ENBIK2025 conference proceedings

Editors

Petr Čech, Daniel Svozil

Prague 2025

ENBIK2025 conference proceedings

Printed by powerprint s. r. o.
Brandejsovo nám. 1219/1, 165 00 Praha 6 – Suchdol

Published by the University of Chemistry and Technology, Prague
Technická 5, 166 28 Praha 6, Czech Republic

ISBN 978-80-7592-292-2

SESSION 1

Sequences

L1-01

Efficient Analysis of Genome Annotation Colocalization

Gafurov A.[1,2], Vinař T.[1], Medvedev P.[3], Brejová B.[1]

[1] *Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava, Slovakia*
[2] *LIRRM, Montpellier, France*
[3] *The Pennsylvania State University, USA*

An annotation is a set of genomic intervals sharing a particular function or property. Examples include genes, conserved elements, and epigenetic modifications. A common task is to compare two annotations to determine if one is enriched or depleted in the regions covered by the other, suggesting some underlying biological mechanism. We study the problem of assigning statistical significance to such a comparison based on a null model representing two random unrelated annotations. To make the problem trackable, we use a Markov chain as a null model. We developed an exact quadratic-time algorithm based on dynamic programming as well as a linear-time algoritm based on a normal approximation. We also incorporate background information into our analyses by differentiating among several genomic contexts. These contexts can capture various confounding factors, such as GC content or assembly gaps. Currently, we are extending the tool to also separately analyze fixed-sized windows of the genome to pinpoint areas contributing to the enrichment the most. We demonstrate the efficiency and accuracy of our algorithms on synthetic and real data sets, including the recent human telomere-to-telomere assembly. The use of genomic contexts to correct for GC-bias resulted in the reversal of some previously published findings. In one striking example, the set of all human exons appears enriched for overlap with copy number losses but enrichment turns into depletion after taking into account assembly gaps and GC content. Our software, called MCDP2, is freely available at https://github.com/fmfi-compbio/mcdp2 under the MIT licence.

This work was partially supported by a grant from the European Union Horizon 2020 research and innovation program No. [872539] (PANGAIA); and grants from the Slovak Research and Development Agency [APVV-22-0144] and the Scientific Grant Agency VEGA [1/0140/25, 1/0538/22].

L1-02

Elastic-Degenerate Strings in Bioinformatics: Motivation and Open Problems

Bohuslavová D.[1]

[1] *Czech Technical Univeristy in Prague*

As the volume of biological data grows, effectively managing and efficiently storing this data becomes a key challenge in bioinformatics. While processing, data structures with different representations are built to acquire better answers in a meaningful amount of time and with minimal memory consumption. One of the promising representations of highly similar sequences is the elastic-degenerate string (EDS), a mathematical concept that allows multiple sequences to be stored in a compact space. Although the concept of EDS has been around for many years now, a number of open problems remain unsolved due to struggles with biological meaning and the computational complexities of the algorithms. The biggest potential and ongoing application of EDS lies in manipulating data for population studies, pangenomics and other studies based on variant callings or sequence alignments. This talk provides an overview of the current state of the art, as well as the main challenges associated with processing such representations.

L1-03

Plasmid Identification Through Graph Neural Networks

Brejová B.[1], Tordová V.[1], Andraščík K.[1], Chauve C.[2], Vinař T.[1]

[1] *Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava, Slovakia*
[2] *Dept. of Mathematics, Faculty of Science, Simon Fraser University, Burnaby, BC, Canada*

Identification of plasmids from sequencing data is an important and challenging problem related to antimicrobial resistance spread. The problem is typically addressed by machine learning methods combining features derived from input contigs. We provide a new architecture for identifying plasmid contigs in fragmented genome assemblies built from short-read data. We employ graph neural networks (GNNs) and the assembly graph to propagate the information from nearby nodes, which leads to more accurate classification, especially for short contigs that are difficult to classify based on sequence features or database searches alone. Our software tool, plASgraph2, either outperforms or performs on par with a wide range of state-of-the-art methods.

Some methods also use additional features based on homology to sequences typical for known plasmids or chromosomes. We propose a method for creating such features using log-odds scores based on ideas similar to those traditionally used in sequence alignment scoring, such as BLOSUM scoring matrices. The framework is flexible as it can handle both close homolog tags derived from a pangenome of training sequences as well as protein domains capturing distant homology. Inclusion of these features into the plASgraph2 graph neural network further significantly improves the accuracy of the predictions.

Availability: Our software is available at https://github.com/cchauve/plasgraph2 and the training and testing data sets are available at https://github.com/fmfi-compbio/plasgraph2-datasets.

Acknowledgements: This work was partially supported by a grant from the European Union Horizon 2020 research and innovation program No. [872539] (PANGAIA); and grants from the Slovak Research and Development Agency [APVV-22-0144] and the Scientific Grant Agency VEGA [1/0140/25, 1/0538/22]. This research was enabled in part by computational infrastructure support provided by Digital Research Alliance of Canada (https://alliancecan.ca).

L1-04

Cryptic binding site detection with protein language models

Škrhák V., Hoksza D.

Protein-ligand binding sites play a vital role in cellular function and drug discovery. A particularly challenging subset, cryptic binding sites (CBSs), only form after significant conformational changes, making them difficult to detect in unbound (apo) structures. Existing structure-based methods for binding site prediction often fail in such cases due to their dependence on particular protein conformations.

In this work, we explore the use of protein language models (pLMs) - deep learning models trained on amino acid sequences - to identify CBSs from sequence alone, as sequence information is inherently conformation-independent. Starting with a transfer learning approach, we construct a baseline predictor and subsequently evaluate multiple fine-tuning methods to enhance predictive power. One such approach involves multitask learning, where the model simultaneously predicts binding site residues and estimates local flexibility, an important factor in the CBS formation.

Our models are fine-tuned primarily on CryptoBench, a large benchmark dataset containing CBSs. However, we also consider additional data sources. Our experiments lead to consistent improvements in predictive performance across standard evaluation metrics, including more than a 2% increase in AUC. We also conduct an error analysis to better understand model limitations and implement a post-processing step to enhance prediction smoothness.

Our work opens a new direction for CBS prediction and contributes to the effort of improving early-stage drug discovery tools.

SESSION 2

Small molecules

L2-01

Handling Compound Promiscuity in CZ-Openscreen

Hanzlík A.[1,2], Voršilák M.[1,2], Škuta C.[2], Bartůněk P.[2]

[1] *UCT Prague*
[2] *CZ-Openscreen*

High Throughput Screening (HTS) is a cornerstone of early-stage drug discovery, enabling rapid evaluation of extensive chemical libraries for biological activity against specific targets. While HTS significantly accelerates the discovery process compared to traditional manual assays, its heightened sensitivity often results in a high incidence of false positives. These artifacts can lead to costly and unproductive follow-up studies. False positives typically stem from assay technology interference or non-specific interactions with biological targets. Over time, problematic compounds exhibiting such behaviors have been categorized using terms such as *nuisance compounds*, *promiscuous compounds*, *PAINS* (Pan-Assay Interference Compounds)(1), *frequent hitters*(2), and *aggregators*(3). To mitigate their impact, researchers have developed filtering strategies including SMARTS-based substructure searches, similarity-based filtering to known nuisance compounds, and machine learning approaches.
In this work, we present a methodology for annotating, visualizing, and estimating the likelihood of compound promiscuity based on primary screening data generated at CZ-Openscreen. Leveraging rich metadata from ScreenX—CZ-Openscreen's laboratory information management system—such as assay format and detection technology (e.g., fluorescence or luminescence), we aim to reduce experimental noise and provide more informative annotations. This methodology is currently being integrated into the ScreenX platform to support the design and evaluation of future HTS campaigns, ultimately enhancing data quality and decision-making in early drug discovery.

References

[1] Baell JB, Holloway GA. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. Journal of Medicinal Chemistry. 2010;53(7):2719-40.
[2] Stork C, Chen Y, Šícho M, Kirchmair J. Hit Dexter 2.0: Machine-Learning Models for the Prediction of Frequent Hitters. Journal of Chemical Information and Modeling. 2019;59(3):1030-43.
[3] Irwin JJ, Duan D, Torosyan H, Doak AK, Ziebart KT, Sterling T, et al. An Aggregation Advisor for Ligand Discovery. Journal of Medicinal Chemistry. 2015;58(17):7076-87.

L2-02

QSPRpred, Spock and DrugEx: Developing an Open Source Ecosystem for Cheminformatics

Šícho M.[1,2], van den Maagdenberg Helle W.[2], Bernatavicius A.[2,3], Svozil D.[1],van Westen G.[2]

[1] CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Department of Informatics and Chemistry, Faculty of Chemical Technology, University of Chemistry and Technology Prague, Technická 5, 166 28, Prague, Czech Republic
[2] Leiden Academic Centre for Drug Research, Leiden University, 55 Einsteinweg, 2333 CC, Leiden, The Netherlands
[3] Leiden Institute of Advanced Computer Science, Leiden University, Niels Bohrweg 1, 2333CA Leiden, The Netherlands.

This talk presents QSPRpred, Spock, and DrugEx, an open-source cheminformatics ecosystem designed to accelerate drug discovery by integrating predictive modeling, molecular docking, and generative chemistry. QSPRpred is a modular Python framework for building robust QSPR/QSAR models with automated preprocessing and hyperparameter optimization. Spock serves as an automated molecular docking framework that generates and stores ligand-protein complexes, enabling structure-based virtual screening to identify promising ligands. Complementing these, DrugEx leverages both Spock pipelines and QSPRpred models to guide de novo molecular generation. Together, these tools facilitate streamlined workflows from property prediction and docking to compound design, improving reproducibility and reducing complexity through interoperable, version-controlled components. Case studies demonstrate how this ecosystem enhances virtual screening campaigns by combining data-driven modeling with structure-based docking, supported by open-source development and community collaboration.

L2-03

Fragment-based de novo design and searching for hit molecules in ultra-large chemical libraries

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

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

A significant limitation of fragment-based structure generation is the poor synthetic accessibility of the structures produced. The previously established CReM framework [1] offers a potential solution to this issue. By integrating this framework with molecular docking, pharmacophore modeling, or machine learning, we have developed a suite of tools capable of addressing various tasks, including de novo design, hit generation, hit/lead optimization, and scaffold hopping. Nevertheless, the structures generated may still be syntheticaly infeasible. To address this issue, we proposed a pipeline that facilitates the rapid identification of hit molecules within ultra-large libraries of synthetically accessible compounds. The fundamental concept involves generating molecules de novo, selecting the most promising candidates, and utilizing these candidates for similarity searches within an ultra-large library. We validated this protocol during the first CACHE challenge, which aimed to identify binders for the WD40 domain of the LRRK2 kinase, a target that previously lacked known binders and for which only the X-ray structure of the domain was available. We employed the proposed strategy and designed promising hits de novo using the CReM-dock tool [2], which were subsequently used to search for similar molecules in Enamine REAL Space, containing approximately 23 billion structures at that time. As a result, out of 82 synthesized compounds, eight exhibited binding affinity with K_d values ranging from 25 to 117 μM, thereby confirming the efficacy of the proposed protocol. The outcomes of all top-performing teams have recently been published in a collaborative article [3].

The work was supported by the Ministry of Education, Youth and Sports of the Czech Republic through INTER-EXCELLENCE II LUAUS23262, the e-INFRA CZ (ID:90140, ID:90254), ELIXIR-CZ (LM2018131, LM2023055), CZ-OPENSCREEN (LM2018130, LM2023052) grants.

References

[1] Polishchuk, P. *J. Cheminf.* **2020**, 12 (1), 28.
[2] Minibaeva, G. et al. *ChemRxiv* **2024** - https://doi.org/10.26434/chemrxiv-2024-fpzqb-v3.
[3] Li, F. et al. *J. Chem. Inf. Model.* **2024**, 64 (22), 8521-8536.

L2-04

Ligand hallucinations in cofolding methods

Dehaen W.[1,2]

[1] Department of Organic Chemistry, Faculty of Chemical Technology, University of Chemistry and Technology Prague, Technická 5, 16 628 Prague 6, Czech Republic;
[2] Department of Informatics and Chemistry, Faculty of Chemical Technology, University of Chemistry and Technology Prague, Technická 5, 16 628 Prague 6, Czech Republic;

Cofolding methods such as the Nobel-prize winning AlphaFold3 and its closely related methods such as Boltz1, Chai and Protenix have revolutionized protein-ligand structure prediction. Unlike traditional pose prediction methods, these deep learning-based methods predict the structure of the protein and the ligand at the same time. On common benchmarks, these data-driven methods outperform or match molecular docking, the most commonly used traditional pose prediction method.

