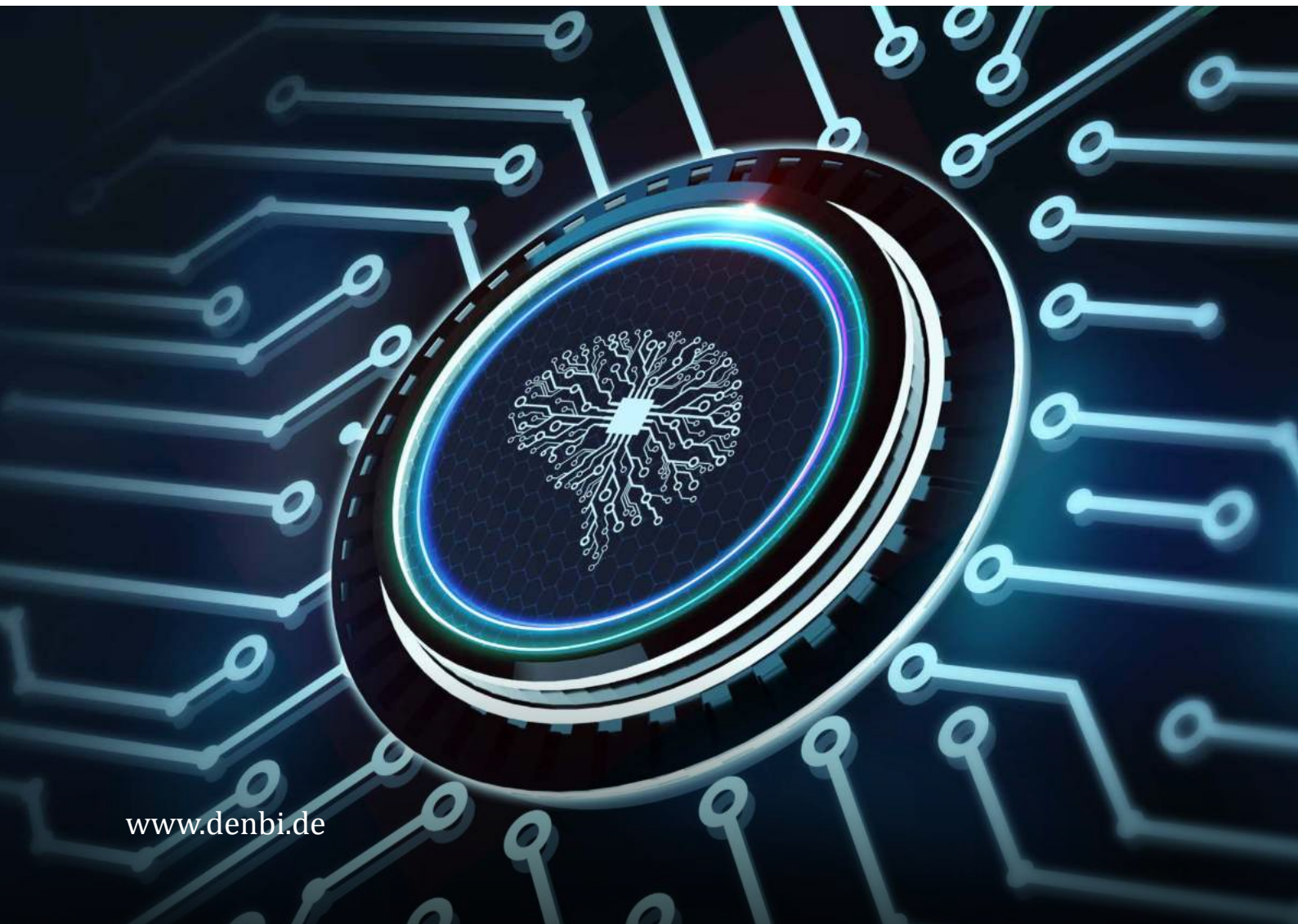


FUTURE-ORIENTED ANALYSIS OF LIFE SCIENCES DATA USING ARTIFICIAL INTELLIGENCE

Contributions of the German Network for
Bioinformatics Infrastructure



FOREWORD

Dear Reader,

Breakthroughs in deep learning led to a great comeback of the topic artificial intelligence (AI) and raises hopes for great innovations. This spectacular progress in the field of AI is only made possible by large amounts of available data for learning. AI can perform complex tasks while simulating human-like levels of intelligence and has stepped in to navigate the scientific community through the enormous ocean of data produced all over the world. And thus, AI will transform research in the life sciences.

This brochure gives an overview about de.NBI and ELIXIR Germany activities in the field of AI, which are carried out by de.NBI members and/or with de.NBI resources. In total, we highlight 16 projects showing various aspects of integration and usage of de.NBI resources in AI projects, starting from prediction and modelling with the support of AI, improvement for research services and the acceleration of science through the application of AI. This booklet demonstrates that the existing diverse bioinformatics infrastructure of the de.NBI network and ELIXIR Germany is aware of the AI potential for the life sciences community.



Prof. Dr. Andreas Tauch

The German Network for Bioinformatics Infrastructure (de.NBI) and the German Node of the European ELIXIR network are aware of the potential of AI and the richness of data and therefore bundle all forces and bioinformatics experts to meet this challenge. To handle, analyze and store Big Data, de.NBI provides bioinformatic tools and infrastructure like the de.NBI cloud, to apply AI to answer biological questions and improve AI algorithms. Moreover, de.NBI makes data available in a transparent, democratically controlled, and directly usable form (FAIR principles).



Prof. Dr. Alfred Pühler

The editorial team and the authors of this brochure, hope that we can provide all interested readers with exciting insights into current research approaches in the field of AI. We wish you enjoy reading these articles.

Andreas Tauch
Head of Node of ELIXIR Germany

Alfred Pühler
de.NBI Coordinator

CONTENT

FOREWORD	3
CONTENT	4
THE GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE (de.NBI)	6
THE GERMAN NODE WITHIN ELIXIR EUROPE	8
de.NBI CLOUD – CLOUD COMPUTING FOR LIFE SCIENCES	10
de.NBI SERVICES – TOOLS, WORKFLOWS, DATABASES, CONSULTING	12
de.NBI TRAINING FOR LIFE SCIENTISTS	13
de.NBI INDUSTRIAL FORUM – INDUSTRY SERVICES, CONSULTING, NETWORKING	14
de.NBI USER MEETINGS	15

PREDICTION REQUIRED – PHENOTYPING AND MODELLING OF CELLULAR ROCESSES USING AI

16

DEEP-iAMR – IDENTIFICATION OF NEW ANTIMICROBIAL RESISTANCE TARGETS BY HIGH-THROUGHPUT DEEP LEARNING	18
DEEP LEARNING FOR ANALYZING MICROSCOPY IMAGES – COMPUTER-BASED IMAGE ANALYSIS AND CELLULAR PHENOTYPING	24
REmatch: AI FOR DRUG DISCOVERY AND REPURPOSING – IMAGE-BASED PROFILING TO CREATE A HIGH-RESOLUTION REFERENCE MAP OF TARGETABLE CELLULAR PATHWAYS	28
DEEP LEARNING-BASED CANCER PATIENT STRATIFICATION	34
MIDAS – MEDICAL IMAGE AND DATA ANALYSIS – WHY THE COMPUTATIONAL INFRASTRUCTURE MATTERS	40
HIDDEN PHENOTYPES – MICROPHENOMICS REVEALS NOVEL DISEASE RESISTANCE GENES USING DEEP LEARNING AND AUTOMATED MICROSCOPY	44

BIOLOGICAL DATA MEETS AI – OMICS, BIG DATA AND MACHINE LEARNING AS TOOLS TO ACCELERATE UNDERSTANDING OF BIOLOGICAL MECHANISMS

50

AI BASED METHODS FOR PLANT PROTEIN FUNCTIONAL INFERENCE – USING LEARNED PROFILE HIDDEN MARKOV MODELS ALLOWS PLANT GENOME COMPARISONS	52
MACHINE LEARNING FOR ELUCIDATING MICROBIOME FUNCTIONS – MACHINE LEARNING APPROACHES FOR THE CHARACTERIZATION OF MICROBIAL SECONDARY METABOLISM AND ASSOCIATIONS WITH HOST TRAITS	56
DEEP LEARNING FOR PROTEIN VARIANTS DETECTION – DeProVIDEO WILL FACILITATE THE IDENTIFICATION OF PROTEIN VARIANTS IN MASS SPECTROMETRY- BASED PROTEOMICS EXPERIMENTS	62
DeepSIVaL: DEEP LEARNING PEPTIDE SPECTRUM IDENTIFICATION VALIDATOR – DEEP LEARNING APPROACHES FOR PEPTIDE-SPECTRUM-MATCH VALIDATION PROTEOMICS EXPERIMENTS	66
ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING IN RNA BIOINFORMATICS	70
EXPLAINABLE TRANSCRIPTOME (?) ANALYSES – THE ROUTE OF BULK, SINGLE-CELL AND SPATIAL TRANSCRIPTOMICS ANALYSES TAKEN BY EXPLAINABLE AI ALGORITHMS	74

BEHIND THE SCENES – AI AS AN ENABLER OF SCIENTIFIC DISCOVERY IN THE LIFE SCIENCES

80

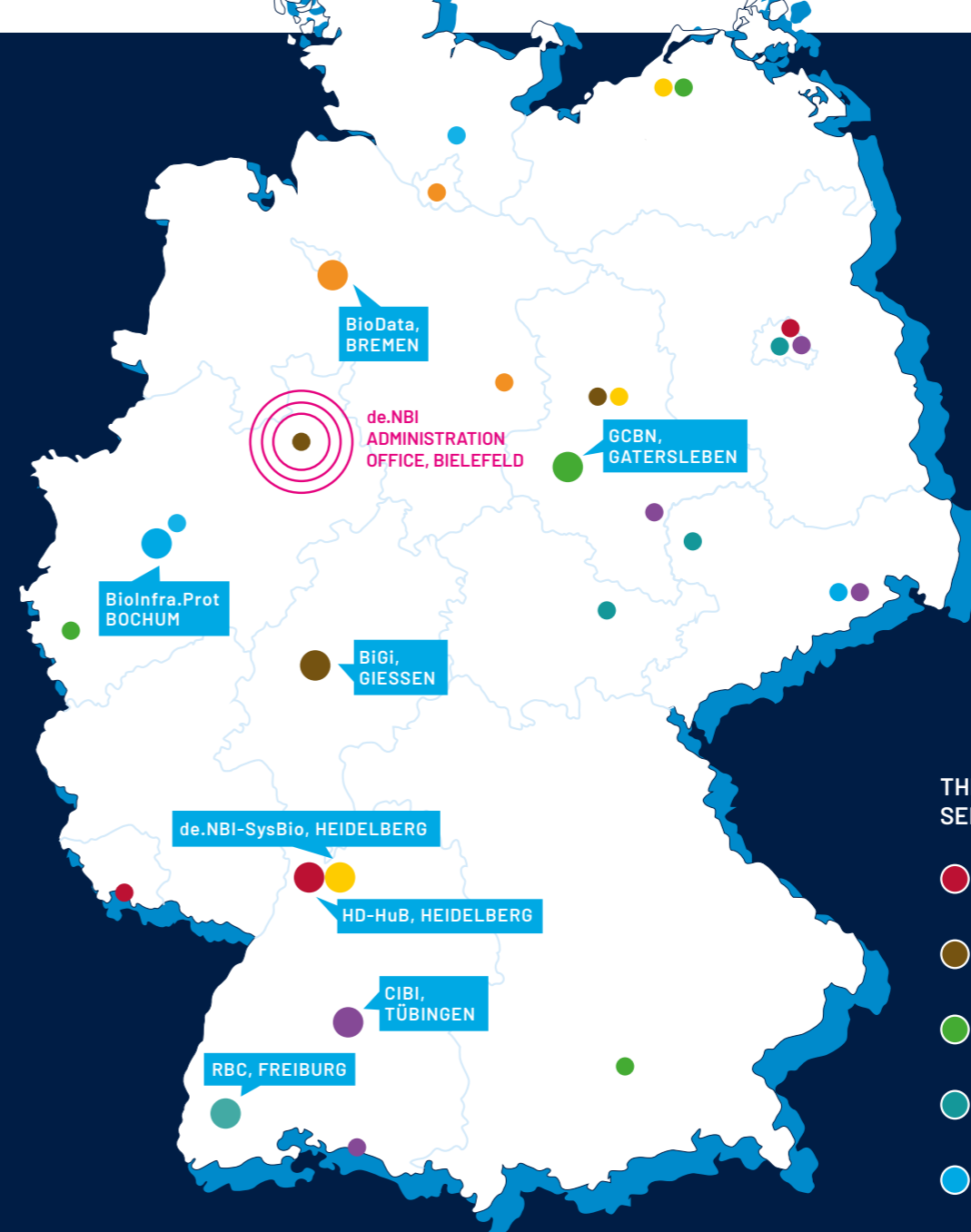
FROM GENOMES TO PHENOTYPES – HOW AI HELPS MOBILIZING AND ANALYZING BIG DATA AND PREDICTING PROPERTIES FOR THE MANY UNCULTURED BACTERIA	82
TOOL RECOMMENDER SYSTEM IN GALAXY USING DEEP LEARNING	88
TOWARDS SMART WAYS TO HELP DATA CURATION – NATURAL LANGUAGE PROCESSING FOR THE LIFE SCIENCES	94
IMPROVING THE SEARCH FOR NEEDLES IN A HAYSTACK – CLASSIFIER- INDEPENDENT OVERSAMPLING FOR IMBALANCED DATA	100

IMPRINT	110
---------	-----



THE GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE – de.NBI

de.NBI is a distributed bioinformatics infrastructure which started in March 2015 as an academic and non-profit initiative of the German Ministry of Research and Education (BMBF). The de.NBI network is aimed to deliver high standards of bioinformatics services, comprehensive training, powerful computing capacities (de.NBI Cloud) as well as connections to industrial companies that assist researchers to more effectively exploit their own data and contribute to the advancement of Life Science research in Germany and Europe.

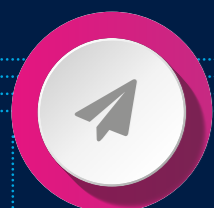


THEMATIC FOCUSES & SERVICE CENTERS:

- **HUMAN BIOINFORMATICS**
HEIDELBERG CENTER FOR HUMAN BIOINFORMATICS (HD-HuB)
- **MICROBIAL BIOINFORMATICS**
BIELEFELD-GIESSEN RESOURCE CENTER FOR MICROBIAL BIOINFORMATICS (BiGi)
- **PLANT BIOINFORMATICS**
GERMAN CROP BIOGREENFORMATICS NETWORK (GCBN)
- **RNA BIOINFORMATICS**
RNA BIOINFORMATICS CENTER (RBC)
- **PROTEOME BIOINFORMATICS**
BIOINFORMATICS FOR PROTEOMICS (BioInfra.Prot)
- **INTEGRATIVE BIOINFORMATICS**
CENTER FOR INTEGRATIVE BIOINFORMATICS (CIBI)
- **BIODATABASES**
CENTER FOR BIOLOGICAL DATA (BioData)
- **DATA MANAGEMENT/SYSTEMS BIOLOGY**
de.NBI SYSTEMS BIOLOGY SERVICE CENTER (de.NBI-SysBio)

● LOCATIONS OF SERVICE CENTERS

● LOCATIONS OF PARTNERS



SERVICE

- TOOLS
- WORKFLOWS
- DATABASES
- CONSULTING



TRAINING

- TRAINING COURSES
- SUMMER SCHOOLS
- HACKATHONS
- WEBINARS



de.NBI CLOUD

- INFRASTRUCTURE
- PLATFORM AND SOFTWARE AS A SERVICE



INDUSTRIAL FORUM

- INDUSTRY SERVICES
- CONSULTING
- NETWORKING



USER MEETINGS

- CONTINUOUS DEVELOPMENT OF TOOLS
- CONSULTING
- EXCHANGE OF OPINIONS AND EXPECTATIONS
- BOTTOM-UP FEEDBACK

CONTACT

www.denbi.de

[@denbiOffice](https://twitter.com/denbiOffice)

[linkedin.com/company/de-nbi](https://www.linkedin.com/company/de-nbi)



THE GERMAN NODE WITHIN ELIXIR EUROPE

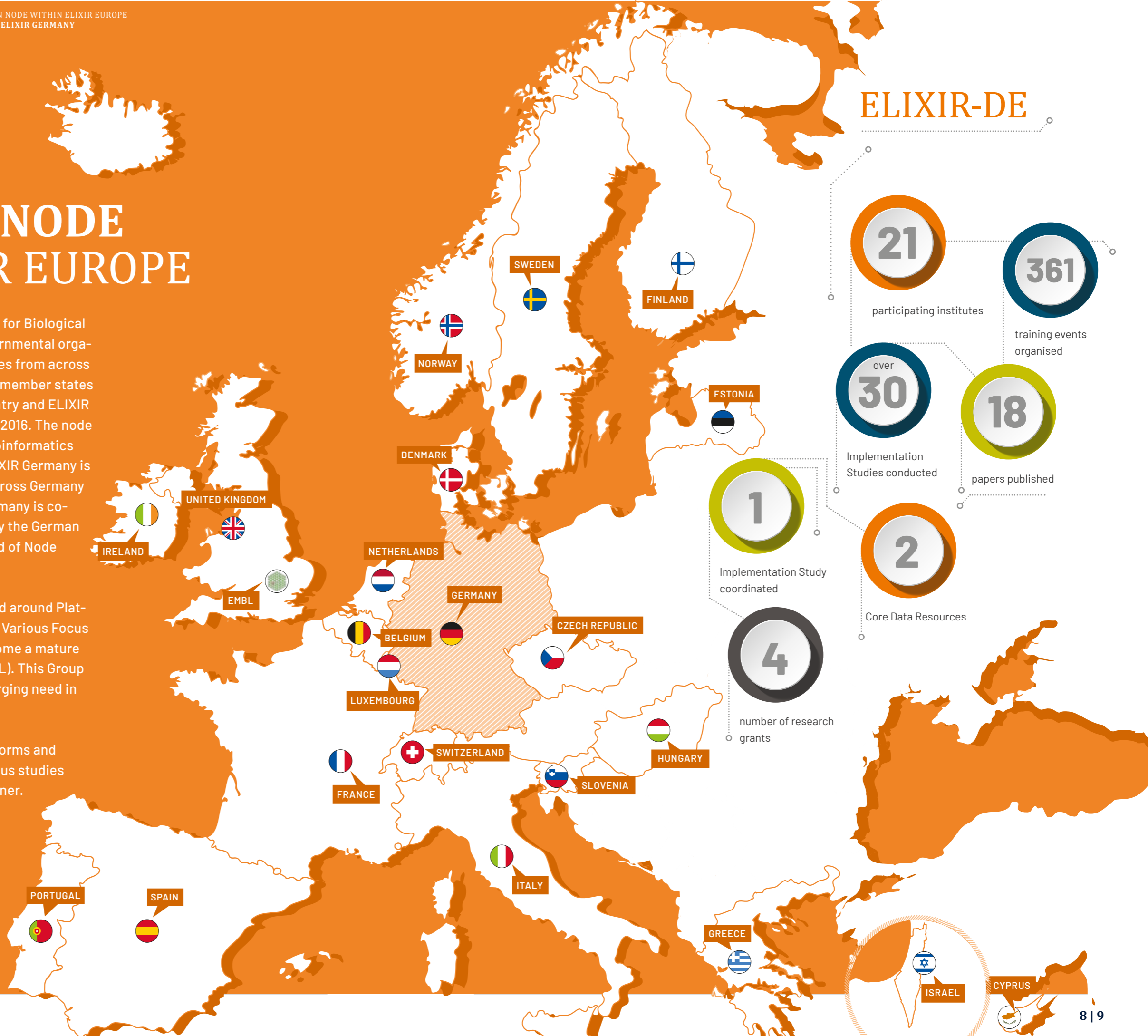
ELIXIR, the European Life Science Infrastructure for Biological Information, was founded in 2014 as an intergovernmental organization and brings together life science resources from across Europe. The consortium currently consists of 22 member states plus EMBL, and Cyprus which is an observer country and ELIXIR Germany being the German Node of ELIXIR since 2016. The node is run by members of the German Network for Bioinformatics Infrastructure (de.NBI). The infrastructure of ELIXIR Germany is represented by eight service units distributed across Germany and the associated EMBL Heidelberg. ELIXIR Germany is coordinated from Bielefeld University and funded by the German government. The National Node is led by the Head of Node Prof. Dr. Andreas Tauch.

The organization of ELIXIR activities is structured around Platforms, Communities and different Focus Groups. Various Focus Groups are currently under consideration to become a mature community – one of which is Machine learning (ML). This Group was initiated in October 2019 to capture the emerging need in Machine learning expertise across the network.

