ISSN 2186-7437

NII Shonan Meeting Report

No. 179

Computational metabolomics and machine learning

Sebastian Böcker
Hiroshi Mamitsuka
Juho Rousu

May 8–11, 2023

National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-Ku, Tokyo, Japan

Organizers

Sebastian Böcker (Friedrich Schiller University, Jena, Germany)
Hiroshi Mamitsuka (Kyoto University, Japan)
Juho Rousu (Aalto University, Finland)

Attendees

Wout Bittremieux (University of Antwerp, Belgium)
Sebastian Böcker (Friedrich Schiller University, Jena, Germany)
Roman Bushuiev (The Czech Academy of Sciences, Prague, Czech Republic)
Jacques Corbeil (Université Laval, Québec, Canada)
Sašo Džeroski (Jozef Stefan Institute, Ljubljana, Slovenia)
Kai Dührkop (Friedrich Schiller University, Jena, Germany)
Markus Heinonen (Aalto University, Finland)
Lukas Käll (Science for Life Laboratory, Sweden)
Fleming Kretschmer (Friedrich Schiller University, Jena, Germany)
Hiroshi Mamitsuka (Kyoto University, Japan)
Lennart Martens (Ghent University and VIB, Belgium)
Grégoire Montavon (FU Berlin, Germany)
Shohei Nakamukai (RIKEN CSRS, Yokohama, Japan)
Dai Hai Nguyen (University of Tsukuba, Japan)
Tomas Pluskal (The Czech Academy of Sciences, Prague, Czech Republic)
Juho Rousu (Aalto University, Finland)
Hiroto Saigo (Kyushu University, Japan)
Michael Stravs (Eawag, Dübendorf, Switzerland; ETH Zürich, Zürich, Switzerland)

Abstract

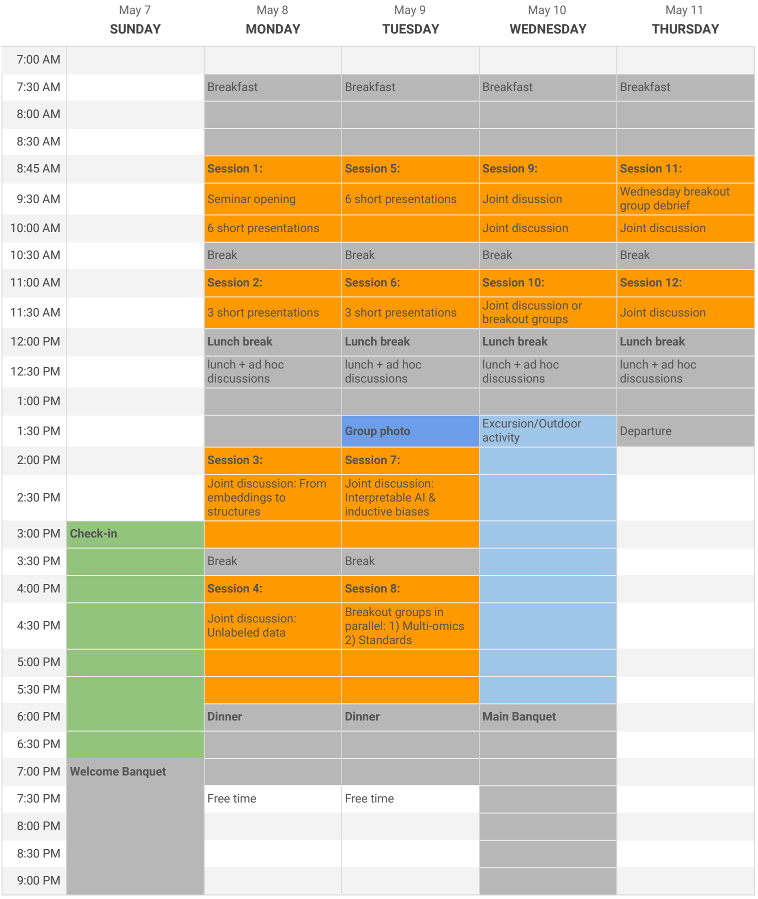
Machine learning is crucial for small molecule research, aiding in structure annotation, understanding protein and small molecule interactions, and uncovering cellular mechanisms. Metabolomics, the study of small molecules, is sometimes considered the pinnacle of omics-sciences as it closely relates to biological phenotypes. Mass spectrometry is the primary method for detecting and identifying metabolites and small molecules in high-throughput experiments. Technological advancements have enabled novel investigations but also led to a significant increase in data complexity, posing challenges in data interpretation. Machine learning has made significant progress in small molecule identification, although the problem is far from being solved. In drug discovery, machine learning has been instrumental in high-throughput anticancer drug screening, guiding precision medicine and drug repurposing efforts. The interpretation of metabolomics data is highly challenging, surpassing methods used in genomics or proteomics. Representing metabolites as graphs for machine learning algorithms requires special considerations, and available training data is often non-representative and lacks uniformity. Generalization beyond existing data is difficult, and overfitting is a common issue. The seminar facilitated idea exchange between bioinformatics and machine learning experts in the small molecule field. A diverse group of scientists met at the workshop, including established researchers and promising talent from machine learning, bioinformatics, metabolomics, and drug discovery fields. Short presentations highlighted the techniques and challenges, while brainstorming sessions and break-out groups encouraged detailed discussions and fostered collaborations among participants. This report has been written in cooperation with all attendees.

