Sessions description

1) The data-driven discovery in Life Sciences (L)

Stanislav Mazurenko, Masaryk University, Brno - stan.mazurenko@gmail.com

Our understanding of biological processes and underlying mechanisms has enabled us to engineer biological systems, design new antibiotics, improve the synthesis of biocompatible materials and renewable bioenergy, discover new enzymes, and address many other challenges facing humanity. Historically, such discoveries were driven by experimentation, and subsequent in-depth data analysis was often a prerogative of quirky scientists. In the last several decades, however, the accumulated experimental data, coupled with increasing computational power, allowed the data analysis to supplant the field- and lab-work. We are now observing an exciting symbiosis of the two approaches, cross-fertilizing and directing each other. In my talk, I will highlight the data types, algorithms, applications, and challenges that data-driven methods encounter in Life Sciences.

Keywords: big data, biotechnology, drug discovery, genes, machine learning, proteins

2) Deep learning as a tool for in silico drug screening (L,D)

Thomas Evangelidis, IOCB Prague, thomas.evangelidis@uochb.cas.cz

* Introduction to Machine Learning
* The Multi-Layer Perceptron (MLP)
* Autoencoders
* Recurrent Neural Networks (LSTM)
* Convolutional Neural Networks

A quick introduction to vector representations (feature vectors) of small molecules in Cheminformatics will be given, followed by a description of the prime neural network (NN) architectures used in Computational Medicinal Chemistry. The Multi-Layer Perceptron will be the model system to describe the basic components of a simple NN, namely the activation function, the loss function, the optimization algorithm, and regularization techniques to prevent overfitting or underfitting. Subsequently an overview of the Autoencoders will be given, followed by the Recurrent Neural Networks (RNN) and Long-Short Term Memory (LSTM) cells, which are used for SMILES string processing, illustrating analogue library generation as an example application. Finally, the anatomy and way of function of the Convolutional Neural Networks (CNN) will be described. The presentation will close with some applications of CNNs in computer-aided drug design.

Keywords: deep learning, neural networks, MLP models,cheminformatics, in silico drug screening,

3) Machine learning: an interdisciplinary inspiration (L,D)

Jan Švec, University of West Bohemia, Pilsen - honzas@ntis.zcu.cz

The progress of machine learning is driven mostly by its successful applications. The rise of deep learning techniques is observable in many research areas like computer vision and natural language processing. The talk will try to clarify different machine learning approaches ranging from pattern recognition to sequence classification and clustering. The talk will include two case studies: the application of predictive models to evaluate the cross-reactivity of allergens, and the very promising use of learned representations for protein sequence classification.

Keywords: clustering, pattern recognition, cross-reactivity of allergens, protein sequence classification

4) Machine learning-based detection of protein-ligand binding sites (L,D)

David Hoksza, Charles University, Prague, david.hoksza@gmail.com

The prediction of active sites of biomolecules is a problem that is frequently solved by machine learning techniques. In this contribution, we will show how to use the Random Forest classifier to predict ligand-binding sites from protein structure and how such a method compares to other approaches, including modern deep learning techniques. We will also show how easily can such an approach be adapted to related problems such as protein-DNA detection. Finally, we will show a software tool that implements the introduced solution.

Keywords: ligand binding, Random Forest classifier, deep learning, protein-dna interaction

5) NMR assignments using a single NOESY spectrum and machine learning

Konstantinos Tripsianes, CEITEC - Central European Institute of Technology, Masaryk University, kostas.tripsianes@ceitec.muni.cz

There is little doubt that machine learning applications are transformative technologies in most areas of our lives (data security, financial trading, healthcare, etc.). Today, image recognition by machines trained via machine learning in some scenarios is better than humans. The NMR assignment problem, if anything else, is a pattern recognition problem. First, I will discuss the workflow, advantages, and key findings of our novel algorithm (4D-CHAINS) that enables fully automated assignments of NMR chemical shifts. Then, I will justify the need for methodological breakthroughs, e.g. machine learning, to leverage the benefits of the next-generation 1.2 GHz NMR spectrometers in the NMR assignment problem. Finally, I will present the performance of new machine learning based methods and specialized graph algorithms in obtaining NMR assignments using a single 4D NOESY spectrum. Besides illustrating the merits of artificial intelligence to the given problem, the pitfalls associated with the development of machine learning models will be discussed.

Keywords: NMR, resonance assignment, automation, random forests, graph algorithms, 3D structure

6) Prediction of protein aggregation using machine learning (L)

Ekaterina Grakova & Antonin Kunka, IT4Innovations, Technical University of Ostrava & Masaryk University - ekaterina.grakova@vsb.cz, 393392@mail.muni.cz

Protein aggregation is an essential biophysical characteristic of natural proteins. It is highly relevant clinically since many critical human diseases (including Alzheimer's Disease, Huntington's Disease, Parkinson's Disease, prion diseases, etc.) are connected directly to protein aggregation. Moreover, protein aggregation is also important for the heterologous production of proteins for biotechnological applications. Currently, there are available different predictors and meta-predictors for the prediction of protein aggregation. These predictors are, however outdated and exhibited low accuracy. We aim to develop a new prediction tool for protein aggregation. The tool will take advantage of the recent success of deep learning. Deep learning models are sensitive to parametrization. Thus selecting the parametrization that will lead in acceptable predictive performance is often a subject of an exhaustive hyper-parameter search resulting in a need to compute many independent computational tasks. Similarly, it is desirable to validate the predictive model performance using cross-validation techniques which again results in a set of relatively independent tasks that need to be computed. There are tools such as HyperLoom that allow users to pipeline such computational tasks into computational workflows and automatise their efficient execution in distributed environments such as high-performance computing clusters. In our contribution, we will discuss our deep-learning based approach, its parametrization and preliminary results on the predictive accuracy.

Keywords: aggregation, Alzheimer's Disease, deep learning, Parkinson's Disease, machine learning, high-performance computing