ELIXIR Germany is represented in almost all platforms and communities and actively participates in numerous studies to further develop them in a future-oriented manner.

CONTACT

Andreas Tauch
Head of Node
tauch@cebitec.uni-bielefeld.de
www.denbi.de/elixir-de



ELIXIR-DE





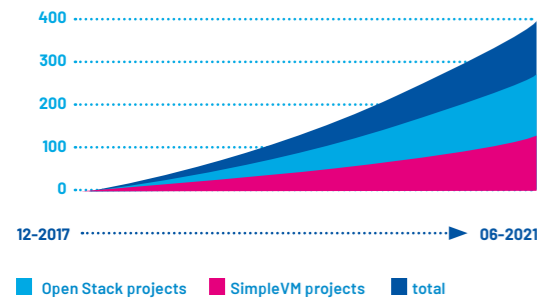
de.NBI CLOUD

Cloud Computing for Life Sciences

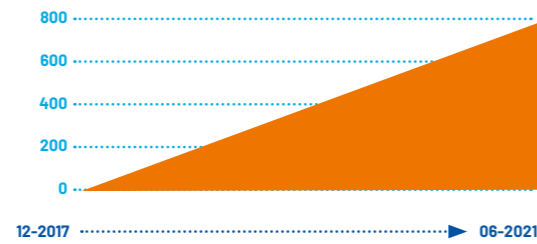
In today's life sciences the handling, analysis and storage of enormous amounts of data is a challenging issue. An appropriate IT infrastructure is crucial to perform analyses with such large datasets and to ensure secure data access and storage. The de.NBI Cloud is an excellent solution to enable integrative analyses and the efficient use of data in research and application. Researchers from the life sciences in Germany can use the de.NBI Cloud free of charge. User meetings are regularly organized to ensure that the requirements of the community are taken into account for the future development of the de.NBI Cloud.

Largest scientific cloud in Germany and one of the leading European academic clouds in life sciences.

de.NBI Cloud Projects



de.NBI Cloud ELIXIR AAI Users



- Full OpenStack Environment per Project
- For fully customizable provisioning and development of VMs and Services/Clusters



- Custom project-type based on OpenStack
- For simple development of VMs and Services/Clusters and integration of e.g. Bioconda

CONTACT

Alexander Sczyrba

de.NBI Cloud Coordinator

asczyrba@cebitec.uni-bielefeld.de

<https://cloud.denbi.de/get-started/>

Peter Belmann

de.NBI Cloud Governance

cloud@denbi.de

<https://cloud.denbi.de/get-started/>

Cloud Access

- principal investigator of German university or research institution applies for cloud resources by proposing a project and describing required resources through the de.NBI Portal
- the project is reviewed by a scientific committee
- after approval of the application, the project is created in the de.NBI Cloud Portal
- project resources are allocated at one of the cloud sites

REGISTER

Register for an ELIXIR account and apply for membership in the de.NBI virtual organisation.

LOGIN

Log in at the de.NBI Cloud portal using your existing ELIXIR account.

SELECT PROJECT

Select a project type in 'New Application'.

SUBMIT

Fill in the application form for the selected project type and submit.

REVIEW

Now the application will be reviewed by the Cloud committee.

APPROVAL

You will be notified as soon as your application is approved.

ALLOCATION

The requested resources are now allocated in the de.NBI Cloud and managed within our portal.

ADD MEMBERS

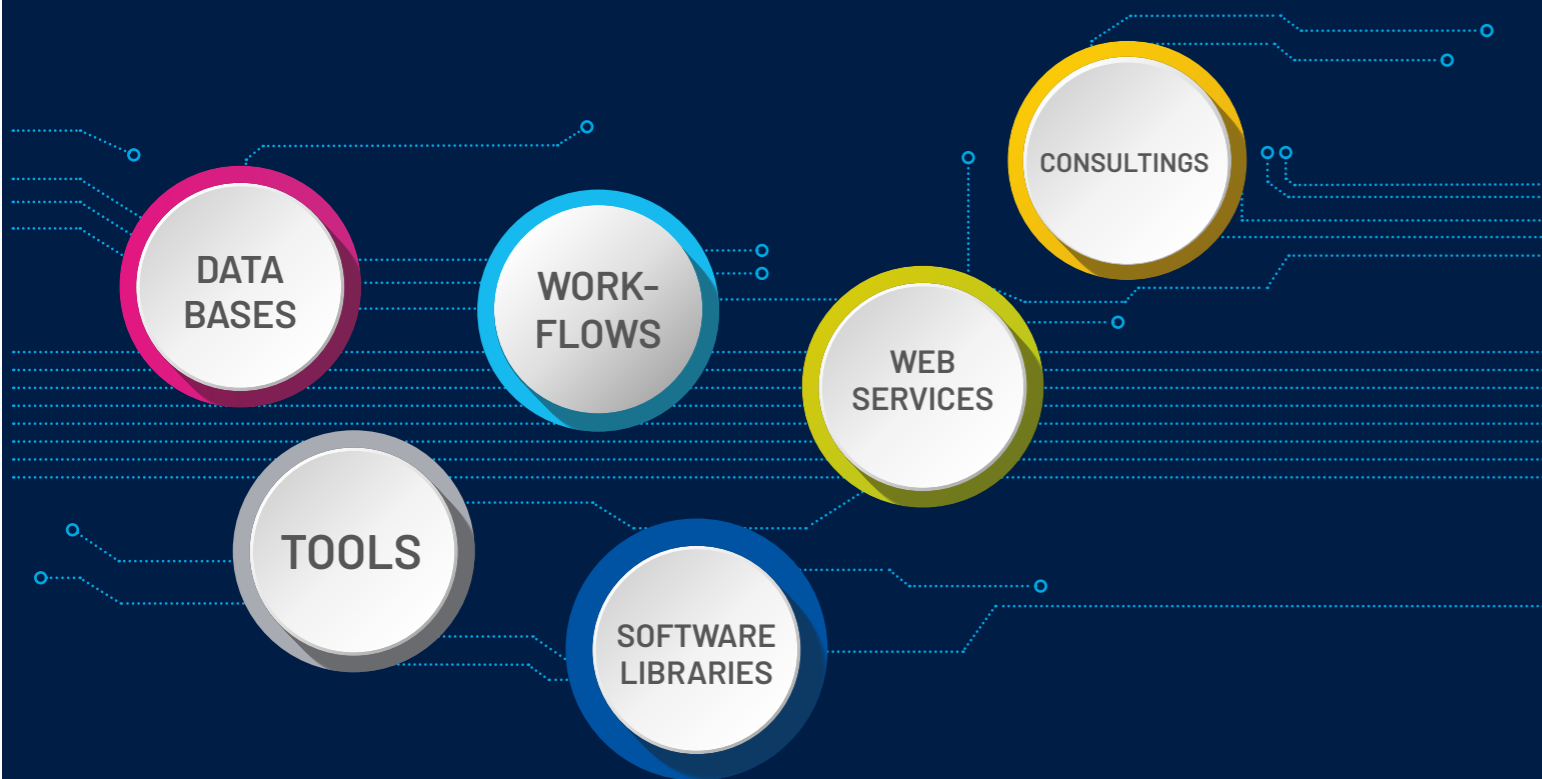
Add members to your project.



de.NBI SERVICES

Tools, Workflows, Databases, Consulting

One of the main tasks of the de.NBI network is the service area. de.NBI offers a diverse portfolio for the analysis of large amounts of data. Services are aimed at application users in life sciences as well as bioinformaticians and developers. The de.NBI services will be unified with regard to standards, interoperability and reproducibility.



CONTACT

Johanna Nelkner
de.NBI Service Coordinator
contact@denbi.de
<https://www.denbi.de/services>



Status: September 2021



de.NBI TRAINING for Life Scientists

The de.NBI network organizes high-quality, coherent, timely, and impactful training events and provides online training materials on a broad range of topics in bioinformatics. Current developments in the field of bioinformatics are also addressed in de.NBI symposia, special workshops and annual summer schools. Life scientists learn how to handle and analyze biological big data more effectively by applying tools, standards and compute services provided by de.NBI.

CONTACT

Daniel Wibberg
de.NBI Training Coordinator
contact@denbi.de
www.denbi.de/training

Fields of activity of the member companies of the Industrial Forum



MEDICINE /
HUMAN DATA



AGRICULTURE /
PLANTS



BIOTECHNOLOGY /
MICROBIOLOGY



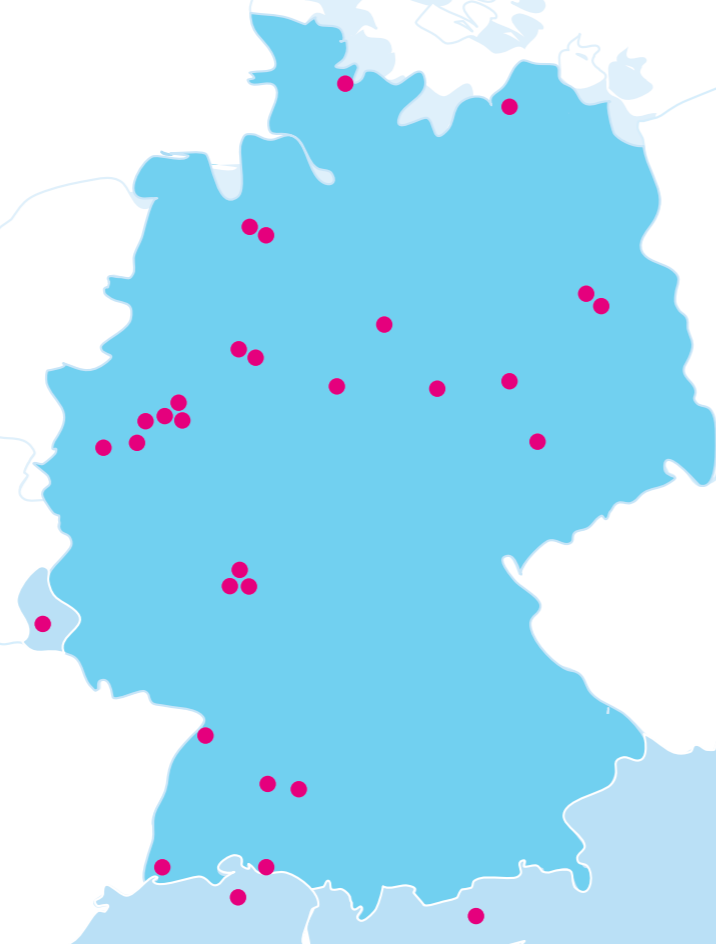
OTHERS



de.NBI INDUSTRIAL FORUM

Industry services, Consulting, Networking

The de.NBI Industrial Forum offers a networking platform for industrial companies that deal with huge amounts of data in the life sciences. Members of the de.NBI Industrial Forum receive access to de.NBI services and training, and are informed about developments in the network.



CONTACT

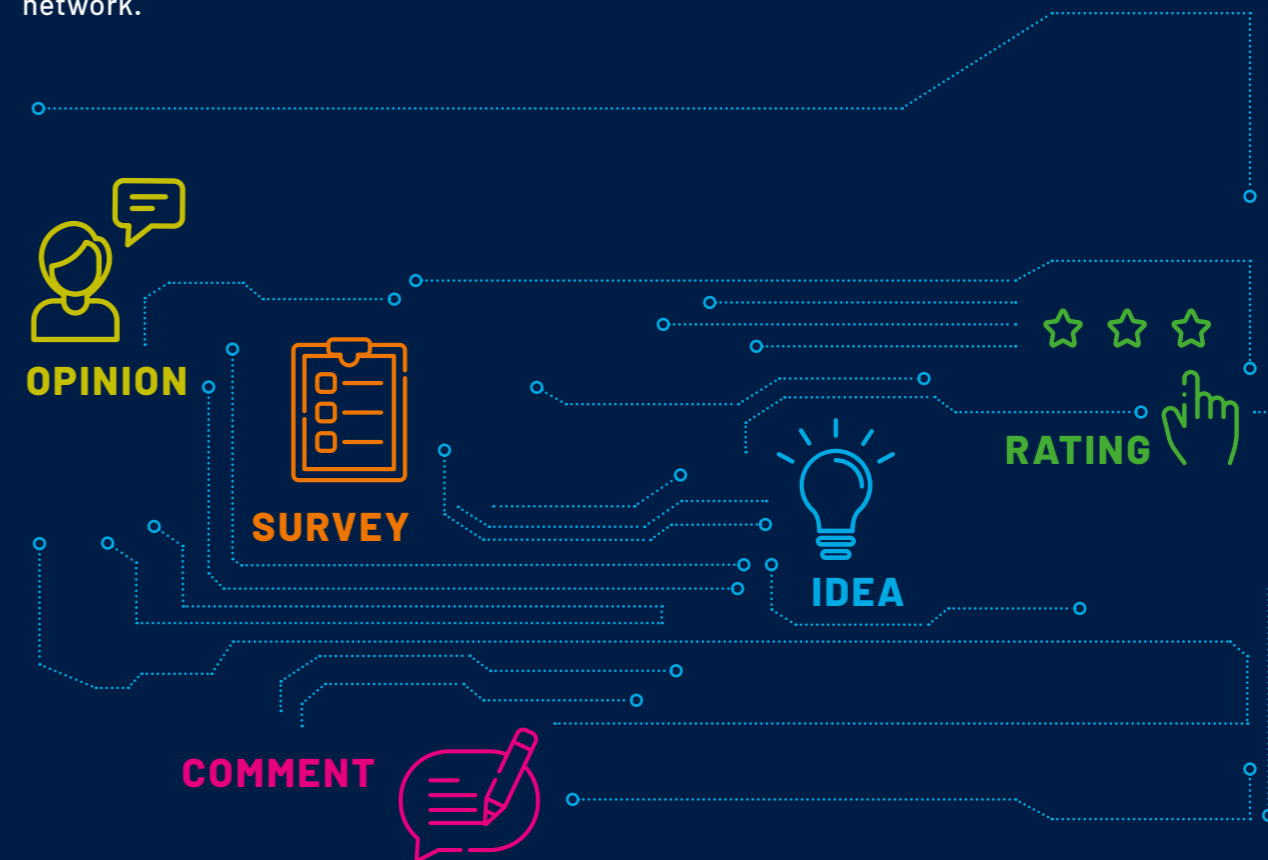
Manuel Wittchen
de.NBI Industrial Forum Manager
contact@denbi.de
<https://www.denbi.de/industrial-forum>



de.NBI USER MEETINGS

The user meetings within the de.NBI network are targeted towards a network-wide framework for user-centered activities. The main aim of those activities is to exchange experiences, opinions and expectations of de.NBI users. The gathered users' feedback should be implemented in the offered services, training, and compute resources of the de.NBI network.

With this, the user meeting efforts support the improvement of the offered bioinformatics service and foster the sustainable development of the bioinformatics infrastructure within the de.NBI network.

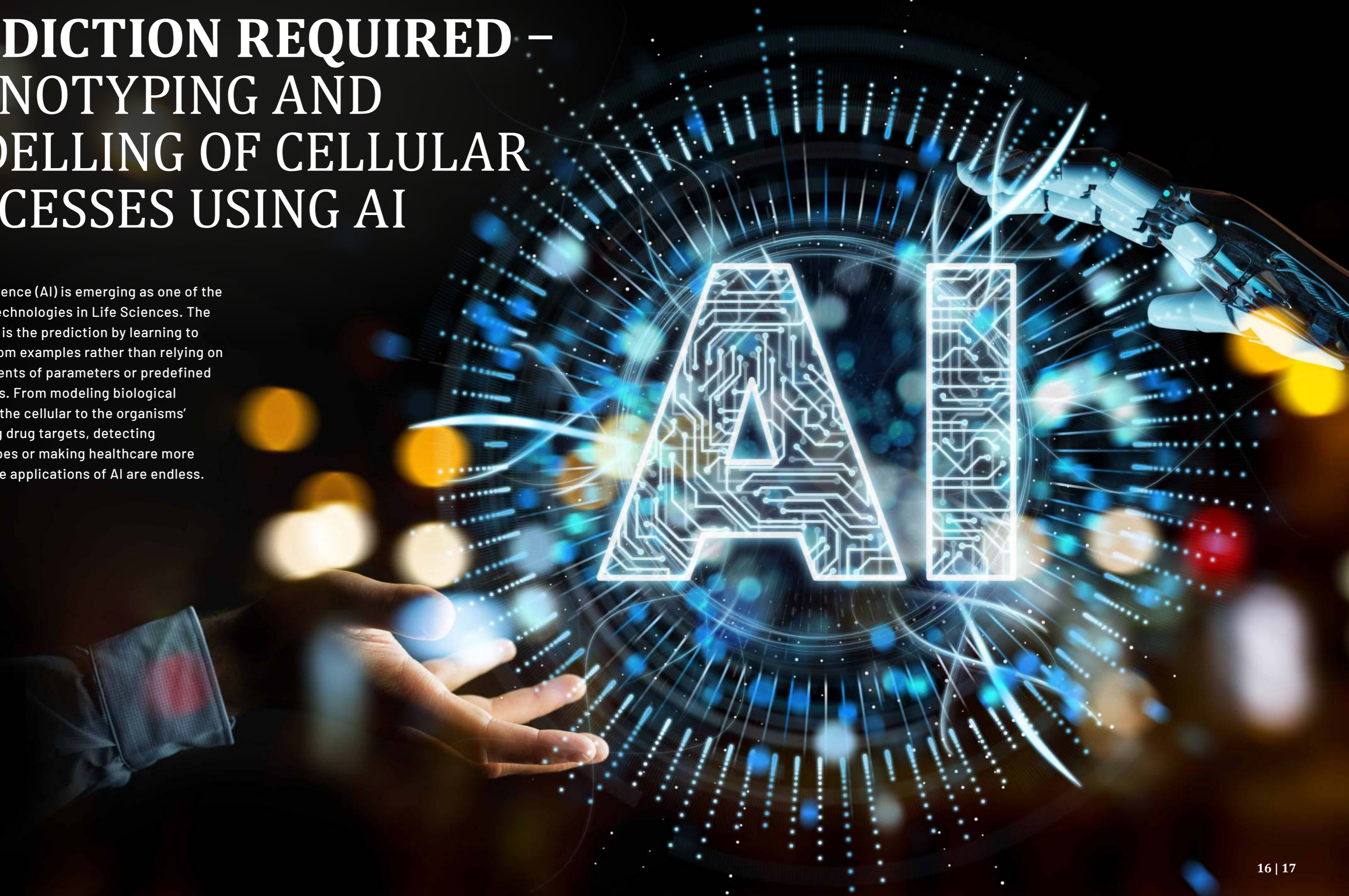


CONTACT

Nils-Christian Lübke
de.NBI Community Coordinator
contact@denbi.de
www.denbi.de

PREDICTION REQUIRED – PHENOTYPING AND MODELLING OF CELLULAR PROCESSES USING AI

Artificial Intelligence (AI) is emerging as one of the key disruptive technologies in Life Sciences. The strong suit of AI is the prediction by learning to process rules from examples rather than relying on manual adjustments of parameters or predefined processing steps. From modeling biological processes from the cellular to the organisms' level, identifying drug targets, detecting hidden phenotypes or making healthcare more personalized, the applications of AI are endless.