Introduction

In recent years, machine learning has emerged an important tool in the research of small molecules, aiding their identification from measurement data, deciphering their interactions with proteins and other small molecules, and helping to elucidate the inner workings of cellular machinery. The research on the small molecule complement of the genome, metabolomics, has been referred to as the apogee of the omics-sciences, as it is closest to the biological phenotype. Metabolites are not only responsible for tasks such as growth, development, and reproduction, but also directly relevant to structure, signaling, and chemical interactions with other organisms. Most pharmaceuticals are small molecules that bind to their targets, thus altering their behavior. In small molecule identification, mass spectrometry is the predominant analytical technique for detecting and identifying metabolites and other small molecules in high-throughput experiments. Huge technological advances in mass spectrometers and experimental workflows during the last decades enable novel investigations of biological systems on the metabolite level. But these advances also resulted in a tremendous increase of both amount and complexity of the experimental data, and "making sense" of the data is among the most pressing issues in high-throughput settings. Machine learning methods for small molecule identification have made great progress during the last decade, however, the identification problem is still far from being "solved". In drug discovery, several high-throughput anticancer drug screening efforts have been conducted, providing drug interaction and response measurements that allow for the identification of compounds that show increased efficacy in specific human cancer types or individual cell lines, therefore guiding both the precision medicine efforts as well as drug repurposing applications. Machine learning methods have shown their potential in these tasks, in e.g. several recent DREAM challenges organized around the theme. Similarly, in functional genomics, the prediction of biosynthetic gene clusters through machine learning is an active topic, concerned with the elucidation of the metabolites associated with a biosynthetic pathway of an organism. During the last decade, metabolomics has seen numerous cooperations between experimental and computational scientists. It turns out that the interpretation of the data is highly challenging, and as soon as one goes beyond the presence or absence of peaks in MS1 experiments, methods which have been developed in genomics or proteomics cannot be applied to metabolomics data. In particular, the application of machine learning techniques is impeded by several issues: For example, metabolites are graphs and representing them for machine/deep learning algorithms requires special care. (In truth, the situation is even worse, as even the representation of molecular structures as graphs is too restricted and often plainly wrong.) Also, available training data for small molecules is usually very far from being a representative, even less so a uniform subsample of the complete space of molecules. Thus, generalization outside the current data is challenging and machine learning methods are prone to overfit. The key goal of this seminar was to foster the exchange of ideas between bioinformatics and machine learning for small molecules. We invited to the workshop a diverse group of scientists working in the fields of machine learning, bioinformatics, metabolomics and drug discovery, including both leading names in their fields as well as young rising talent. State-of-the-art methods from computer science, statistics, analytical and biological experiments were presented through short presentations, along with problems arising from these techniques. Brainstorming sessions and break-out groups were used to discuss individual topics in greater detail, to initiate new collaborations between participants who have not yet worked together. The organizers are confident that the exchange of expertise in the seminar will result in several scientific advances in the coming years, which will push forward computational metabolomics as a field.

Schedule Of The Seminar

On Monday morning (Sessions 1 and 2) and Tuesday morning (Sessions 5 and 6), participants gave short introductory talks, either about their work or about open questions and future directions. The other sessions were used for open discussions, with topics chosen by the participants before and during the seminar. Some sessions were held in parallel, in case not all of the participants were interested in the same topic; this also allowed us to cover a more comprehensive selection of topics. An outing on Wednesday complemented the scientific program. The two sessions on Thursday were used for some short scientific discussions, a general wrapup of the seminar, a discussion of future seminars, as well as a discussion on the general direction the field is heading and its perspectives.



