VŠCHT PRAHA

FCHT

Sborník anotací

Studentská vědecká konference

19. 11. 2026

Ústav informatiky a chemie

(143)

ÚSTAVNÍ KOORDINÁTOR

Ing. Jiří Znamenáček

SEZNAM SEKCÍ

– Bioinformatika a chemická informatika

Bioinformatika a chemická informatika

MÍSTO: B1322

KOMISE

prof. Mgr. Daniel Svozil, Ph.D. (předseda)

MSc. Wim Dehaen, Ph.D.

Ing. Ivan Čmelo, Ph.D.

Mgr. Jan Pačes, Ph.D.

Ing. Martin Šícho, Ph.D.

PROGRAM

14:00 **zahájení**

14:05 **Marek Cuker** (M2; doc. Ing. Filip Lankaš, Ph.D.)
Exploring the mechanical code of DNA duplexes

14:20 **Barbora Čurdová** (M2; MSc. Wim Dehaen, Ph.D.)
Peptide Hypervector Encodings and Tsetlin Machines

14:35 **David Kolář** (B3; Ing. Martin Šícho, Ph.D.)
Ligand-based de novo drug design using deep neural networks

14:50 **Ester Machová** (M2; Ing. Martin Šícho, Ph.D.)
Předpověď biologické aktivity z dat kapalinové chromatografie a hmotnostní spektrometrie pomocí strojového učení

15:05 **Lenka Sixtová** (M1; Ing. Martin Šícho, Ph.D.)
Chiroptical Properties of Helicenes: Data Acquisition, Analysis, and Modelling

15:20 **Dominika Soukupová** (M2; Ing. Martin Šícho, PhD.)
ML-Driven Design of Antimicrobial Peptides

15:35 **Tereza Sujová** (M1; MSc. Wim Dehaen, Ph.D.)
Efficiency metrics in total syntesis

15:50 **Jáchym Urban** (M1; Ing. Ivan Čmelo Ph.D.)
Virtual screening for potential O-methyltransferase inhibitors

vyhlášení výsledků

SPONZOŘI – BIOINFORMATIKA A CHEMICKÁ INFORMATIKA



Exploring the mechanical code of DNA duplexes

Bc. Marek Cuker (M2)

školitel: doc. Ing. Filip Lankaš, Ph.D.

Sequence-specific structure and stiffness of DNA has been implicated in many biological phenomena, including transcription regulation, chromatin compaction, and even in the field of DNA nanostructure design.

We have developed a rigid base model and parametrized it with hexamer-scale data from all-atom molecular dynamics simulations. Here we use this model together with Monte Carlo simulations to predict elastic rod parameters characterising mechanical properties of an arbitrary sequence.

This model has been used to generate a large dataset, which is then employed to finetune DNABERT-2, a pre-trained DNA language model, enriching it with mechanical information. This finetuned DNABERT-2 model can be used to create mechanics-aware embeddings, and allows for the exploration and mapping of elastic properties for an even larger set of sequences. Using these modified embeddings can improve performance on mechanics-related downstream tasks as illustrated on cyclisability, without catastrophic forgetting of the original representation.

Peptide Hypervector Encodings and Tsetlin Machines

Bc. Barbora Čurdová (M2)

školitel: MSc. Wim Dehaen, Ph.D.

Most machine learning methods require fixed-size numerical input, but peptides are sequences of variable length. Methods used for small molecules, such as fingerprints, are not designed to preserve residue order, so peptides with the same amino acids in a different order might end up with a very similar encoding. Sequence based encodings usually accept only the 20 coded amino acids. That excludes peptides with modified residues, which are of considerable interest in drug design.

This work introduces a peptide encoding into hypervectors: high-dimensional vectors constructed through simple operations on a predefined set of vectors. Modified residues are encoded from their physicochemical properties, and noise tolerance allows the retrieval of encoded components. Several variants were developed, along with a similarity metric inspired by protein sequence alignment.

Simple vector operations can train a classifier in a few passes. The encodings can also serve as input to Tsetlin Machines, which require binary data and learn interpretable logical rules. Both were evaluated on established benchmarks and on a dataset assembled from public databases and literature, including negative samples and modified residues, which existing datasets lack.

Ligand-based de novo drug design using deep neural networks

David Kolář (B3)

školitel: Ing. Martin Šícho, Ph.D.

Many promising bioactive molecules fail to advance through drug development stages because of unsuitable physicochemical properties, and traditional incremental lead optimization is often insufficient. In such scenarios, deep generative neural networks can be used to discover structurally novel molecules that preserve fundamental bioactive properties. Here, we introduce a workflow relying on the DrugEx generative model for creation of an analogue data set from a single example compound using shape matching.

Specifically, this work implements a multi-objective reinforcement learning pipeline that guides a SMILES-based recurrent neural network toward high 3D shape similarity, favorable drug-likeness (QED), and synthetic accessibility. To enable this, we expand DrugEx with new capabilities, including flexible 3D scoring functions, automated reference structure handling, and fine-tuning regularizers. Finally, model performance is evaluated by comparing the top 95th percentile of generated molecules against the baseline prior.

Předpověď biologické aktivity z dat kapalinové chromatografie a hmotnostní spektrometrie pomocí strojového učení

Bc. Ester Machová (M2)

školitel: Ing. Martin Šícho, Ph.D.