DEEP-iAMR

Identification of new antimicrobial resistance targets by high-throughput deep learning

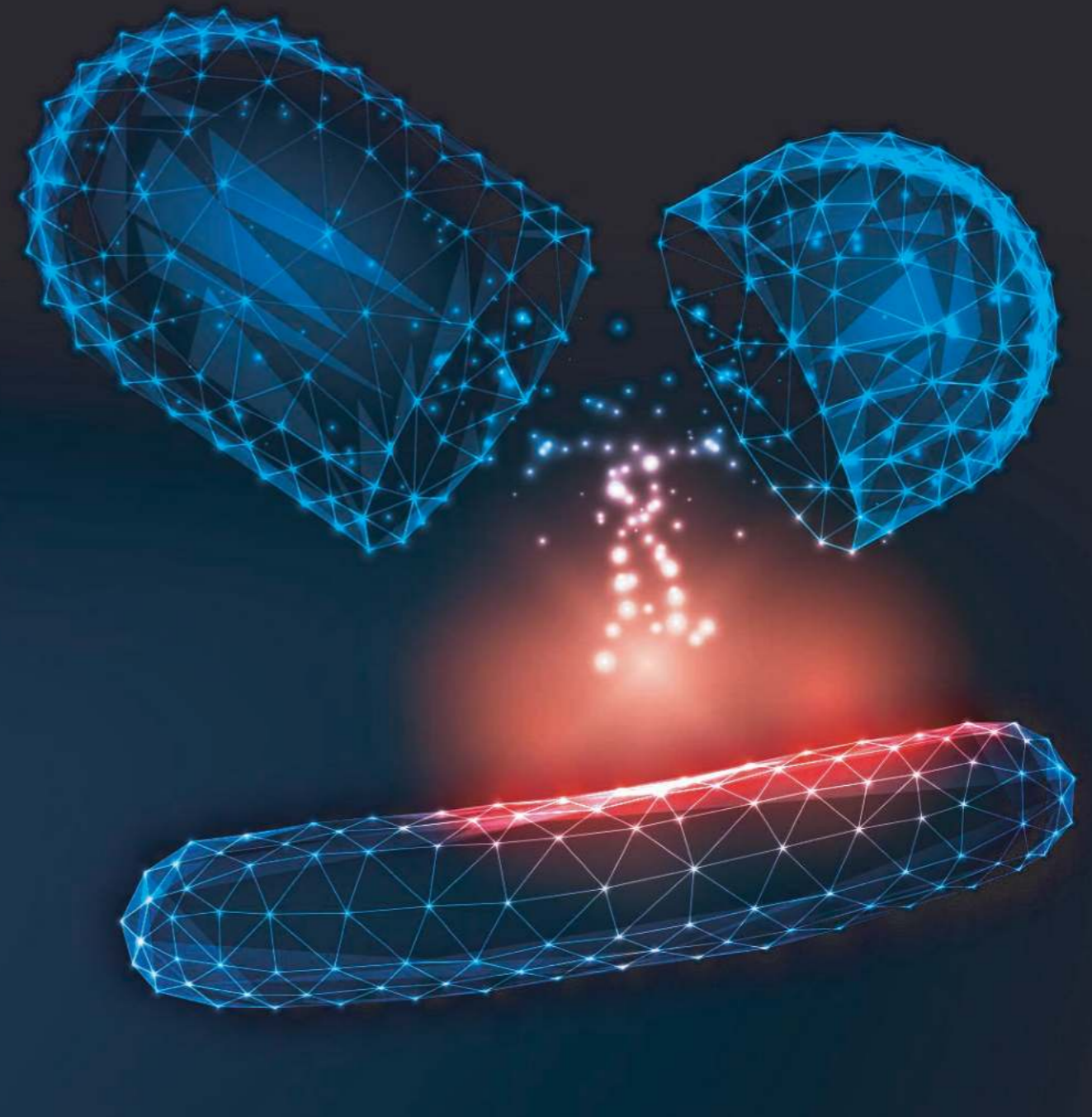
Antibiotic-resistant bacteria are increasingly common in hospitals, farm animals, food and the environment. The continued rise in resistance even against last-resort antibiotics and the lack of new compounds make treatments challenging and inevitably lead to a return to the pre-antibiotic era. The recent increase in bacterial genomic data provides a promising source to explore approaches that tackle this problem and address so far unexplained antibiotic resistance. The BMBF funded project Deep-iAMR aims at developing automatic scalable bioinformatics workflows. New deep learning approaches are applied to predict antimicrobial resistance rapidly and reliably and to uncover hitherto hidden pathways and novel targets for the development of new antibiotics.

ANTIBIOTIC RESISTANT BACTERIA – A GLOBAL THREAT FOR PUBLIC HEALTH WORLDWIDE

Antibiotic resistance is one of the biggest threats to global health, food security and development today. Antimicrobial resistance (AMR) threatens the effective prevention and treatment of a constantly increasing range of infections caused by bacteria, parasites and fungi. The presence of clinically relevant AMR has signifi-

cantly increased worldwide resulting in expensive and difficult-to-treat infections in humans. While some new antibiotics are in development, none of them are expected to be effective against highly antibiotic-resistant bacteria. For example, resistance in *Escherichia coli* to one of the most common drugs used to treat urinary tract infections (fluoroquinolone antibiotics) is very widespread. This treatment is already ineffective in more than half of the patients in many countries. The ad-

vent of affordable and high throughput genome sequencing technologies has opened new avenues to address the problem of AMR. With the resulting availability of large-scale genome data sets, comparative and genome-wide studies have revealed associations with known and novel genetic AMR determinants (genes or single nucleotide variations (SNVs)). Current AMR predictions are generally based on the detection of the presence or absence of previously recognized genetic deter-



minants. Regardless, this 'presence-absence' approach does not adequately account for the plethora of AMR phenotypes that bacteria exhibit. Notably, the extent and the varying degree of resistance, which is commonly indicated by the minimum inhibitory concentration, could not be elucidated. The AMR profile is a cumulative result of contributions from more than one genetic determinant, in which each genetic determinant imparts a different weightage [1-3]. A strategy that

considers genes or SNVs irrespective of previous knowledge individually or in combination with varying weightage of each genetic determinant is required to predict the qualitative and quantitative profile of AMR. This is a computationally time-consuming and expensive strategy that takes into account a large number of parameters (genetic differences between bacterial strains) and is essentially much larger than the total number of samples. For the bacterium *E. coli*, in silico detec-

tion of single antibiotic resistance and its level of resistance could result from one or more or combinations of any of five million bases of the genome. In summary, analyzing and interpreting such a large scale of information has been a limiting factor. With advances in scalable, federated and nearby ('cloud') infrastructures, effective machine learning approaches and more complex analyses can now be used to accommodate the huge increases in data.

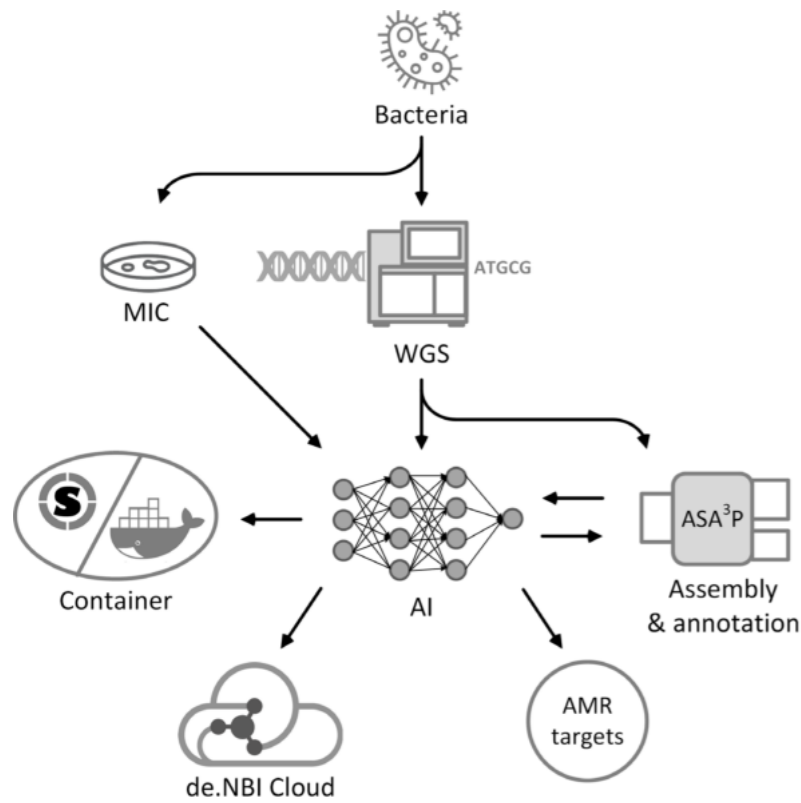


FIGURE 1: The flow of data within the Deep-iAMR project. The project aims to combine various omics data sets with clinical and phenotypic information for a large well-characterized set of multi-drug resistant *E.coli* isolates. This data is then used to train deep neural networks. Finally, our software will be deployed via user-friendly containerization techniques and scalable cloud solutions.

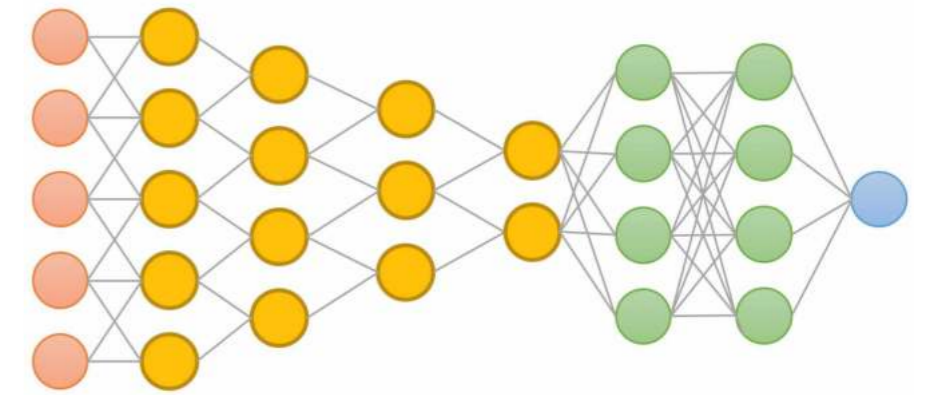


FIGURE 3: Schematic structure of a CNN. Red: input neurons; yellow: convolutional or pool neurons; green: feed forward neurons; blue: output neuron.

In the joint Deep-iAMR project, we apply deep learning approaches for modelling AMR that satisfy the demand for combining a large number of parameters. These models will enable rapid and accurate diagnostics, enhance surveillance and allow an exploration of the role of genetic bacterial determinants in treatment failures beyond the classical well-studied resistance genes. The high dimensionality of the data required for genotype-phenotype predictions tends to hinder generalizations and challenges the scalability of most learning algorithms. Therefore, this project aims at combining various omics data sets together with clinical and phenotypic information available for a large well-characterized set of multi-drug resistant *E. coli* isolates. As illustrated in Figure 1, this data will be used to train deep neural networks (DNNs). Clinical samples for this project are collected at the Institute of Medical Microbiology at the Justus Liebig University (JLU) hospital Giessen headed by Prof. Dr. Trinad Chakraborty. By this means, a contemporary set of multi-drug resistant bacterial pathogens are collected, sequenced and a phenotypic AMR profile is determined. During this project, this set of input data will be extended by higher-level information from detailed feature annotations that will be generated by automated bioinformatics analysis pipelines provided by Prof. Dr. Alexander Goesmann and his

team from the Systems Biology group at JLU Giessen. Ultimately, it is our goal to use the DNNs developed by Prof. Dr. Dominik Heider and his group from the Department of Mathematics and Computer Science at the University of Marburg for sophisticated prediction and classification of AMR mechanisms and patterns in newly sequenced genomes. In addition, we will extract relevant elements from the DNNs and validate whether they indicate potentially new targets for AMR.

ARTIFICIAL INTELLIGENCE FOR DRUG RESISTANCE PREDICTION

Artificial intelligence (AI), and particularly deep learning (DL), is well-suited for the development of predictive models in many different areas, especially for image data, e.g., magnetic resonance imaging or computer-assisted tomography scans for medical diagnostics. In the current project, we focus on genomic data, either from microbial communities or single bacterial genomes, which is typically not provided as image data. Different approaches exist to incorporate and encode genomic data into images for further analyses. One very promising approach is the Chaos Game Representation (CGR). CGR can be used to visualize sequential data as a fractal. As DNA molecules can be represented as sequences of characters (namely A, C, G and T), DNA

can also be encoded using CGR leading to a fractal of squares (Figure 2). It has already been demonstrated that models based on CGR-encoded data are highly accurate and very fast [4,5], too, for instance in enabling phylogenetic analyses of bacterial genomes.

In our project, we will encode the bacterial genomic data with CGR and use the resulting fractals as input for Convolutional Neural Networks (CNNs). CNNs are a class of deep neural networks, namely regularized versions of multilayer perceptrons. A CNN typically uses a tensor as input and consists of convolutional layers, which learn filters and thus features of the data as well as pooling layers, which perform a non-linear down-sampling of the input space, e.g., by using the max function (Figure 3).

We will analyze different network topologies, i.e. different numbers of layers, different numbers of neurons per layer, etc. to find the best working model for the prediction of AMR. Moreover, we will be able to identify novel resistance mechanisms by differential CGRs and gain insights into the underlying biology of mutations in multi-resistant pathogens, which will lead to better treatment of the patients.

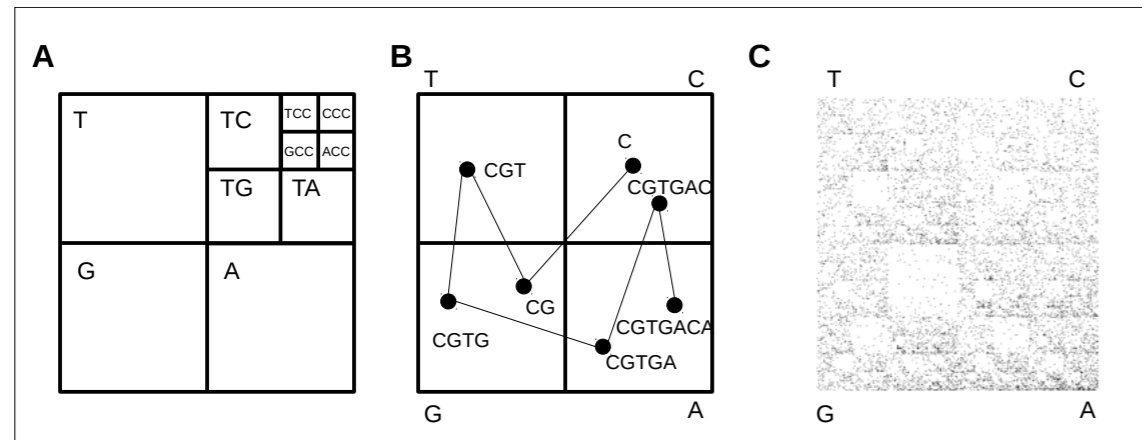


FIGURE 2: Chaos Game Representation for DNA. A) Sub-quadrants of the CGR. B) Way walked to draw points. C) CGR of the HIV genome (NCBI Reference Sequence: NC_001802.1). Here we reuse Figure 1 from 'Deep learning on chaos game representation for proteins' by Löchel et al., 2020, reproduced by permission of Oxford University Press.

SCALABLE DATA PROCESSING AND USER-FRIENDLY BIOINFORMATICS APPLICATIONS IN THE de.NBI CLOUD

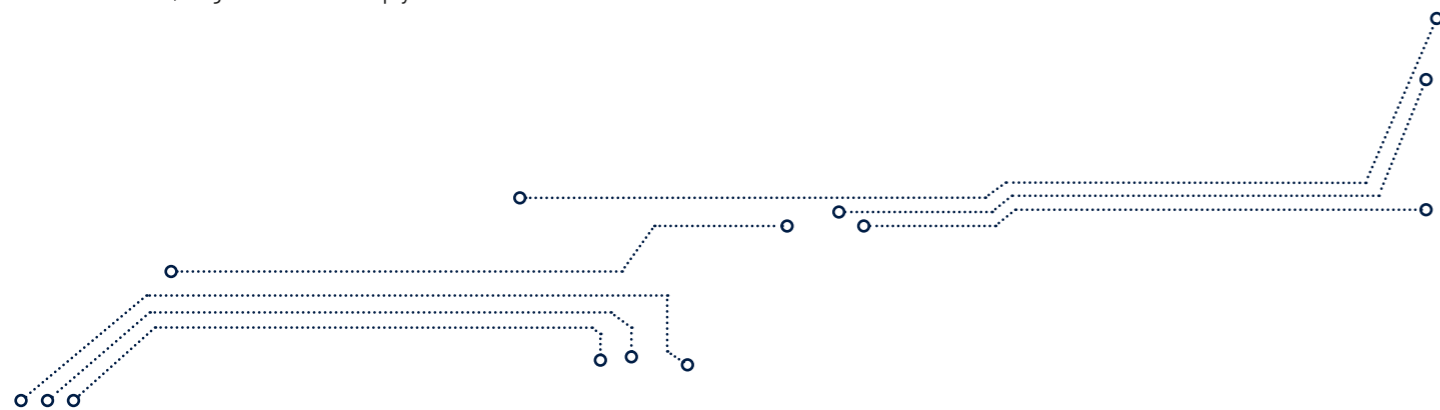
Due to the tremendous progress in the field of DNA sequencing, characterizations of pathogenic bacteria have changed considerably over the last decades. Today, bacterial genomes can be deciphered in many laboratories worldwide within a few hours. As a result, large amounts of raw sequencing data need to be analyzed in scalable and automatic manners to fully exploit this genetic treasure trove and extract all information encoded therein. Moreover, the analysis of such sequence data via modern DL approaches requires vast and standardized training data sets of the highest quality. Therefore, large numbers of collected and sequenced pathogenic bacterial genomes are automatically processed in the de.NBI Cloud to apply strict quality controls and finally transform raw sequencing data into higher-level genome characterizations usable as input features for different DL approaches. Hence, as a first data processing step, raw sequencing data are filtered and revised to meet strict quality requirements and to streamline the subsequent assembly process resulting in bacterial genomes. In a second step, these genomes are then annotated to assign genomic features, e.g. genes and regulatory elements. Furthermore, all genomes are deeply charac-

terized by various *insilico* analyses as for example the detection of AMR genes and virulence factors. High standards of curation enable comparison to high-quality reference genomes to detect individual mutations, e.g. SNVs. To distribute the computational workload of the analysis of these large datasets and to exploit the vast capacities of modern cloud computing infrastructures as, for instance the de.NBI Cloud, all data processing and analysis workflows are implemented using Nextflow - a state-of-the-art workflow management system [6]. Finally, this information is automatically collected and used as standardized input features for DL models.

As soon as successful DL models have been sufficiently trained and validated they will be used to implement reusable bioinformatics software tools for improved AMR predictions. In order to conduct reproducible analysis workflows and to provide these tools to the scientific community, resulting software tools will therefore be packaged and distributed via modern containerization techniques like, for instance, Docker and Podman. By doing so, researchers are enabled to scale out their analysis within high-performance or cloud computing infrastructures in order to meet the growing computational requirements of ever increasing amounts of data.

CONCLUSION & OUTLOOK

The huge increase of genome-based data from bacterial genome sequencing studies represents a scientific treasure trove for developing robust, rapid, and validated approaches to predict antimicrobial resistance. However, much information remains hidden in the data due to its sheer amount and the implied requirements for suitable data analysis strategies and IT infrastructures. Modern DL approaches are a promising tool to address these issues. Within the project Deep-iAMR we will use deep neural networks to exploit the hitherto unknown genetic information hidden in the genomic data to improve the prediction of AMR resistances *insilico*. In combination with our genotype-phenotype studies it will potentially help to identify new targets for the development of new antibiotic drugs and provide new insights for the rational assessment of treatments against resistant bacterial pathogens.



REFERENCES: [1] Geisinger E et al., 2018, PLoS Pathogens 14(5) e1007030. DOI: 10.1371/journal.ppat.1007030. [2] Hwang S et al., 2016, Scientific Reports 19;6:26223. DOI: 10.1038/srep26223. [3] Kröger C et al., 2018, Nucleic Acids Research 2;46(18) 9684-98. DOI: 10.1093/nar/gky603. [4] Hoang T et al., 2016, Genomics 108(3-4) 134-42. DOI: 10.1016/j.ygeno.2016.08.002. [5] Löchel HF et al., 2020, Bioinformatics 36(1) 272-9. DOI: 10.1093/bioinformatics/btz493. [6] Di Tommaso P et al., 2017, Nature Biotechnology 35(4) 316-9. DOI: 10.1038/nbt.3820.

AUTHORS: Oliver Schwengers^{1,2}, Karina Brinkrolf¹, Swapnil Doijad³, Trinad Chakraborty^{2,3}, Dominik Heider⁴ and Alexander Goesmann^{1,2}

¹ Bioinformatics & Systems Biology, Justus Liebig University Giessen, Heinrich-Buff-Ring 58, 35392 Giessen

² German Center for Infection Research (DZIF), Partner Site Giessen-Marburg-Langen, Schubertstraße 81, 35392 Giessen

³ Institute of Medical Microbiology, Justus Liebig University Giessen, Schubertstraße 81, 35392 Giessen

⁴ Department of Mathematics and Computer Science, Philipps-University of Marburg, Hans-Meerwein-Str. 6, 35043 Marburg

DEEP LEARNING FOR ANALYZING MICROSCOPY IMAGES

Computer-Based Image Analysis and Cellular Phenotyping

Automated analysis of microscopy image data is important to elucidate cellular processes. However, analyzing such data poses a number of challenges. Recently, deep learning methods within the field of artificial intelligence emerged which yield superior results compared to classical methods. Deep learning methods use deep neural networks and are typically trained from example data. We describe deep learning methods for computer-based image analysis of cell microscopy data.