	May 7 **SUNDAY**	May 8 **MONDAY**	May 9 **TUESDAY**	May 10 **WEDNESDAY**	May 11 **THURSDAY**
7:00 AM					
7:30 AM		Breakfast	Breakfast	Breakfast	Breakfast
8:00 AM					
8:30 AM					
8:45 AM		**Session 1:**	**Session 5:**	**Session 9:**	**Session 11:**
9:30 AM		Seminar opening	6 short presentations	Joint disussion	Wednesday breakout group debrief
10:00 AM		6 short presentations		Joint discussion	Joint discussion
10:30 AM		Break	Break	Break	Break
11:00 AM		**Session 2:**	**Session 6:**	**Session 10:**	**Session 12:**
11:30 AM		3 short presentations	3 short presentations	Joint discussion or breakout groups	Joint discussion
12:00 PM		**Lunch break**	**Lunch break**	**Lunch break**	**Lunch break**
12:30 PM		lunch + ad hoc discussions	lunch + ad hoc discussions	lunch + ad hoc discussions	lunch + ad hoc discussions
1:00 PM					
1:30 PM			**Group photo**	Excursion/Outdoor activity	Departure
2:00 PM		**Session 3:**	**Session 7:**		
2:30 PM		Joint discussion: From embeddings to structures	Joint discussion: Interpretable AI & inductive biases		
3:00 PM	**Check-in**				
3:30 PM		Break	Break		
4:00 PM		**Session 4:**	**Session 8:**		
4:30 PM		Joint discussion: Unlabeled data	Breakout groups in parallel: 1) Multi-omics 2) Standards		
5:00 PM					
5:30 PM					
6:00 PM		**Dinner**	**Dinner**	**Main Banquet**	
6:30 PM					
7:00 PM	**Welcome Banquet**				
7:30 PM		Free time	Free time		
8:00 PM					
8:30 PM					
9:00 PM					

Session Abstracts

Monday 14:00-15:30: From embeddings to structures

We first discussed different types of embeddings: Namely, spectrum to embedding to molecular structure, protein to embedding to molecular structure, or two embedding layers for mass spectra and structure information. The various generative modeling approaches were discussed at length. The utility of stable diffusion for generating molecules from spectra was discussed (e.g. DiffLinker). The validity of the generated molecules is an important issue (proper bonds, valencies etc.). Various variants such as conditional diffusion and classifier-free guidance were discussed as potential methods. The (lack of) invariance with respect to permutations of the adjacency matrix was noted as an issue for discrete diffusion models which is not present in the coordinate-based diffusion models. Modular flows were mentioned as a good method to generate high-validity molecules, but not good for molecular property optimization. Methods for assessing the aptness of a molecule generation was also discussed at length. Another discussion thread centered around the use of SMILES representations, which allow the use of the generative models developed in the NLP camp. The limitations of SMILES and their practical relevance were discussed, their observed tendency to frequently generate molecules that don't look realistic to chemists and the lack of smoothness of the embedding space being the issues. How to quantify the "unhappiness" of the chemists to the generated molecules was also brought forward as a question. Beyond the validity of the structures, also their stability could be taken into account (e.g. through Group Contribution Theory).

