401036.
4. Shamji, A. F., et al., A dataset of images and morphological profiles of 30,000 small-molecule treatments using the Cell Painting assay. GigaScience Database, **2017**, https://doi.org/10.5524/100351.
5. Skuta, C., et al., *Probes & Drugs portal.* **2017**. Available at: https://www.probes-drugs.org/compounds/standardized#compoundset=353@AND

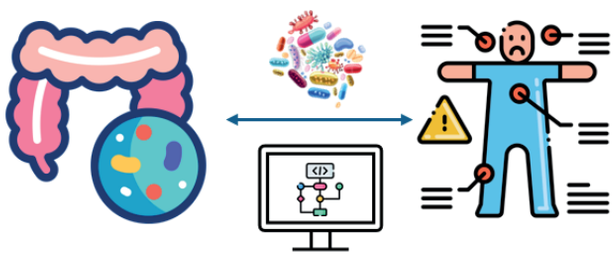
P68: Gut Microbiota Metabolic Mimicking Drugs for Autoimmune/ Infectious Diseases

Shayma El-Atawneh, Oliver Koch

Institute of Pharmaceutical and Medicinal Chemistry, Universität Münster, Germany

The human microbiota, particularly the gut microbiome, plays a pivotal role in maintaining host physiology and preventing disease, ranging from metabolic, cardiovascular, gastrointestinal, neurodegenerative diseases to cancer [1]. Recent studies have highlighted the significance of microbiota-host interactions in the development of infectious and autoimmune disorders [2]. The gut microbiome produces metabolites that induce a series of physiological and pathological functions on hosts and other bacteria, such as modulation of energy metabolism, nutrition absorption, and regulation of gut microbiota composition. These metabolites interact with host receptors, including G protein-coupled receptors (GPCRs), to modulate immune and inflammatory responses [3,4].

The GPCRs are a large family of membrane proteins that are targeted by numerous drugs, making them an attractive target for therapeutic intervention. Several GPCRs like GPR41, GPR43 and GPR109A (activated mainly by Short-chain fatty acids (SCFAs)) are involved in regulating inflammation and maintaining gut barrier integrity [1]. Designing drugs that mimic microbial metabolites activity is an innovative therapeutic modality in medicinal chemistry.



Our aim is to identify novel gut-metabolite mimicking (GMM) molecules -single or multitargeting (that can simultaneously bind and affect multiple relevant GPCRs). Using different classification *in silico* methods (ISE, NN and Random Forest) and through virtual screening and scoring of molecular libraries, we will identify top-ranked GMM candidates for further experimental validation

References

1. Liu J. et al. Functions of Gut Microbiota Metabolites, Current Status and Future Perspectives. Aging Dis. **2022** Jul 11;13(4):1106-1126. doi: 10.14336/AD.2022.0104. PMID: 35855347; PMCID: PMC9286904
2. Holmes, E. et al. Understanding the role of gut microbiome–host metabolic signal disruption in health and disease. Trends Microbiol 19, 349–359 (**2011**).
3. Colosimo, D. A. et al. Mapping Interactions of Microbial Metabolites with Human G-Protein-Coupled Receptors. Cell Host Microbe 26, 273-282.e7 (**2019**).
4. Chen, H. et al. A Forward Chemical Genetic Screen Reveals Gut Microbiota Metabolites That Modulate Host Physiology. Cell 177, 1217-1231.e18 (**2019**).

Poster Session BLUE:
Integrative Structure-Based Drug Design

P70: Molecular Dynamics and Experimental Insights into the Fungistatic Mechanism of Mutanobactin D

Patricia Brandl[1], Lukas Lüthy[2], Felix Pultar[1], Moritz Hansen[2], Erick M. Carreira[2], Sereina Riniker[1]

[1] *Department of Chemistry and Applied Biosciences, ETH Zürich, Switzerland*

[2] *Laboratory of Organic Chemistry, ETH Zürich, D-CHAB, Switzerland*

Mutanobactins, an intriguing class of cyclic peptides, are produced by the bacterium *Staphylococcus mutans*, a member of the human oral microbiome, and were found to have a fungistatic effect on *Candida albicans* by inhibiting its yeast-to-hyphen transition while leaving the planktonic state unaffected (1,2). The total synthesis of the most active member of the family, mutanobactin D, enabled complete configurational assignment and bioactivity studies (3), forming the basis of this work: While the enantiomer and various modifications in the ring scaffold retain bioactivity, stereoinversion of the lipid tail residue renders the peptides inactive (Figure 1). Inspired by membrane-related modes of action of other antimicrobial peptides (4), we employed molecular dynamic (MD) simulations to study how that stereoinversion affects the peptides' behavior at a polar/apolar interface.

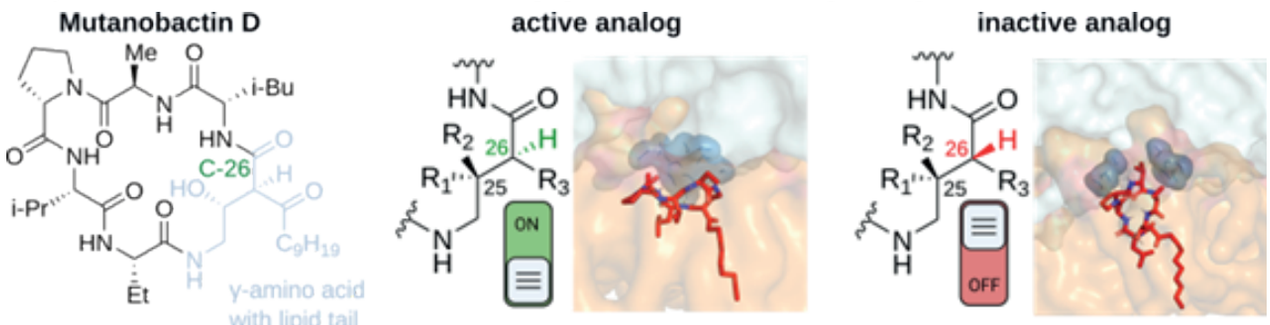


***Figure 1**: Structure of mutanobactin D, the stereo-switch of bioactivity and examples of membrane insertions for bioactive and inactive analogs with associated waters.*

We ruled out pore formation and permeability differences as main factors, suggesting more subtle differences explain the fungistatic effect. Simulations show that in isotropic environments, backbone conformations are similar among all analogs. However, in anisotropic environments, only bioactive analogs can adopt a unique and stable insertion conformation, which aligns with the ambiphilic environment and is corroborated by micellar NMR experiments. The findings are confirmed in simulations with an explicit membrane model. More stable insertion in conjunction with slightly altered membrane properties due to increased water uptake (5) may help explain macroscopic observables like the yeast-to-hyphae transition inhibition. Our investigations open new avenues for probing the biological mechanisms of mutanobactin D.