Nonetheless, some unusual issues plague these new methods. The authors of AlphaFold3 already point out that asymmetric carbons are frequently inverted in the output of their method, but otherwise they manage to pass standard validity checks. Surprisingly, we have found several examples where the input molecules (given as SMILES) did not match the output molecules (given as CIF). Some common problem motifs are: cyclohexanes (aromatized to benzenes), tetrahydrofurans (aromatized to furan), allenes, alkynes and nitriles (all 3 with non-linear structures). These can be considered an example of the “hallucination” phenomenon which is well known in other generative AI methods such as LLMs. These ligand hallucinations are also interesting because they subvert pose validation methods, being able to both pass PoseBusters as well as pass RMSD based checks compared to a reference.

We propose a method to systematically identify instances of cofolding ligand hallucinations by generating ligand structures, inferring their connectivity and comparing it to the input connectivity. We use this to compile a list of problematic functionalities. This will pave the way for correcting these issues (e.g. through data augmentation), which is of high priority to enhance the efficacy of the applications of cofolding methods on novel ligands.

L2-05

Quantum-Mechanics Based Multiscale Modeling and Rational Design of Insulin Analogs with Improved Binding Affinities

Yurenko Y.[1], Pecina A.[1], Řezáč J.[1], Fanfrlík J.[1], Lepšík M.[1]

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

Accurate modeling of electronic effects in protein–protein interactions (PPIs) remains a significant challenge in computational biophysics. Classical force fields fail to describe polarization and charge transfer, while full quantum mechanical (QM) methods are often computationally prohibitive. To address this, we developed a multiscale protocol that integrates molecular dynamics, system fragmentation, semiempirical quantum mechanics (PM6-D3H4S/COSMO2), and virtual glycine scanning (VGS) to dissect residue-level contributions to protein interface energetics [1].

Using the insulin–insulin receptor complex as a model, we applied this approach to evaluate the impact of point mutations at seven key positions within the receptor-binding region. The calculated changes in binding free energy (ΔΔG) show strong correlation with experimental binding data and clearly outperform classical Molecular Mechanics/Generalized Born (MM/GB) approaches. Notably, our scoring protocol is conceptually linked to the recently published SQM2.20 scoring function for protein–ligand systems, which yields DFT-quality predictions with practical speed [2].

Building on these insights, we designed a series of novel insulin analogs predicted to exhibit enhanced receptor binding. Several promising candidates—such as A19Phe, B16Arg, and B26Lys— are being synthesized and are currently undergoing experimental evaluation. These analogs were selected based on favorable QM-predicted interaction profiles, taking into account both individual residue contributions and the overall binding energetics. Preliminary experimental results confirm improved receptor binding for several variants, supporting the predictive power of our quantum-informed design strategy.

References

[1] Yurenko, Y *et al*. *ChemRxiv* 2024, DOI:10.26434/chemrxiv-2024-4jjrc

[2] Pecina, AY.; Fanfrlík, J; Lepšík, M.; Řezáč, J. *Nat. Commun.* **2024**, *15*, 1127

SESSION 3

Sequences

L3-01

Crazy avian genome, stuttering genes, and why we love nanopore

Hron T.[1], Miklík D.[1], Pačes J.[1], Nehyba J.[1], Elleder D.[1]

[1] *Institute of Molecular Genetics, Academy of Sciences of the Czech Republic, Vídeňská 1083, 14220, Prague, Czech Republic*

Avian genomics is full of mysteries. One of them concerns the apparent absence of many well-known vertebrate genes. This year marks a decade since we first reported that most of these avian “missing” genes do, in fact, exist, yet their sequences are so unusually problematic that they elude detection by standard sequencing technologies. The issue lies in their extreme sequence composition, characterized by a massive accumulation of short repeats and strongly biased nucleotide content. Intriguingly, these genes are found almost exclusively on the so-called avian dot chromosomes,yet another of birds' genomic curiosities.

Subsequent research has revealed that some genes — otherwise conserved across vertebrates — have indeed been lost during avian evolution. And once again, the trail leads back to the avian dot chromosomes, where these genes were originally located.

Apparently, avian dot chromosomes undergo a form of dynamic genetic instability that has not yet been characterized, and whose underlying mechanism remains unknown. So far the only viable approach to studying this phenomenon appears to be nanopore sequencing, which offers at least partial resistance to the extreme sequence bias present in these regions.

L3-02

Structural variation in closely related songbird species

Halenková Z.[1,2], Rídl J.[1,2], Cakl L.[1], Schlebusch S.[1,2], Albrecht T.[1,3], Reif J.[4], Reifová R.[1]

[1] Department of Zoology, Faculty of Science, Charles University, Prague, Czech Republic
[2] Institute of Molecular Genetics, Czech Academy of Sciences, Prague, Czech Republic
[3] Institute of Vertebrate Biology, Czech Academy of Sciences, Studenec, Czech Republic
[4] Institute of Environmental Studies, Faculty of Science, Charles University, Prague, Czech Republic

Structural variants, i.e. mutations altering the position, orientation, or number of copies of longer (> 50 bp) regions within the genome, represent a rich source of variation in the evolution of species. These genomic rearrangements can lead to the development of novel adaptations and act as key factors in the establishment and maintenance of reproductive barriers between species, although direct evidence linking structural variants to speciation remains scarce. In our research, we focus on structural variation in two closely related songbird species, the common nightingale and the thrush nightingale, which diverged approximately 1.8 Mya, but still occasionally hybridize in their secondary contact zone spanning Central Europe. We employed a combination of bioinformatic approaches to detect structural variants in these species, comparing and integrating their results to gain an insight into the rearrangements that differentiate the nightingales' genomes. Through bioinformatic analyses, we were able to identify events that remained undetected in previous karyotype analyses. Among those, there were also 13 species-specific inversions ranging in length from 100 kbp to 2.4 Mbp, many located in the pericentromeric region. We hypothesize that these pericentric inversions might amplify the recombination suppression effect of centromeres and thereby reduce the rate of interspecific recombinant genotypes. This mechanism could facilitate speciation despite ongoing gene flow between the species.

L3-03

Plasmid-Mediated Antibiotic Resistance Dynamics in Broiler Chickens Revealed by Long-Read Sequencing

Ryšavá M.[1,2,3], Palkovičová J.[2,3], Schwarzerová J.[4,5,6], Čejková D.[4], Jakubíčková M.[4], Dolejská M.[1,2,3,7,8]

[1] Department of Biology and Wildlife Diseases, Faculty of Veterinary Hygiene and Ecology, University of Veterinary Sciences Brno, Brno, Czech Republic
[2] Central European Institute of Technology, University of Veterinary Sciences Brno, Brno, Czech Republic
[3] Department of Microbiology, Faculty of Medicine in Pilsen, Charles University, Pilsen, Czech Republic
[4] Department of Biomedical Engineering, Faculty of Electrical Engineering and Communication, Brno University of Technology, Czech Republic
[5] Molecular Systems Biology (MOSYS), Department of Functional and Evolutionary Ecology, Faculty of Life Sciences, University of Vienna, Vienna, Austria
[6] Department of Molecular and Clinical Pathology and Medical Genetics, University Hospital Ostrava, Ostrava, Czech Republic
[7] Department of Chemistry and Biochemistry, Faculty of AgriSciences, Mendel University in Brno, Brno, Czech Republic
[8] Division of Clinical Microbiology and Immunology, Department of Laboratory Medicine, University Hospital Brno, Brno, Czech Republic

The widespread use of antibiotics in human and veterinary medicine has accelerated antimicrobial resistance (AMR), primarily through horizontal gene transfer. This process allows bacteria to exchange antibiotic resistance genes (ARGs) via mobile genetic elements like plasmids. Antibiotic use creates ARG reservoirs in bacterial populations, promoting the spread of resistance and contributing to multidrug-resistant pathogens. Intensive poultry farming has notably contributed to AMR, with excessive antibiotic use promoting the accumulation of ARGs in the chicken gut microbiome. These ARGs, often carried on plasmids, can transfer between bacteria, including pathogens, posing a risk to human health via food or environmental exposure.

In this study, long-read sequencing was used to analyze the plasmidome and resistome in 12 fecal samples collected from three chicken barns on a commercial farm. All chickens received enrofloxacin during the first days of life, with one barn was additionally treated with sulfamethoxazole/trimethoprim. For comparison, short-read metagenomic sequencing was also performed on the same samples.

The analysis revealed ARGs associated with resistance to 26 antibiotic classes, highlighting the diversity of the poultry resistome. Strong genetic links were observed between specific plasmids and ARGs, with MOBP plasmids frequently associated with fluoroquinolone resistance. Temporal trends indicated progressive mobilization of resistance genes, suggesting an increasing potential for horizontal gene transfer. While fluoroquinolone resistance expanded, diaminopyrimidine resistance remained stable despite treatment. Most ARGs were carried on small plasmids (2.6 to 47.6 kb); several exhibited high similarity to plasmids from other bacterial species, indicating potential for cross-species gene transfer, with possible transmission from poultry to humans through shared agricultural or environmental spaces. Most ARGs were found on small plasmids ranging from 2.6 to 47.6 kb. Several plasmids showed high sequence similarity to those from other bacterial species, indicating a potential for cross-species transfer and possible transmission from poultry to humans through shared agricultural or environmental pathways.

These findings provide new insights into the dynamics of the poultry gut microbiome and antimicrobial resistance, emphasizing the central role of plasmids in gene mobilization. The results underscore the importance of sustained monitoring and targeted strategies to mitigate the spread of antimicrobial resistance in both agricultural and clinical environments.

The study has been supported by the Czech Science Foundation (22-16786S).

L3-04

Analyzing Retrotransposons in RNA Data: Can I Hand It Over to AI or Do I Still Have to Work?

Ostašov P.[1,2], Klieber R.[1,3], Dekojová T.[1,3], Holubová M.[1,3,4]

[1] *Charles University, Faculty of Medicine in Pilsem, Laboratory of Tumour Biology and Immunotherapy*
[2] *Charles University, Faculty of Medicine in Pilsem, Department of Histology and Embryology*
[3] *University Hospital Pilsen, Department of Hematology and Oncology*
[4] *Consortium for iNKT Research and Therapy*

Background
With the growing popularity of AI in the form of LLMs for coding, we tested whether such models could simplify complex bioinformatics tasks sufficiently for non-experts to generate not only basic analysis code but also scripts for less common tasks such as retrotransposable element (TE) analysis.
Aims
We evaluated whether LLMs could set up the environment and perform TE analysis with a single prompt: link samples to FASTQ files, align reads, quantify gene and TE family expression, and compare TE expression between two groups of patients, where TE expression could be altered by a disease status and affect immune response.
Methods
Cloud-based models (GPT-o3, GPT-4o, Sonnet, Gemini) and local models (Qwen3, Gemma3, DeepSeek-r1) were given a one-shot TE analysis task. The results were evaluated for workflow design, code quality, and whether a non-expert could identify and correct errors. Analyzed data were represented by RNA-seq of invariant NKT cells from two groups of myeloma patients with partial or complete remission.
Results
All models produced errors, including missing dependencies, broken data links, and flawed logic. GPT-o3 even tried to implement its own method for differential expression analysis. Cloud models generated compact scripts; local models provided more stepwise instructions. RNA-seq alignment and basic differential expression were mostly correct, but TE analysis and environment setup often failed. No model created a fully functional workflow. Reasoning ability and model type did not consistently predict better performance.
Conclusion
LLMs can assist in generating analysis code but cannot yet replace experts, especially in niche or emerging fields where limited training data is available. Dividing the task into smaller subtasks would improve accuracy but still risk guiding users into dead ends.

Funding
Supported by the Ministry of Health of the Czech Republic in cooperation with the Czech Health Research Council under project No. NW24-03-00079.