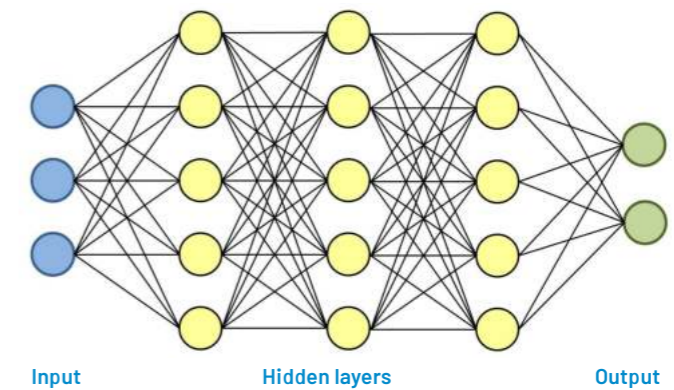


FIGURE 1: General architecture of a deep neural network.

Analyzing high-throughput and high-content microscopy image data is important for elucidating cellular processes to better understand diseases and to find suitable medical treatments. Typically, enormous amounts of digital image data are generated in biological experiments. However, accurate and efficient computer-based analysis of microscopy image data poses a number of challenges. Recently, deep learning methods within the field of artificial intelligence emerged which have a high potential to improve automated image analysis.

DEEP NEURAL NETWORKS

Deep learning is a subfield of machine learning within artificial intelligence. The basis of deep learning methods are deep neural networks, which consist of multiple layers: An input layer, multiple hidden layers, and an output layer (Figure 1). These artificial neural networks model the function of the human brain using multiple connected artificial neurons (also denoted as perceptrons). An artificial neuron has multiple inputs, and computes the weighted sum over the input

followed by a non-linear activation function (Figure 2). The weights are learned during training from sample images using backpropagation, which is a gradient-based optimization method.

Multiple artificial neurons build one network layer. The fundamental layer type is a fully connected layer, where every neuron in one layer is connected to all neurons in the previous and the next layer. The layers are stacked to generate a multi-layer neural network (also denoted as multi-layer perceptron). Networks with multiple (hidden) layers are called deep neural networks in comparison to shallow networks.

Besides fully connected layers, there exist other types of layers. Important are convolutional layers, which perform a convolution on the input data using multiple kernels to generate different feature maps. A network which contains convolutional layers is called Convolutional Neural Network (CNN). CNNs are powerful in processing multi-dimensional data such as images since they learn a hierarchical representation of features. In addition to

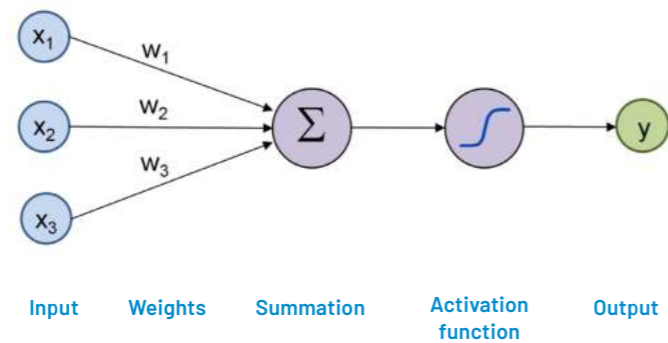


FIGURE 2: Artificial neuron.

feedforward neural networks which contain only forward connections, there also exist Recurrent Neural Networks (RNNs) which are suited to process sequential data. RNNs contain blocks with loop connections to represent information from previous sequential steps.

Training of deep neural networks can be performed supervised, semi-supervised, or unsupervised. For supervised learning, annotated (labeled) data is required. Since labeling of data is generally difficult and time consuming, semi-supervised and unsupervised learning methods have been introduced. A general problem of network training is overfitting, which occurs when the complexity of the model is high compared to the amount and variety of labeled data. Then the model can fit the training data very well, but is not able to generalize well to unseen data. To improve the generalization capability of a neural network, regularization can be used. An often used technique is dropout, where single neurons are temporarily removed during training.

A main advantage of deep learning methods is that the features are learned automatically, whereas classical machine learning methods employ hand-crafted features. This is important for difficult tasks such as automated analysis of cell microscopy images.

DEEP LEARNING FOR COMPUTER-BASED IMAGE ANALYSIS

Deep learning methods can be used for different kinds of data. In particular, such methods have been applied to analyze natural video images and medical images. Deep learning methods have been shown to outperform classical methods and they partially exceed the performance of human annotation [1, 2].

A central task of computer-based image analysis is segmentation. The aim is to partition an image into a set of meaningful regions. In the case of microscopy images, it is often important to identify cells and to distinguish them from the background. Cell segmentation is a prerequisite to quantify cell properties such as size, shape, and signal intensity. This is required in many applications and denoted as cellular phenotyping. However, cell segmentation in microscopy images poses a number of challenges such as high image noise, low image contrast, inhomogeneous image intensities, and high variation of cell size and shape.

The Biomedical Computer Vision group at Heidelberg University headed by PD Dr. Karl Rohr is developing deep learning methods for accurate computer-based analysis of cell microscopy images. The aim is to improve automated quantification of cellular phenotypes at the single

cell level as well as to efficiently process large scale microscopy data. In particular, deep learning methods for cell segmentation have been developed. The methods combine different types of neural network architectures such as convolutional neural networks and recurrent neural networks. This enables exploiting information at different image scales as well as performing iterative refinement of the segmentation result [3, 4, 5]. A deep neural network with an hourglass shape and an encoder-decoder structure has been developed. The network comprises densely connected blocks, gated recurrent neural networks, pooling and unpooling blocks, and residual blocks.

Network training is performed end-to-end using example images. For the loss function, an extension of the cross-entropy is used to deal with class imbalance, and stochastic gradient descent is employed for optimization. To reduce the amount of manually annotated training images, data augmentation is employed. With this technique, the available annotated image data is enlarged by applying different image transformations such as rotation, flipping, scaling, and image intensity changes. In addition, transfer learning methods can be used. This means that the network is pre-trained on available annotated images from other domains, and then fine-tuned using images from the considered application.

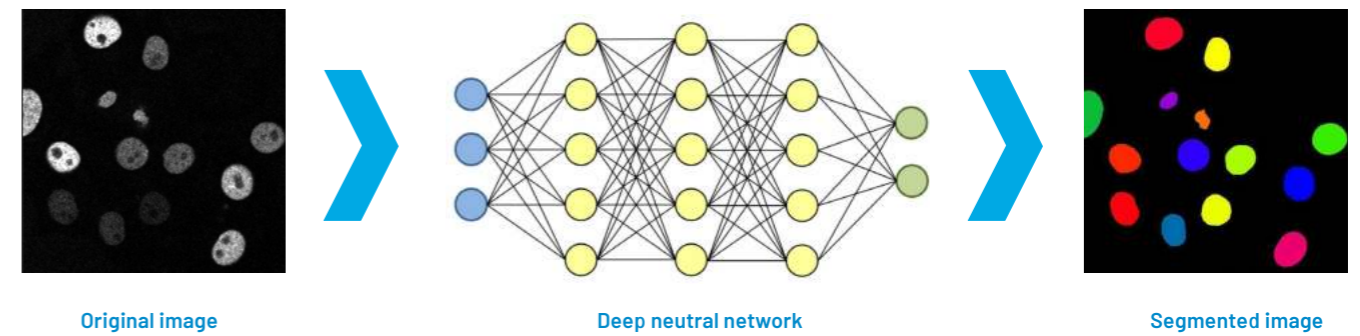


FIGURE 3: Deep neural network for segmentation of cell microscopy images.

The developed deep learning methods for cell segmentation have been applied to analyze high-throughput and high-content microscopy image data. Cell fluorescence microscopy images and tissue images have been used to extract cellular phenotypes. An experimental comparison with classical image analysis methods such as local thresholding, k-means clustering, and random forest classifier showed that the deep learning methods yield superior segmentation results. An example segmentation result for a cell microscopy image is displayed in Figure 3. It can be seen that cells with high and low image contrast can be well segmented. The developed methods have been ap-

plied, for example, to segment cells in tissue images for subsequent quantification of the length of telomeres (end of chromosomes) which can be exploited for medical diagnosis. This work has been carried out within the BMBF project CancerTelSys. The development and application of the deep learning methods has been benefitting from the de.NBI computing infrastructure and cloud.

INFORMATION

For further information, please visit:
<http://www.bioquant.uni-heidelberg.de/bmcbv>
<https://www.hd-hub.de/galaxy-image-analysis/>

CONCLUSION

We describe deep learning methods for computer-based analysis of microscopy images, which is important to elucidate cellular processes. Deep learning methods are based on deep neural networks. These methods improve the image analysis results compared to classical methods and are well suited to cope with the challenges of cell microscopy data. Automated image analysis and extraction of cellular phenotypes from microscopy image data is important for biological research and to identify relevant targets for medical diagnosis and therapy.

REFERENCES: [1] LeCun Y et al. 2015 Nature, 521, 436–444. DOI: 10.1038/nature14539. [2] Litjens G et al. 2017 Med Imag Anal, 42, 60–88. DOI: 10.1016/j.media.2017.07.005. [3] Wollmann T et al. 2019 Med Imag Anal, 56, 68–79. DOI: 10.1016/j.media.2019.04.011. [4] Ritter C et al. 2019 Int J Comput Assist Radiol Surg, 14, 1847–1857. DOI: 10.1007/s11548-019-02010-3. [5] Wollmann T and Rohr K 2021 Med Imag Anal, 70, 102019. DOI: 10.1016/j.media.2021.102019.

AUTHORS: Carola Krug¹ and Karl Rohr¹

¹ Heidelberg University, BioQuant, IPMB, Biomedical Computer Vision Group, Im Neuenheimer Feld 267, 69120 Heidelberg



REmatch: AI FOR DRUG DISCOVERY AND REPURPOSING

Image-based profiling to create a high-resolution reference map of targetable cellular pathways

New image-based methodologies that enable to measure phenotypic effects of perturbations are increasingly being used to identify and characterize drug candidates early in the drug development process. These methodologies promise to generate deep biological profiles about intended and unintended effects of pharmaceutical agents and aid the decisions about their further development. However, such increasingly deep data sets pose new challenges in their analysis and our ability to learn meaningful biological information. Here, we summarize recent approaches in the field and our efforts to create an integrated platform for the generation and analysis of image-based drug screens. This platform can predict a drug candidate mode of action, learn from drug profiles for predicting their targets and off-target toxicities and evaluate opportunities for drug repurposing. We describe how artificial intelligence (AI) plays a key role in the analysis of large image data sets and supports analysis to derive target and biological profiles for pharmacological agents.

CHALLENGES IN DRUG DISCOVERY

Though new technologies and methodological breakthroughs enable drug discovery research at an increasing pace, relatively few new drugs reach the market. In drug discovery this trend is known as 'Eroom's law' [1]. Concordantly, the potential to deploy approved drugs in multiple disease areas often remains unused. New strategies for drug development are therefore required to (1) shorten the development time for new chemical or biological entities, (2) reduce the failure rate of candidates and (3) uncover re-purposing potential.

One strategy broadly discussed has been to gain more comprehensive information on the biological effects of drug candidates early in the drug development process. The information gathered about a drug candidate's biochemical properties and its effects on biologically relevant model systems can form comprehensive profiles to identify nonspecific or ineffective chemicals and guide further drug development. Cell-free biochemical assays to identify pharmacologically active agents are widely used in drug development. However, unlike testing chemicals on biological models, such as cells, biochemical assays lack informa-

tion on biological characteristics of the drug candidate that are important for understanding their impact on a biological system. Furthermore, each functional property of a new chemical is tested using a specific biochemical assay, leading to narrow functional profiles that lack the depths to comprehensively identify a drug candidate's characteristics.

IMAGE-BASED PROFILING DELIVERS INFORMATIVE PROFILES

Image-based profiling of candidate compounds promises to address key challenges of classical biochemical assays

by generating information-rich profiles in single assays. These phenotypic profiles can be used to quantitatively assess changes in shape, texture and staining intensity of cells and are derived from measuring a cell's reaction to the drug based on microscopy images. AI aided computer vision algorithms are used to extract numeric measurements from images of fluorescently stained cells that are later combined into profiles describing the different facets of cellular reactions. Thanks to modern laboratory automation this can be done by testing thousands of drugs in parallel.

These approaches, however, come with three major challenges: First the collected data often comprise millions of microscopy images. A screen of 2000 chemicals in four cell lines can amount to more than 2 TB of raw data. Second the speed of available algorithms limits rapid analysis of screening data sets. Analyzing this 2 TB of raw imagery would for example take up to 8000 CPU hours of pure computation. Furthermore, the lack of suitable models and reference profiles for data interpretation and exploitation complicate the wide-spread use of large-scale image-based profiling for routine drug discovery pipelines.

Extracted profiles often comprise hundreds of data points, each describing a different feature measured on a single cell and having different levels of information content towards the characteristics of the investigated chemical. If integrated with prior information on drug candidates and reference drugs the profiles can be used to establish predictive models for drug re-purposing and models for off-target and mode-of-action prediction [2,3]. Together, these approaches can be employed early in the drug development process and facilitate the cost-efficient early filtering of successful and failing candidate compounds.

A PROOF-OF-CONCEPT KNOWLEDGE BASE OF DIVERSE PROFILES

In an ERC funded Proof-of-Concept project we developed REmatch, a technology platform that integrates an optimized image-based profiling assay [3], a diverse set of relevant biological cell line models [4,5], and an optimized pre-processing pipeline with AI models for improved predictions. The goal of this project has been to create a diverse reference database of phenotypic profiles allowing to chart a comprehensive map of the drug target space. One idea to achieve this goal is to test a large collection of chemicals in a biologically diverse set of cellular models in an image-based screening campaign (Figure 1A). With the help of the de.NBI Cloud to scale our prototype as a cloud infrastructure millions of images could then be processed through a feature extraction pipeline leading to a collection of thousands of profiles in as little as seven days of computational processing time (Figure 1B). In a next step, new algorithms are used that are capable of extracting the most relevant information from the profile collection. This makes it possible to find the most active chemicals and reduce data complexity while conserving a maximum of information (similar to filtering a clear voice in a noisy phone call). Furthermore, novel approaches are used to collect a reference set of profiles for well annotated chemicals as an anchor point for teaching an AI to distinguish different drugs according to useful characteristics (Figure 1C).

A MAP OF HIGHLY ACTIVE REFERENCE COMPOUNDS

To chart a map of signalling pathways affected by the drug within a cell the extracted profiles of all drugs can be reduced to the most informative parts and arranged in a clustered heat map of activities. Here, signalling pathways

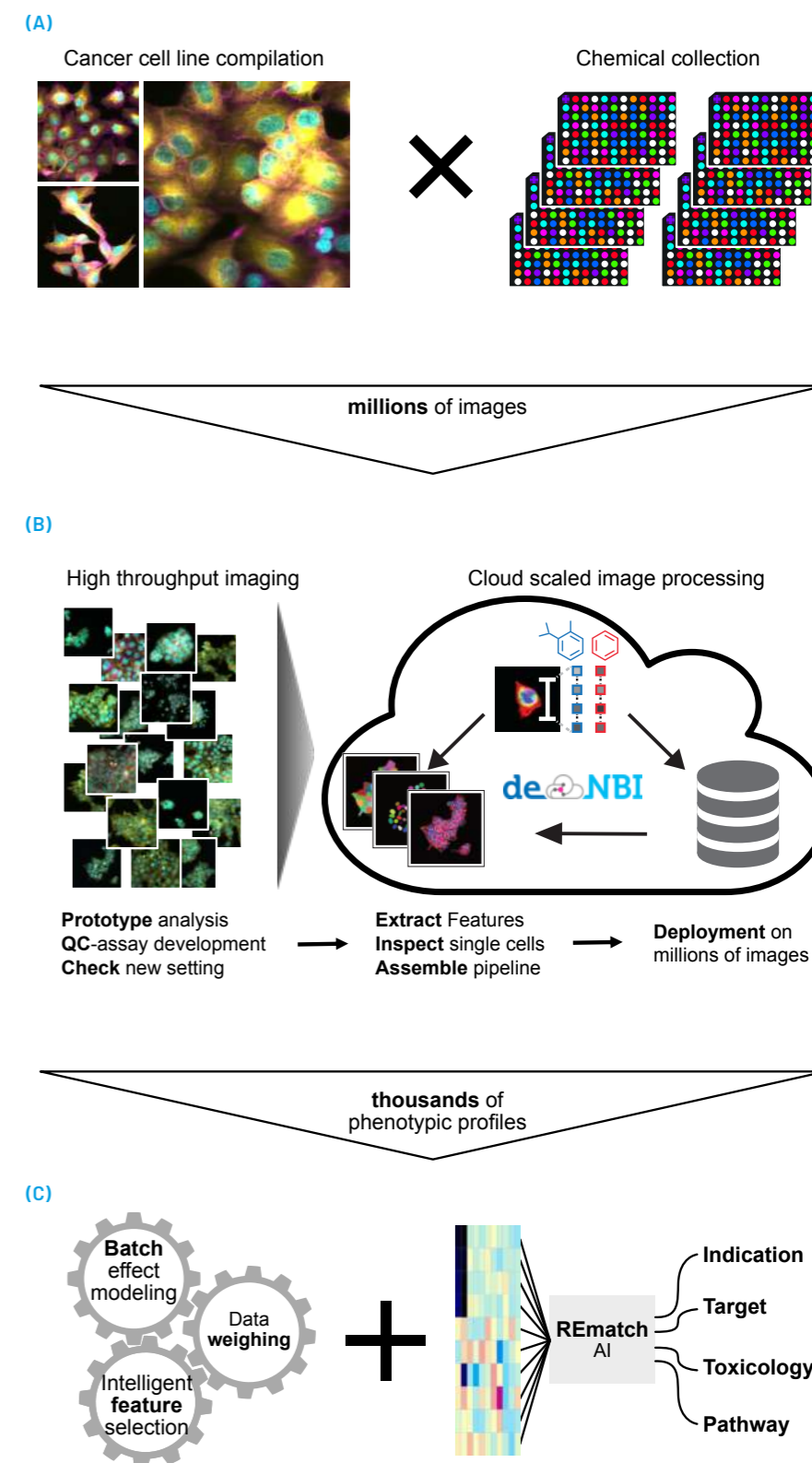


FIGURE 1: Image based drug profiling for drug discovery. (A) By image-based profiling diverse sets of cell lines are screened against large chemical compound collections. (B) Each treatment is then imaged using fluorescent microscopy and images are analyzed using specialized platform technologies such as the cloud-based REmatch. (C) Novel data integration methods maximize information content while reducing data dimensionality prior to machine learning model training.