References

1. Joyner, P.M., et al., Mutanobactin A from the human oral pathogen Streptococcus mutans is a cross-kingdom regulator of the yeast-mycelium transition. Org. Biomol. Chem., **2010**, 8, 24, 5486-5489, 10.1039/C0OB00579G
2. Wang, X., et al., Fungal biofilm inhibitors from a human oral microbiome-derived bacterium. Org. Biomol. Chem., **2012**, 10, 10, 2044-2050, 10.1039/C2OB06856G
3. Pultar, F., et al., Mutanobactin D from the Human Microbiome: Total Synthesis, Configurational Assignment, and Biological Evaluation. J. Am. Chem. Soc., **2021**, 143, 27, 10389-10402, 10.1021/jacs.1c04825
4. Menlo, M., et al., Antimicrobial peptides: linking partition, activity and high membrane-bound concentrations. Nat. Rev. Microbiol., **2009**, 7, 3, 245-250, 10.1038/nrmicro2095

5. Yao, Y., et al., Water–lipid interface in lipidic mesophases with excess water. Faraday Discussions, **2024**, 249, 0, 469-484, 10.1039/D3FD00118K
6. Sudbery, P., et al., Growth of Candida albicans hyphae. Nat Rev Microbiol, **2011**, 9, 10, 737-748, 10.1038/nrmicro2636

P72: Benchmarking state-of-the-art *In silico* peptide design methods and evaluating peptide-optimization for fluorescent probes

Mark Fonteyne[1,3], Peter J.K. Kuppen[1], Alexander L. Vahrmeijer[1], Willem Jespers[2*], Gerard J. P. van Westen[3*]

[1] Department of Surgery, Leiden University Medical Center, The Netherlands

[2] Department of Pharmacy, University of Groningen, The Netherlands

[3] Medicinal Chemistry, Leiden University (LACDR), The Netherlands

** These authors contributed equally as last authors.*

Fluorescence-guided surgery (FGS) is a promising and rapidly growing field, where fluorescence imaging offers real-time intraoperative guidance, promoting favorable treatment prognosis.(1) Peptide-based probes are of great interest, providing high affinity and good specificity to targets, while maintaining relatively cheap production costs.(2, 3) With the rise of AlphaFold and RoseTTAFold, the way we perform computational peptide-probe design transformed significantly. Hence there is a need to assess to what extend structure-based computational methods can work in tandem with AI based approaches to generate novel fluorescent peptide-probes with higher confidence. In this benchmark, a wide range of computational methods are reviewed based on three main categories, target binding site identification, peptide virtual screening, and peptide-probe optimization. Every category is benchmarked based on current state-of the-art methods, and promising novel methods are investigated that could improve prediction capabilities. The nonredundant PepPro database is used to evaluate the most promising in silico methods, containing 58 experimentally obtained apo- and holo-protein structures with peptide sequence lengths ranging from 5-30 amino acids.(4) The tools AutoDock-CrankPep, HADDOCK3, AlphaFold, and RoseTTAFold are benchmarked to evaluate the best approach for each of the three main categories. Furthermore, the results will be combined into a state-of the-art methodology for peptide-probe design. This benchmark provides more insights on what the state-of-the-art methods are for peptide binding site identification, virtual screening, and optimization.

References

1. Ullah Z, Roy S, Muhammad S, Yu C, Huang H, Chen D, et al. Fluorescence imaging-guided surgery: current status and future directions. Biomater Sci. **2024**;12(15):3765-804.
2. Wang L, Wang N, Zhang W, Cheng X, Yan Z, Shao G, et al. Therapeutic peptides: current applications and future directions. Signal Transduct Target Ther. **2022**;7(1):48.
3. Vlieghe P, Lisowski V, Martinez J, Khrestchatisky M. Synthetic therapeutic peptides: science and market. Drug Discov Today. **2010**;15(1-2):40-56.
4. Xu X, Zou X. PepPro: A Nonredundant Structure Data Set for Benchmarking Peptide-Protein Computational Docking. J Comput Chem. **2020**;41(4):362-9.

P74: Benchmarking AI-Driven Protocols for Cyclic Peptide Conformer Prediction and Ranking

Rodrigo Ochoa[1], Jovan Damjanovic[2]

[1] Novo Nordisk A/S, Måløv, Denmark

[2] Novo Nordisk US R&D, Lexington, United States of America

Generative AI is being used in peptide design to optimize binders from early research stages. One notable application is with respect to cyclic peptides, wherein AI assists in studying conformers in bound or solution states. This work benchmarks state-of-the-art protocols' ability to predict conformers and rank cyclic peptide binders using sequence or all-atom representations. The datasets consist of cyclic peptides with experimental structures available in public repositories. Various tools were included in the benchmark: HighFold[1], RoseTTAFold All-Atom[2], EvoBind2[3], CycPepDock[4], Boltz-1[5], and Chai-1[6]. For conformers in solution, the impact of classic and enhanced molecular dynamics (MD) as a complementary approach to refine the predictions was also assessed. The implementation process was conducted with the BioLib benchmarking infrastructure to ensure reproducibility and accuracy of predictions. This study provides insights applicable to the design and optimization of cyclic peptides, contributing to advancements in peptide-based therapeutics.

References

1. Zhang, Chenhao, et al. HighFold: accurately predicting structures of cyclic peptides and complexes with head-to-tail and disulfide bridge constraints. *Briefings in Bioinformatics,* **2024**, 25.3: bbae215.
2. Krishna, Rohith, et al. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science,* **2024**, 384.6693: eadl2528.
3. Li, Qiuzhen, Efstathios Nikolaos Vlachos, and Patrick Bryant. Design of linear and cyclic peptide binders of different lengths only from a protein target sequence. *bioRxiv,* **2024**: 2024-06.
4. Kosugi, Takatsugu, and Masahito Ohue. Design of Cyclic Peptides Targeting Protein–Protein Interactions Using AlphaFold. *International Journal of Molecular Sciences,* **2023**, 24.17: 13257.
5. Wohlwend, Jeremy, et al. Boltz-1: Democratizing Biomolecular Interaction Modeling. *bioRxiv,* **2024**: 2024-11.
6. Chai Discovery, et al. Chai-1: Decoding the molecular interactions of life. *bioRxiv,* **2024**: 2024-10.