Monday 16:00-18:00: Using unlabeled data for training

This session covered the possibilities of using large unlabeled datasets of small molecules and mass spectra. We discussed our experience and the most promising future directions based on results from other domains such as natural language processing (NLP) and protein representation learning. We agreed that the most valuable source of raw LC-MS/MS data is the GNPS part of MassIVE. More precisely, Kai's filtering of GNPS resulted in 500K to millions of unlabeled molecules, while Roman's filtering produced datasets containing 2M to 700M spectra with varying degrees of quality and redundancy. Comparatively, parsing the MetaboLights repository proved to be challenging. Kai investigated weak supervision with labels from COSMIC for subsequent structure prediction, but it didn't work. Roman experimented with self-supervised masking of peaks (m/z ratios, intensities, or both values), and the masking of two m/z ratios seems to be a reasonable pre-training objective. By validating the training after each epoch, the neural network gradually derives structural properties of small molecules. However, in Wout's analysis of a similar approach for proteomics, masking objectives were too simplistic for the model. Furthermore, we aimed to understand why self-supervised pre-training does not work as effectively for small molecules as for protein sequences. Our overall conclusion was that while protein space is sparse when deleting an amino acid, replacing a functional group or an atom in a small molecule can be achieved through various options. Although this conclusion does not hold true consistently for all molecular classes, on average, it may be a severe bottleneck for self-supervision. To address this issue, ”state-of-the-art” methods experiment with augmenting masking objectives with various other tasks, such as predicting molecular properties derived from molecular structures. However, the prediction of more complex properties (yielding richer representations of molecules) often relies on experimental annotations and, therefore, has limitations for pre-training. Finally, we briefly discussed practical considerations for experimental scientists when collecting new LC-MS/MS data. Our first observation highlighted the homogeneity of GNPS in terms of molecules, such as the prevalence of bile acids and the insufficient representation of other compound classes. Ideally, the GNPS dataset should include a more balanced distribution of underrepresented metabolites. Our second observation emphasized the difficulty for machine learning to learn from “dirty” data, such as chimeric spectra from dissolved organic matter (DOM) samples. Therefore, it is important to explicitly mark or perform algorithmic analysis on such datasets, simplifying their utilization as pre-training data.

Tuesday 14:00-15:30: Inductive bias & explainable AI

The session covered different ways of introducing prior knowledge into the model (inductive bias), how this can be done a priori (by particular choices of data/models/objectives) or a posteriori (via Explainable AI). The section also covered more generally the question of explainability of ML models. The meaning of inductive bias was first discussed, and we adopted a broad interpretation covering prior knowledge in the Bayesian sense, knowledge introduced by the human into the model, regularization, bias in the sense of bias/variance tradeoff. Distinctions have also been made, e.g. between language bias and preference bias. Afterwards, the practical motivations for introducing bias were discussed, as well as some potential disadvantages compared to fully relying on the data. A primary reason for introducing bias is to resolve limitations intrinsic to the available data (e.g. imposing a preference between different spuriously correlated features). When data coverage is comprehensive, however, the inductive bias can become a burden. Then, various forms of biases/prior knowledge have been described for specific types of data and applications. This includes biases specific to SMILES and SELFIES data, inductive biases on the level of the molecule (e.g. MSNovelist, feature engineering to specify a molecular formula) and inductive biases on the level of spectra. These biases on the spectra include replacing mass with formulas, isotopes in the spectra, forcing representation similarity of different spectra from the sample, including features from related spectra (e.g. adding chemical structure of the nearest neighbor spectra in the similarity group, possibly including other spectra types). Additionally, possible refinements of the similarity measure between different spectra was discussed, for example, based on a chain of instances with strong mutual similarity. Further examples of inductive biases have been discussed for the task of protein prediction, where it was mentioned that a huge boost of performance is obtained when including evolutionary structure (e.g. flavonoids). A second part of the session has focused on the way bias should be introduced into the model for a maximum benefit. The presence of spurious correlations in medical data (e.g. caused by parameters of the data acquisition being correlated to the task) and the resulting “Clever Hans” classifiers was stressed. The fact that not only the data but also the ML algorithm may cause or influence the emergence of such flaws was mentioned. Among methods for model improvement, a distinction was made between proactive approaches (e.g. feature selection, regularization) and reactive approaches (e.g. based on explainable AI). A potential advantage of the latter approach is that it allows to only introduce prior knowledge when the model needs it (thereby not overriding what can be learned readily from the data). On the other hand, it requires a user-in-the-loop, which would work well only in certain scenarios (e.g. with willing cooperating partners). The usefulness of interpretability for finding flaws in the model was discussed. It was mentioned that different types of interpretability (e.g. attribution vs. counterfactuals, vs. support vectors) may be needed for different types of data in order for the explanation to be informative for the user. Lastly, the benefit of interpretability for enhancing acceptability of a ML model (independent on whether it allows to improve the model or not) was mentioned.