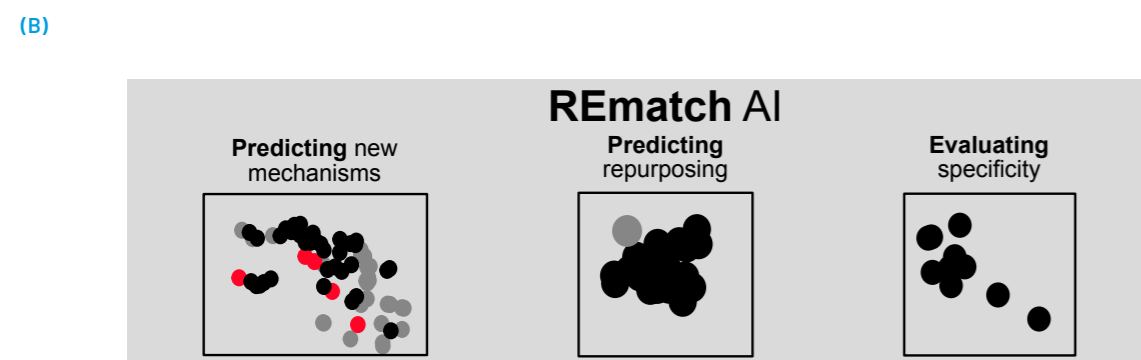
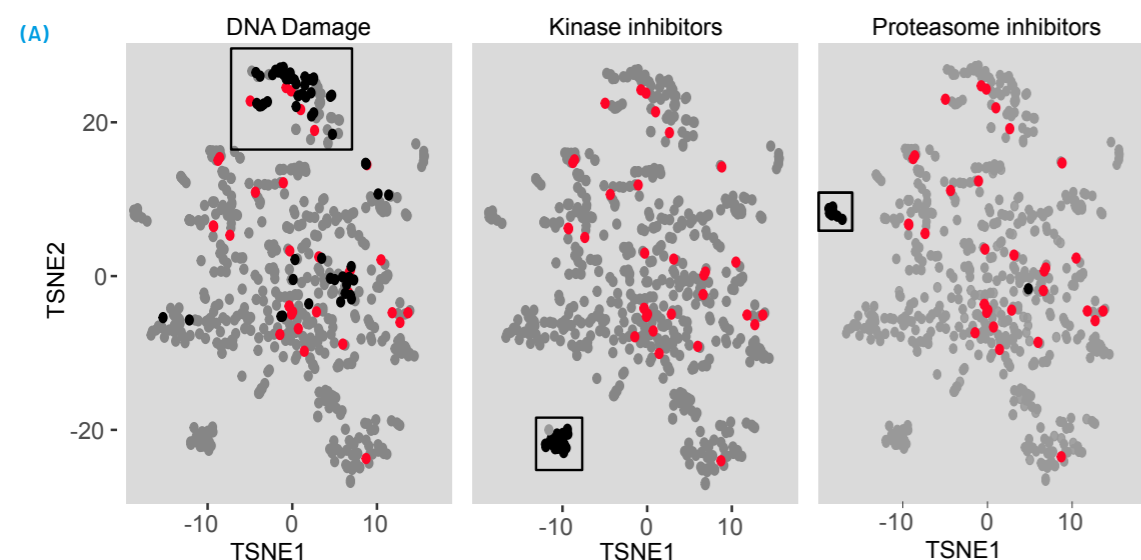


FIGURE 2: Image-based profiling creates a reference of rich profiles for compound characterization. (A) In an optimized visual layout of active compounds, clusters of compounds with shared targets form. Thereby a map of compounds with different target classes is charted and serves as the optimal input for predictive modelling. (B) These models can be used to predict targets of yet unseen compounds, highlight potential re-purposing and evaluate target specificity. Red = candidate drugs, Black = known drugs, grey = other

can be understood like the wirings and switches that run the inner workings of a cell. Clustering the chemicals by the ways they change the profiles of the cells then firstly separates chemicals that show a visible effect on the cells versus chemicals that had no impact on the look and shape of cells. Secondly, clustering also

sorts chemicals into smaller separated groups, spatially laid out on a kind of map (Figure 2A). If two chemicals locate next to each other on this map and by that fall into the same cluster this is indicative of shared properties of the chemicals in that cluster. Thus, the layout can also be understood as a map of shared char-

acteristics among grouped chemicals. In the shown example, compounds targeting, for example, different kinds of molecular switches like specific kinases, DNA damage response genes or the proteasome were among the groups that occupy specific areas of the map.

A POWERFUL MODEL TO CHARACTERIZE NEW CANDIDATE DRUGS

Artificial intelligence trained on the basis of this reference map can help to solve three specific tasks in drug discovery (Figure 2B). One can predict modes-of-action (which molecular switches it mainly affects) for formerly unseen compounds. This guides decisions on where to lead the development of a drug next in a precise and unbiased way. Furthermore, models based on a reference map can be used to sort old and new compounds across a broad spectrum of molecular targets. By showing, for example, how drugs used to treat depression act like chemo therapeutics when applied on cancer cells this highlights drugs with potential uses in more than one specific medical indication. This is widely referred to as re-purposing. Lastly, these analyses reveal opportunities to assess the target specificity of many compounds simultaneously and thus allow fast filtering of compounds for their specificity. One can picture this process analogues to an email spam filter that browses through thousands of emails and filters out those that are unspecifically sent to many unrelated recipients.



CONCLUSION & OUTLOOK

Image-based profiling is one of the most promising avenues to accelerate drug discovery, detect off-target effects early and enable re-purposing of drugs. The ability to collect large amounts of image data in a fast manner, opens new avenues to collect rich information profiles on chemical compounds within biologically relevant contexts. This however comes with significant challenges based on the size, complexity and high dimensionality of the data produced. With the support of an ERC Proof-of-Concept grant we developed REmatch, a fully integrated technology platform for screening perturbations of cellular phenotypes, extracting information from large image data sets and interpretation

of resulting phenotypic profiles for prediction of drug characteristics. We created a reference map of compound target pathways and developed a new pre-processing pipeline. AI models provide solutions for compound characterization, specificity assessment and re-purposing. Applied to large chemical compound collections such AI models will reduce required resources throughout the pre-clinical phases of drug development by predicting biological properties based on 'learned' phenotypic profiles. Future directions include developing platforms for additional applications such as context-dependent profiles of drug effects on genome-edited cellular contexts.



REFERENCES: [1] Ringel MS et al., 2020 Nat. Rev. Drug Discov. 19, 833–834. DOI: 10.1038/d41573-020-00059-3. [2] Simm J et al. 2017 bioRxiv 108399, DOI:10.1101/108399. [3] Caicedo JC et al. 2017 Nat. Methods 14, 849–863. DOI: 10.1038/nmeth.4397 [4] Rauscher B et al. 2018 Mol. Syst. Biol. 14, e7656. DOI: 10.15252/msb.20177656. [5] Breinig M et al. 2015 Mol. Syst. Biol. 11, 846. DOI: 10.15252/msb.20156400.

AUTHORS: Florian Heigwer^{1,2}, Christian Scheeder^{1,2}, Maria Boulougouri^{1,2} and Michael Boutros^{1,2}
¹ Division Signaling and Functional Genomics, Deutsches Krebsforschungszentrum (DKFZ), Heidelberg,
² Department Cell and Molecular Biology, Medical Faculty Mannheim, Heidelberg University, Mannheim

A microscopic view of several large, reddish-orange cancer cells with prominent nuclei and radiating filopodia, set against a blue background. The cells are the central focus of the image, with one in the foreground and others slightly behind it.

DEEP LEARNING-BASED CANCER PATIENT STRATIFICATION

Cancer is a disease of the genome, however, drug development and diagnostics do not make full use of genomic technologies and machine learning methods for decision making processes such as patient stratification and biomarker discovery. With our project we are aiming to change this and bring multi-omics and state-of-the-art deep learning methods to diagnostics and drug development for oncology. We show that this approach is more performant than others and works for multiple cancer types.

Cancer is a major public health and economic issue and its burden is ever increasing. Only in the US, it accounts for \$90 billion in direct medical costs. It costs \$2.7 billion to develop a cancer drug. About 9.5 million people die of cancer every year [1]. This is higher than the population of some of the biggest cities in the world. Our Arcas project aims to improve this situation by employing state-of-the-art data integration and analysis methods on top of decades worth of genomics know-how.

WHY IS CANCER BURDEN SO HIGH?

Developing drugs and getting them to patients is a long and time consuming process. First, a compound that needs to get to patients, has to go through pre-clinical research and different phases of clinical trials. Next comes the approval process, and finally doctors have to prescribe and believe that it could help the patient. Sometimes such drugs need to be listed in the guidelines in order to be prescribed - even if they are approved. It's safe to say that there are many inefficiencies in each step of this process.

The most recurring or impactful reason for these inefficiencies is an overly simplistic and narrowly focused approach in different parts of the process chain. Here is an example: We say cancer is a disease of the genome, however, cancer patients rarely get their genome sequenced, and interpreted. When they do, only a couple of hundred genes are sequenced. Information, such as imaging, and histopathological staging etc. is useful, but provides a limited picture of the disease. As a result, if you do not take the whole genome into account, a lot of information goes missing. This is the case for therapeutic decisions, and also in drug development. It is now clear that response to therapy, especially for targeted drugs, is strongly

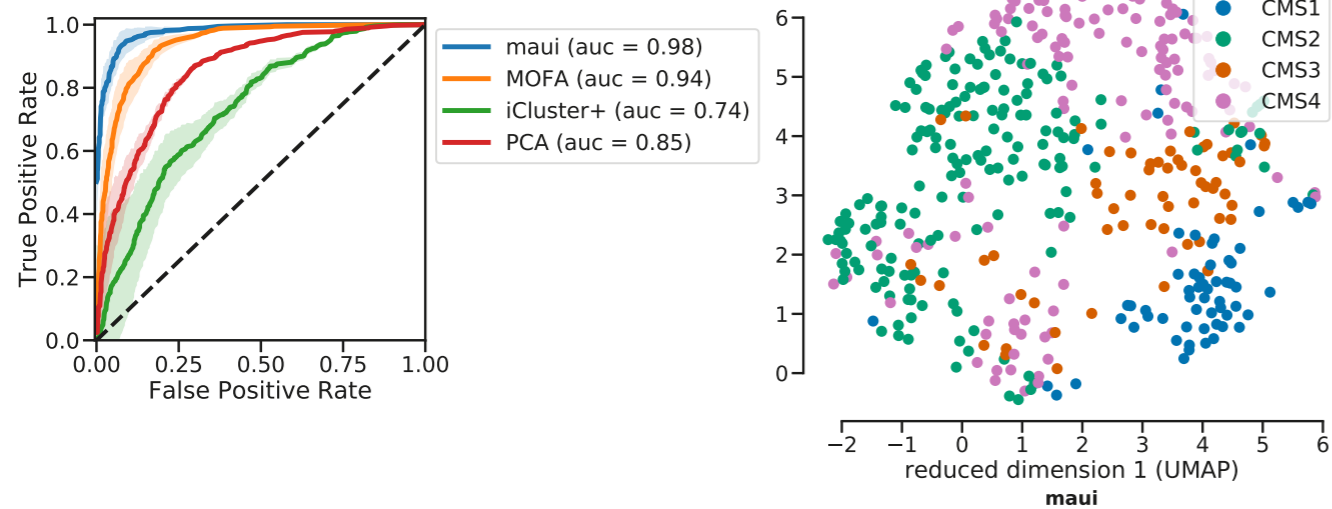


FIGURE 1: Predicting subtypes using latent factors obtained via deep learning is more accurate. Left, accuracy in comparison to other tools. Right, Representation of colorectal tumors by reducing latent factors to 2D. Plot obtained from [3].

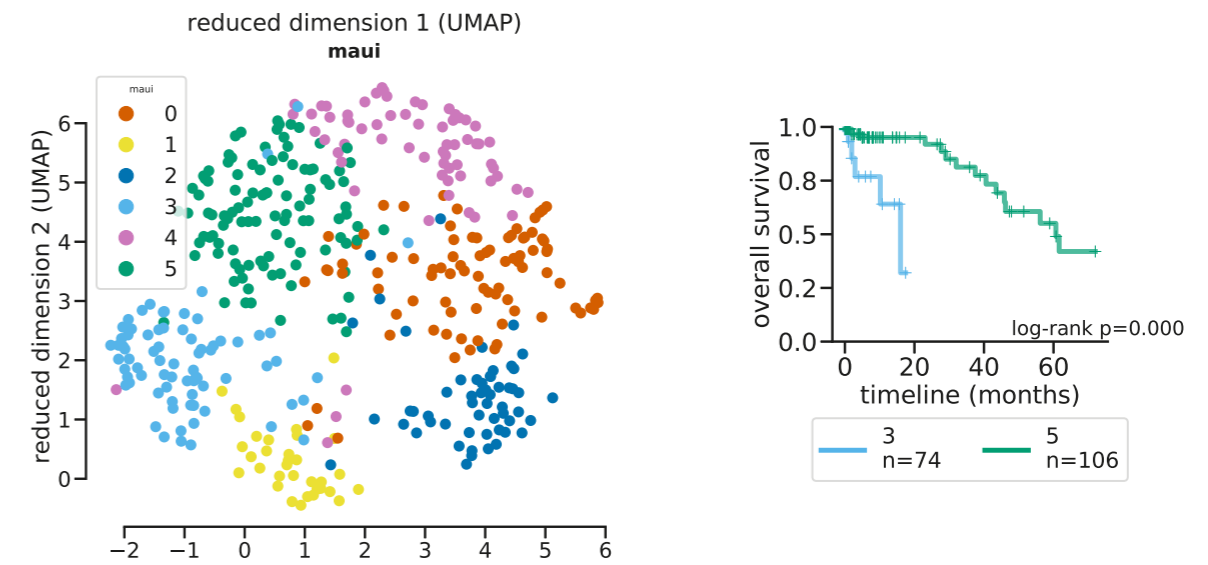


FIGURE 2: Refined subtypes for colorectal cancer. Green and blue dots represent CMS2 subtype, however separating this subtype into 2 subtypes as suggested by latent factor clustering makes more sense in terms of different survival characteristics of these two groups. The plot obtained from [3].

dependent on cancer's genetic, epigenetic, transcriptomic makeup of the tumor, as well as tumor microenvironment (Example: WINTHER Trial [2]). All of these vary substantially between different cancers, even from the same tissue. So, drug responses must be evaluated in relation to a cancer's genotype/epigenotype/transcriptome. Moreover, many drugs will fail simply because they are effective only on a subset of cancers, which was not initially recognized at the time of the trial.

MULTI-OMIC PATTERNS FOR CLINICAL VARIABLE MODELING

One way to eliminate these problems is to accept that clinical variables, such as drug response, are driven by genome/transcriptome/epigenome patterns and not just by mutations of single genes. Once we accept that, we have to have efficient methods for analyzing multilevel data sets, such as multi-omics, from cancer biopsies or tumor models. Efficient integration of multi-omics data will

provide an assumption-free or assumption-sparse, data-driven and integrative approach for modelling clinical variables. For our Arcas translational project, we have developed such a framework which uses deep learning to integrate any kind of omics data and discover molecular patterns, or so-called latent factors. Latent factors can be used for 1) clustering/subtype detection or mapping disease models and primary tumors - analogous to a biological search engine. Imagine you can input multi-omics data for your disease models and we can tell you which primary cancers are best represented by those models. In addition, 2) we can model variables such as survival and drug response. Furthermore, 3) we can also interpret the latent factors, and understand which molecular mechanisms, or pathways they correspond to.

PATIENT STRATIFICATION USING DEEP LEARNING

Molecular patterns or latent factors can stratify patients based on prognosis or

response to drugs, or any other clinical variable. We have applied part of Arcas technology on colorectal cancers [3]. Colorectal cancers have four subtypes defined using mainly gene expression profiles. These subtypes are known as consensus molecular subtypes or CMS. The consortium defined four subtypes that have different molecular characteristics that correlate with survival to some degree, however 19% of the patients could not be assigned to a particular subtype.

If these molecular patterns, i.e. latent factors, contain relevant information, we should be able to predict the CMS from them. In Figure 1 (left panel) we are looking at a receiver operating characteristic curve, or ROC curve, which shows how good a classifier is. In our case, the ROC curve shows that latent factors are able to predict CMS subtypes. We compare our method with other methods that also produce latent factors or the related principal components. In all cases, our prediction accuracy is higher.

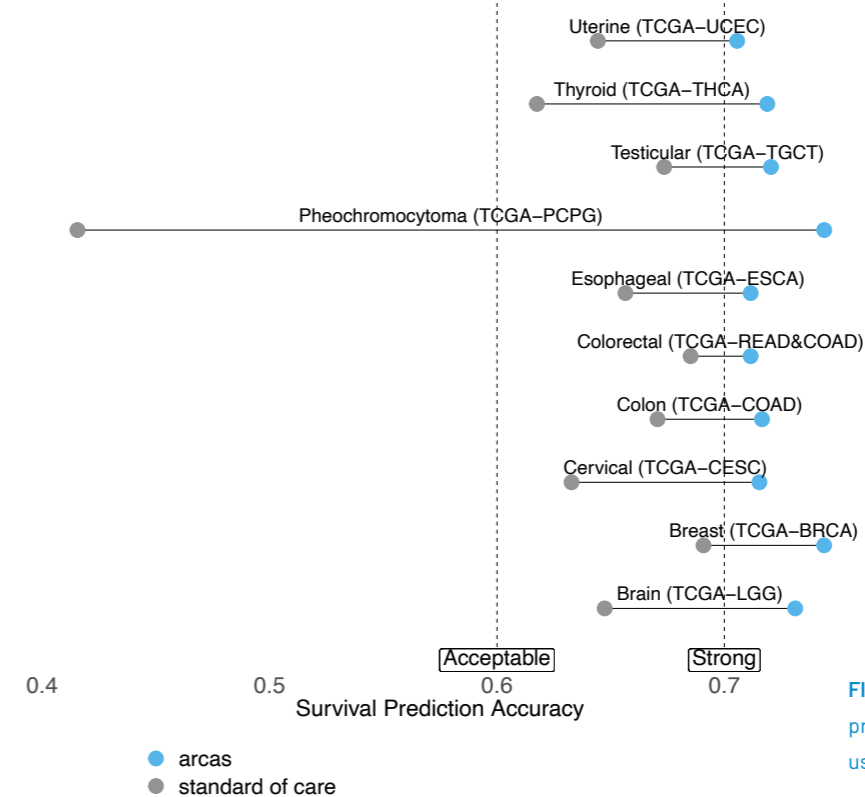


FIGURE 3: Arcas platform improves survival prediction over using clinical features.

In Figure 1 on the right side, we color-coded the 2D projection of latent factors based on the CMS status. Each dot is a primary tumor colored by the CMS status. You can see separation of colors, which means there is information in latent factors about CMS status.

We can further refine the subtypes using this technology. You might have noticed that there are more clusters than the colors based on CMS in Figure 1. If we apply a clustering algorithm, we can find six clusters. The biggest difference is that we separate CMS2 into two clusters. In terms of survival, this actually makes a lot of sense. In Figure 2 (right panel), we show survival curves of these new two clusters, which are very different. It could be, therefore, justified to break-up CMS2 to two subtypes.

One of the things we should emphasize is that this method is not limited to a specific cancer type. It works in any data set that has multi-omics information, including tumor models, such as cell lines, PDX or organoids. In fact you can integrate cell lines, PDX, and primary tumors using our method - something that is hard to do normally.

We have run this on the cancer genome atlas data sets that have at least 100 samples and in this plot, we are showing how much we improve the C-index. C-index is a measure of survival prediction accuracy. In Figure 3, we show what happens when we try to predict survival just by using clinical variables, such as age, gender, and tumor stage - in comparison to clinical variables + latent factors. As you can see, in many cancers, when using latent factors, we push this accuracy metric to a higher level.



CONCLUSION & OUTLOOK

As sequencing prices drop, the data needed to build and run our models are getting easier to generate. In the near future, liquid biopsies and biopsies will be routinely assayed by multi-omics methods. Integrating and making sense of such datasets is the key to improve drug development and diagnostic processes. Our Arcas platform provides actionable insights from multi-omics datasets from tumor biopsies or disease models.

REFERENCES: [1] Global Cancer Observatory, <https://gco.iarc.fr/today/home> [2] Rodon J et al. 2019 Nature Med, 25, 751-758. DOI: 10.1038/s41591-019-0424-4. [3] Ronen J et al. 2019 Life Sci Alliance, 2;2(6):e201900517. DOI: 10.26508/lsa.201900517.

AUTHORS: Jonathan Ronen¹, Bora Uyar¹, Vedran Franke¹ and Altuna Akalin¹
¹ Bioinformatics and Omics Data Science Platform, Berlin Institute for Medical Systems Biology, Max Delbrueck Center for Molecular Medicine, Hannoversche Straße 28, 10115 Berlin,

MIDAS – MEDICAL IMAGE AND DATA ANALYSIS – WHY THE COMPUTATIONAL INFRASTRUCTURE MATTERS

Automated analysis of medical image data is a challenging task requiring advanced data analysis techniques including machine learning methods as well as large and complex data sets. A prerequisite for tackling medical image analysis tasks is a dedicated and optimized computational infrastructure that enables secure data storage, flexible access to data for researchers as well as highly performing computing hardware for machine learning applications. This article provides a short overview of how the Medical Image and Data Analysis (MIDAS) research group of the University Hospital Tübingen utilizes the de.NBI infrastructure for research projects in the field of medical image analysis.

MEDICAL IMAGE ANALYSIS: A CHALLENGING TASK

Acquisition, reconstruction and analysis of medical imaging data is a highly challenging task due to rapidly growing amounts of acquired data, increasing complexity of generated images and broadening clinical demand for quantitative analyses. These developments put strain on medical organizations and clinical specialists who often struggle to meet the expectations. Already today, large parts of medical image data can only be analyzed partially with an often narrow focus.

Recent developments in the field of machine learning, especially for computer vision, have opened a perspective for computer-assisted analysis of medical images that could contribute to more efficient clinical workflows and potentially more accurate diagnostic results.