P76: Active learning FEP using 3D-QSAR for prioritizing bioisosteres in medicinal chemistry

Venkata K. Ramaswamy, Matthew Habgood and Mark D. Mackey

Cresset, New Cambridge House, Litlington, Cambridgeshire, UK

Bioisostere replacement is a powerful and popular tool used to optimize the potency and selectivity of candidate molecules in drug discovery. Selecting the right bioisosteres to invest resources in for synthesis and subsequent optimization is key to an efficient drug discovery project. In this retrospective study, we used human aldose reductase inhibitors to demonstrate an active learning workflow that prioritizes molecules from a large pool of bioisostere replacements generated by Spark[1]. This workflow combines two rigorous computational approaches: 3D-quantitative structure activity relationships (3D-QSAR) with shape and electrostatic descriptors[2], and free energy perturbation (FEP) for binding free energy calculations in Flare[1]. This workflow can rapidly locate the strongest-binding bioisosteric replacements with a relatively modest computational cost (a total of only 16% of the candidate pool was processed with FEP requiring 20% or even less GPU hours than if FEP were to include all candidates). The ROC-AUC for selection of known actives in 80 top-ranked candidates improved to 0.88 from 0.64, and the top picks were enriched with highly potent ALR2 inhibitors, including the well-known clinical candidate Zopolrestat developed by Pfizer[3,4].

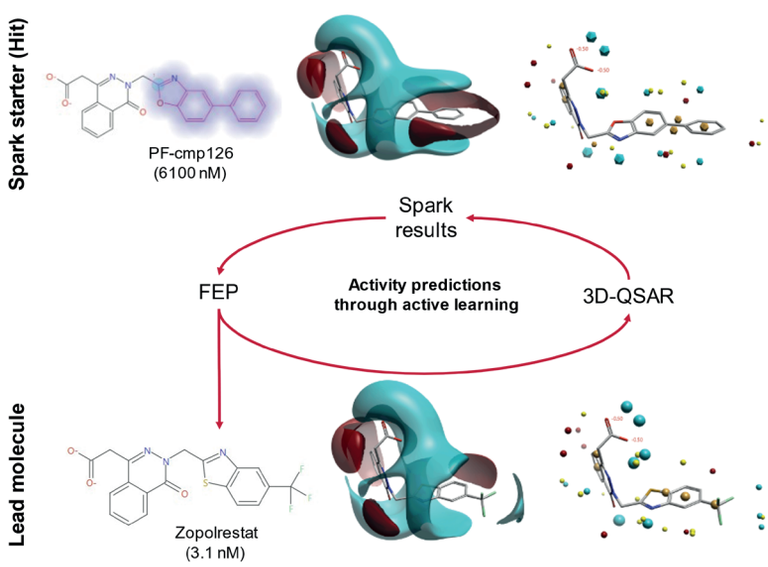


***Figure 1**: Graphical overview of the active learning workflow showing the starter molecule (PF-cmp126 with the region selected for R-group replacement highlighted) and one of the most potent lead molecules (Zopolrestat) identified by this workflow. 2D representation, 3D molecular electrostatic potential and field points of the molecules are shown (colour code: Red for positive, blue for negative, yellow for shape and orange for hydrophobic).*

References

1. Flare™, Cresset®, Litlington, Cambridgeshire, UK, https://www.cresset-group.com/software/flare/; Spark™. https://www.cresset-group.com/software/spark/
2. Cheeseright, T., et al., Molecular field extrema as descriptors of biological activity: definition and validation. J Chem Inf Model., **2006**, 46, 2, 665-676. DOI: 10.1021/ci050357s.
3. Mylari, B.L., et al., Novel, potent aldose reductase inhibitors: 3,4-dihydro-4-oxo-3-[[5-(trifluoromethyl)-2-benzothiazolyl] methyl]-1-phthalazineacetic acid (zopolrestat) and congeners. J Med Chem., **1991**, 34, 1, 108-122. DOI: 10.1021/jm00105a018.
4. Mylari, B.L., et al., Potent, orally active aldose reductase inhibitors related to zopolrestat: surrogates for benzothiazole side chain. J Med Chem., **1992**, 35, 3, 457-465. DOI: 10.1021/jm00081a006.

P78: PIEZO1: *In-Silico* Investigation of Channel Activation and Deactivation

Joana Massa[1], Benedikt Frieg[2], Christian Tyrchan[2], Oliver Koch[1]

[1] *Institute of Pharmaceutical and Medicinal Chemistry, University of Münster, Germany*

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

PIEZO1 is a cation channel that is primarily activated through mechanical force. Its activation has been shown to be influenced by membrane curvature. Furthermore it is ubiquitously expressed and plays critical roles in various patho-/physiological processes such as inflammatory diseases or red blood cell (RBC) volume regulation. In RBCs, PIEZO1 activation induces Ca^{2+} influx, which subsequently triggers $K_{Ca}3.1$ and leads to RBC dehydration. This process is impaired by various gain of function (GOF) mutations such as R2456H.[1,2]

Although prior studies have explored the activation mechanism of PIEZO1 at the molecular level through Molecular Dynamics simulations, comprehensive investigations linking the full-length human channel to ion conductance are missing. Furthermore, recently released cryo-EM structures of wild-type and GOF-mutation human PIEZO1 facilitate the investigation of this channel.[2] However, these structures omit essential terminal helices similar to previously released mouse PIEZO1 structures.[3] To overcome this limitation, a full-length homology model of PIEZO1 was developed based on the experimental structure of complete mouse PIEZO2 with all nine transmembrane helical bundles.[4] The results give insights into PIEZO1's activation and deactivation mechanism, enabling correlation of its structural state, ion conduction and membrane curvature. In the future, this will enable the investigation of GOF mutations in a similar fashion and aid in the development of new molecules targeting PIEZO1.

References

1. Xiao, B., Mechanisms of mechanotransduction and physiological roles of PIEZO channels. Nat. Rev. Mol. Cell. Biol., **2024**, 25, 886–903, https://doi.org/10.1038/s41580-024-00773-5.
2. Shan, Y., et al., Structure of human PIEZO1 and its slow inactivating channelopathy mutants. eLife, **2024**, 13, RP101923, https://doi.org/10.7554/eLife.101923.2.
3. Yang, X., et. al., Structure deformation and curvature sensing of PIEZO1 in lipid membranes. Nature, **2022**, 604, 377–383, https://doi.org/10.1038/s41586-022-04574-8.
4. Wang, L., et al., Structure and mechanogating of the mammalian tactile channel PIEZO2. Nature, **2019**, 573, 225–229, https://doi.org/10.1038/s41586-019-1505-8.