SESSION 4

Tools

L4-01

5 years of Alphafoldology

Berka K.[1]

[1] *Department of Physical Chemistry, Faculty of Science, Palacký University Olomouc, karel.berka@upol.cz*

AlphaFold 2 was first presented at CASP 14 in December 2020. It was a major success in protein structure prediction, leading to a Nobel Prize in Chemistry last year. Especially since AlphaFoldDB and AlphaFold 2 source code opened in July 2021, the field of structural biology has moved towards the full structural description of all possible proteins with known and, to some extent, even unknown sequences. The speed and breadth of post-AlphaFold tools and services development, for which Marian Novotný and I coined the term „Alphafoldology“ in September 2021, has been breathtaking. In the lecture, I will review the current state-of-the-art of the Alphafoldology field and what usage it enables for biology and chemistry.

L4-02

MARTS-DB: The Mechanisms And Reactions of Terpene Synthases DataBase

Engst M.[1,2], Brokeš M.[1,2], Čalounová T.[1], Tajovská A.[1], Perković M.[1], Samusevich R.[1], Chatpatanasiri R.[3], Pluskal T.[1]

[1] Institute of Organic Chemistry and Biochemistry of the Czech Academy of Sciences, Flemingovo náměstí 2, Prague, Czech Republic
[2] Faculty of Chemical Technology, University of Chemistry and Technology Prague, Technická 5, Prague, Czech Republic
[3] Czech Institute of Informatics, Robotics and Cybernetics (CIIRC), Czech Technical University, Jugoslávských partyzánů 1580/3, 160 00 Prague, Czech Republic

Terpene synthases (TPS) are enzymes that catalyze some of the most complex reactions in nature, the cyclizations of terpenes, the carbon backbones to the largest group of natural products, terpenoids. On average, more than half of the carbon atoms in a terpene scaffold undergo a change in connectivity or configuration during these enzymatic cascades. Understanding TPS reaction mechanisms remains challenging, often requiring intricate computational modeling and isotopic labeling studies. Moreover, the relationship between TPS sequence and catalytic function is difficult to decipher, and data-driven approaches remain limited due to the lack of comprehensive, high-quality data sources. To address this gap, we introduce the Mechanisms And Reactions of Terpene Synthases DataBase (MARTS-DB)—a manually curated, structured, and searchable database that integrates TPS enzymes, the terpenes they produce, and their detailed reaction mechanisms. MARTS-DB includes over 2,600 reactions catalyzed by 1,334 annotated enzymes from across all domains of life. Where available, reaction mechanisms are mapped as stepwise cascades. Accessible at https://www.marts-db.org, the database provides advanced search functionality and supports full dataset download in machine-readable formats. It also encourages community contributions to promote continuous growth. By enabling systematic exploration of TPS catalysis, MARTS-DB opens new avenues for computational analysis and machine learning, as recently demonstrated in the prediction of novel terpene synthases.

L4-03

mutation_scatter_plot: Tackling codon usage analysis from a different angle

Mokrejš M.[1], Zahradník J.[1]

[1] *First Faculty of Medicine, Charles University, BIOCEV, Prumyslova st. 595, Vestec 25250, Czechia*

Codon usage is commonly used in phylogenetics to unleash evolution of protein-coding regions, in particular, as the ratio between synonymous and non-synonymous changes. We wanted to study multiple sequence alignment of the Covid19 S protein in a vast set of raw NGS reads and determine which codons out of the theoretical 64 are used in every amino acid position of the encoded protein. To our surprise we could not find a tool to achieve that. Polishing the multiple sequence alignment spanning dozens of millions of entries is not possible with conventional tools although mostly just some gaps would need to be introduced here and there. The major obstacle is introducing padding gaps into the reference sequence to facilitate recognition of INSertion events in the sample reads. First, we developed rather simple program *calculate_codon_frequencies.py* to count the codons occurring in three columns of the DNA alignment while moving along the reference sequence and keeping ribosome reading-frame of the CDS region and output TSV files with their frequencies. Alternatively, the same tool can provide frequencies of the encoded amino acid residues. Second, we developed *mutation_scatter_plot.py* to display the frequencies as scatter plots with interactive bubbles upon mouse hover(). The changes can be color-coded according to e.g. physicochemical properties of the amino acid residues (PAM matrices) or their evolutionary conservation (BLOSUM matrices) or any other color-palette. However, such efforts are a bit naïve as the weights for each amino acid are not within the same minimum-maximum range and thus are not directly comparable. The software is available at https://github.com/host-patho-evo/mutation_scatter_plot.

This was in part supported by Czech Science Foundation Grant No. 25-17643M: „Unveiling Divergence and Convergence Points in Coronavirus Evolution for Host Receptor Recognition“ and by National Institute of virology and bacteriology (Programme EXCELES, ID Project No. LX22NPO5103).

L4-04

Predicting Gene Regulatory Networks with Augusta

Sedlar K.[1], Musilova J.[1,2], Hermankova K.[1], Vafek Z.[2,3], Zimmer R.[4], Helikar T.[2]

[1] *Brno University of Technology, Department of Biomedical Engineering, BioSys lab*
[2] *University of Nebraska-Lincoln, Department of Biochemistry*
[3] *Brno University of Technology, Institute of Forensic Engineering*
[4] *Ludwig-Maximilians-Universität München, Department of Informatics*

Gene Regulatory Network (GRN) inference using transcriptomic data became a relatively common procedure in model organisms, where high-quality single-cell sequencing data, required by bioinformatics tools, are available. Unfortunately, similar data cannot be produced for non-conventional bacteria due to the lack of single-cell culturing techniques. We addressed this limitation by developing the Augusta tool specifically designed for bulk RNA-Seq and non-model organisms. Augusta performs several polishing steps that help remove inherent bias in bulk sequencing data including database searches and sequence motifs predictions. Optionally, Augusta transforms the static GRN into dynamic Boolean Networks (BN) suitable for following dynamic analyses.

GRN inference requires time series data and the precision is highly dependent on sampling that needs to be dense enough to cover all regulatory changes. Nevertheless, there are potentially many applications where the most important regulations happen in relatively long periods and can be inferred from bulk data as the majority of cells in culture are metabolically synchronized. Here, we demonstrate that using an example of *Caldimonas thermodepolymerans*, a non-conventional bacterium and a potent producer of PolyHydroxyAlkanoates, biologically produced polymers that could be used as plastics. Even non-uniformly sampled time series are sufficient to capture important changes coupled with PHA synthesis.

Augusta was developed as an open-source Python package and is available from github.com/JanaMus/Augusta along with documentation, examples, and tutorials.

This project has been supported by grant project GACR Junior Star 25-17459M.

L4-05

Negative Reference Extraction for Machine-Assisted Flow Cytometry Immunophenotyping

Koza A.[1], Fišer K.[1]

[1] *Department of Bioinformatics, Second Faculty of Medicine, Charles University, Prague, Czech Republic*

Flow cytometry is a method for quantitatively measuring multiple parameters on millions of individual cells. This capability is widely used in both research and diagnostics. For example, in hemato-oncology, identifying malignant cell populations by flow cytometry is a crucial part of diagnosis and disease monitoring. Beyond identifying cell populations, their description, known as the immunophenotype, is essential for clinical decision-making.

Currently, immunophenotypes are often reported based on individual expert judgment. We have previously developed a system for automatically deriving immunophenotypes from data with labeled cells of interest and cells serving as a negative reference. However, in most cytometric data, the appropriate negative reference population is unknown. This is either because it is unclear which of the labeled populations should serve as negative, or whether each dimension of the data requires its own reference, or the negative reference is not among the labeled populations at all.

Here, we present possible approaches to identify the negative reference in flow cytometry data. In particular: 1) selection of one of the labeled populations as a negative reference using maximal distance to other populations or by the absolute position of its mode; 2) extraction of a reference for each dimension of the data separately, either for the constitution of an artificial joined population or for subsequent per-dimension analysis; and 3) an attempt at extracting Gaussian mixtures from individual dimensions on unlabeled data under the assumption that flow cytometry data parameters are normally distributed within cell populations.

Overall, we demonstrate different approaches for the automatic detection of negative cell populations to facilitate automatic immunophenotyping of flow cytometry data.

SESSION 5

Sequences

L5-01

Modeling changes in cell populations between groups of samples of scRNA-seq data

Modrák M.[1], Niederlová V.[2,3], Neuman V.[4], Šumník Z.[4], Štěpánek O.[2]

[1] Department of Bioinformatics, Second Faculty of Medicine, Charles University, Prague, Czech Republic
[2] Laboratory of Adaptive Immunity, Institute of Molecular Genetics of the Czech Academy of Sciences, Prague, Czech Republic.
[3] Department of Cell Biology, Faculty of Science, Charles University in Prague, Czech Republic.
[4] Department of Pediatrics, Second Faculty of Medicine, Charles University in Prague & Motol University Hospital, Prague, Czechia

Once cell populations have been identified in single cell sequencing data, it may be desirable to understand how the abundance of those populations changes between two or more groups of samples. Taking the assignment of cells to populations as given, two interesting issues arise: First, we need to address the compositional nature of the data, i.e. that the total number of cells sequenced is determined by the experimental design and we thus obtain only relative abundances. Second problem is that cell populations typically form a hierarchical structure. This needs to be reflected in modeling and forces us to ask whether it is more meaningful to look at abundance relative to the total or relative to parent population.

We argue that the compositional structure could be relatively easily bypassed as measuring the total number of cells per sample is often possible and even when it is not, the problem is not huge. We further show that it is natural to model the hierarchical structure in a hierarchical model (i.e. nested random effects) – this model can then answer questions about both absolute abundance and abundance relative to a parent population. Finally, we take note of an – as far as we are aware – unsolved problem of disentangling discrete population structure and continuous latent states of cells within population (cell cycle, proliferation).

We illustrate the issues and possible solutions with an example of comparing populations of peripheral blood T-cells from children newly diagnosed with type 1 diabetes mellitus and healthy donors.

L5-02

Ancient DNA Computational Genomics: Tracing Human Origins and Mobility in Prehistoric Europe

Ehler E.[1]

[1] *Laboratory of Genomics and Bioinformatics, Institute of Molecular Genetics of the Czech Academy of Sciences, Vídeňská 1083, 142 00 Prague 4, Czech Republic, email: edvard.ehler@img.cas.cz*

Ancient DNA (aDNA) research has been at the forefront of prehistory studies for the past decade. It has shed light on various aspects of human history, from the formation of our species and interactions with other members of the genus Homo, to the spread of Paleolithic hunter-gatherer populations. Additionally, aDNA has provided insights into the advent and expansion of agriculture, the development of metal processing technologies, the rise and fall of empires, and the life stories of significant individuals. Furthermore, aDNA has revealed patterns of human migrations and the expansion of steppe populations, enriching our understanding of how these movements shaped societies. Through aDNA, we can now tell these stories with greater accuracy and detail.

In my talk, I will briefly summarize the topic aDNA sampling, processing, and the challenges in interpreting the results, along with the main findings of aDNA research. I will focus on Central Europe, particularly the Late Bronze Age and Early Iron Age. During this period, the widespread practice of cremation—characterized by the Urnfield phenomenon—created significant gaps in our understanding of the genetic history of the Lusatian, Knovíz-Štítary, and Hallstatt cultures. Our research project has assembled a unique collection of over 150 human skeletal samples from archaeological sites in Upper Silesia (Poland), and the Moravia and Bohemia regions (Czech Republic). These samples are being sequenced using the Illumina NovaSeq X platform. Our aim is to assess the genetic backgrounds of populations associated with these cultures and the gene flow between populations favoring inhumation practices. Kinship analysis among densely sampled archaeological sites will provide further insights into marriage patterns, social structures, and inheritance practices. To achieve these objectives, we employ a multidisciplinary approach, integrating aDNA analysis, radiocarbon dating, and isotopic measurements (δ15N, 87Sr/86Sr, δ18O, and δ13C). I will present the initial results from sequencing screening of our sample set and discuss our bioinformatics approach, including authenticity and contamination measurement of the aDNA samples, and insights into planned population genomics analyses.

This study is supported by Czech Science Foundation (GAČR), project No. 24-14385L and by Ministry of Education, Youth, and Sport (MŠMT), project No. LM2023055.