Ve své práci se zabývám zpracováním dat z kapalinové chromatografie spojené s iontovou mobilitou a tandemovou hmotnostní spektrometrií malých molekul. Pomocí těchto dat se snažím vytvořit modely strojového učení, které předpovídají, zda je daná molekula biologicky aktivní a případně jak moc. Zaměřuji se především na rozličné možnosti reprezentace výstupu z tandemové hmotnostní spektrometrie (MS/MS spektra) pomocí embeddingů vhodných pro predikční modely. Dále hledám nejvhodnější sestavu a parametry vlastních modelů. Predikovaná hodnota odpovídá schopnosti dané molekuly blokovat některý krok v bakteriální proteosyntéze.

Chiroptical Properties of Helicenes: Data Acquisition, Analysis, and Modelling

Bc. Lenka Sixtová (M1)

školitel: Ing. Martin Šícho, Ph.D.

Helicenes are polycyclic aromatic systems with considerable potential for circularly polarized luminescence (CPL), a useful property in the development of chiroptical materials. Luminescence (g_{lum}) and absorption dissymmetry (g_{abs}) factors are among the main parameters that drive the chiroptical behaviour of these compounds.

No systematic analysis between the structure of helicenes and their chiroptical properties exists yet, the main hindrance being the availability of structured and reliable data. Therefore, we aimed to develop a comprehensive data extraction workflow and conduct for the first time a large-scale data-driven structure–property analysis.

For initial analysis, we manually extracted 111 molecules from 3 review articles. A random forest regressor was trained to predict g_{lum}. The results were encouraging but still limited by the small dataset size. Therefore, we also built an app combining optical chemical structure recognition and large language models to improve data collection efficiency. Being the first systematic structure-property study of helicenes' chiroptical properties, this analysis should yield valuable new insights.

ML-Driven Design of Antimicrobial Peptides

Bc. Dominika Soukupová (M2)

školitel: Ing. Martin Šícho, PhD.

Due to the ever-growing antimicrobial resistance, the search for new antimicrobial agents is essential. Many ML tools for AMPs (antimicrobial peptides) prediction are already available and can be utilized in a generalized approach for the design of new AMPs with bounded physico-chemical properties.

To address the need for new antibiotics, we implement a new computational-based method to search for novel AMPs. For generation of new potent sequences, we used both EvoGrad algorithm and our implementation of hill-climbing, both integrated with physico-chemical filters. The obtained sequences were then filtered with online Macrel and DBAASP tools. In order to preserve diversity, the final candidates were semi-manually chosen according to the predicted activity and shape (PepFold) corresponding to the intended mechanism of action.

We predicted de novo AMPs as well as AMPs from a starting point sequence and provided evaluation using publicly available ML prediction tools (AMPlify, CAMPR4_RF, AI4AMP). Some of the designed peptides were synthesized and show promising results in vitro against specific bacterial strains (S. aureus, S. epidermidis, B. subtilis, E. coli). This in-silico approach was significantly more effective at enhancing antimicrobial properties than a previously used manual design.

Efficiency metrics in total syntesis

Bc. Tereza Sujová (M1)

školitel: MSc. Wim Dehaen, Ph.D.

Total synthesis is one of the important concepts in synthetic organic chemistry. The efficiency metrics are used to compare and rank molecules, reactions, and pathways, as they can provide a helpful insight into their complexity. In order to determine and evaluate these metrics, a computational approach is often useful or, in some cases, even necessary. Here, a comparison of complexity metrics was conducted on a selection of both well-known and newer metrics of different categories. In total, eight metrics were used. Exploratory data analysis on a dataset of more than 1.4 million reactions was conducted. The data were converted into a graph where the reactions were nodes and connectivity and network analysis were performed. This dataset was also used as a training set for a sink-or-source reaction classifier. Finally, all eight metrics alongside the classifier were applied to 18 strychnine synthesis pathways in order to determine their ability to predict the key step of a reaction.

Virtual screening for potential O-methyltransferase inhibitors

Bc. Jáchym Urban (M1)

školitel: Ing. Ivan Čmelo Ph.D

Virtual screening is a set of computational methodologies often used to discover novel drug candidates within the vast chemical space of all accessible chemical compounds. In practice, this usually involves screening extensive virtual libraries of possible compounds, from catalogues provided by vendors or computationally generated libraries, against a biological target of interest.

This work involves structure-based virtual screening of O-methyltransferase from the pathogenic bacterium *Francisella tularensis*, the cause of tularemia in humans and animals both. The target protein was isolated and structurally resolved by the research group led by Evžen Bouřa. Methyltransferases are a compelling pharmacological targets, as they play a pivotal enzymatic role in a number of biological processes, including epigenetic regulation and metabolic control. Affecting these would affect the bacterium as a whole.

The developed virtual screening pipeline hinges on molecular docking, a two-step process comprising pose prediction and scoring. In this study, three docking approaches are compared and combined: AutoDock Vina (an open-source standard), Gnina (Vina enhanced with deep convolutional neural network scoring), and Protenix-Dock (an approach based on a combination of empirical knowledge and force fields). The study employs a consensus docking strategy with the objective of optimizing the reliability of virtual screening outcomes. The most promising candidate structures will be manually evaluated for viability and novelty and potentially purchased for in vitro testing.