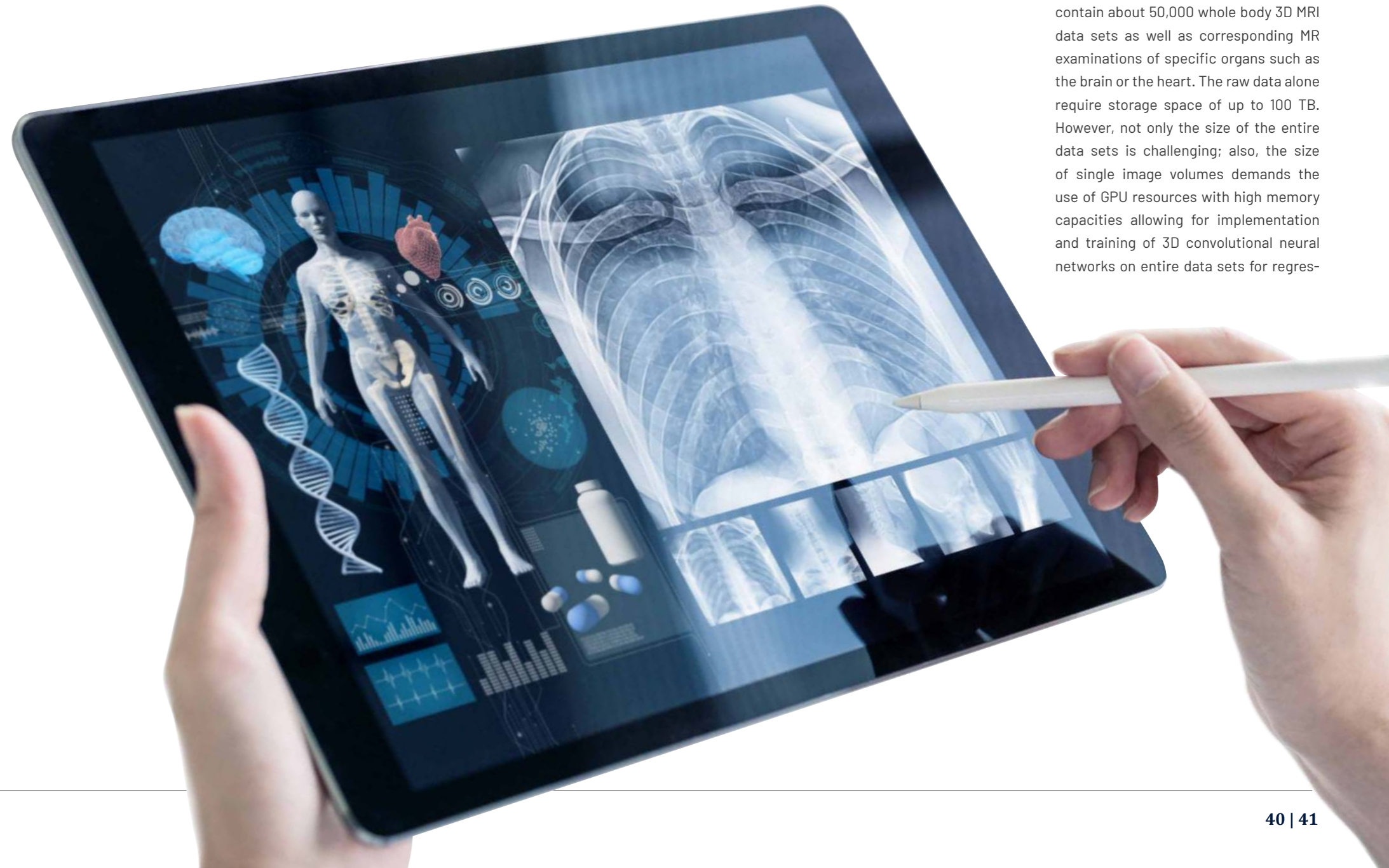
While the methodological developments of the past decade, mainly based on deep learning techniques, have been impressive, the translation of these tools to clinical application still poses challenges that are associated with algorithm robustness, estimation of algorithm uncertainty, as well as ethical and legal questions.

The overarching goal of the Medical Image and Data Analysis Lab (MIDAS lab) at the University Hospital in Tübingen is to develop, implement and validate machine learning methods for reliable and efficient reconstruction and analysis of medical image data. Our focus lies on analyzing multiparametric clinical imaging data from Magnetic Resonance Imaging (MRI), Computed Tomography (CT) and Positron Emission Tomography (PET) as well as large-scale data from epidemiological imaging studies.

INFRASTRUCTURAL AND COMPUTATIONAL DEMANDS

Besides scientific and medical aspects, the main challenge in this field of research lies in the high demand for computational resources for (i) safe storage of medical image data (ii) efficient prototyping of algorithms and (iii) training of large deep learning models. The de.NBI Cloud provides the necessary infrastructure that allows us to flexibly and efficiently plan and execute our research projects.

The following example illustrates the specific design of a typical project in medical imaging research: As part of a DFG-funded project ('Assessment of organ-specific biological age based on whole body MR data from the German National Cohort MR Study' project number 428219130), we aim to establish a framework for quantification of the biological age of organs and tissues based on phenotypes drawn from whole body imaging data. The underlying data is provided by two large epidemiological imaging studies, the German National Cohort Study (NAKO) and the UK Biobank study. Together, these growing data sets already contain about 50,000 whole body 3D MRI data sets as well as corresponding MR examinations of specific organs such as the brain or the heart. The raw data alone require storage space of up to 100 TB. However, not only the size of the entire data sets is challenging; also, the size of single image volumes demands the use of GPU resources with high memory capacities allowing for implementation and training of 3D convolutional neural networks on entire data sets for regres-



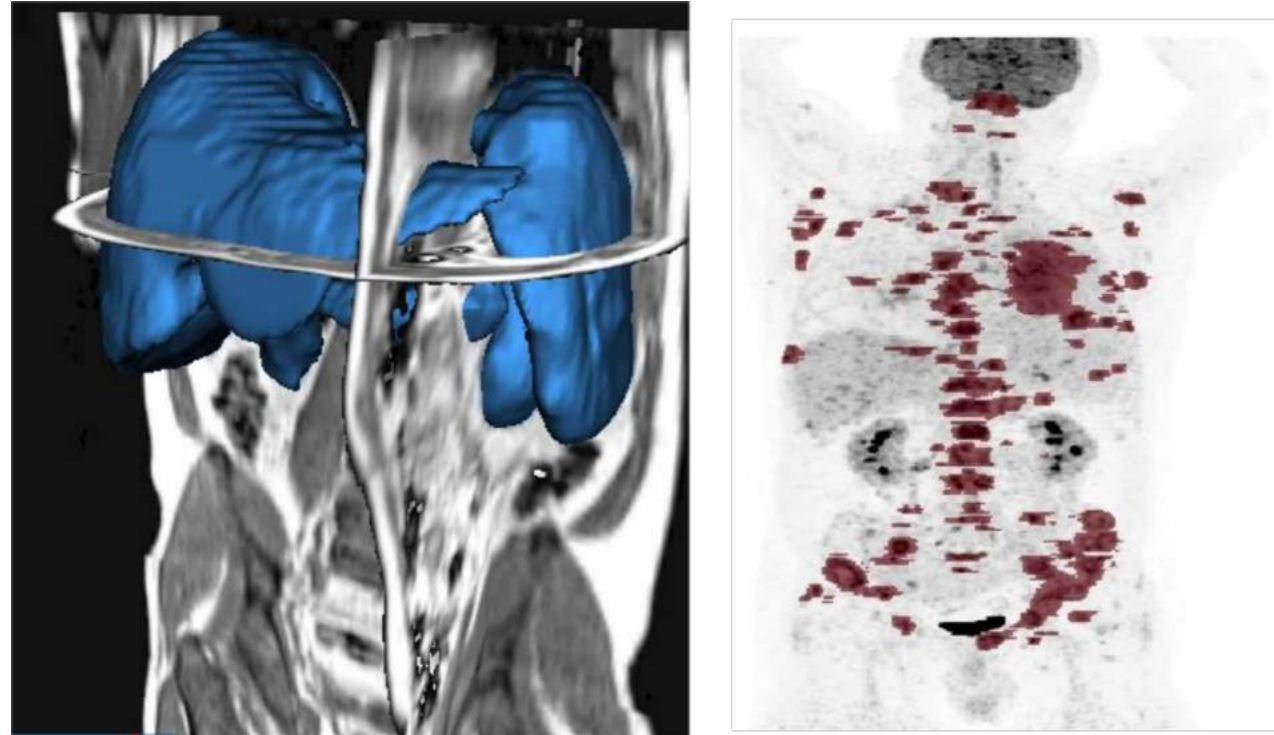


FIGURE 1: Examples for automated medical image analysis tasks: Automated segmentation of abdominal organs on MRI scans (left, masked in blue). Automated detection and segmentation of tumor metastases on PET scans (right, masked in red)

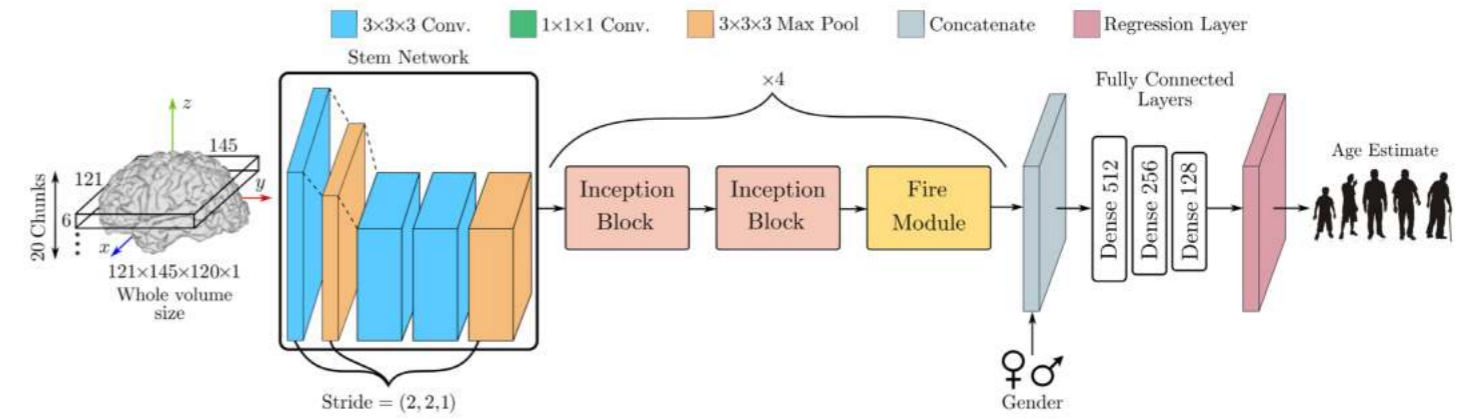


FIGURE 2: Example of a deep learning architecture for age estimation on entire 3D MRI data sets. The large size of image volumes requires the use of GPU-based computing [1].

sion, classification and segmentation tasks within this project.

At the moment, members of our research group, affiliated students as well as cooperation partners can simultaneously work on the de.NBI Cloud infrastructure and thus contribute to advancing research on medical image and data analysis. Access to virtual machines running on the de.NBI infrastructure is provided by the Secure Shell Protocol (SSH) enabling direct terminal access in a comfortable working environment with predefined permission. Thus, the de.NBI infrastructure has become an integral part of our research environment.

Compared to our previous infrastructure consisting of separate workstations and scattered storage, this environment has not only substantially increased the effi-

ciency of research and cooperation but has made larger projects feasible that had not been possible before.

The combination of the de.NBI infrastructure that allows for training of complex machine learning models together with local computation infrastructure within the secure hospital environment that can be used for deployment of trained models offers both, a high degree of safety and flexible, high performing computational resources.

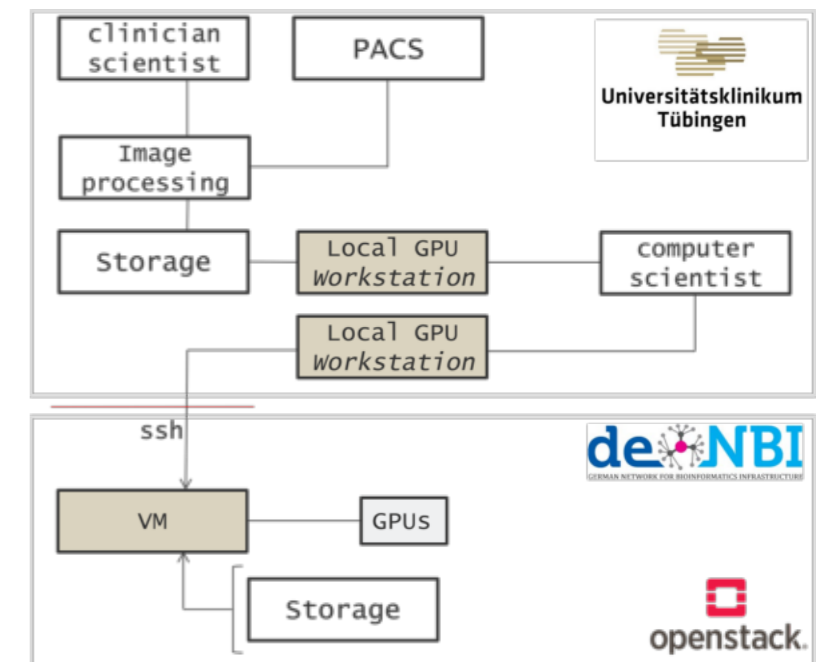
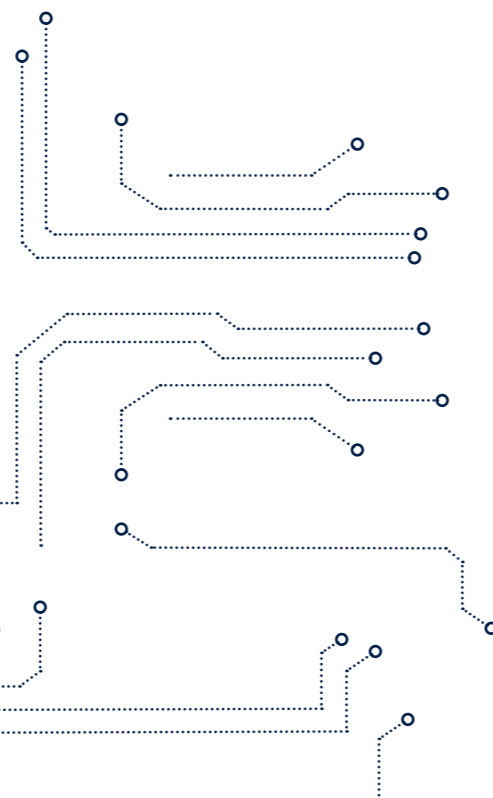


FIGURE 3: Simplified depiction of our computational infrastructure. Anonymized clinical data can be processed within the secure hospital environment using machine learning models trained on the de.NBI Cloud environment (PACS = Picture Archiving and Communication System, VM = Virtual Machine, GPU = Graphics Processing Unit).

REFERENCES: [1] Armanious K et al. 2021 IEEE Trans. Med. Imag. 40(7):1778-1791, DOI: 10.1109/tmi.2021.3066857.

AUTHOR: Thomas Küstner and Sergios Gatidis¹

¹University Hospital Tübingen, Diagnostic and Interventional Radiology, Geissweg 3, 72076 Tübingen

HIDDEN PHENOTYPES

Microphenomics reveals novel disease resistance genes using deep learning and automated microscopy

Crops are humankind's vital source of food, feed, oil, and fibers, and their performance was deliberately improved ever since the early years of plant domestication. Modern phenotyping provides impartial computer vision-based and molecular tools for crop improvement. Visualizing previously imperceptible features enables the next level of understanding the relationship between genome, environment, and plant performance. An example is Microphenomics that helps to dissect the resistance mechanisms involved in the early microscopic stages of plant-pathogen interactions. Deep learning tools enable extracting valuable phenotypic information from the enormous amount of complex microscopy image data.

USING PLANT BIOINFORMATICS FOR HIGH-THROUGHPUT ANALYSES OF PHENOTYPES

The German Crop BioGreenformatics Network (GCBN) service center is part of the German Network for Bioinformatics Infrastructure (de.NBI) and provides tailored services and training for plant research. With the specific focus on crops, the GCBN activities are closely linked to practical utilization, e.g., in essential breeding programs for feed and food.

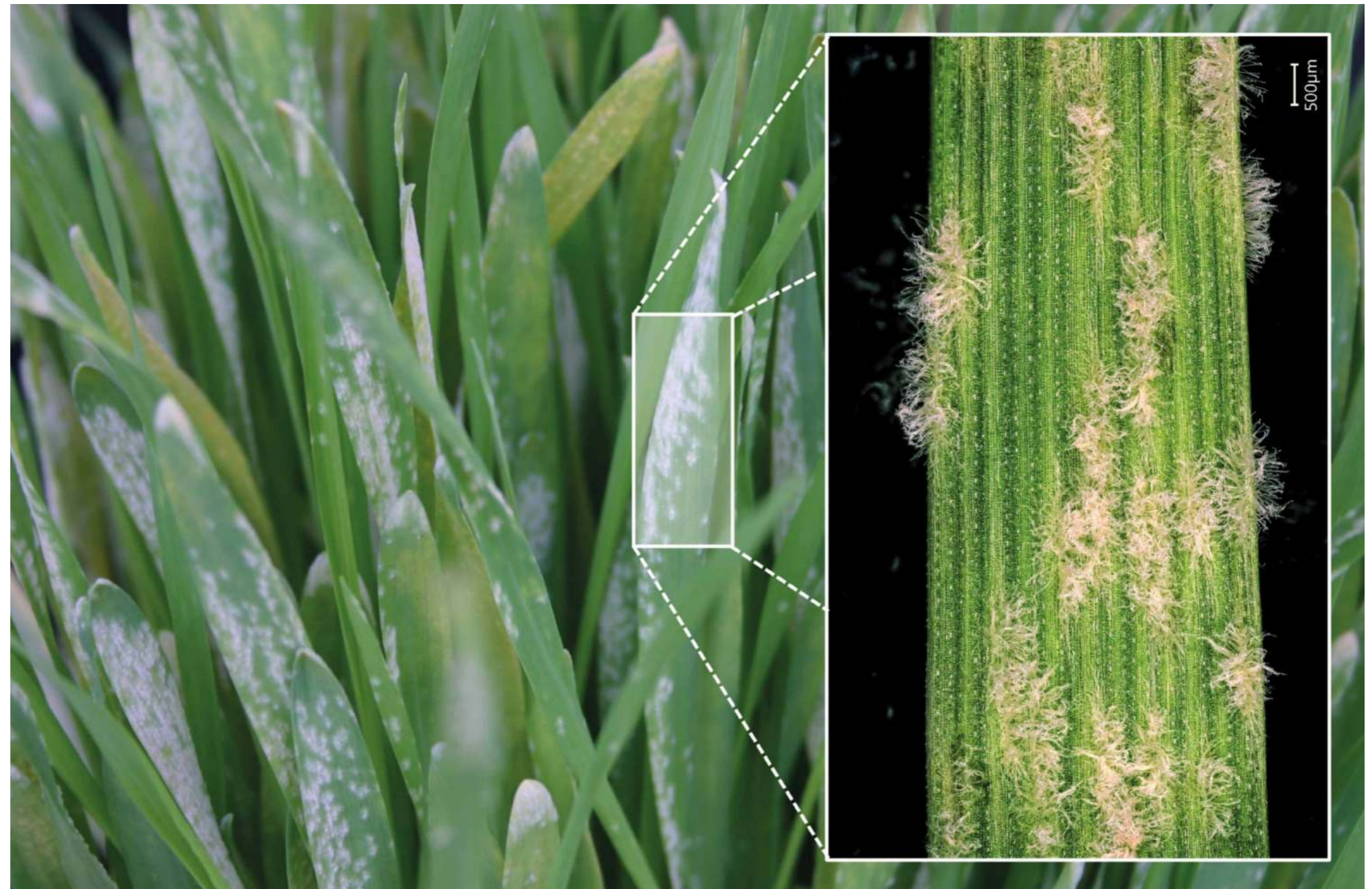
The continuous progress in sequencing technologies enormously accelerated genotype unlocking. In contrast, the acquisition of high-dimensional data on valuable organism characteristics

(phenotypes) involves specific challenges and often requires tailor-made developments. Phenotyping is required to better understand the pathways that connect genotypes to phenotypes and identify the genetic basis of complex traits. Advancement of automation of phenotyping raised a new field called phenomics. As a consequence, more and more data that needs to be adequately analyzed is being generated. New methods focusing on making the flood of data manageable and analyzable are being developed. Artificial intelligence (AI) methods like deep learning are being applied with growing success. Phenotyping of interspecies relations, such as host-pathogen interactions or microbiomes, adds another

level of complexity and new challenges. In plant-pathogen interactions, the initial stages of the infection are microscopic events, and their phenotyping was strongly hampered by the effort required for manual microscopy and the lack of high-throughput phenomics tools. To meet this challenge, we have developed the microphenomics approach

that combines high-throughput automated microscopy with artificial neural network tools for extracting phenotypic information from complex microscopy images of plant-pathogen interactions. The challenges are to analyze such phenotyping data and to integrate it into de.NBI services for plant researchers.