Tuesday 16:00-18:00: Where is the data? Standardized datasets, competitions

One of the primary issues highlighted was the dispersed nature of the available data. Numerous online repositories and datasets house relevant information, making it difficult for researchers to identify relevant data resources. Various data types were identified, including MS data, retention time data, ion mobility data, and molecular data. Noteworthy online resources for these data types were mentioned, such as GNPS, MoNA, MassBank, MetaboLights, NIST, METLIN SMRT, RepoRT, PredRET, CCSbase, COCONUT, HMDB, PubChem, LOTUS, CheBI, and the boeckerlab biodatabase. To address this issue, we propose collecting representative data and presenting it in a format that is easily usable by ML practitioners without requiring extensive domain expertise. A previously compiled dataset from GNPS, which was employed in the development of CANOPUS, will be updated and shared with the scientific community through a dedicated website. This dataset will be divided into appropriate training, validation, and testing data subsets. This approach aims to enable researchers to develop and evaluate novel bioinformatics tools consistently, ensuring comparable results across studies. In addition to the MS data, the dataset will also incorporate a biomolecule dataset serving as a structural database for searching purposes. Drawing inspiration from other fields that have greatly benefited from structured competitions, such as ImageNet and CASP, we will organize a new CASMI challenge, led by Tomas Pluskal. This competition will be divided into two categories: (i) cases where the correct answer is available in PubChem and (ii) cases where the correct answer is missing from PubChem. In the latter category, participants will be evaluated based on the structural similarity of their proposed molecule with the ground truth molecule, rather than a binary 0/1 evaluation solely based on an exact match. This discussion highlighted the urgent need to address the challenges associated with data accessibility and integration in the field of computational metabolomics and machine learning. By streamlining data access and establishing evaluation benchmarks, we aim to facilitate innovation and drive progress in computational metabolomics and machine learning.

Tuesday 16:00-18:00: Multi-OMICS

The session covered two specific multi-omics use cases: spatial omics, and metabolic rare diseases. The first topic, spatial omics, concerns the combination of two relatively recent innovations: imaging mass spectrometry (iMS), and spatial transcriptomics . Both approaches create a very high dimensional 'pixel map' of a slice of a tissue, an organism, or (a set of) bacterial colonies. An interesting approach to combine the different data modalities is likely CCA. CCA enables time courses, which could be adapted to a spatial dimension rather than a time dimension. Moreover, sparse-nonlinear CCA holds the promise of picking up small numbers of features which might induce non-linear correlations (e.g. gradKCCA, SCCA-HSIC). Another suggestion was spatial mixture models, to capture and elucidate a possible genotype or phenotype mixture across the image. This approach appears to mesh well with the diversity use cases (e.g., differential drug uptake across cells in a tissue). The second topic concerned a connection between proteomics data-driven protein-functional association networks, genomic sequencing data, and identified metabolites, as together these can reveal fine-grained information on the dysregulation of metabolic pathways in metabolic rare diseases. These diseases, which often remain poorly or incompletely diagnosed, tend to have very few cases and several matched controls, yet genomic sequencing (which is the standard approach to investigate these diseases) usually reveals far too many variations to elucidate a clear point of failure in the patient's metabolism. Suggestions to cope with this situation included explainable AI to visualize the induced 'explanations' for different individuals, and anomaly detection to find deviations from the norm across the available data. L1 or group LASSO regularization was also seen as relevant, as the disease cause tends to be a single, or very few, mutation(s). Flux balance analysis could help pick up metabolic fluxes through the pathways, but not concentrations as enzyme kinetics are likely not known.

Wednesday 8:45-10:30: Hammers in search for nails

This session was an attempt to brainstorm novel applications for interesting algorithms. We discussed the application of Explainable AI (XAI), which uses different strategies to highlight particular features in a model input that are the main determinants particular model prediction. Potential applications in computational metabolomics include highlighting peaks in an input that are responsible for the ranking of two candidates. Similarly, counterfactual explanations may contrast two input spectra and explain why there is a model response in one but not another spectrum. Frequently, the input data to such a model is not directly interpretable. A solution for this is an additional invertible transform (autoencoder) as first layers of the model, which maps the model input to an interpretable counterpart. For proper results, it is mandatory that this transform have negligible reconstruction loss. This is particularly important if trying to use XAI to find flaws in a model, possibly less relevant if using XAI as a user-facing interpretability tool. Finding interpretable explanations on surrogate models should be avoided, since those might in fact learn different algorithms to arrive at the same solution; for the case of "clever Hans" phenomena, they may be present in both, neither, or only one of the full versus the surrogate model. We discussed autoencoders to detect anomalies in time series, where an increasing reconstruction loss shows an overall drift. In clinical settings, where mass spectrometry will increasingly monitor larger numbers of metabolites, such anomalies of interest could be similar to drift, as in slowly deteriorating towards a medical problem. Similarly, in laboratory settings such models may indicate the need for preventive maintenance. Finally, we considered Bloom filters, a probabilistic data structure to test for set membership. In the context of mass spectrometry, this could have applications in a database search setting if combined with an appropriate discrete representation of spectra, such as locality-sensitive hashing methods.