L5-03

Identifying intercellular patterns in Vestibular schwannoma

Pfeiferová L.[1,2], Kolář M.[1,2]

[1] *Laboratory of Genomics and Bioinformatics, Institute of Molecular Genetics of the Czech Academy of Sciences, Prague, Czech Republic*
[2] *Department of Informatics and Chemistry, Faculty of Chemistry and Technology, University of Chemistry and Technology Prague, Prague, Czech Republic*

Cell–cell interactions orchestrate key processes in tumor development and immune regulation. In this study, we investigate the spatial architecture of vestibular schwannoma (VS) using a combination of spatial transcriptomics and advanced computational modeling. We applied the MISTy (Multiview Intercellular SpaTial modeling) framework to identify intercellular signaling patterns and to assess how local cellular environments influence gene expression profiles. To complement this, we employed Ficture, a segmentation-free approach, to define spatial domains based on transcriptional similarity rather than predefined cell types. This allowed us to detect functionally distinct regions within the tumor that may reflect structural or signaling compartments relevant to VS biology. By integrating these analyses, our approach provides a comprehensive view of the spatial signaling landscape in VS. This work highlights the power of combining segmentation-free methods with spatially informed intercellular modeling to uncover the organizational of the tumor microenvironment.

SESSION 6

Sequences

L6-01

Bioinformatics approaches to studying the relationships between structure, function and evolution of amylolytic enzymes – the case of the alpha-amylase family GH57

Janeček Š.[1,2]

[1] Laboratory of Protein Evolution, Institute of Molecular Biology, Slovak Academy of Sciences, Dubravska cesta 21, SK-84551 Bratislava, Slovakia; e-mail: Stefan.Janecek@savba.sk
[2] Institute of Biology and Biotechnology, Faculty of Natural Sciences, University of SS Cyril and Methodius, Nam. J. Herdu 2, SK-91701 Trnava, Slovakia

In the sequence-based classification of carbohydrate-active enzymes, the CAZy database (https://www.cazy.org/), four glycoside hydrolase (GH) families – GH13, GH57, GH119 and GH126 – are known as the alpha-amylase families. The family GH57, created in 1996 and named as the second alpha-amylase family has recently been classified within the clan GH-T with family GH119. It counts more than 5,800 sequences from Bacteria and Archaea, but only 44 its members have already been experimentally characterized as various amylolytic enzymes. The catalytic domain adopts an incomplete TIM-barrel succeeded by a bundle of alpha-helices with the catalytic machinery formed by a glutamic acid and aspartic acid at the strands, respectively, beta-4 and beta-7 of the barrel domain. The first in silico analysis of the family GH57 was performed in 2004 when the five conserved sequence regions (CSRs) characteristic for the family were defined. Later, in 2012, the five well-established family GH57 CSRs were described as the so-called "sequence fingerprints" containing the unique sequence features typical for the individual enzyme specificities of the family. Subsequent 2018 bioinformatics study of ~1,600 GH57 members delivered clusters in the phylogenetic tree reflecting the eight distinct enzyme specificities. Finally, in 2025, based on a detailed in silico analysis of a dataset of ~5,000 sequences, the family GH57 has been officially divided into ten subfamilies. Importantly, each GH57 subfamily can be characterized by its sequence fingerprints, i.e. the logo of the five GH57 CSRs.

L6-02

Streamlined workflow for bacterial methylation analysis using nanopore data

Jakubíčková M.[1], Šabatová K.[1], Bezdíček M.[2,3], Lengerová M.[2,3], Vítková H.[1]

[1] Department of Biomedical Engineering, Faculty of Electrical Engineering and Communication, Brno University of Technology, Brno, Czechia
[2] Department of Internal Medicine – Hematology and Oncology, University Hospital Brno, Brno, Czech Republic
[3] Department of Internal Medicine – Hematology and Oncology, Masaryk University, Brno, Czech Republic

Recent advantages in sequencing technologies enable direct detection of DNA methylation such as N^4-methylcytosine, 5-methylcytosine or N^6-methyladenine without further chemical treatment or specific protocol usage. Therefore, efficiently studying bacterial epigenetics at the whole-genome level is possible. Although obtaining methylation data is straightforward, their further postprocessing remains challenging as there is a lack of comprehensive tools that connect the methylation information with the functional genomic context. Therefore, we proposed a comprehensive computational pipeline for genome-wide methylation profiling using nanopore sequencing data. The proposed approach consists of basecalling and methylation detection using Dorado, followed by quality filtering. Then, *de novo* assembly of genomes is performed, and reads with detected methylation are mapped. Methylation calls are exported as bedMethyl files and further processed using our designed tool MethylomeMiner to link them with genomic features such as genes and intergenic regions. Additionally, it is possible to assign functional categories to methylated genes via COG classification. The pipeline was tested on 10 *Klebsiella pneumoniae* genomes sequenced using the ONT P2S platform. Over 60,000 methylated positions were detected per genome, with N^6-methyladenine being the most abundant type. Approximately 47% of methylation sites were located in coding regions, and many affected genes were involved in key cellular functions such as transcription and metabolism. Our solution provides a practical framework for routine exploration of bacterial methylomes, enabling insights into epigenetic regulation and its potential impact on gene expression, adaptation, and pathogenicity.

Acknowledgement: This work was supported by GA23-05845S.

L6-03

Rapid genetic change in the passerine germline restricted chromosome

Schlebusch S. A.[1], Rídl J.[1,2], Poignet M.[1,3], Ruiz-Ruano F. J.[4,5,6,7], Reif J.[8,9], Pajer P.[10], Pačes J.[2], Albrecht T.[1,11], Suh A.[4,5,6], Reifová R.[1]

[1] Department of Zoology, Faculty of Science, Charles University, Prague, Czech Republic
[2] Institute of Molecular Genetics, Czech Academy of Sciences, Prague, Czech Republic
[3] Present address:Department of Biology, Insititute of Life-Earth-Environment, Universirty of Namur, Namur, Belgium
[4] Present address: Centre for Molecular Biodiversity Research, Leibniz Institute for the Analysis of Biodiversity Change, Adenauerallee 127, 53113, Bonn, Germany
[5] School of Biological Sciences, University of East Anglia, Norwich, UK
[6] Department of Organismal Biology – Systematic Biology, Evolutionary Biology Centre, Science for Life Laboratory, Uppsala University, Norbyvägen 18D, 752 36, Uppsala, Sweden
[7] Institute of Evolutionary Biology and Ecology, University of Bonn, An der Immenburg 1, 53121, Bonn, Germany
[8] Institute for Environmental Studies, Faculty of Science, Charles University, Prague, Czech Republic
[9] Department of Zoology, Faculty of Science, Palacky University, Olomouc, Czech Republic
[10] Military Health Institute, Military Medical Agency, Tychonova 1, 160 01, Prague 6, San Antonio, Czech Republic
[11] Institute of Vertebrate Biology, Czech Academy of Sciences, Brno, Czech Republic

The passerine germline-restricted chromosome (GRC) represents a taxonomically widespread example of programmed DNA elimination. This chromosome's apparent ubiquity in the order suggests that it is indispensable, but we know little about the GRC's genetic composition, function, and evolutionary significance. We sequenced the testis (with the GRC) and kidney (without the GRC) of the closely related common and thrush nightingale and compared the sequencing libraries to identify GRC derived reads and assemble the two GRC. In total we identify 192 different genes across the two GRC, with many of them present in multiple copies and often appearing as pseudogenized fragments. Interestingly, the genetic content of the GRC differs dramatically between the two species, despite only 1.8 million years of species divergence. Only one gene, cpeb1, has a complete coding region in all examined individuals of the two species and shows no copy number variation.

The acquisition of this gene by the GRC corresponds with the earliest estimates of the GRC origin. Altogether, this suggests that the GRC is under little selective pressure, with rapid changes in genetic content observed and many genes potentially being non-functional pseudogene fragments. The standout nature of cpeb1, a gene known to play a function during oocyte maturation and early embryonic development, makes it a good candidate for the functional indispensability of the passerine GRC.

Poster session

Monday, 9. June

Poster session is sponsored by the
National Infrastructure for Chemical Biology.

CZ OPENSCREEN
National Infrastructure for Chemical Biology

P-01

In-silico design towards MAO-B selective covalent inhibitors based on Rasagiline

Aksu A.[1,2], Dehaen W.[2,3], Sicho M.[2], Svozil D.[2], Yurtsever M.[1] , Sungur Fethiye A.[4]

[1] Chemistry Department, Istanbul Technical University, Maslak, Istanbul, Turkey
[2] Department of Informatics and Chemistry, Faculty of Chemical Technology, UCT Prague, Czechia
[3] Department of Organic Chemistry, Faculty of Chemical Technology, UCT Prague, Czechia
[4] Computational Science and Engineering Division, Informatics Institute, Istanbul Technical University, Istanbul, Turkey

Parkinson's disease (PD) is a neurodegenerative disorder that results from oxidative stress in the central nervous system (CNS). It is known that the B isoform of the monoamine oxidase enzyme (MAO-B) contributes to this stress, so inhibiting it is essential for treating PD. While several drugs known as MAO-B inhibitors are currently available on the market, they can cause significant side effects and also engage the other isoform of the MAO enzyme, known as MAO-A. [1]

In this study, we aim to develop novel MAO-B inhibitors using in-silico methods, including molecular docking, virtual screening and molecular generation. In the crystal structure of the covalently bound rasagiline-MAO-B complex (pdb:1s2q) the 4-position of rasagiline overlaps with the entrance cavity[2] of the MAO-B active site and thus, modification at this position offers promising potential for the design of ligands. Using molecular generators as well as manual design based on structural knowledge, we explore diverse substitutions in this position, and evaluate them using covalent and non-covalent molecular docking and molecular dynamics simulations.

This results in a set of de novo designed new potential MAO-B ligands that will be prioritized for synthesis and testing in the future.

References

[1] del Pozo, J. S. G., et al (2013). Rasagiline meta-analysis: A spotlight on clinical safety and adverse events when treating Parkinson’s disease. In Expert Opinion on Drug Safety (Vol. 12, Issue 4, pp. 479–486). https://doi.org/10.1517/14740338.2013.790956

[2] Milczek, E. M., et al (2011). The “gating” residues Ile199 and Tyr326 in human monoamine oxidase B function in substrate and inhibitor recognition. FEBS Journal, 278(24), 4860–4869. https://doi.org/10.1111/j.1742-4658.2011.08386.x

P-02

Probing the interactions between Ipragliflozin with RAGE for treating Alzheimer's disease: An *in-silico* drug repurposing approach

Bhogal I.[1], Pankaj V.[1], Roy S.[1]

[1] *Department of Biomedical Engineering, Faculty of Electrical Engineering and Communication, Brno University of Technology, Brno, 616 00, Czech Republic*

Alzheimer's disease (AD), a progressive neurodegenerative condition, is characterized by the accumulation of amyloid beta peptides and neurofibrillary tangles. The receptor for advanced glycation end products (RAGE) is a multiligand receptor of the immunoglobulin superfamily that has emerged as a crucial target for various complex diseases, including AD. Repurposing existing drugs offers a reliable and cost-effective approach in drug discovery. The current study aims to investigate the possibility of repurposed drugs for treating AD via targeting RAGE. A library of 4,541 repurposed molecules from ChemDiv was used for structure-based virtual screening. Compounds were then filtered using molecular docking, molecular dynamics (MD) simulations, molecular mechanics generalized born surface area (MM/GBSA) binding free energy calculations, principal component analysis (PCA), and absorption, distribution, metabolism, excretion, and toxicity (ADMET) to assess binding affinity and stability of the top-ranked compounds. Five repurposed drugs were predicted as potential RAGE inhibitors based on docking scores ranging from -5.79 to -8.61 kcal/mol. The MD simulation and MM/GBSA studies elucidated the conformational dynamics and stability of predicted repurposed drugs, and Ipragliflozin emerged as a significant binder of RAGE. These findings suggest that Ipragliflozin may offer therapeutic potential for treating neurodegenerative diseases, especially Alzheimer's disease, after further validation through *in vivo* and *in vitro* studies.