Artificial neural networks were initially developed to simulate the function of human brain cells inside a computer algorithm to learn and autonomously make decisions. The learning process requires feeding the network with labeled training data. A special feedback algorithm responds to the prediction precision and adjusts the network's weighted associ-



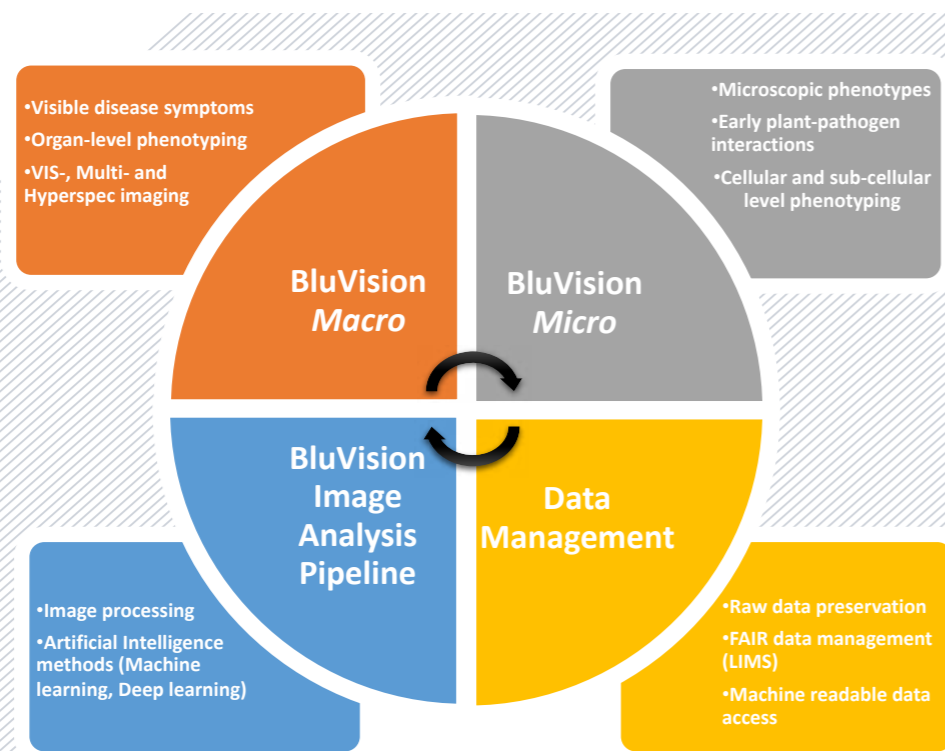


FIGURE 1: The BluVision Micro- and Macrophenotyping framework consisting of two modules for phenotyping on micro- and macro-scale, image analysis pipeline, and FAIR-compliant data management.

ations according to the learning rules until an optimum is achieved. A properly trained neural network model can predict objects in complex images or understand text documents with very high accuracy. Convolutional Neural Networks (CNN) are a class of deep neural networks commonly applied to analyze images. A typical CNN has a multilayer architecture with input and output layers and several hidden layers, which are important for learning specific abstract features. For achieving high accuracy, a neuronal network needs a large amount of training data. To develop some of our models, we have used about 10,000 images of each class.

Automated microscopy applications were significantly accelerated in recent years by releasing on the market several so-called digital slide scanners. One of the most advanced devices is the Zeiss Axio Scan.Z1, used in our system, that can digitalize several hundred samples per day with a resolution sufficient to recognize subcellular structures on mul-

multiple Z-levels. However, such high productivity is coupled with generating a massive amount of primary image data, which requires a very efficient software algorithm for image processing and analysis. Our system uses the advantages of GPU computing, which employs the graphic processing units (GPU) as a co-processor, enormously accelerating the computing process.

OUTLINE – THE BLUVISION FRAMEWORK

By combining the methods described above, we have designed a Microphenomics framework for high-throughput and precise phenotyping on microscopic and macroscopic levels of plant-pathogen interactions named BluVision (Figure 1).

The framework is aimed primarily at phenotyping powdery mildew disease of barley and wheat. The phenotypes delivered by the BluVision platform, combined with genomics data (Figure 2), allow the

discovery of novel disease-resistant genes of high potential interest to plant pathologists and breeders. The high level of automation and throughput of the system allows the phenotyping of large collections of genotypes for genetics and genomics studies. Moreover, the system enables phenotypes that were hardly accessible with manual methods, such as precise quantification of the fungal hyphae area and finding rare infection events in ‘near-nonhost’ interactions (Figure 3).

The data is being organized according to the FAIR principles, with machine-readable metadata, assessable via a Web interface, reusable, and interoperable. The framework allows complete pipelines for complex data analysis like building statistical models and Genome-wide association studies (GWAS). The results of the described approaches will be integrated continuously into de.NBI services and enables the users to access results from deep learning data analysis.

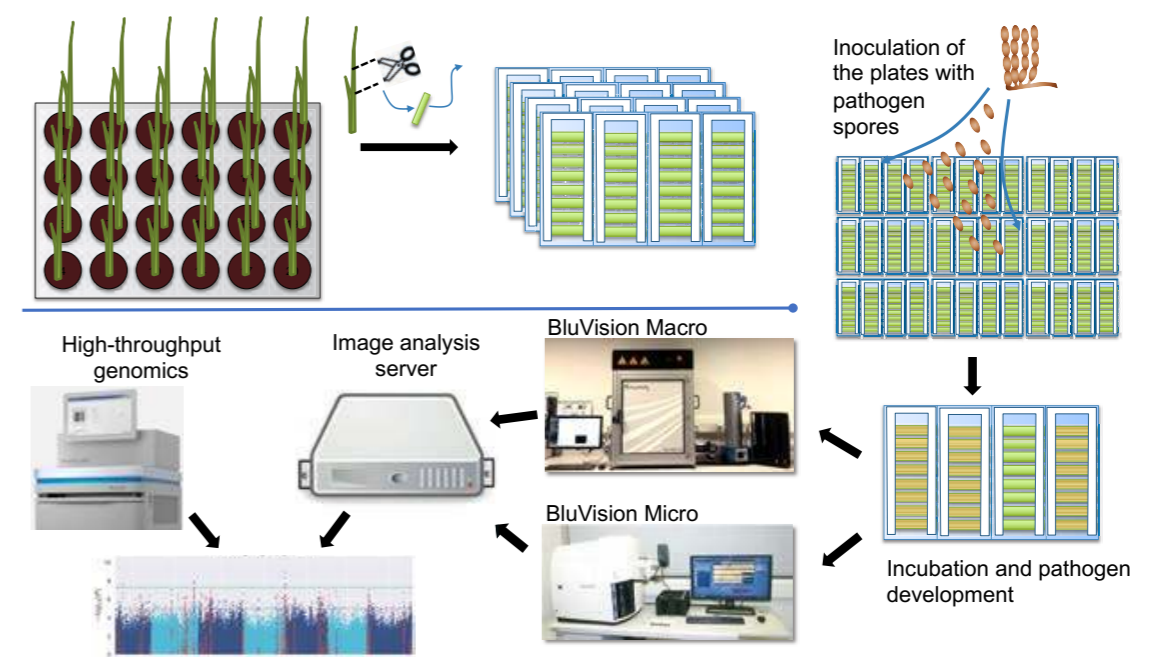


FIGURE 2: The BluVision Micro- and Macrophenomics pipeline. The whole experiment is performed in a highly controlled environment to minimize the influence of non-genetical variations. Detached leaf segments are mounted on special agar plates and in-

oculated with a controlled amount of pathogen spores. After incubation, the leaves are scanned on one or more of the phenotyping modules. The obtained quantitative phenotype, combined with high-throughput genomics data, can be used, for instance,

for performing genome-wide associations studies (GWAS) for discovering genes and genomic regions associated with specific resistant genotypes.

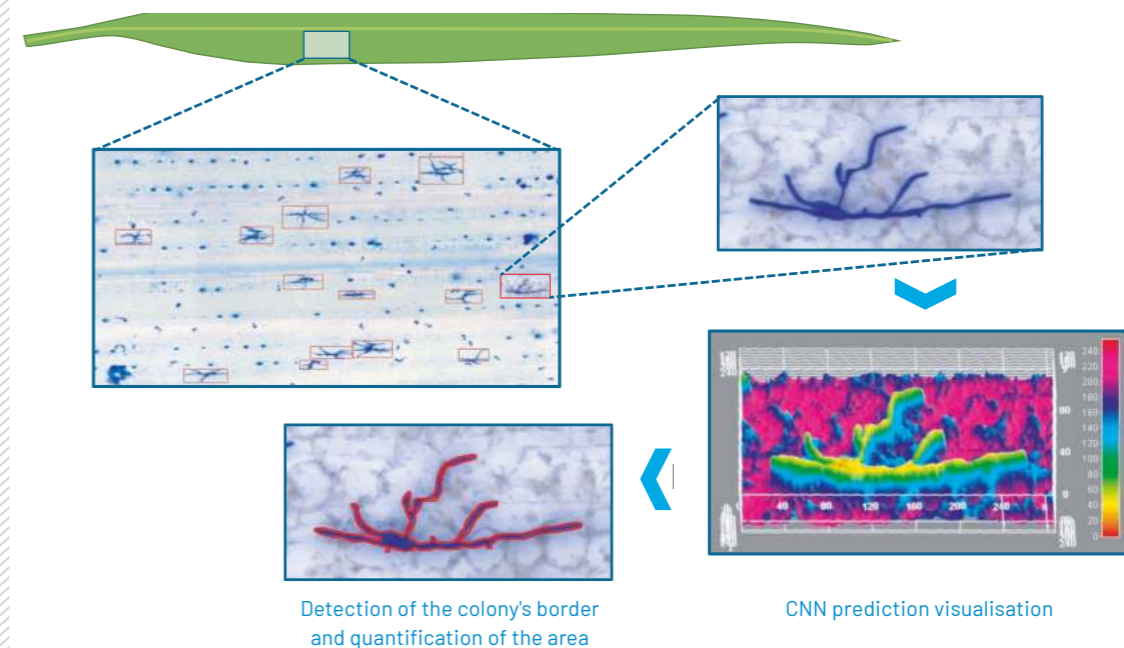


FIGURE 3: Detection and quantification of barley powdery mildew (*Blumeria graminis f.sp. hordei*) colonies. The infected leaves are scanned on the BluVision micro module,

which detects the colony number and precisely measures every colony's area using Convolutional Neural Networks (CNN)-based software.

**MODEL SYSTEM
 POWDERY MILDEW**

Crop protection is still mainly provided by applying chemical agents, many of which are strongly suspected to be harmful to the environment, biodiversity, and non-target organisms, including humans (EU directive 2009/128/EC). Achieving independence from chemical pesticides in the following decades while maintaining the economic balance requires a more profound knowledge of the biology of plant-pathogen interactions and deeper employment

of the natural and engineered plant disease resistance mechanisms. Biological models provide the intellectual frameworks needed to transform data into knowledge. The powdery mildew of barley and wheat are among the best-studied powdery mildews and an essential model for understanding plant-pathogen interactions, biotrophy, and plant immunity. The powdery mildew plant diseases are caused by a large group of obligate biotrophs fungi of the order Erysiphales with over 400 species that can infect more than 10,000 plant species. They represent a

significant threat in agriculture, affecting both the quality and quantity of the food, feed, technical and ornamental crops. Powdery mildew colonies have fast and synchronous growth and can accomplish their life cycle and produce a massive amount of spores only within one week (Figure 4). The light spores can be spread by the wind on distances of hundreds of kilometers. The ability for sexual reproduction and high genotype diversity assign the powdery mildews to the highest pathogen risk class.

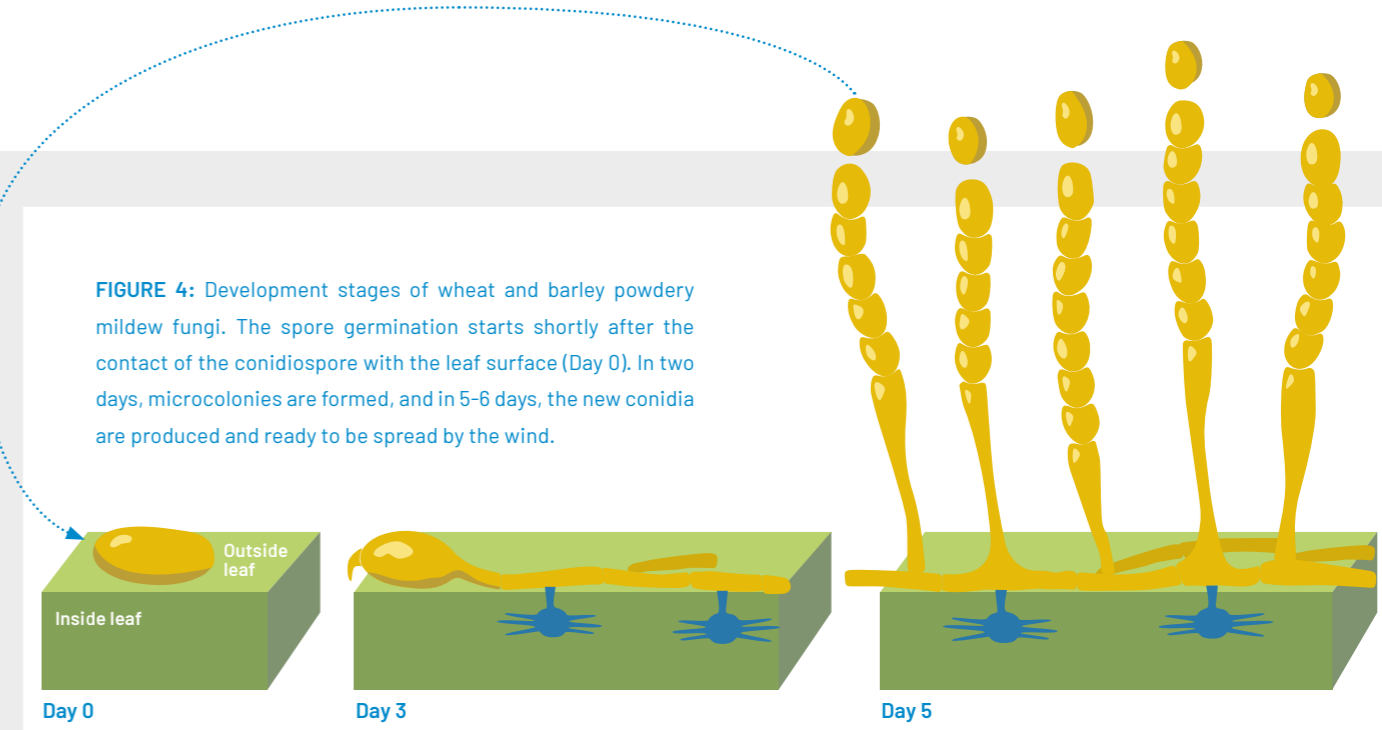


FIGURE 4: Development stages of wheat and barley powdery mildew fungi. The spore germination starts shortly after the contact of the conidiospore with the leaf surface (Day 0). In two days, microcolonies are formed, and in 5-6 days, the new conidia are produced and ready to be spread by the wind.

AUTHORS: Stefanie Lück^{1,2}, Uwe Scholz¹ and Dimitar Douchkov²

¹ Bioinformatics and information technology group, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Corrensstr. 3, D-06466 Seeland

² Biotrophy and immunity group, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Corrensstr. 3, 06466 Seeland.





BIOLOGICAL DATA MEETS AI – OMICS, BIG DATA AND MACHINE LEARNING AS TOOLS TO ACCELERATE UNDERSTANDING OF BIO- LOGICAL MECHANISMS

The strengths of AI to tackle a specific problem based on data with a certain intelligence leads to a new renaissance of the field. AI techniques applied to scientific discovery will quickly enable complex research challenges to be addressed that could never be solved by humans alone working within feasible time limits and resources. In the Life Sciences, it will allow us to understand and explore new cellular processes of a disease or a certain condition even on single cell level.



AI BASED METHODS FOR PLANT PROTEIN FUNCTIONAL INFERENCE

Using learned profile Hidden Markov Models allows plant genome comparisons

Genome sequences of several hundred plant species have already been published and are tracked in www.plabipd.de. However, due to the ground-breaking improvements in sequencing technologies, this number is rapidly increasing. To keep abreast with analysis of these genomes, it is necessary to be able to quickly and deterministically annotate protein functions. To this aim, we have developed the tool Mercator. Mercator v4 is based on more than 6000 specifically trained profile Hidden Markov Models that have been trained to specifically recognize land plant protein families. Here, we report on the use of these families for protein functional inference.

The recent introduction of long read sequencing technologies and their ever-improving outputs and decreasing error rates have made it possible to decipher more and more plant genomes [1]. Indeed, the assembly problem to generate one consistent genome out of the reads, can be considered as good as solved for most homozygous diploid plant species when using latest molecular techniques. With this breakthrough in mind, the plant community is faced with an unprecedented challenge: What to do with all these genomic data sets and how to compare and integrate genome data consistently?

While human-readable annotation pipelines such as AHRD exist, they are focused on human readability of protein descriptions (e.g. 'glucokinase' instead of 'similar to hypothetical protein from *E. coli*') and did not provide a machine-readable classification nor semantic similarity based on broad plant biological concepts.

Other endeavours like the GO ontology provide consistent ontological terms, but these suffer from often being generated *ad hoc* and are thus inconsistently assigned between different plant genomes. Their assignment can be achieved using tools such as BLAST2GO [2] which, despite its name, uses and cleverly combines multiple information sources. However, GO terms show redundancy, which can complicate inference and interpretation tasks.

We had developed the MapMan framework for biological inference and machine

learning tasks for plants by introducing redundancy reduced terms. Indeed, an independent assessment had shown that these terms outperformed the GO ontology for some learning tasks [3]. However, similar to GO term assignments, it was necessary to develop a performant and deterministic system to classify novel proteins into this ontology.

FROM SIMPLE SEQUENCE SIMILARITY TO PROFILE HIDDEN MARKOV MODELS

To this aim, we had initially developed Mercator v3 [4], which was using simple sequence similarity and domain assignments to determine protein function which informs about the role of individual proteins using similar ideas as BLAST2GO. However, this approach suffered from well-known problems of partial sequence similarities and hence wrong functional assignments as well as the problem of which sequences to use as templates and how to deal with conflicting annotations. We initially addressed this by generating protein clusters, but have since come up with a machine learning based solution which has resulted in the Mercator v4 line of sequence classification pipeline [5].

To this aim, firstly the known protein space was partitioned into protein sets using unsupervised sequence-based clustering. These clusters were then further split or merged using multiple sequence alignments and by phylogenetic inference. The clusters were filtered by a

known function assignable to the individual proteins of the cluster (preferably as published in a peer-reviewed article). Afterwards, protein sequences in optimized clusters were used to train individual sequence profile HMMs that capture the signature of a functionally distinct protein family e.g. all those sequences likely encoding for glucokinases, but not those encoding for other sugar kinases. These generated profiles were then subjected to a performance assessment and were further improved until accuracy converged. In certain cases, it was therefore necessary to have different HMMs representing the same biological function.

Finally, the learned ensemble of HMM classifiers was used to annotate several plant genomes to assess their performance and potential overlaps. Subsequently, performing sets were further improved in a supervised step. Given that each HMM was tuned by the multi-step learning process and since each HMM comprises its own detection thresholds, the results for each HMM ensemble are completely deterministic. An additional advantage is that on the level of each HMM, one can compare sub-families across a wide range of plant genomes. This is because the computation intensive steps of many to many sequence comparisons to infer clusters of orthologous proteins are foregone, by assigning a protein to a specific HMM. For example, if a protein from tomato is assigned to the HMM of glucokinase and another protein from watermelon is assigned to the same glucokinase profile, it follows that

these two proteins likely have the same function and that they must belong to the same protein family. Hence, it is possible to assign protein families by iteratively investigating new genomes, instead of recomputing protein family assignments between a set of genomes.

This allows to directly investigate genomes for the loss of protein families by simple comparisons as the machine learning output has greatly simplified comparison tasks.

As an example, we have shown that the parasitic plant *Cuscuta campestris* seems to have lost multiple proteins involved in plastidial gene editing [5].

CURRENT STATE AND LATEST DEVELOPMENTS

This new pipeline is continuously updated, and new releases are made available on a yearly basis. Currently, it is possible to deterministically classify and thus functionally annotate about 50 % of all proteins in a land plant genome. As the functional classification is completely deterministic, it is easily possible to determine potentially missing protein functions as well as too low classification rates (indicating e.g. the inclusion of too many pseudogenes) also on a broad level in a genome. This is reflected in the web interface to Mercator v4 which reports a ‘fill-rate’ indicating the number of classes that were found at least once and thus informs about potential general completeness.