Wednesday 8:45-10:30: Retention Time

Fleming and Sebastian, together with the group of Michael Witting (Munich), have created a repository for retention times (https://github.com/michaelwitting/RepoRT). The important point is that for retention time (RT) and order predictions, you need to know several things beyond the molecular structure itself: This includes the stationary phase (column), the mobile phase (eluents, gradient) and even the temperature. All of these things do affect retention time and even retention order. There was a short discussion that even columns of the same type (e.g, C18) columns can have a (very) different retention order. The repository contains fewer than 10k unique structures, but Tomas offered to measure about 15k compounds. This would substantially increase the covered chemical space. In the best case, multiple measurements on different columns, but that would require major amounts of time. We then had a longer discussion about stereochemistry: Much information is still missing, so the way to go might be to first drop stereochemistry information altogether and only consider it later on when a foundation has been laid. Other cases like mesomerism and tautomerism were also discussed; tautomers could create two peaks in the chromatogram (so two RTs) but it's also possible to just have a broader peak. There is an unresolved issue that in some cases, there are multiple RT values for a single compound. It is possible that these are actually mixtures of different stereoisomers, or possibly something else. We discussed how to handle those, and the feeling was: If those are indeed different stereoisomers, then take the average. Predicting the average would already be helpful. We also discussed special purpose columns and how a model could help users to select a column best suited for their application. Moving to mass spectra in the context of stereochemistry, we decided that it is probably not an issue right now (fragmentation similar; what happens in the gas stage is really different than what happens in solution anyways). There was a short discussion on ClassyFire vs. NPClassifier regarding pros and cons; it looks like there is no optimal solution right now.

Wednesday 11:00-12:00: Large scientific models

"Large language models" (LLMs), Transformer-based models pre-trained with vast amounts of natural language data and a very large parameter space, have found widespread application for any task related to natural language. LLMs serve catch-all foundation model which can be fine-tuned for specific tasks with limited amounts of data, frequently outperforming models designed for a specific purpose. Herein, we explored the concept of a "large chemical model" or "chemical foundation model", pre-trained on very large chemical datasets, as a basis for fine-tuning for tasks of interest specific for computational metabolomics. In proteomics, protein sequence data has been used to train LLMs analogously to natural language LLMs, e.g. by masking parts of the sequence. The latent representation of the resulting foundation model (ProteinBERT) was used as an input to train peptide MS2 prediction, outperforming the state of the art, and showed promising performance in other protein-related tasks seemingly unrelated to the input data. An analogous attempt for small molecules would entail masking parts of a molecule, e.g. subgraphs. However, masking in molecules is not a promising choice, as there are hardly any meaningful restrictions on valid subgraphs. Possibly, masking bonds rather than atoms could be a more suitable task. More suited tasks for a base model could be translation tasks (between chemical representations) or property prediction tasks. Some properties (such as fingerprints) are probably insufficiently complex to learn a foundation model. Promising features could be results from density-functional theory (DFT) calculations, for which multiple datasets are available. Further discussion revolved around suitable datasets, e.g. MD17 and QM7-X. We suggest avoiding experimental data related to biology, such as toxicity. Finally, since chemical property prediction is a key topic in drug development, we discussed whether existing models in these communities might be unknown to the computational mass spectrometry community. A quick literature survey reveals multiple BERT-inspired chemical models, though with unclear scope and performance.

Wednesday 11:00-12:00: Metabolomics responses of drugs

We discussed the problem of learning from metabolomic time-series that result as a response of administering a drug. Discussions focused on the discovery of change points (e.g. the onset and probable path of a disease) and event sequences from multiple time series, as well as feature extraction for predictive models. A major bottleneck is the lack of datasets in the public domain that contain this type of information.

Thursday 11:00-11:30: Do current spectral libraries cover the metabolite space well or poorly?