P80: Pharmacophore-based Discovery of Novel Cytochrome P450 (CYP) 4A11 Inhibitors to Combat Non-alcoholic Fatty Liver Disease (NAFLD)

Clemens A. Wolf[1], Matthias Bureik[2], Gerhard Wolber[1]

[1] *Pharmaceutical and Medicinal Chemistry (Computer-Aided Drug Design), Institute of Pharmacy, Freie Universität Berlin, Germany*

[2] *School of Pharmaceutical Science and Technology, Tianjin University, China*

Cytochrome P450 (CYP) enzymes are a superfamily of monooxygenases with 57 known members in humans [1] [2]. CYP enzymes oxidize a variety of substrates, including many xenobiotics as well as steroids and fatty acids [2]. A less studied [3] [4] member of this superfamily, CYP4A11, has been found to be a catalyst of hepatic ω-hydroxylation of fatty acids [4]. Therefore, CYP4A11 is thought to be involved in the development of non-alcoholic liver disease (NAFLD) [4].

In NAFLD, increased ω-hydroxylation of fatty acids promotes the production of reactive oxygen species (ROS) [4]. Being a common comorbidity of NAFLD, obesity results in an excess of free fatty acids (FFA) in the liver released from adipose tissue [4], which are subject to ω-oxidation catalyzed by CYP4A11 [4]. ROS, in turn, drive inflammation mainly via NFκB, tumor necrosis factor α (TNFα), and interleukin 6 [4]. This further aggravates chronic inflammation. Thus, CYP4A11-catalyzed ω-hydroxylation of fatty acids contributes to progressive remodeling of liver tissue paving the way to cirrhosis, and liver cancer. [4]. As there is no CYP4A11 inhibitor with market authorization, we present here an approach to design novel, potent CYP4A11 inhibitors as a strategy to combat liver degeneration.

Using known CYP4A11 inhibitors, we identified the essential interactions between CYP4A11 and any small molecule inhibitor and synthesized them into a pharmacophore model with desirable selectivity and sensitivity. We used the pharmacophore to virtually screen the Enamine screening collections and docked the retrieved hits into the binding site of CYP4A11. We selected promising candidates based on their conformation and subjected them to molecular dynamics (MD) simulations to validate the stability of their interactions in the binding site. The 15 most stable compounds were purchased and tested for CYP4A11 inhibition in a luminescence-based assay in yeast cells. Of the 15 compounds selected, three show reasonable inhibition. We are now conducting further measurements to determine their IC_{50}.

References

1. Durairaj, P., et al., Functional expression and activity screening of all human cytochrome P450 enzymes in fission yeast. FEBS letters, **2019**, 593, Jg., Nr. 12, S. 1372-1380, doi.org/10.1002/1873-3468.13441
2. Guengerich, F. P., Common and uncommon cytochrome P450 reactions related to metabolism and chemical toxicity. Chemical research in toxicology, **2001**, 14, Jg., Nr. 6, S. 611-650, doi.org/10.1021/tx0002583op
3. Machalz, D., et al., Structural insights into understudied human cytochrome P450 enzymes. Drug discovery today, **2021**, 26, Jg., Nr. 10, S. 2456-2464, doi.org/10.1016/j.drudis.2021.06.006
4. Gao, H., et al., CYP4A11 is involved in the development of nonalcoholic fatty liver disease via ROS-induced lipid peroxidation and inflammation. International journal of molecular medicine, **2020**, 45, Jg., Nr. 4, S. 1121-1129, doi.org/10.3892/ijmm.2020.4479

P82: Database AutoPH4: Pharmacophore Analysis of Multiple Protein Structures

Chris Williams[1], Andrew Henry[1], Steve Maginn[1], Guido Kirsten[1], Markus Kossner[1], Miklos Feher[2]

[1] Chemical Computing Group, Montreal, Canada

[2] D.E. Shaw Research, New York, USA

Drug discovery projects often involve determination, assessment and analysis of multiple protein ("apo") or protein-ligand ("holo") structures – these may have been produced by X-ray crystallography or cryo-EM measurements, or may be conformational ensembles from MD simulations. Such studies can reveal structural factors which can be used to find new potential binding sites, binders, or advance a medicinal chemistry campaign in some other way. Here, we present extensions and further developments of the AutoPH4 method(1) in the Molecular Operating Environment (MOE) software system which, in conjunction with the application of MOE's Site Finder and other tools, enable the analysis of such multiple structures to produce consensus pharmacophores, binding site information and classification and new ways to assess docking results. The utility of these methods is illustrated by an analysis of a database of Abl kinase structures.

References

1. S. Jiang, M. Feher, C. Williams, B. Cole, D.E. Shaw; J. Chem. Inf. Model. 60 (2020) 4326–4338

P84: From 45 billion small molecules to potential new ligands for the intracellular binding site of CCR2, using High Throughput Virtual Screening, Molecular docking and Relative Free Energy Perturbation

D.J.M. van Pinxteren[1,2], H. Gutiérrez-de-Terán[2], E. Gibert[3], F. Martin Garcia[3], and W. Jespers[1,2]

[1] *Medicinal Chemistry, Photopharmacology and Imaging group, Rijksuniversiteit Groningen, The Netherlands*

[2] *MODSIM Pharma AB, Uppsala, Sweden*

[3] *Pharmacelera, PCB, Torre R, Barcelona, Spain*

With around 45 billion commercially available compounds and the rapid expansion of chemical space in drug discovery, efficient screening and optimization methods to identify potential therapeutic candidates to find potential ligands for druggable targets. [1–3] In this study, a combination of ligand-based and structure-based high-throughput virtual screening (HTVS) methods was employed to discover novel small molecule inhibitors targeting the intracellular binding site of the CC Chemokine Receptor 2 (CCR2). Utilizing ExaScreen for initial hit identification from the Enamine REAL Space dataset, followed by rDock and Glide for precision docking, the screening process narrowed down 12,000 candidates to 2,186 for further analysis. QligFEP was subsequently applied for relative free energy perturbation (RFEP) calculations to optimize lead compounds. [2,3] Despite the identification of several promising candidates, no compound demonstrated superior binding affinity compared to the reference ligand. Nonetheless, the study did show potential in using a combination of ExaScreen, docking techniques, and QligFEP in a computational workflow to identify new ligands for a durable target. With potential improvements to the software, the results could be significantly improved in the coming research. Also, this study emphasizes the importance of lead optimization in drug discovery. With minor adjustments to a ligand, affinity could be increased/decreased. Overall, this research shows the strengths and limitations of combining ligand-based and structure-based computational methods.