Keywords
RAGE, Drug repurposing, Alzheimer's disease, Ipragliflozin, Virtual screening, Molecular dynamics simulation

P-03

Detection of positive selection in songbird spermatozoa

Cakl L.[1], Stopka P.[1], Albrecht T.[1,2], Reif J.[3], Schlebusch S.[1], Reifová R.[1]

[1] *Department of Zoology, Faculty of Science, Charles University, Prague*
[2] *Institute of Vertebrate Biology, Czech Academy of Sciences*
[3] *Institute of Environmental Studies, Faculty of Science, Charles University, Prague*

Spermatozoa show very high variability in morphology and performance between species. Those differences are mostly driven by postcopulatory sexual selection and/or sexual antagonism. The molecular mechanisms underlying this variation are, however, still poorly understood. Here, we have investigated the molecular evolution of proteins expressed in spermatozoa across passerines, the largest group of birds consisting of over 6500 species. Passerine spermatozoa show a unique helical shape and swim by rapidly rotating around their longitudinal axis. Their morphology is also highly diverse among the species. Out of the 940 genes whose protein products have been detected in spermatozoa, we have found 22 which are evolving under positive selection. Gene ontology analysis has revealed an overrepresentation of biological process and molecular functions related to microtubule-based movement and microtubule motor activity respectively. Our results bring the first insight into the molecular mechanisms which might drive sperm evolution in passerines.

P-04

Explainable AI for Pharmacophore-Based Drug Activity Prediction

Ceklarz J.[1], Waniová K.[2], Dehaen W.[1], Danel T.[3], Šícho M[1]

[1] CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Department of Informatics and Chemistry, Faculty of Chemical Technology, University of Chemistry and Technology Prague, Technická 5, 166 28 Prague, Czech Republic
[2] Faculty of Mathematics and Computer Science, Jagiellonian University, Kraków, Poland
[3] Faculty of Chemistry, Jagiellonian University, Kraków, Poland

Pharmacophore representations are commonly used by medicinal chemists to identify and visualize structures necessary for biological function. Using them as representations of molecules for Graph Neural Network (GNN) training has yet untapped potential in deep learning. Deep learning architectures, to which GNNs belong, despite their excellent performance in many areas, come with one critical disadvantage - reduced or non-existent comprehensibility on *how* they reach their results. Many GNN-specific methods have been developed to answer questions about both feature and structural importance. When combined with the proposed pharmacophore representations, these methods could provide valuable insights to model users. Their application to chemical data, however, remains largely unexplored. We have developed two GNN models, a Graph Convolutional Network and a Graph Isomorphism Network, trained on 2D pharmacophore representations of small molecules, for drug activity prediction. We compare the results against shallow models, and against GNNs trained on traditionally used atomic representations of molecules. Using a selection of techniques, we aim to explain results of such models. We plan to obtain both local (molecule-level) and global (model-level) explanations, allowing us to analyse individual predictions as well as overarching model behaviour, to help identify the sources of errors and refine our models accordingly.

P-05

Challenges (and solutions?) in Xenium Spatial Transcriptomics Data Analysis.

Delattre M.[1], Kubovciak J.[1], Kolář M.[1]

[1] *Laboratory of Genomics and Bioinformatics, Institute of Molecular Genetics of the Czech Academy of Sciences, 14200 Prague, Czech Republic*

Our laboratory was recently equipped with the new Xenium instrument. Like other spatial transcriptomics platforms, Xenium can detect transcripts within a tissue sample. However, unlike many other methods, it is able to capture the exact position of transcripts, allowing to explore transcripts in subcellular spatial context. After our first Xenium run we encountered several challenges, such as an unexpectedly low transcript count. In this poster, we present the protocol we followed to analyze the data, how we tried to adapt to these challenges, and other key lessons we learned in the process. We are eager to discuss our experience with Xenium data, and hope to share practical insights and reusable code strategies that may benefit others working with spatial omics data in real-world research contexts.

P-06

Advancing Evaluation: Recall-Based Metrics for Molecular Generators

Fil V.[1], Svozil D.[1,2]

[1] Department of Informatics and Chemistry & CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Faculty of Chemical Technology, University of Chemistry and Technology, Technická 5, 16628, Prague, Czech Republic.
[2] CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Institute of Molecular Genetics of the Czech Academy of Sciences, Prague, Czech Republic.

Molecular generators enable the systematic exploration of chemical space to identify novel compounds with desirable properties. However, assessing their performance remains challenging due to the structural diversity and volume of the generated molecules. Common evaluation metrics, focused on chemical validity and novelty, do not fully align with the primary goal of molecular generation: the discovery of new biologically active compounds. To address this limitation, we introduce scaffold-based recall metrics that evaluate a generator's ability to recover biologically relevant scaffolds absent from the input set. Using these metrics, we assessed several molecular generators, including Molpher and DrugEx. The DrugEx Graph Transformer demonstrated the highest scaffold recall and scaffold hopping potential. The proposed recall-based metrics thus provide a more biologically meaningful framework for evaluating molecular generators and optimizing their performance to enhance the design of virtual libraries for drug discovery.

P-07

De Novo Drug Design: Combining Pharmacophore Modeling, 3D Shape Matching, and Generative AI

Fülöp J.[1], Šícho M.[1], Svozil D.[1]

[1] *Department of Informatics and Chemistry & CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Faculty of Chemical Technology, University of Chemistry and Technology, Technická 5, 16628, Prague, Czech Republic*

Recent work shows that shape-based generative reinforcement-learning models enable scaffold hopping and diverse lead discovery [1], while voxel-based shape autoencoders can uncover novel scaffolds from 3-D inputs [2]. Pharmacophore modeling that exploits shape-and-colour Tanimoto scores (e.g., ROCS) reliably detects compounds sharing key 3-D features beyond close analogues [3]. Building on this, we introduce an open-source pipeline that (i) aligns ligands by shape and pharmacophoric "colour," (ii) filters for high Tanimoto shape/colour similarity and synthetic accessibility, and (iii) uses DrugEx to propose new chemotypes meeting both criteria [4]. Using the chemokine receptor CCR2, an allosteric target involved in inflammation and metastasis, as a case study [5], we benchmark ROCS/OMEGA against RDKit/CDPKit [6, 7] and find speed-versus-optimisation trade-offs on large libraries, yet comparable overlay accuracy. This integrated workflow therefore provides a practical open-source alternative for next-generation lead discovery.

References

[1] Papadopoulos et al., *Bioorg. Med. Chem.*, 2021
[2] Skalic et al., *J. Chem. Inf. Model.*, 2019
[3] Kearnes & Pande, *J. Comput. Aided Mol. Des.*, 2016
[4] Šícho et al., *J. Chem. Inf. Model.*, 2023
[5] Zheng et al., *Nature*, 2016
[6] RDKit: www.rdkit.org
[7] CDPKit: github.com/molinfovienna/CDPKit

P-08

TRIPLE: Transforming RDF Interoperability with Solid Pods for Next Level Experience

Bolleman J.[1], Crum E.[2], Dragan I.[1], Galgonek J.[3], Ibberson M.[1], Mendes de Farias T.[1], Moos M.[3], Pagni M.[1], Sima Ana C.[1], Taelman R.[2], Vondrášek J.[3]

[1] *SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland*
[2] *Ghent University, Gent, Belgium*
[3] *Institute of Organic Chemistry and Biochemistry of the Czech Academy of Sciences, Prague, Czech Republic*

The Resource Description Framework (RDF) provides a powerful way to access data resources, but it is currently underexploited due to significant barriers limiting its use to a select few researchers. RDF's knowledge graphs are generally not sufficiently described, making it very difficult to ascertain what to query and how to do so. In addition, execution times for complex queries are often very long and difficult to optimise. Finally, the query results are often difficult to integrate with private data. The TRIPLE project will address these challenges to bring existing and future RDF resources to a broader group of researchers by developing innovative solutions on four fronts. 1) We will store private (unpublished) data in Solid Pods, an emerging technology enabling decentralised private vaults to host private RDF endpoints, execute federated SPARQL queries and cache data and results. 2) We will optimise federated queries spanning public and private SPARQL endpoints allowing users to query multiple resources from within their Solid Pod. 3) We will adapt state-of-the-art RDF documentation tools and make them available for all SPARQL endpoints, including Solid Pods. 4) We will develop data model visualisation, sets of standardised federated queries, and advanced query analysis and evaluation tools to help new users understand the data sources and run efficient federated queries. Finally, a demonstrator will show the impact of these advances when applied to a technically challenging use case of scientific relevance: the search for suitable organisms for bioremediation.

P-09

Bioinformatics multi-omics approach for data integration from various diagnostic types in pediatric oncology

Jurásková K.[1,2], Bystrý V.[1], Pokorná P.[1,3]

[1] *Central European Institute of Technology, Masaryk University, Brno, Czech Republic*
[2] *National Centre for Biomolecular Research, Faculty of Science, Masaryk University, Brno, Czech Republic*
[3] *Department of Biology, Faculty of Medicine, Masaryk University, Brno, Czech Republic*

Pediatric tumor diagnostics produce heterogeneous molecular data, yet these datasets are often analyzed in isolation, which limits their interpretive potential. We have developed a Shiny-based application for multi-omics data integration in pediatric oncology that unifies results from somatic and germline variant calling, fusion gene detection, and RNA expression profiling. The platform enhances data exploration through integrated visualization tools, including IGV for variant inspection and Cytoscape.js for pathway visualization. A key innovation is the simultaneous display of expression data, fusion events, and small variants within relevant signaling pathways, revealing molecular interactions that might remain undetected when analyzing isolated datasets. The comprehensive reports generated by the application are specifically designed to meet the needs of clinical geneticists, facilitating efficient data exploration and clinical decision-making.

P-10

Mapping Protein Language: Exploring Amino Acid Functionalitythrough Machine Learning

Kapral F.[1], Hanžl V.[1], Spiwok V.[1]

[1] *Department of Biochemistry and Microbiology, University of Chemistry and Technology, Prague, Technická 5, Prague 6, 166 28, Czech Republic*

Language is a structured system that enables humans to exchange information through words and sentences, where meaning depends on both vocabulary and context. Similarly, proteins can be viewed as a universal "language" with amino acids acting as words whose functions are determined by their sequence and structural environment. Just as natural language contains homonyms – words like bow (a weapon) and bow (to bend) that share spelling but differ in meaning - amino acids such as histidine can adopt distinct roles depending on their context. For instance, histidine may function as a charged surface residue, a metal-chelating site, or a catalytic component in an enzyme.

Advances in machine learning, particularly transformer-based architectures, have revolutionized our ability to decode such biological and linguistic patterns. Originally developed for natural language processing (e.g., translation models like ChatGPT), transformers now power breakthroughs in protein science, including structure prediction tools like AlphaFold, OpenFold, and ESMfold. ESMfold employs a large language model to "read" amino acid sequences, treating each residue as a word embedded in an 1280-dimensional vector. As these vectors pass through the transformer, they evolve to encode not just the identity of the amino acid but also its structural and functional context - mirroring how word meanings shift in sentences.

Building on this, we use ESM-2 (Evolutionary Scale Modeling) to map residue-level functionality across human proteins. Each amino acid is transformed into a vector representation, visualized in 2D and linked to 3D protein structures for deeper analysis. Our goal is to create an interactive atlas of amino acid residues, enabling researchers to explore the diverse roles of residues in proteins. By integrating protein language modeling with structural biology, this work bridges computational and experimental insights, offering a new lens to study protein function.

This work was supported by the Ministry of Education, Youth and Sports (reg. No. LUC24136, LM2023055), ELIXIR CZ and COST (ML4NGP, CA21160).

P-11

Tissue-restricted antigen-like expression of endogenous retroviruses in the thymus supports their role in negative selection

Klianitskaya M.[1,2], Sinitsyn S.[3], Reiniš J.[2], Vincendeau M.[4], Frishman D.[3], Filipp D.[5], Pačes J.[1,2]

[1] Department of Informatics and Chemistry, Faculty of Chemical Technology, University of Chemistry and Technology, Prague, Czech Republic
[2] Laboratory of Genomics and Bioinformatics, Institute of Molecular Genetics of the Czech Academy of Sciences, Prague, Czech Republic
[3] Associate Professorship of BioInformatics, TUM School of Life Sciences, Technical University of Munich
[4] Endogenous retroviruses group, Institute of Virology, Helmholtz Munich
[5] Laboratory of Immunobiology, Institute of Molecular Genetics of the Czech Academy of Sciences, Prague, Czech Republic

Endogenous retroviruses (ERVs) are remnants of ancient viral infections that have been part of the host genome for millions of years. Despite their viral origin, ERV-derived proteins and RNA are found in healthy tissues without triggering an immune response. This raises a fundamental question: why does the immune system tolerate these foreign elements? A likely explanation is that ERVs are incorporated into immune tolerance mechanisms, preventing autoimmunity against their expression products. Since immune tolerance is established in the thymus through the negative selection of autoreactive T cells, we hypothesize that ERVs are expressed in medullary thymic epithelial cells (mTECs) and presented as self-antigens.