Different participants have come to different conclusions on whether the available mass spectrometry data is covering the "space of biomolecular structures" well or poorly. We discussed that Tanimoto distances are not well-suited to measure chemical similarity or dissimilarity, despite their undoubted advantage of allowing ultrafast screening. A better choice might be MCES distances (Maximum Common Edge Subgraph distance, graph edit distance). See the MCES paper which claims that MS data are actually covering the space rather well. There, also Jupyter Notebooks are provided, and one can interactively explore the UMAP embeddings. Disadvantage of MCES is that a single instance requires between 0.1 and 1 seconds; so, downsampling is necessary. Roman's thesis (available soon) claims mostly the opposite: namely, that current MS training data form small cluttered islands within the otherwise rather empty "sea of biomolecules". We will have to see what statement is closer to the truth; potentially both claims are true at the same time.

Participant Abstracts

Bittremieux

There is a key need to develop novel bioinformatics tools that better communicate uncertainty in the data analysis results of untargeted metabolomics experiments. For example, currently isomeric alternatives for spectrum annotations are not or not sufficiently considered, even though these can likely not be differentiated using mass spectrometry. Similarly, there is an urgent need to provide statistical confidence estimation of the spectrum annotation results. Additionally, this should be combined with a community education effort. Acknowledgements: Wout Bittremieux was supported by a travel grant from the Research Organization – Flanders (FWO).

Böcker

Large-scale datasets for small molecule structure may pose severe problems for machine learning models: This includes bias in the selection of training data, as well as incompletely labeled training data due to mesomerism, tautomerism, and racemates. It is easy for a machine learning model to pick up this bias, resulting in evaluation statistics much better than what we will observe in practice.

Bushuiev

Mass spectral libraries are limited to known molecules that can be easily acquired or isolated. To address this limitation, we are developing a self-supervised deep learning model capable of extracting knowledge from millions of raw, unannotated mass spectra. We collected a dataset of 700 million experimental mass spectra from diverse LC-MS/MS measurements and used it to train a large neural network in a self-supervised manner. By training the network on artificial tasks, such as predicting masked sections of the input spectra, we observed the emergence of rich molecular features derived directly from the experimental mass spectra.

Corbeil

Our team identifies disease signatures using metabolomics to inform diagnostics, monitor treatments and develop new drugs. We combined high-throughput mass spectrometry and machine learning to identify these signatures. For the drug development approach, we use the signature as an objective function to inject biological knowledge into a generative flow network to drive the process of finding new compounds of interest.

Dührkop

My research involves developing machine learning methods for small molecule annotation. Next to model architecture, the choice of input/output features is crucial. While many methods use binned vectors for spectral input representation, they fail to capture high mass accuracy details and the combinatorial nature of small molecule fragmentation. Fourier features combined with transformer methods show promise, but their ability to model mass deltas and mass defects remains uncertain. Kernel methods are well-suited for combinatorial data like trees and currently outperform transformer models in small molecule annotation. However, training times for kernel methods increase cubically with training data size. The Nyström approximation offers a solution by enabling linear scaling of running times with additional training data and facilitating the combination of deep neural network architectures with kernel methods.

Dzeroski

My research is best described by the keyphrase “Artificial Intelligence for Science”. It covers topics such as the analysis of complex data (relational learning, structured output prediction, semi-supervised learning), automated modeling of dynamic systems, and ontologies for (computer) science. The connection to metabolomics is through existing work on the analysis of spectrometry data, including mass spectrometry and Raman spectrometry data, and planned work on relating mass spectrometry data (MALDI-TOF, FTIR) to pathogen information (identification, predicting antibiotic resistance profiles).

Heinonen

My research interests are in probabilistic deep learning, dynamical systems and in generative modeling, with examples in molecular and protein complex generative models. Current molecular generative models are largely isolated for one task and still struggle to understand simple concepts such as valencies. There is a need for chemistry-aligned embeddings that understand chemical structures, energy landscapes, molecular properties and interactions; conceptually similar to the ‘foundation’ models of language and images.

Käll

My group is developing methods for interpreting high-throughput experiments, particularly for mass spectrometry-based proteomics. In my talk I discussed a novel type of multiomics experiment I have been obtaining data from, a spatial co-analysis of transcripts and metabolites in tissue. The data is obtained by first applying MALDI Imaging mass spectrometry, and subsequently performing Visium analysis for the same tissue. We see this as a promising resource for determining covariation between metabolites and transcripts (or lack thereof).