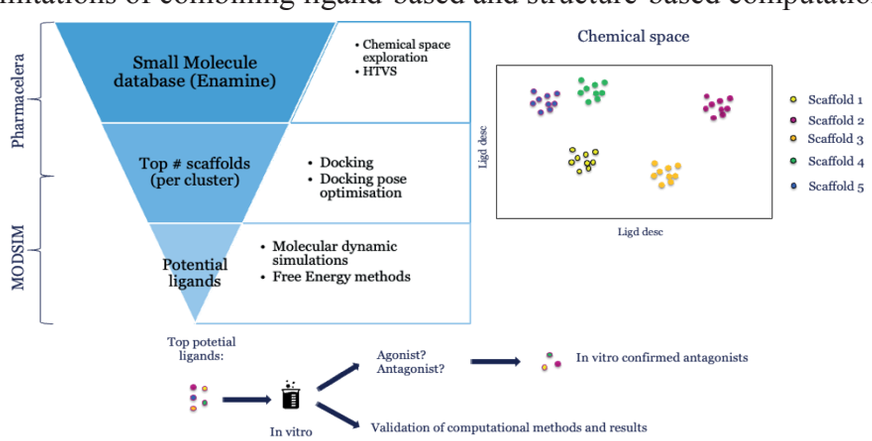


***Figure 1**: Graphical abstract: From 45 billion small molecules to potential new ligands for the intracellular binding site of CCR2, using High Throughput Virtual Screening, Molecular docking and Relative Free Energy Perturbation.*

References

1. Bender BJ, Gahbauer S, Luttens A, et al. A practical guide to large-scale docking. *Nat Protoc*. 2021;16(10):4799. doi:10.1038/S41596-021-00597-Z
2. Jespers W, Esguerra M, Åqvist J, Gutiérrez-De-Terán H. Qligfep: An automated workflow for small molecule free energy calculations in Q. *J Cheminform*. 2019;11(1):1-16. doi:10.1186/S13321-019-0348-5/TABLES/4
3. Vázquez J, Deplano A, Herrero A, et al. Development and Validation of Molecular Overlays Derived from Three-Dimensional Hydrophobic Similarity with PharmScreen. *J Chem Inf Model*. 2018;58(8):1596-1609. doi:10.1021/ACS.JCIM.8B00216/ASSET/IMAGES/LARGE/CI-2018-002168_0013.JPEG

Poster Session BLUE:
New Modalities and Large Chemical Data Sets

P86: Scaffold-Based Library Design vs. Make-on-Demand Space: A Comparative Assessment of Chemical Content

Leonard Bui, Teodora Djikic-Stojsic, Guillaume Bret, Esther Kellenberger

Laboratoire d'Innovation Thérapeutique, UMR 7200 CNRS-Université de Strasbourg, Illkirch-Graffenstaden, France

Chemical libraries are essential in drug discovery, providing a vast variety of compounds for screening and exploration [1,2]. Previously, through collective efforts of the chemoinformaticians and chemists, our group has created two libraries: the essential eIMS containing 578 in-stock compounds on plates ready for High Throughput Screening and a companion virtual library vIMS, containing 821.069 compounds generated from the scaffolds of the eIMS compounds, and decorated with substituents from customized collection of R-groups[3].

In this study, we aim at validating this library design approach, which builds on scaffold-based structuring and decoration guided by chemists' expertise. Specifically, we evaluate its effectiveness in comparison to the widely adopted reaction- and building block-based approach. Using REAL space data[4], we developed scaffold-focused libraries and systematically compared them to the make-on-demand chemical space containing the same scaffolds. The results showed a notable degree of similarity between the two, but with limited strict overlap. Interestingly, a significant portion of the R-groups were not identified as such in the make-on-demand library. SAscore analysis of two compound sets indicated overall low to moderate synthetic difficulty, with slightly better scores for highly similar compounds. These findings confirm the value of the scaffold-based method for generating focused libraries, offering high potential for lead optimization in drug discovery.

References

1. Saldívar-González, F.I, et al., Chemoinformatics-based enumeration of chemical libraries: a tutorial. Journal of Cheminformatics **2020**, 12, 64, doi.org/10.1186/s13321-020-00466-z
2. Hartenfeller, M., et al., A Collection of Robust Organic Synthesis Reactions for In Silico Molecule Design. Journal of Chemical Information and Modeling, **2011**, 51, 12, 3093–3098, doi.org/10.1021/ci200379p
3. Djikic-Stojsic, T., et al., The IMS Library: from IN-Stock to Virtual. ChemMedChem, **2024**, 19, doi.org/10.1002/cmdc.202400381.
4. Grygorenko, O., et al., Generating Multibillion Chemical Space Of Readily Accessible Screening Compounds. iScience, **2020**, 23,11, 101681, doi.org/10.1016/j.isci.2020.101681

Poster Session BLUE:
Open Science, Omics, and Natural Products

P88: Biosynfoni: A *Biosynformatic* Molecular Descriptor for Natural Product Research

Lucina-May Nollen[1,2], David Meijer[1], Maria Sorokina[3], Justin J.J. van der Hooft[1,4]

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

[2] *Current address: Leiden Academic Centre of Drug Research, Leiden University, The Netherlands*

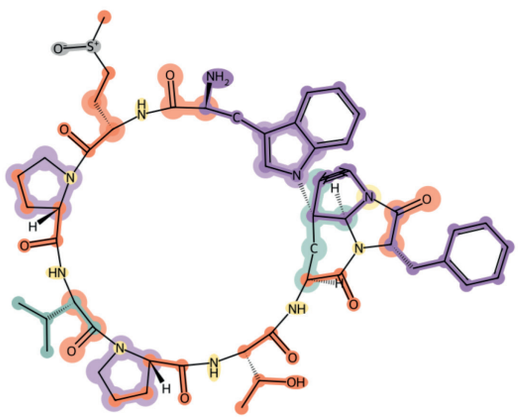
[3] *DS&AI, Bayer Pharmaceuticals, Germany*