To test this, we analyzed single-cell RNA sequencing (scRNA-seq) data from human and mouse thymic tissues, focusing on mTECs—specialized cells that train T cells by expressing a broad set of tissue-restricted antigens (TRAs). Our results show that ERVs are actively transcribed in mTECs following a pattern similar to TRAs: the number of expressed ERV loci is significantly higher in mTECs compared to other thymic cells. We also found that younger ERV subfamilies are more transcriptionally active, and many ERV loci overlap with long non-coding RNAs (lncRNAs), some of which are linked to different pathological conditions. These findings were consistent in both humans and mice, pointing to an evolutionarily conserved role for ERVs in thymic selection.

This study provides the first locus-specific analysis of ERV expression in mTECs and suggests that ERVs may actively contribute to immune tolerance. By exposing developing T cells to ERV-derived antigens, the thymus might prevent autoimmunity

against these elements later in life. Our findings challenge the view of ERVs as inactive genomic fossils and highlight their potential role in shaping immune regulation, with implications for autoimmune diseases and immune responses to ERV reactivation in disease states.

P-12

Investigation of vertebrate head evolution using single cell RNA-Seq of amphioxus embryos

Markos A.[1], Kubovciak J.[1], Mikula Mrstakova S.[1], Zitova A.[1], Paces J.[1], Machacova S.[1], Kozmik Z.-Jr.[1], Kozmik Z.[1], Kozmikova I.[1]

[1] *Institute of Molecular Genetics of the Czech Academy of Sciences*

To shed light on the enigmatic origin of the vertebrate head, our study employs an integrated approach that combines single-cell transcriptomics, perturbations in signaling pathways, and cis-regulatory analysis in amphioxus. As a representative of a basal lineage within the chordate phylum, amphioxus retains many traits thought to have been present in the common chordate ancestor. Through cell type characterization, we identify the presence of prechordal plate-like, pre-migratory, and migratory neural crest-like cell populations in the developing amphioxus embryo. Our findings provide evidence that the key features of vertebrate head development can be traced back to the common ancestor of all chordates. Research was backed up by utilization of single-cell transcriptomics data (generated by 10x Chromium platform) from multiple timepoints of amphioxus embryonic development. Presented poster aims to emphasize analytical approaches chosen for processing of these data, including cell type annotation in each timepoint and subsequent pseudotime analysis along with computation of cell fate transition graph to explore focal developmental trajectories. Furthermore, transcriptional similarity-based homology between our data and already annotated zebrafish dataset was quantified in order to confirm shared evolutionary history of the cell types involved in embryonal head development.

P-13

Genomic Benchmarks QC: Automated Quality Control for Genomic Machine Learning Datasets

Maršálková E.[1,2], Grešová K.[3,4], Šimeček P.[1], Alexiou P.[3,4]

[1] *Central European Institute of Technology, Masaryk University, Brno, Czech Republic*
[2] *National Centre for Biomolecular Research, Faculty of Science, Masaryk University, 61137 Brno, Czech Republic*
[3] *Centre for Molecular Medicine and Biobanking, University of Malta, Msida, Malta*
[4] *Department of Applied Biomedical Science, Faculty of Health Sciences, University of Malta, Msida, Malta*

As machine learning becomes increasingly prominent in genomics, the quality and composition of training data become crucial to developing robust and generalizable models. While AlphaFold2's success in structural bioinformatics was made possible by a carefully curated benchmark of experimentally determined protein structures, genomics still lacks comparable, high-quality datasets for many of its key challenges. When training machine learning models, they can become only as good as the datasets they are built upon—if the training dataset contains biases, the models will learn those biases instead of capturing the biological features. This results in models that perform well on internal validation but fail to generalize across datasets. For instance, through simulated data, we can show that when GC-content bias exceeds a certain threshold, convolutional neural networks begin to learn this bias instead of recognizing actual sequence motifs.

To address this issue, we present a tool for automated quality control of genomic datasets. It evaluates properties such as nucleotide composition, GC-content, sequence length distribution, and duplication rates, producing a detailed report that helps researchers check, detect and mitigate misleading data patterns. The tool is easily accessible through GitHub [1] and as a Python package, designed to reduce overhead and promote the development of biologically meaningful, high-quality genomic datasets. We have applied this tool to the datasets from our previous work in Genomic Benchmarks v1 [2] to evaluate their quality and identify potential candidates for inclusion in a release of an improved and unbiased version 2 of this benchmarking dataset collection.

References

[1] https://github.com/katarinagresova/GenBenchQC
[2] Grešová, K., Martinek, V., Čechák, D. et al. Genomic benchmarks: a collection of datasets for genomic sequence classification. BMC Genom Data 24, 25 (2023). https://doi.org/10.1186/s12863-023-01123-8

Poster session

Tuesday, 10. June

Poster session is sponsored by the
National Infrastructure for Chemical Biology.

CZ OPENSCREEN
National Infrastructure for Chemical Biology

P-14

MolMeDB - Molecules on Membranes Database

Juračka J.[1,2], Storchmannová K.[1], Martinát D.[1], Bazgier V.[1], Berka K.[1]

[1] Department of Physical Chemistry, Faculty of Science, Palacký University Olomouc, Czech Republic
[2] Department of Computer Sciences, Faculty of Science, Palacký University Olomouc, Czech Republic

Biological membranes are natural barriers to cells. They play a key role in cell life and the pharmacokinetics of drug-like small molecules. A small molecule can pass through the membranes in two ways: via passive diffusion or actively via membrane transport proteins. There is a huge amount of data available about interactions of the small molecules with membranes and about interactions among the small molecules and the transporters.

MolMeDB (molmedb.upol.cz) is a 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 the 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, PubChem, The IUPHAR/BPS Guide to PHARMACOLOGY[3]). Data in the MolMeDB are fully searchable and browsable by name, SMILES, membrane, method, transporter, or dataset, and we offer collected data openly for further reuse. 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.; Šrejber, M.; Melíková, M.; Bazgier, V.; Berka, K. MolMeDB: Molecules on Membranes Database. *Database* **2019**, *2019*, baz078. https://doi.org/10.1093/database/baz078

[2] Schwöbel, J. A. H.; Ebert, A.; Bittermann, K.; Huniar, U.; Goss, K.-U.; Klamt, A. 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. https://doi.org/10.1021/acs.jpcb.9b11728

[3] Harding, S. D.; Armstrong, J. F.; Faccenda, E.; Southan, C.; Alexander, S. P. H.; Davenport, A. P.; Spedding, M.; Davies, J. A. The IUPHAR/BPS Guide to PHARMACOLOGY in 2024. *Nucleic Acids Research* **2024**, *52* (D1), D1438–D1449. https://doi.org/10.1093/nar/gkad944

P-15

Determination of ADP/ATP translocase isoform ratios in malignancy and cellular senescence

Liblova Z.[1$], Maurencova D.[1$], Salovska B.[1+], Kratky M.[1], Mracek T.[2], Korandova Z.[2#], Pecinova A.[2], Vasicova P.[1], Rysanek D.[1], Andera L.[1], Fabrik I.[3], Kupcik R.[3], Kashmel P.[1], Sultana P.[1], Tambor V.[3], Bartek J.[1,4,5], Novak J.[1], Vajrychova M.[3], Hodny Z.[1]

[1] Laboratory of Genome Integrity, Institute of Molecular Genetics of the Czech Academy of Sciences, Prague, Czech Republic
[2] Laboratory of Bioenergetics, Institute of Physiology of the Czech Academy of Sciences, Prague, Czech Republic
[3] Biomedical Research Center, University Hospital Hradec Kralove, Hradec Kralove, Czech Republic
[4] Danish Cancer Society Research Center, Copenhagen, Denmark
[5] Department of Medical Biochemistry and Biophysics, Division of Genome Biology, Science for Life Laboratory, Karolinska Institute, Stockholm, Sweden
[$] equal contribution

Current affiliation:
+ Yale Cancer Biology Institute, Yale University, West Haven, CT, USA
 # Department of Adipose Tissue Biology, Institute of Physiology of the Czech Academy of Sciences, Prague, Czech Republic

Cellular senescence has recently been recognized as a significant contributor to the poor prognosis of glioblastoma, one of the most aggressive brain tumors. Consequently, effectively eliminating senescent glioblastoma cells could benefit patients. Human ADP/ATP translocases (ANTs) play a role in oxidative phosphorylation in both normal and tumor cells. Previous research has shown that the sensitivity of senescent cells to mitochondria-targeted senolytics depends on the level of ANT2. In this study, we systematically mapped the transcript and protein levels of ANT isoforms in various types of senescence and glioblastoma tumorigenesis. We employed bioinformatics analysis, targeted mass spectrometry, RT-PCR, immunoblotting, and assessment of cellular energy state to elucidate how individual ANT isoforms are expressed during the development of senescence in non-cancerous and glioblastoma cells. Notably, we consistently observed an elevation of ANT1 protein levels across all tested senescence types, while ANT2 and ANT3 exhibited variable changes. The alterations in ANT protein isoform levels correlated with shifts in the cellular oxygen consumption rate. Our findings suggest that ANT isoforms are mutually interchangeable for oxidative phosphorylation and manipulating individual ANT isoforms could be a potential target for senolytic therapy.

P-16

Somatic Mutation Profiling in Head and Neck Squamous Cell Carcinoma: Early Insights from an Ongoing Cohort Study

Melichar V.[1,2], Galiyeva L.[1,4], Netušil J.[1,3], Kolář M.[1,2], Vomastek T.[5], Plzák J.[4], Šáchová J.[1]

[1] Laboratory of Genomics and Bioinformatics, Institute of Molecular Genetics of the Czech Academy of Sciences
[2] Department of Informatics and Chemistry, University of Chemical Technology Prague
[3] Faculty of Science, Charles University, Prague
[4] Department of Otorhinolaryngology and Head and Neck Surgery, First Faculty of Medicine, Charles University Prague, Faculty Motol Hospital
[5] Laboratory of Cell Signalling, Institute of Microbiology of the Czech Academy of Sciences

Head and neck squamous cell carcinoma (HNSCC) is a genetically heterogeneous disease with a complex mutational landscape. In this study, we aim to identify recurrent somatic mutations in a cohort of 70 HNSCC patients using high-throughput sequencing. Currently, variant calling has been performed on an initial subset of 18 matched tumour-normal samples. Surprisingly, the number of high-confidence somatic variants passing standard filters according to GATK's Best Practices is relatively low. The most frequent causes of variant exclusion include normal artifact, strand bias and low base quality. The basis for this phenomenon is unclear. Preliminary analysis indicates that *TP53, KMT2D* and *FAT1* are the most commonly mutated genes in the cohort so far, which is consistent with previous studies. As additional samples are processed, we aim to refine the mutational landscape of HNSCC and explore correlations with clinical and pathological features. Transcriptomic data from the same patients are generated and analysed in parallel. This work highlights the technical challenges of somatic variant calling in HNSCC and the importance of rigorous filtering to ensure high-quality and validity of called genomic variants.

P-17

CReM-dock: de novo design of chemical reasonable compounds guided by molecular docking

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

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

The objective of this study is the development of a tool to design compounds de novo and decorate co-crystallized ligands with the special focus on synthetic accessibility. This tool employs the CReM approach [1] to generate ligand structures and control their synthetic feasibility and molecular docking by EasyDock [2] to assess their binding to a target protein. The developed tool offers two modes: i) iterative growth of a fragment co-crystallized with a protein preserving the position of the parent part of a molecule and ii) de novo compound generation starting from a custom set of fragments, which are subsequently docked and iteratively expanded. CReM-dock offers an optional augmentation of docking score and bias physicochemical properties of generated compounds, e.g. the fraction of sp^3 carbon atoms. CReM-dock was benchmarked on de novo generation of ligands of targets belonging to different families to investigate dependency of diversity, synthetic accessibility, docking score and other properties of generated structures from chosen settings. We compared CReM-dock with REINVENT4 [3] and demonstrated that both tools result in comparable docking and synthetic accessibility scores of generated molecules, while CReM-dock compounds have higher novelty. The developed tool offers predictable control over synthetic feasibility of generated molecules and great flexibility to perform pure de novo generation as well as fragment expansion or scaffold decoration. This work was funded by the Ministry of Education, Youth and Sports of the Czech Republic through INTER_EXCELLENCE II grant LUAUS23262, the e-INFRA CZ (ID:90254), and CZ-OPENSCREEN project (LM2023052).