Kretschmer

A generalizable model for retention time (better: order) prediction requires both a large collection of diverse datasets in terms of chromatographic setups used and structures measured. A repository of suitable datasets with a numeric description of the chromatographic setup, easily pluggable into machine learning, is now available (RepoRT), but important challenges to make the most out of training data still have to be overcome. How to handle wrongly labeled data, stereoisomers and tautomers are among the most pressing issues.

Mamitsuka

Machine learning on graphs, higher-dimensional data (tensor and hypergraph) and data integration, and bioinformatics on sequences, networks, biomedical text mining and a variety of drug-related issues are the current major research topics of my group. A recent topic, a hypergraph neural network-based sparse stochastic block model for predicting drug-drug interactions (adverse drug effects caused by drug pairs) was briefly presented.

Martens

My group is looking into the interface between metabolomics and proteomics, as evidenced by metabolites (or metabolic intermediates) that are found as modifications on proteins. This promising avenue of research, for which we already have some first indications of biological and biomedical relevance, is likely to prove to be a fruitful endeavor over the next years. But it is clear that any decent effort in this area will benefit greatly from interactive involvement of both the proteomics and the metabolomics communities.

Montavon

My research interests are in the development of Explainable AI approaches targeted for practical applications. This includes methods to systematically inspect a ML model, e.g. to detect the potential use of spurious features (aka. Clever Hans effects) by the model, and remove them from the model. Another set of applications are to gain novel insights into complex systems of interest (seen through a ML model), for example, for predicting proteomic networks.

Nakamukai

The retention time in LC/MS is one of the information for compound identification. However, different measurement conditions result in giving different retention time values. Therefore, the objective of my research is to explore the use of retention order in LC/MS for compound identification, especially peptide natural products whose database information is limited. To address the issue of limited training data in the database, I plan to use peptide data derived from the peptidomics to train the amino acid portion and to use small molecules from the databases to train modifications.

Nguyen

With the rapid development of machine learning (ML), there is much potential for data-driven biological knowledge discovery. However, it is not straightforward due to the complexity of the domain knowledge inherent in the data and also there are rich interdependencies among biological components such as atoms, molecules, cells or organisms. ML approaches in this domain usually involve analyzing such interdependence structures encoded by graphs. My research focuses on developing new ML methods for such kinds of graph structured data with theoretical guarantees and improved performances compared to the state-of-the-art methods.

Pluskal

Although plants are an incredibly rich source of pharmaceutically relevant specialized metabolites, biosynthetic pathway elucidation in plants has proven challenging. My lab is developing workflows for connecting biosynthetic gene sequences (RNAseq data) to their downstream metabolites (LC-MS data). For this, we designed a "top-down" approach based on correlating expression levels of enzymes with metabolite abundance across a large plant family, and a "bottom-up" approach based on predicting the substrate specificity and catalytic function of individual biosynthetic enzymes directly from their sequences using self-supervised deep learning. We are also developing large-scale foundational deep learning models for mass spectra that could be used for predictions of chemical structures and for assessing the novelty of detected natural products.

Rousu

The goal of my research group is to develop principled machine learning methods for predicting structured, non-tabular data arising in biomedicine, drug discovery and systems biology. Our current methodological focus includes tensor-based models, sparse kernel models and representation learning for structured objects. Applications include drug combination prediction, complex biomarker discovery, and enzyme function prediction.

Saigo

My research interests are in machine learning and its application to bio/chem informatics. My first topic was about supervised learning of small compounds, in which kernel methods and GNN are the popular choices, I have introduced an interpretable alternative that makes a linear model out of the space of all the subgraphs. My second topic was about prediction of chemical reactions, in which the metabolic network is regarded as a graph, where nodes and edges are compounds and reactions, respectively, and the task is to fill missing reaction categories.

Stravs

Current methods for structure elucidation of small molecules rely on finding similarity with spectra of known compounds, but do not predict structures de novo for unknown compound classes. Existing methods (CSI:FingerID) predict molecular fingerprints from MS2 spectra, and search for matching chemical structures in databases. MSNovelist uses the rich structural information in predicted molecular fingerprints as an input for molecule generation. For this purpose, a LSTM neural network was enhanced with feature engineering to favor the formation of compounds with specified molecular formulas, including a self-supervised LSTM to learn implicit hydrogen atoms on SMILES. The resulting model is able to reproduce 60