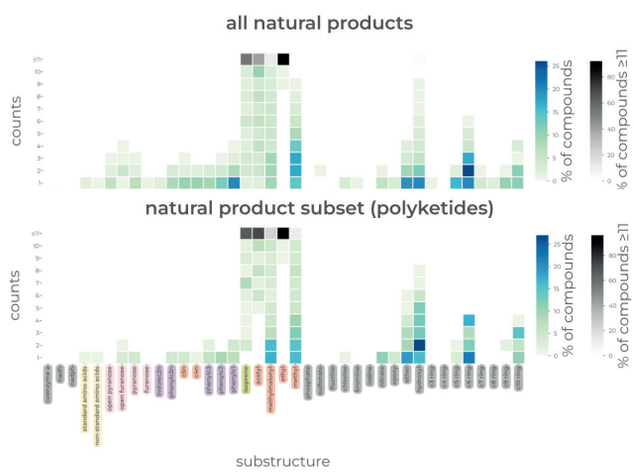
[4] *Department of Biochemistry, University of Johannesburg, South Africa*

Natural products are a vast source of diverse and evolutionarily-optimised bioactive molecules for pharmaceutical and agricultural applications. Molecular fingerprints are fundamental for large-scale structural analysis and serve as key input for machine learning models. However, existing fingerprints fail to explicitly capture characteristic biosynthetic features of natural products, limiting the interpretability of predictive models.

Here, we introduce Biosynfoni, a natural product-tailored molecular fingerprint derived from a curated set of 39 biosynthetically relevant substructures inspired by Dewick's 2009 book on principles in natural product biosynthesis. Compared to conventional fingerprints such as MACCS, Morgan, and RDKit fingerprints, Biosynfoni better captures biosynthetic distance through Tanimoto scores and maintains competitive accuracy in natural product classification.

The compact design of Biosynfoni, coupled with its well-defined substructure representation, enables clear and novel visualisations of detected features in both single and multiple molecules. This enables more transparent predictions and deeper insights into feature importance within machine learning models. Our findings highlight the potential of a tailored, biosynthetically grounded fingerprint to provide lightweight machine learning models with improved explainability in natural product cheminformatics.





References

1. Dewick, P. M. Medicinal Natural Products: A Biosynthetic Approach. (Wiley, **2009**). doi:10.1002/9780470742761.
2. Boldini, D. et al. Effectiveness of molecular fingerprints for exploring the chemical space of natural products. J Cheminform **16**, 35 (**2024**).

P90: Drug and clinical candidate drug data in ChEMBL

Fiona Hunter, Harris Ioannidis, Melissa F Adasme, James Blackshaw, Nicolas Bosc, Sybilla Corbett, Marleen de Veij, Eloy Felix, Tevfik Kizilören, Emma Manners, Juan F. Mosquera, Ines Smit, Barbara Zdrazil and Noel O'Boyl

European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge, United Kingdom

ChEMBL[1,2] is a manually-curated, large-scale, open-access, FAIR[3] database of bioactive molecules with drug-like properties. The first public launch of the ChEMBL database (www.ebi.ac.uk/chembl) in 2009 was a milestone in the recent history of chemical biology and drug discovery because it provided unprecedented free access to large amounts of high-quality, curated data on bioactive molecules. ChEMBL has grown significantly since then and now impacts a wide range of areas that include drug discovery, data science and the development and validation of AI, machine learning and other in silico methods.

Curation of data for approved drugs, and drugs that are progressing through the clinical development pipeline ('clinical candidate drugs'), has formed an integral part of the core offering of the ChEMBL database since its inception. The curated drug data enable the scientific community to answer important and practical science questions in drug discovery and chemical biology. Examples include the large-scale assessment of drug and ligand physicochemical properties and ligand efficiencies[4], the identification of drug repurposing opportunities for COVID-19[5], identifying potential drugs to treat the neglected tropical disease, schistosomaisis[6], or the use of drug indication data as a reference set to test emerging Large Language Models[7].

This poster focuses on the curation of drug and clinical candidate drug data, although it is recognised that many of these drugs will also display bioactivity data in ChEMBL. The key areas of curation for drug and clinical candidate drug data include: the drug name, synonym(s) and trade name(s), chemical structure or biological sequence, data source(s), drug indication(s), drug mechanism(s), drug warning(s) and drug properties such as maximum phase of development, molecule type, prodrug status, year of approval.

Overall, the drug and clinical candidate drug data in ChEMBL provide a useful, publicly available, searchable record of approved and clinical drugs and their characteristics, supporting drug discovery research.

References

1. Zdrazil, B. et al. The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. Nucleic Acids Research 52, D1180–D1192 (**2024**). doi: https://doi.org/10.1093/nar/gkad1004 .
2. Gaulton, A. et al. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 40, D1100-7 (**2012**). doi: https://doi.org/10.1093/nar/gkr777 .
3. Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data (**2016**). doi: https://doi.org/10.1038/sdata.2016.18.
4. Leeson, P. et al., Target-Based Evaluation of "Drug-Like" Properties and Ligand Efficiencies, Journal of Medicinal Chemistry, Vol 64, Issue 11 (**2021**). doi: https://doi.org/10.1021/acs.jmedchem.1c00416.
5. Bouhaddou, M. et al., The Global Phosphorylation Landscape of SARS-CoV-2 Infection, Cell, Volume 182, Issue 3, 6 August **2020**, Pages 685-712.e19, doi: https://doi.org/10.1016/j.cell.2020.06.034.
6. Padalino et al., Using ChEMBL to Complement Schistosome Drug Discovery, Pharmaceutics, 15(5):1359 (**2023**).doi: https://doi.org/10.3390/pharmaceutics15051359.
7. Oniani D. et al., Emerging opportunities of using large language models for translation between drug molecules and indications, Sci Rep., 14, 10738 (**2024**). doi: https://doi.org/10.1038/s41598-024-61124-0.