References

[1] Polishchuk, P., *J Cheminf.* **2020**, 12, 28.
[2] Minibaeva G., et al, *J Cheminf.* **2023**, 15, 102.
[3] Loeffler, H.H., et al. *J Cheminf.* **2024**, 16, 20.

P-18

FAIRification of Augusta, a Python package for RNA-Seq-Based Inference of Gene Regulatory and Boolean Networks

Musilova J.[1,4], Vafek Z.[2,4], Hermankova K.[1], Zimmer R.[3], Helikar T.[4], Sedlar K.[1]

[1] *Brno University of Technology, Department of Biomedical Engineering, BioSys lab*
[2] *Brno University of Technology, Institute of Forensic Engineering*
[3] *Ludwig-Maximilians-Universität München, Department of Informatics*
[4] *University of Nebraska-Lincoln, Department of Biochemistry*

Gene regulation plays a central role in controlling biological processes. RNA Sequencing (RNA-Seq) has become a widely adopted technique to measure gene expression across the genome under various conditions, providing comprehensive data for studying these regulatory mechanisms. Computational models such as Gene Regulatory Networks (GRNs) and Boolean Networks (BNs) offer valuable approaches for capturing gene interactions in silico and uncovering the regulatory logic underlying gene expression. Despite their importance, accurate inference of these networks on a genome-wide scale remains a major challenge in bioinformatics.

Augusta is an open-source Python package developed to reconstruct computational models from time-series RNA-Seq datasets, supporting the inference of both GRNs and BNs at the genome scale. The workflow in Augusta encompasses several key stages: it begins with normalization of gene expression count tables, proceeds to GRN inference via mutual information computation, and incorporates a two-step validation process. This validation involves the *de novo* identification of transcription factor binding motifs (TFBM), followed by integration of curated database evidence to refine the resulting network. For BN inference, Augusta combines curated knowledge with logical rule assignments, enabling a more holistic analysis of regulatory interactions.

The ongoing development of Augusta is focused on adhering to the FAIR-RS (Findability, Accessibility, Interoperability, and Reusability for Research Software) principles. Backed by the FAIR-IMPACT initiative under the "Assessment and improvement of Research Software" support action, Augusta is committed to promoting transparent, reproducible methods in the analysis of gene regulatory systems.

Augusta is available from github.com/JanaMus/Augusta along with documentation, examples, and tutorials.

This project has been supported by grant project GACR Junior Star 25-17459M.

P-19

An integrated computational strategy to identify selective HDAC6 inhibitors against breast cancer

Pankaj V.[1], Bhogal I.[1], Roy S.[1]

[1] *Department of Biomedical Engineering, Faculty of Electrical Engineering and Communication, Brno University of Technology, Brno, 61600, Czechia.*

Breast cancer is considered one of the leading causes of cancer-related mortality in women worldwide, with its progression often driven by aberrant estrogen receptor (ER) signaling and epigenetic dysregulation. Histone deacetylase 6 (HDAC6), a cytoplasmic class IIb enzyme, has emerged as a critical epigenetic regulator that modulates ER activity and promotes breast cancer development and metastasis. Despite the therapeutic potential of HDAC6 inhibition, currently available inhibitors have shown limited clinical efficacy due to poor selectivity, toxicity, and undesirable off-target effects. This study employed an integrated computational pipeline to identify novel, selective HDAC6 inhibitors with improved pharmacological profiles. A comprehensive virtual screening of 355,289 molecules was conducted, from which 2,107 virtual hits were shortlisted based on their initial binding potential. These hits underwent molecular docking experiments, leading to the top seven hit compounds prioritized based on their binding affinity to the HDAC6 catalytic domain. These selected compounds were further analyzed through 100 ns molecular dynamics simulations to evaluate their structural stability and interaction dynamics within the HDAC6 active site. Hit-1 emerged as a potent candidate as HDAC6 inhibitor, demonstrating a highly favorable MM/GBSA binding free energy of -131.12 kcal/mol, exceeding the binding performance of reference inhibitor Trichostatin A (-114.24 kcal/mol). Hit-1 demonstrated robust stability, minimal structural fluctuations, and maintained consistent interactions with key catalytic residues ASP649, HIS651, and ASP742 throughout the simulation period. *In-silico* ADMET profiling confirmed Hit-1 compounds' high oral bioavailability, non-mutagenic nature, low hepatotoxicity, and favorable metabolic stability. In conclusion, Hit-1 represents a potent and selective HDAC6 inhibitor with better binding and pharmacokinetic properties. This study demonstrates the efficacy of a multi-stage computational approach to accelerate drug discovery and supports further experimental validation for clinical development.

Keywords
Breast cancer, HDAC6 inhibitor, molecular docking, molecular dynamic simulation, MM/GBSA binding energy, *in-silico* ADMET

P-20

User-friendly web tool for typing and characterization of ESKAPEE pathogens

Šabatová K.[1], Jakubíčková M.[1], Dzurčaninová N.[1], Bezdíček M.[2,3], Lengerová M.[2,3], Vítková H.[1]

[1] Department of Biomedical Engineering, Faculty of Electrical Engineering and Communication, Brno University of Technology, Czechia
[2] Department of Internal Medicine – Hematology and Oncology, University Hospital Brno, Brno, Czechia
[3] Department of Internal Medicine – Hematology and Oncology, Masaryk University, Brno, Czechia

Multidrug resistant bacteria pose a significant threat in hospital environments, particularly pathogens from the so-called ESKAPEE group. Identifying and characterizing individual strains of a given species is crucial for tracking the spread of infections within hospital units and for selecting appropriate treatments. Multilocus sequence typing (MLST) differentiates strains by comparing selected housekeeping gene sequences with existing allelic profiles to assign a sequence type, while also enabling prediction of associated resistance and virulence factors. Although tools exist for these analyses, their use typically requires bioinformatics expertise to run analyses and interpret results coherently.
Here, we present a comprehensive web-based tool for rapid bacterial typing and characterization using MLST, designed for use by hospital personnel without programming skills. The sequence type of a pathogen is determined either from assembled contigs using BLAST or directly from raw reads using KMA, referencing allelic profiles in PubMLST database. In addition, the presence of virulence genes is predicted using hits from the BIGSdb-Pasteur and PubMLST databases, while antibiotic resistance genes and single nucleotide polymorphisms are identified using ResFinder tool. The tool was developed as a web application using Python Flask framework, offering submission of input data, selection of requested analysis and providing comprehensive result summaries in the form of tables and heatmaps to visualize similarities with the local result database and assist in tracking of possible outbreaks.

This work was supported by NW24-09-00126.

P-21

ELNX: An Electronic Laboratory Notebook for Seamless Integration into Daily Research

Škuta C.[1], Müller T.[1], Voršilák M.[1], Bartůněk P.[1]

[1] *CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Institute of Molecular Genetics of the Czech Academy of Sciences, Vídeňská 1083, 14220 Prague 4, Czech Republic*

ELNX is a universal electronic laboratory notebook (ELN) developed to seamlessly integrate into everyday research routines, combining strict compliance with standard laboratory documentation requirements and a user-friendly interface. It offers essential features such as time-stamped entries, version history, immutable records, role-based access control, supervisory signing, and hierarchical organization at user group/laboratory level.

ELNX enables users to effortlessly input a wide range of content, including formatted text, images, tables, and arbitrary attachments (PDFs, office documents, etc.) via direct input, copy-paste, or mobile capture. Both desktop and mobile interface support in-app image annotation. The automatic record saving ensures uninterrupted workflow with the ability to resume work at any time. Records within individual notebooks can be organized into collections using a hierarchical structure, re-ordered based on user preference, labelled with custom tags and searched for through the full-text search functionality. To ensure the FAIRness of the notebook content and simplify the creation of structured data entry (e.g., form-based records), the users can design custom metadata templates enhanced with the connection to underlying ontologies (utilizing the EBI's Ontology Lookup Service).

Beyond the documentation of day-to-day work, ELNX features an integrated booking system with calendars for tracking time on projects, managing personnel availability, and (soon) booking lab equipment with role-based decision workflows. ELNX thus serves as a comprehensive digital platform supporting both scientific documentation and operational coordination across diverse research domains.

P-22

Methylome Profiling Using Third-Generation Sequencing: A Comparison of PacBio and ONT in a PHA-Producing Bacterium

Umair M.[1], Jakubickova M.[1], Buchtikova I.[2], Bezdicek M.[3,4], Obruca S.[2], Vitkova H.[1], Sedlar K.[1]

[1] Department of Biomedical Engineering, Faculty of Electrical Engineering and Communication, Brno University of Technology, Brno, Czechia
[2] Department of Food Chemistry and Biotechnology, Faculty of Chemistry, Brno University of Technology, Brno, Czechia
[3] Department of Internal Medicine – Hematology and Oncology, University Hospital Brno, Brno, Czechia
[4] Department of Internal Medicine – Hematology and Oncology, Masaryk University, Brno, Czech Republic

There are two major third-generation sequencing technologies that allow DNA methylation detection on a genome-wide scale: Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT). Since methylation detection is performed directly from raw sequencing signals, specialized tools for particular sequencing platforms need to be used. Here, we used kineticsTools and Modkit to detect methylation in PacBio and ONT, respectively. In this case study, we used *Caldimonas thermodepolymerans*, a thermophilic bacterium that produces polyhydroxyalkanoate (PHA), a naturally occurring biodegradable polymer potentially usable to design bioplastics. Understanding its epigenetic regulation, particularly DNA methylation, will play a crucial role in optimising its biotechnological and industrial potential. Initial comparison of all the reported methylation positions revealed that ONT identified significantly more sites (~4.4 million) than PacBio (~300,000), with all the sites detected by PacBio being identified by ONT, suggesting a difference in their baseline sensitivity or signal reporting between platforms. To get more confident methylation sites, we applied stringent filters (methylation percentage ≥95%) and found that PacBio retained ~29000 sites while ONT retained only ~11000 sites, with ~8000 high-confidence sites concordantly detected by both. This indicated that ~74% of ONT high-confidence calls were supported by PacBio. However, PacBio identified a larger set of filtered sites meeting its criteria; this likely arises primarily due to distinct algorithmic approaches used to assign confidence scores, thus reflecting how each platform processes and thresholds methylation signals. Motif analysis showed stronger agreement for major shared methylation motifs (e.g., CTGCAG 6mA, GAGCTC 4mC, GAYAN...GTG 6mA) with more than 99.6% site-level concordance. However, notable platform-specific differences were observed with ONT showing several uniquely identified motifs with a high fraction likely involving 5mC modification, while PacBio reported only a few low-fraction motifs.

Our study shows that while both platforms reliably capture methylation patterns, algorithm sensitivity and filtering threshold differences notably influence detection outcome, especially for non-canonical motifs. Understanding these platform-specific differences is essential for studying epigenetic regulation of genes involved in PHA biosynthesis. It guides future efforts to optimise industrial microbes through methylome engineering and synthetic biology approaches.

This study was funded by the Czech Science Foundation (GACR) (Project No. GM25-17459M)

P-23

Refining scRNAseq Analysis for Homogeneous Stem Cell Data: Exploring Clustering Approaches and Non-Negative Matrix Factorization to Uncover Dynamic Infection Responses

Večerková K.[1,2], Ribeiro Bas I.[3], Kolář M.[1,2], Alberich Jordà M.[3]

[1] Laboratory of Genomics and Bioinformatics; Institute of Molecular Genetics of the Czech Academy of Sciences
[2] Department of Informatics and Chemistry; University of Chemistry and Technology, Prague
[3] Laboratory of Haematooncology, Institute of Molecular Genetics of the Czech Academy of Sciences

Single-cell RNA sequencing (scRNAseq) is a powerful tool for exploring cellular heterogeneity, but standard analysis pipelines often fail when applied to homogeneous cell populations. In this study, we analyzed a time-series scRNAseq dataset from hematopoietic stem cells responding to infection across five timepoints and a control. Standard workflows predominantly captured variation due to cell cycle effects, obscuring relevant biological signals. To address this, we refined the clustering process through iterative exploration, aiming for exclusive and biologically meaningful marker expression. In parallel, we applied non-negative matrix factorization (NMF) to identify interpretable gene expression programs reflecting condition- and timepoint-specific responses. Our results demonstrate that adapting analysis workflows to data characteristics can reveal otherwise hidden transcriptional programs in homogeneous single-cell datasets.