List of Participants

Full Name	Organization / Company	Country
Aya Abdelbaky	Pangea Bio	DE, Germany
Uday Abu Shehab	University of Vienna	AT, Austria
Melissa F Adasme	EMBL's European Bioinformatics Institute (EMBL-EBI)	UK, United Kingdom
David Alencar Araripe	Leiden University	NL, Netherlands, The
Carla Araya-Cloutier	Wageningen University	NL, Netherlands, The
Emma Sarah Armstrong	University of Sheffield	UK, United Kingdom
Rafał Adam Bachorz	Simulations Plus	US, United States of America
Nada Badr	Leiden University Medical Centre	NL, Netherlands, The
Bernd Beck	Boehringer Ingelheim Pharma GmbH&Co.KG	DE, Germany
Andrius Bernatavicius	Leiden University	NL, Netherlands, The
Marc Bianciotto	Sanofi R&D	FR, France
Akash Deep Biswas	Dompé - Exscalate	IT, Italy
Michael Kiran Blakey	NextMove Software	UK, United Kingdom
Gerd Blanke	StructurePendium Technologies GmbH	DE, Germany
Marc Boef	Leiden University	NL, Netherlands, The
Patricia Brandl	ETH Zuerich	CH, Switzerland
Hans Briem	Bayer AG	DE, Germany
Helena Brinkmann	Eindhoven University of Technology	NL, Netherlands, The
Leonard Bui	University of Strasbourg	FR, France
Shu-Yu Chen	ETH Zurich	CH, Switzerland
Ya Chen	University of Vienna	AT, Austria
Kostiantyn Chernichenko	J&J Innovative Medicine	BE, Belgium
Alex Michael Clark	Collaborative Drug Discovery	CA, Canada
Wei Dai	Queen Mary University of London	UK, United Kingdom
Andrew Peter Dalke	Dalke Scientific	SE, Sweden
Maedeh Darsaraee	University of Bern	CH, Switzerland
Bernardo de Souza	FACCTs GmbH	DE, Germany
Wim F. Dehaen	UCT Prague	CZ, Czech Republic
Ivan Derbenev	Astex Pharmaceuticals	UK, United Kingdom

Niklas Piet Doering	Freie Universität Berlin	DE, Germany
Thomas Doerner	Dr. Thomas Doerner - Discovery Informatics Information	DE, Germany
Uschi Dolfus	Bayer Ag	DE, Germany
Nico Domschke	Bioinformatics Group, Leipzig University	DE, Germany
Terence Egbelo	University of Sheffield	UK, United Kingdom
Anatol Ehrlich	University of Vienna	AT, Austria
Shayma Elatawneh	University of Münster	DE, Germany
Thomas Engel	LMU Munich	DE, Germany
Eloy Felix Manzanares	EMBL-EBI	UK, United Kingdom
Alejandro Flores Sepulveda	University of Bern	CH, Switzerland
Mark Fonteyne	Leiden University Medical Center	NL, Netherlands, The
Hosein Fooladi	University of Vienna	AT, Austria
Zelimir Galjanic	D. E. Shaw Research	US, United States of America
Helga Gerets	UCB BioPharma SRL	BE, Belgium
Val Gillet	University of Sheffield	UK, United Kingdom
Daria Goldmann	Sanofi	DE, Germany
Andreas Göller	Bayer AG	DE, Germany
Carmen Gratteri	Light Center & Dompé - Exscalate	IT, Italy
Lora Gržin	National Institute of Chemistry	SI, Slovenia
Sebastien Guesne	Lhasa Limited	UK, United Kingdom
Judith Günther	Bayer AG	DE, Germany
Torben Gutermuth	University of Hamburg	DE, Germany
Satoshi Hamano	JAPAN TOBACCO INC.	JP, Japan
Hisashi Handa	Kindai University	JP, Japan
Thierry Hanser	Lhasa Limited	UK, United Kingdom
Alan Kai Hassen	Pfizer Inc. / Leiden University	DE, Germany
Alec Heckert	D. E. Shaw Research	US, United States of America
Samuel Kurt Robert Homberg	University of Münster	DE, Germany
Fiona Hunter	EMBL-EBI	UK, United Kingdom
Idil Ismail	ETH Zurich	CH, Switzerland
Roxane Axel Jacob	University of Vienna	AT, Austria
Célien Jacquemard	University of Strasbourg	FR, France

Philipp Janssen	University Of Muenster	DE, Germany
Dóra Jávorszky	Chemaxon Kft	HU, Hungary
Chiel Jespers	Leiden University	NL, Netherlands, The
Willem Jespers	Groningen University	NL, Netherlands, The
Daan Antoon Jiskoot	Rijksinstituut voor Volksgezondheid en Milieu	NL, Netherlands, The
Johannes Kaminski	University of Münster	DE, Germany
Anna Kapeliukha	Chemspace	UA, Ukraine
Lisa Sophie Kersten	Heinrich Heine University	DE, Germany
Bola Khalil	Leiden University and Johnson&Johnson	BE, Belgium
Guido Kirsten	Chemical Computing Group	DE, Germany
Oliver Koch	University of Münster	DE, Germany
Florian Koensgen	Greenpharma S.A.S.	FR, France
Aleksejs Kontijevskis	Amgen	DK, Denmark
Essi Venla Elina Koskela	Certara	NL, Netherlands, The
József Kozma	Chemaxon	HU, Hungary
Lara Kuhnke	Bayer AG	DE, Germany
Rosan Kuin	Leiden University (LACDR)	NL, Netherlands, The
Anupriya Kumar	OpenEye, Cadence Molecular Sciences	DE, Germany
Antoine Michel Lauder Lacour	Saarland University	DE, Germany
Ingvar Lagerstedt	NextMove Software Limited	UK, United Kingdom
Kevin Lam	Freie Universität Berlin	DE, Germany
Greg Landrum	ETH Zurich	CH, Switzerland
Federico Lazzari	Scuola Superiore Meridionale	IT, Italy
David LeBard	OpenEye, Cadence Molecular Sciences	US, United States of America
Daniil Lepikhov	Technische Universitat Berlin	DE, Germany
Pierre-Yves Libouban	Johnson & Johnson	BE, Belgium
Clayton Lim	Home Team Science and Technology Agency	SG, Singapore
Sijie Liu	Freie Universität Berlin	DE, Germany
Christian Loeffeld	Alipheron AG	CH, Switzerland
James Lumley	GSK	UK, United Kingdom
Filippo Lunghini	Dompé - Exscalate	IT, Italy
Peter Maas	peter.maas@emolecules.com	NL, Netherlands, The

Niels Maeder	ETH Zürich	CH, Switzerland
Candida Manelfi	Dompé - Exscalate	IT, Italy
Veselina Marinova	Cresset	UK, United Kingdom
Joana Massa	University of Münster	DE, Germany
John Wilkinson Mayfield	NextMove Software	UK, United Kingdom
Hans-Peter Meulekamp	Evolvus Inc.	NL, Netherlands, The
Wijnand Mooij	Dotmatics Ltd	NL, Netherlands, The
Pietro Morerio	Istituto Italiano di Tecnologia	IT, Italy
Francesco Moriello	dsm-firmenich	CH, Switzerland
Joerg Muehlbacher	Novartis	CH, Switzerland
András György Németh	Chemaxon Kft	HU, Hungary
Christos A. Nicolaou	Novo Nordisk	US, United States of America
Lucina-May Nollen	Leiden University	NL, Netherlands, The
Kian Noorman van der Dussen	Rijksuniversiteit Groningen	NL, Netherlands, The
Noel Michael O'Boyle	EMBL-EBI	UK, United Kingdom
Leon Moritz Obendorf	Freie Universität Berlin	DE, Germany
Rodrigo Ochoa	Novo Nordisk A/S	DK, Denmark
Floriane Stephanie Christelle Odje	Universität des Saarlandes	DE, Germany
Frank Oellien	AbbVie	DE, Germany
Wiktor Olszowy	dsm-firmenich	CH, Switzerland
Afra Panahi	California Stat University, San Marcos	US, United States of America
Raquel Parrondo-Pizarro	Chemotargets, S.L.	ES, Spain
Lagnajit (Lucky) Pattanaik	D. E. Shaw Research	US, United States of America
David A Pearlman	QSimulate	US, United States of America
Tieu Long Phan	Interdisciplinary Center for Bioinformatics, University Leipzig	DE, Germany
Boris Piakillia	NUVISAN ICB GmbH	DE, Germany
Rachael Mary Elizabeth Pirie	NextMove Software	UK, United Kingdom
Jonathan Pletzer-Zelgert	University of Hamburg	DE, Germany
Pavel Polishchuk	Palacky University	CZ, Czech Republic
Cristian-Catalin Pop	University of Medicine and Pharmacy Iuliu Hatieganu	RO, Romania
Domen Pregeljc	ETH Zurich	CH, Switzerland
Jonny Proppe	TU Braunschweig	DE, Germany

Kristina Sophie Puls	PROSION GmbH	DE, Germany
Emma Louise Pye	Lhasa Limited	UK, United Kingdom
Matthias Rarey	Universität Hamburg	DE, Germany
Max Rausch-Dupont	Saarland University	DE, Germany
Steffen Renner	Novartis Pharma AG	CH, Switzerland
Ben (Bernado) Retamal	Collaborative Drug Discovery, Inc.	US, United States of America
Stefano Ribes	Chalmers University of Technology and University of Gothenburg	SE, Sweden
Christoph Riplinger	FACCTs GmbH	DE, Germany
Alejandro Rodríguez-Martínez	Universidad Católica de Murcia UCAM	ES, Spain
Daniel Rose	University of Vienna	AT, Austria
Luke Rossen	Eindhoven University of Technology	NL, Netherlands, The
Grazia Rovelli	Italfarmaco	IT, Italy
Enrico Martin Ruijsenaars	ETHZ	CH, Switzerland
Thomas Sander	Alipheron AG	CH, Switzerland
Thomas Sanderson	Simulations Plus	US, United States of America
Delia Sayle	NextMove Software Limited	UK, United Kingdom
Roger Sayle	NextMove Software	UK, United Kingdom
Kay Schaller	Novo Nordisk A/S	DK, Denmark
Jenke Scheen	ASAP Discovery / OMSF	NL, Netherlands, The
Carl Christophorus Gerardus Schiebroek	ETH Zurich	CH, Switzerland
Stefan Schoenbichler	Bionorica research GmbH	AT, Austria
Linde Schoenmaker	Leiden University (LACDR)	NL, Netherlands, The
Vincent-Alexander Jean-Luc Christopher Mortimer Scholz	Univeristy of Vienna	AT, Austria
Ansgar Schuffenhauer	Novartis	CH, Switzerland
Paul Selzer	Novartis Pharma AG	CH, Switzerland
Carlos J. V. Simoes	VIB	BE, Belgium
Gellért Sipos	Chemaxon Kft.	HU, Hungary
Delaney Amati Smith	Stanford University	US, United States of America
Riccardo Solazzo	ETH Zurich	CH, Switzerland
Dora Sribar	Nuvisan ICB	DE, Germany
Sanjay Srivastava	Neurocrine Biosciences	US, United States of America
Jess Stacey	MedChemica Ltd	UK, United Kingdom

Katarina Stanciakova	OpenEye, Cadence Molecular Sciences	DE, Germany
Christoph Steinbeck	Friedrich-Schiller-University Jena	DE, Germany
Kateřina Storchmannová	Palacky University in Olomouc	CZ, Czech Republic
Conrad Stork	BASF SE	DE, Germany
Pieter Stouten	Dompé - Exscalate	IT, Italy
Afnan Sultan	Saarland university	DE, Germany
Márk Szabó	Chemaxon	HU, Hungary
Valerij Talagayev	Freie Universitaet Berlin	DE, Germany
Hanz Tantiangco	University of Sheffield	UK, United Kingdom
Olga Tarkhanova	Chemspace LLC	UA, Ukraine
Marvin Taterra	University of Münster	DE, Germany
Valerio Tazzari	Dompé - Exscalate	IT, Italy
Rhea Singh Thakur	D. E. Shaw Research	US, United States of America
Valery Tkachenko	Science Data Experts	US, United States of America
Christian Tyrchan	AstraZeneca	SE, Sweden
Ulrike Christine Uhrig	EMBL	DE, Germany
Remco Leendert Van Den Broek	Leiden University (LACDR)	NL, Netherlands, The
Tim Van den Bulcke	Deltamine	BE, Belgium
Helle van den Maagdenberg	Leiden University (LACDR)	NL, Netherlands, The
Donald Jonannes Marius van Pinxteren	Rijksuniversiteit Groningen	NL, Netherlands, The
Derek van Tilborg	Eindhoven University of Technology	NL, Netherlands, The
Gerard van Westen	Leiden University (LACDR)	NL, Netherlands, The
Mariana Vaschetto	Collaborative Drug Discovery, Inc.	US, United States of America
Jonas Verhellen	University of Copenhagen	DK, Denmark
Modest von Korff	Alipheron AG	CH, Switzerland
Thi Ngoc Lan Vu	University of Vienna	AT, Austria
Markus Wagener	Grünenthal GmbH	DE, Germany
Moritz Walter	Boehringer Ingelheim	DE, Germany
Anne Mai Wassermann	Johnson&Johnson	DE, Germany
James Webster	University of Dundee	UK, United Kingdom
Jordan Wells	D. E. Shaw Research	US, United States of America
Matthias Welsch	University of Vienna	AT, Austria

Henriette Willems	ALBORADA DDI, University of Cambridge	UK, United Kingdom
Egon Willighagen	Maastricht University	NL, Netherlands, The
Sarah Witzke	CCG	UK, United Kingdom
Clemens Alexander Wolf	Freie Universität Berlin	DE, Germany
Huanni Zhang	University of Vienna	AT, Austria

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)