P-24

Consensus-Based Detection of Biosynthetic Gene Clusters with Application to RiPPs from Antarctic Bacteria

Vítková H.[1], Jakubíčková M.[1], Bezdíček Králová S.[2]

[1] Department of Biomedical Engineering, Brno University of Technology, Brno, Czech Republic
[2] Department of Chemistry and Biochemistry, Faculty of AgriSciences, Mendel University in Brno, Brno, Czech Republic

Bacterial secondary metabolites represent a rich source of bioactive compounds, yet much of their biosynthetic potential remains unexplored. One particularly promising group are ribosomally synthesized and post-translationally modified peptides (RiPPs), which form a structurally diverse and pharmaceutically promising class of natural products. Therefore, our study focuses on uncovering novel RiPP biosynthetic gene clusters (BGCs) from extremophilic bacteria inhabiting Antarctic environments, which offer access to unique and largely uncharacterized microbial diversity. A new computational pipeline is being developed to enable robust and comprehensive RiPP detection. The proposed pipeline integrates several specialized tools for detecting RiPPs specifically and BCGs in general, including DeepRiPP, RRE-Finder, decRiPPter, antiSMASH, and DeepBGC. Selected tools employ various algorithmic approaches, primarily utilizing machine learning methods such as bidirectional long short-term memory recurrent neural networks, support vector machines, or hidden Markov models, resulting in varying outputs. These outputs are then combined through a consensus-based approach to generate unified predictions, improving the accuracy and reliability of RiPP identification. A key part of our workflow is the use of long-read sequencing to enable high-quality reconstruction of microbial genomes directly from complex metagenomic samples. This strategy allows us to recover biosynthetic pathways from previously inaccessible and uncultivated microbial lineages, thus broadening the scope for natural product discovery. Identified candidate clusters will undergo comparative genomics and functional annotation to prioritize targets for downstream characterization and synthetic biology applications. Our approach aims to expand the known chemical space of RiPPs and provide a robust resource for future biotechnological exploitation.

This work was supported by GA25-17343S.

P-25

European Chemical Biology Database

Voršilák M.[1,2], Müller T.[1], Škuta C.[1], Bartůněk P.[1]

[1] CZ-OPENSCREEN National Infrastructure for Chemical Biology, Institute of Molecular Genetics of the ASCR v. v. i., Vídeňská 1083, 142 20, Prague 4, Czech Republic
[2] Department of Informatics and Chemistry, Faculty of Chemical Technology, University of Chemistry and Technology Prague, Technická 5, 166 28, Prague, Czech Republic

The European Chemical Biology Database (ECBD, https://ecbd.eu) serves as the central repository for data generated by the EU-OPENSCREEN (EU-OS) research infrastructure consortium, European research infrastructure for chemical biology founded in 2018 to support chemical probe and drug discovery projects. ECBD is developed according to FAIR principles, which emphasize findability, accessibility, interoperability and reusability of data. These data are made available to the scientific community following open access principles. The ECBD stores both positive and negative results from the entire chemical biology project pipeline, including data from primary or counter-screening assays. The assays utilize a defined and diverse library of over 108 000 compounds, the annotations are continuously enriched by external user supported screening projects and by internal EU-OS bioprofiling efforts. As of May 2025, these compounds were screened in 115 currently deposited datasets (assays), with 71 already being publicly accessible, while the remaining will be published after a publication embargo period of up to 3 years. Together these datasets encompass ~4.7 million experimental data points. All public data within ECBD can be accessed through its user interface, API or by database dump under the CC-BY 4.0 license.

Aksu Ali *(aksua18@itu.edu.tr)*
UCT Prague, Praha 6

Bartůněk Petr *(bartunek@img.cas.cz)*
Institute of Molecular Genetics of the Czech Academy of Sciences, Prague

Berka Karel *(karel.berka@upol.cz)*
Palacky University Olomouc, Olomouc

Bhogal Inderjeet *(234190@vut.cz)*
Brno University of Technology, Brno

Blavet Nicolas *(nicolas.blavet@ceitec.muni.cz)*
CEITEC, Brno

Bohuslavová Dominika *(draesdom@cvut.cz)*
Czech Technical University in Prague, Faculty of Information Technology, Prague 6

Bojić Milan *(milan.bojic@img.cas.cz)*
CZ-OPENSCREEN, Praha 4

Brejová Broňa *(bbrejova@gmail.com)*
Univerzita Komenského v Bratislave, Bratislava

Bulantová Lucie *(474241@mail.muni.cz)*
CEITEC Masarykova Univerzita, Brno

Cakl Lukáš *(cakll@natur.cuni.cz)*
Přírodovědecká fakulta Univerzity Karlovy, Praha 2

Ceklarz Joanna *(joanna.ceklarz@vscht.cz)*
VŠCHT, Praha

Čech Petr *(cechp@vscht.cz)*
VŠCHT Praha, Praha

Čejková Darina *(cejkovad@vut.cz)*
Brno University of Technology, Brno

Dehaen Wim *(dehaenw@vscht.cz)*
UCT Prague, Praha 6

Delattre Mathys *(mathys.delattre@img.cas.cz)*
Institute of Molecular Genetics of the Czech Academy of Sciences, Prague 4

Ehler Edvard *(edvard.ehler@img.cas.cz)*
Ústav molekulární genetiky AV ČR, v. v. i., Praha 4

Engst Martin *(engst@uochb.cas.cz)*
Ústav organické chemie a biochemie AV ČR, v. v. i., Praha 6

Epp Trevor *(trevor.epp@img.cas.cz)*
Institute of Molecular Genetics of the CAS, Prague 4

Fil Valeriia *(filv@vscht.cz)*
Vysoká škola chemicko-technologická v Praze, Praha

Fišer Karel *(karel.fiser@lfmotol.cuni.cz)*
Univerzita Karlova, 2. lékařská fakulta, Praha

Fülöp Jozef *(fulopj@vscht.cz)*
University of Chemistry and Technology, Prague, Praha 6

Galgonek Jakub *(jakub.galgonek@uochb.cas.cz)*
Institute of Organic Chemistry and Biochemistry of the Czech Academy of Sciences, Praha 6

Halenková Zuzana *(halenkova.zuzana@gmail.com)*
Faculty of Science, Charles University, Praha 2

Hanzlík Adam *(hanzlikc@vscht.cz)*
Ústav molekulární genetiky AV ČR, v.v.i., Praha 4 – Krč

Hoksza David *(david.hoksza@matfyz.cuni.cz)*
Univerzita Karlova, Matematicko-fyzikální fakulta, Praha

Hron Tomas *(tomas.hron@img.cas.cz)*
Institute of Molecular Genetics, AV CR, v.v.i., Prague

Jakubíčková Markéta *(jakubickova@vut.cz)*
Brno University of Technology, Department of Biomedical Engineering, Brno

Janeček Štefan *(Stefan.Janecek@savba.sk)*
Institute of Molecular Biology, Slovak Academy of Sciences, Bratislava

Jurásková Kateřina *(juraskovakaterina@seznam.cz)*
CEITEC Masarykova Univerzita, Brno

Kaňovský Aleš *(ales.kanovsky@img.cas.cz)*
Institute of Molecular Genetics of the Czech Academy of Sciences, Prague 4

Kapral Filip *(kapralf@vscht.cz)*
VŠCHT Praha, Praha

Klianitskaya Marharyta *(klianitm@vscht.cz)*
Ústav molekulární genetiky AV ČR, Praha 4

Kubovčiak Jan *(kubovcij@img.cas.cz)*
Institute of Molecular Genetics of the Czech Academy of Sciences, Prague

Lukeš Stanislav *(enbik.cz@exyi.cz)*
Ústav molekulární genetiky AV ČR, v. v. i., Praha

Maršálková Eva *(469217@mail.muni.cz)*
CEITEC Masarykova Univerzita, Brno

Martinát Dominik *(dominik.martinat@upol.cz)*
Palacký University Olomouc, Olomouc

Maurencová Dominika *(dominika.maurencova@img.cas.cz)*
Institute of Molecular Genetics of the Czech Academy of Sciences, Prague 4

Medaglia-Mata Alejandro *(amedagliamata2@gmail.com)*
CEITEC MUNI, Brno

Melichar Vojtěch *(vojtech.melichar@img.cas.cz)*
Ústav molekulární genetiky AV ČR, v. v. i., Praha 4

Minibaeva Guzel *(guzel.minibaeva@upol.cz)*
Faculty of Medicine and Dentistry, Palacký University Olomouc, Olomouc

Modrák Martin *(modrak.mar@gmail.com)*
Second Faculty of Medicine, Charles University in Prague, Praha 5

Mokrejš Martin *(martin.mokrejs@lf1.cuni.cz)*
Charles University, First Faculty of Medicine, Prague

Moreno Hugo *(morenoh@natur.cuni.cz)*
Přírodovědecká fakulta UK, Praha

Müller Tomáš *(tomas.muller@img.cas.cz)*
Institute of Molecular Genetics of the Czech Academy of Sciences, Praha

Musilová Jana *(musilovajana@vut.cz)*
Brno University of Technology, Department of Biomedical Engineering, Brno

Ostašov Pavel *(pavel.ostasov@lfp.cuni.cz)*
Faculty of Medicine in Pilsen, Pilsen

Pačes Jan *(hpaces@img.cas.cz)*
Ústav molekulární genetiky AV ČR, Praha

Pankaj Vaishali *(234084@vut.cz)*
Brno University of Technology, Brno

Pfeiferová Lucie *(pfeiferoval@seznam.cz)*
Ústav molekulární genetiky AV ČR, v. v. i., Praha 4

Polishchuk Pavel *(pavlo.polishchuk@upol.cz)*
Palacky University, Faculty of Medicine and Dentistry, Olomouc

Reifova Radka *(radka.reifova@natur.cuni.cz)*
Department of Zoology, Faculty of Science, Charles University, Praha 2

Safiulina Sofya *(sofya.safiulina@email.cz)*
Ústav molekulární genetiky AV ČR, Praha 1

Schlebusch Stephen *(stephen.schlebusch@gmail.com)*
Charles University, Prague 1

Sedlář Karel *(sedlar@vut.cz)*
Vysoké učení technické v Brně, Brno

Stočes Štěpán *(stepan.stoces@seqme.eu)*
SEQme s.r.o., Dobříš

Svozil Daniel *(svozild@vscht.cz)*
UCT Prague, Prague 6

Šabatová Kateřina *(sabatovak@vut.cz)*
Brno University of Technology, Brno

Šícho Martin *(martin.sicho@vscht.cz)*
VŠCHT Praha, Prague

Škrhák Vít *(vit.skrhak@matfyz.cuni.cz)*
Charles University, MFF, Praha 1

Škuta Ctibor *(ctibor.skuta@img.cas.cz)*
Ústav molekulární genetiky AV ČR, Prague 4

Umair Mohammad *(253834@vutbr.cz)*
Brno University of Technology, Brno

Večerková Kateřina *(katerina.vecerkova@img.cas.cz)*
Institute of Molecular Genetics CAS, Prague

Vinař Tomáš *(tomas.vinar@fmph.uniba.sk)*
Univerzita Komenského v Bratislave, Bratislava

Vítková Helena *(vitkovah@vut.cz)*
Brno University of Technology, Department of Biomedical Engineering, Brno

Voller Jiří *(jirivoller@gmail.com)*
Universita Palackého, Olomouc

Voršilák Milan *(milan.vorsilak@img.cas.cz)*
Ústav molekulární genetiky AV ČR, Praha 4

Yurenko Yevgen *(yevgen.yurenko@gmail.com)*
Institute of Organic Chemistry and Biochemistry, Prague, Prague