For a general classification of land plant proteins, we currently develop unsupervised methods to provide a full deterministic classification of the evolutionary conserved part of plant proteomes where a function can't be assigned yet.

FROM GENOMES TO ANALYSIS

As Mercator v4 allows high throughput data analysis of genomes, it was applied to a selection of approx. 50 dicot plants and a red alga (*Chondrus crispus*) as an outgroup control, which were downloaded from the ENSEMBL Plants resource [6]. For the outgroup control only about 25 % of the proteins could be classified in line with expectation given that Mercator v4 was designed for land plant species. On the other hands more than 55 % of the proteins could be classified and assigned to families for species as diverse as *Arabidopsis thaliana*, kiwi and wild cotton in line with the predicted performances.

Using the direct comparison between the genomes allows to quality control different genome datasets. Only investigating presence absence matrices, revealed wide spread apparent absence of genes related to photosynthesis across multiple but not all species. However, this is simply reflecting plastid encoded genes as the plastid genome was included in some reference genomes but wasn't in others (Table 1).

In addition, the Mercator v4 tree view, which puts classes into their biological context flagged the loss of individual genes in the potato annotation. In detail, a cytosolic glucotransferase (DPE2) as well as an isoamylase (ISA3) and a pullulan-6-glucanohydrolase (PU1) all seemed to be lacking from the potato proteome annotation. However, given the importance of starch metabolism this seemed less likely. Indeed, all missing genes were present when analysing one of the most recent potato genome releases [7] which was based on newer sequencing technologies (Figure 1) highlighting the importance of new sequencing technologies and genome annotation quality control.

FIGURE 1: Total protein number content of the original potato genome (blue), the latest analysis (red) and for comparison content number in tomato (green).



TABLE 1: Analysis of a selection of ENSEMBL Plants Genomes for occurrence of proteins. The table shows the number of proteins (corrected for splice isoforms) found in each genome.

CODE	NAME	Chloroplast	<i>A. chinensis</i>	<i>A. trichopoda</i>	<i>A. halleri</i>	<i>A. lyrata</i>	<i>A. thaliana</i>	<i>A. alpina</i>	<i>B. napus</i>	<i>B. oleracea</i>
1.1.1.1.1	component LHCb1/2/3		13	8	9	8	9	11	25	11
1.1.1.1.2	component LHCb4		2	1	3	3	3	3	6	3
1.1.1.1.3	component LHCb5		2	1	1	1	1	1	4	4
1.1.1.1.4	component LHCb6		2	1	1	1	1	1	5	1
1.1.1.1.5	component LHCq		1	1	1	1	1	1	2	1
1.1.1.2.1.1	component D1/PsbA	YES	1	3	0	0	1	1	1	0
1.1.1.2.1.2	component D2/PsbD	YES	0	0	0	3	1	0	0	1
1.1.1.2.1.3	component CP47/PsbB	YES	1	1	0	2	1	1	0	1
1.1.1.2.1.4	component CP43/PsbC	YES	0	0	0	2	1	1	4	1
1.1.1.2.1.5	component alpha/PsbE	YES	0	1	0	2	1	0	2	0
1.1.1.2.1.5	component beta/PsbF	YES	0	0	0	0	1	0	0	0
1.1.1.2.1.6	component PsbI	YES	0	0	0	0	1	0	1	0
1.1.1.2.2.1	component OEC33/PsbO		3	1	3	3	2	2	10	5
1.1.1.2.2.2	component OEC23/PsbP		2	1	2	2	2	2	8	3
1.1.1.2.2.3	component OEC16/PsbQ		2	1	2	2	2	2	4	2
1.1.1.2.3	component PsbH	YES	0	0	0	2	1	1	2	0
1.1.1.2.4	component PsbJ	YES	0	0	0	0	1	0	0	0
1.1.1.2.5	component PsbK	YES	0	0	0	1	1	0	0	0
1.1.1.2.6	component PsbL	YES	0	0	0	2	1	0	0	0
1.1.1.2.7	component PsbM	YES	0	0	0	0	1	0	0	0
1.1.1.2.8	component PsbR		1	1	1	1	1	1	6	1
1.1.1.2.9	component PsbTc	YES	0	0	0	0	1	0	0	0
1.1.1.2.10	component PsbTn		5	1	2	3	2	2	9	3
1.1.1.2.11	component PsbW		3	1	1	1	1	1	8	7
1.1.1.2.12	component PsbX		6	1	1	1	1	2	6	2
1.1.1.2.13	component PsbY		2	1	1	1	1	1	5	3
1.1.1.2.14	component PsbZ	YES	0	0	0	3	1	0	0	0
1.1.1.3.1	assembly factor (LPA1)		1	1	1	1	1	1	2	1
1.1.1.3.2	assembly factor (LPA2)		1	1	1	1	1	1	2	1
1.1.1.3.3	assembly factor (LPA3)		1	1	1	1	1	1	2	1
1.1.1.3.4	assembly factor (HCF106)		2	1	1	1	1	0	6	2
1.1.1.3.5	assembly factor (HCF136)		2	1	1	1	1	0	3	1
1.1.1.3.6	assembly factor (HCF243)		5	1	1	1	1	1	3	2
1.1.1.3.7.1	scaffold component HCF244		2	1	1	1	1	1	2	1

CONCLUSION & OUTLOOK

We have shown the capability of Mercator v4 to deterministically annotate plant proteins in a plethora of land plant genomes drawing on unsupervised and supervised machine learn-

ing techniques and to even use it for quality control purposes. The creation of each profile HMM requires a large amount of supervised data resulting in a highly accurate model for the prediction of a protein function. Mercator v4 is restricted to annotate protein

sequences, which belong to a protein family of known function. In addition, we are developing unsupervised methods to classify evolutionary conserved land plant proteins independent of the knowledge of a function.

REFERENCES: [1] Dumschott K et al., 2020 Journal of Experimental Botany 71:5313–5322. DOI: 10.1093/jxb/eraa263. [2] Götz S et al., Nucleic acids research 2008 36:3420–35. DOI: 10.1093/nar/gkn176. [3] Klie S and Nikoloski Z, 2012 Front Genet 3:115 DOI: 10.3389/fgene.2012.00115. [4] Lohse M et al., 2014 Plant Cell Environ. 37:1250–8. DOI: 10.1111/pce.12231. [5] Schwacke A et al., 2019 Molecular Plant 12:879–892 DOI:10.1016/j.molp.2019.01.003. [6] Bolser D et al., 2017 Methods in molecular biology, 1533:1–31 DOI: 10.1007/978-1-4939-6658-5_1. [7] Pham GM et al., 2020, Gigascience. 9:giaa100. DOI: 10.1093/gigascience/giaa100.

AUTHORS: Rainer Schwacke¹, Marie Bolger¹, Asis Hallab¹, Jan P. Buchman² and Björn Usadel^{1,2}

¹ IBG-4 Bioinformatics, Wilhelm Johnen Str, Forschungszentrum Jülich, Jülich

² Biological Data Science, Universitätsstrasse 1, Heinrich Heine University Düsseldorf, Düsseldorf



MACHINE LEARNING FOR ELUCIDATING MICROBIOME FUNCTIONS

Machine learning approaches for the characterization of microbial secondary metabolism and associations with host traits

Human-associated microbial communities, such as the gut microbiome, are characterized by a large diversity of organisms interacting with their host in complex ways. However, to date, the functions of individual microbial organisms or genes, let alone of the community as a whole in the context of the human superorganism, are rarely known. Researchers thus want to predict them - and machine learning-methods are ideally suited for recognizing patterns in the complex data generated to study the gut microbiome. In this article, we will outline two machine learning applications, SIAMCAT and GECCO, developed for the discovery of microbial biomarkers and secondary metabolites, respectively.

The human body is colonized by a stunning diversity of hundreds of microbial species: prokaryotes, small eukaryotes, and viruses. Collectively, they are called the microbiome. Maturing sequencing technologies allowed us to study the human microbiome in new ways, circumventing the need for cultivation, which is challenging for many microbial species. In metagenomic sequencing, researchers are directly analysing DNA fragments from the microbiome. This technique has been particularly successful for the study of microbes in the human digestive system, the gut microbiome. Even though we are only beginning to understand, how the gut microbiome interacts with its host, its impact on numerous host physiological processes underlying health and

disease is clearly emerging. Gut microbes for instance regulate our immune system and influence how we respond to diet and medication.

Due to its fundamental and complex influences on disease, there is increasing interest in exploring the diagnostic and therapeutic potential of the microbiome. In clinical studies researchers have recently shown that the microbiome can be modulated to cure recurrent *Clostridium difficile* infections or improve the outcome of cancer therapies using fecal microbiota transplants. However, as researchers are beginning to translate findings from microbiome studies into biomedical applications, the need for robust and rigorous biostatistics meth-

odology is growing, as data analysis is complicated by the fact that microbiome composition and function is shaped by many host and environmental factors, which increases the danger of confounding in clinical microbiome studies.

Metagenomic sequencing of DNA extracted directly from microbial communities does not only enable researchers to address the question of 'who's there?' (that is to determine the taxonomic community composition), but also investigate what these microorganisms are doing, as far as this is possible by studying the gene functions present in metagenomic data. The enormous diversity of microbial enzymes and the secondary metabolites they produce have been a

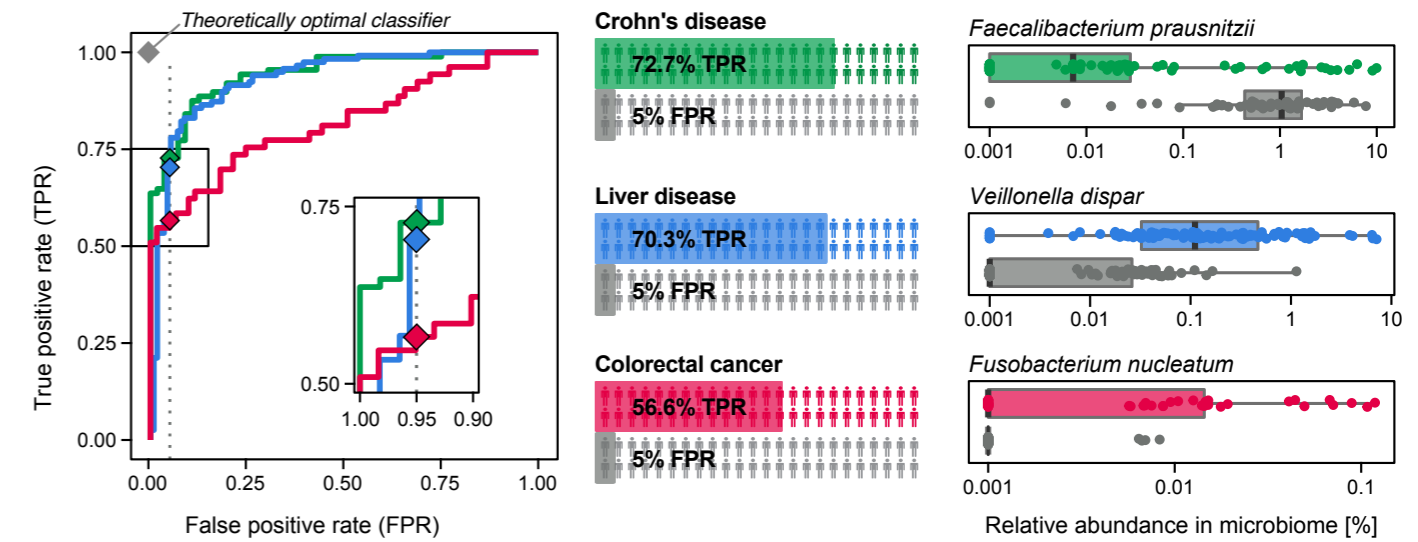


FIGURE 1: Machine learning models for classification of three exemplary diseases from metagenomic data trained and evaluated using SIAMCAT. Left and middle panels show an evaluation of their prediction accuracy, right panel the abundance of a microbiome biomarker from these models in patients (colored according to disease) and (healthy) control individuals (gray).

treasure trove for natural product chemists. Their study has led to the discovery of countless medically and biotechnologically relevant compounds over the last decades. Despite this ongoing success story, the vast majority of enzymes from microbial communities have never been characterized and still constitute a largely untapped source of new molecules to address current and future challenges in medicine – such as the antibiotic crisis – and biotechnology, e.g. for the biological degradation of waste and pollutants.

Given the complexity of microbial communities, machine learning methods are ideally suited to recognize patterns in the data. These learning algorithms can be trained on a subset of data that is well characterized by experiments (also referred to as labeled data), the resulting models can be used to make predictions on the, typically much larger amount of, data that is not well understood. Prior to

training, the labeled data is usually first subdivided into one part for model fitting and one part for the evaluation of the predictions the model makes. This way, it is possible to get an idea of its accuracy, e.g. to estimate the sensitivity (also called true-positive rate) and specificity (inversely related to the false positive rate) in diagnostic applications.

In the following, we will present two machine learning applications for the discovery of microbial biomarkers and secondary metabolites.

SIAMCAT

Multiple studies have shown that the microbiome plays a critical role in many host physiological processes. For many common human diseases, changes in microbiome composition have been linked with disease initiation or progression. These changes can be a basis for devel-

oping biomarkers, even in cases when the causal relationship between microbiome alterations and disease status remains to be elucidated.

Colorectal cancer is an exemplary disease in which this is being explored with the aim of complementing the currently most widely used diagnostic screening approach, colonoscopy, with additional non-invasive options. The key question here is whether biomarkers with specificity and sensitivity suitable for diagnostic applications can be extracted from the fecal metagenomes of colorectal cancer patients in comparison to tumor-free individuals as a control group.

For biomarker identification, machine learning approaches hold a distinct advantage over statistical tests, since machine learning models can make predictions on new data, such as microbiome profiles of new patients. This allows one

to estimate how sensitive and specific these predictions will likely be when used in a diagnostic setting.

Despite the promises for clinical application, extensively validated machine learning workflows for microbiome data analysis are still missing. To close this gap, we developed a R package called SIAMCAT [1, 2]. This toolbox includes machine learning, statistical testing, and advanced visualization approaches tailored to the challenges encountered in metagenomic data analysis.

To validate our method, we conducted a large-scale analysis of 50 microbiome disease association studies including a total of >10,000 samples (see Figure 1 for three examples from this set of studies). Our results show that relatively simple machine learning approaches (so called generalized linear models based on microbial species abundance) are produc-

ing accurate models of many human diseases using the currently available metagenomic data. Given their internal structure, these models are easily interpreted to find out, which microbial species contribute the most to the predictions. These can subsequently be prioritized for further development as clinical biomarkers (Figure 1).

Lastly, SIAMCAT also allows users to conduct meta-analyses, in which data from many independent studies is jointly re-analysed. This way, one can investigate how well machine learning models can be transferred across studies and how well they generalize to patient populations from a different hospital, potentially with different demographics, lifestyle, and co-morbidities. For the example of colorectal cancer, we found in a recent machine learning meta-analysis that microbiome biomarkers and machine learning models generalise with high accuracy

across eight different studies from three continents and would therefore theoretically be globally applicable to non-invasively detect colorectal cancer [1].

GECCO

With the ever-increasing rate at which microbial genomes and metagenomes are sequenced, there are tremendous opportunities to mine these data for microbial enzymes producing secondary metabolites with interesting functions in microbe-microbe and microbe-host interactions.

The fact that many of these enzymes colocalize in genomes in so called biosynthetic gene clusters (BGCs) can be taken advantage of by machine learning approaches. The key idea here is to learn which encoded protein domains are characteristic of BGCs and apply the resulting model to scan genomic sequences for a

local enrichment of such domains. This computational problem is also referred to as sequence segmentation and is commonly encountered in many other domains of computational research, for instance in natural language processing. To solve it, AI researchers have developed many machine learning algorithms, which have continuously improved the accuracy of sequence segmentation - as many of us have experienced for instance when using Google Translate. However, in computational biology, a rather traditional algorithm called Hidden Markov models (HMMs) is still most commonly used and until recently computational tools for BGC mining were based solely on HMMs. While these have led to the discovery of new BGCs, experimental validation of their predictions is made difficult by the large proportion of false-positives.

To address this shortcoming, we developed a high-precision tool for BGC identification, which we called GECCO [3]. It employs a modern, discriminative sequence segmentation approach called Conditional Random Fields, which has been demonstrated to outperform HMMs in many applications. Additionally, we assembled a large database of experimen-

tally characterized BGCs for training and evaluation of this method in comparison to previously developed BGC identification tools (see Figure 2).

Our evaluations showed that all machine learning based tools for BGC prediction performed substantially better after training on a more comprehensive data set. Yet, the predictions by GECCO were a lot more accurate than those made by the other tools, including by a very recently published approach based on Deep Learning (see Figure 2, right panel). In particular the fraction of false positive BGC predictions could be substantially reduced by GECCO, which greatly facilitates currently ongoing follow-up experiments aiming to characterize the chemical properties and biological functions of the encoded secondary metabolites. Finally, our evaluations also showed that GECCO is capable of detecting BGCs with novel gene and domain arrangements (not represented in the training data), which makes it a great tool to explore the uncharted biosynthetic potential of microbial communities including those inhabiting the human body.

CONCLUSION & OUTLOOK

These two examples highlight the potential of machine learning for gaining a better understanding of the complex functions of the microbiome, e.g. for studying how species interact with each other or with the host. The volume of high-throughput data on microbial communities has been quickly growing, so that the experimental characterization of newly discovered organisms and genes has not been able to keep pace. As simultaneously data complexity has increased, now ranging from genes to RNA transcripts, metabolites and images of microbes in host tissues, the relevance of AI in microbiome research will inevitably grow further in the near future to fuel biological discoveries.

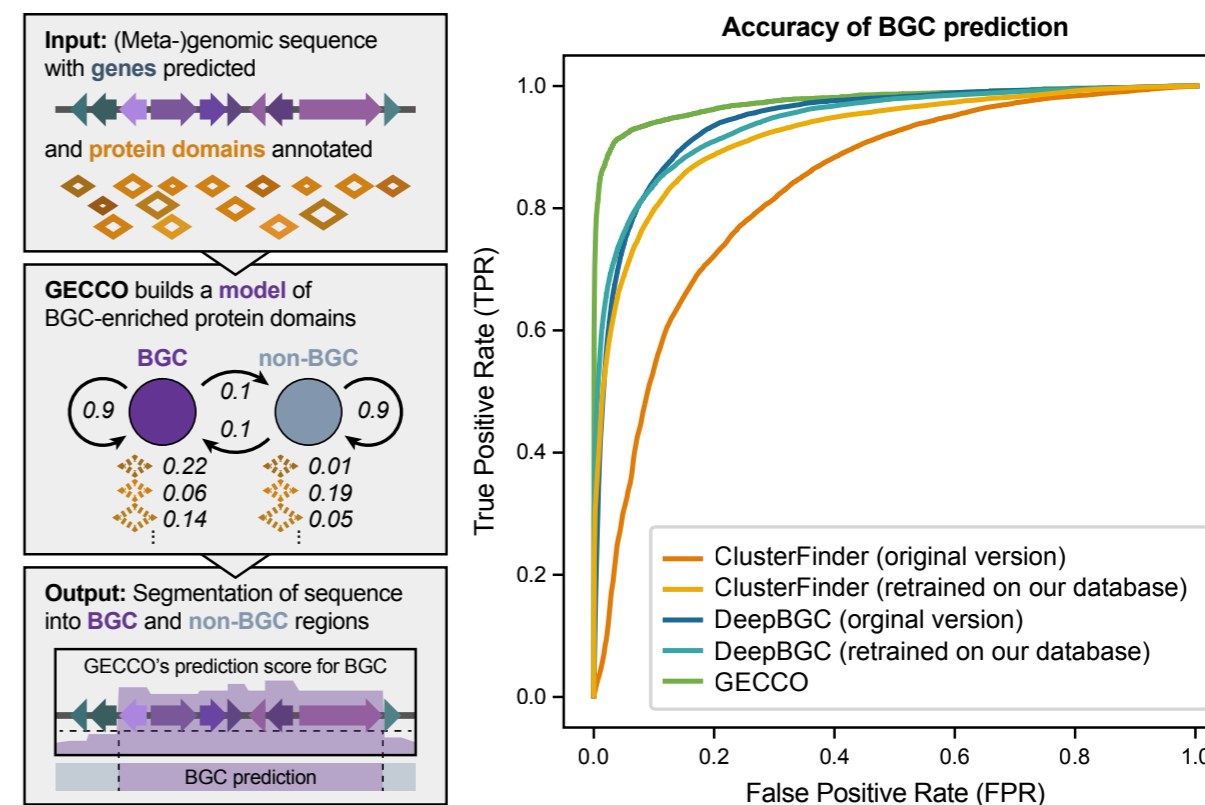
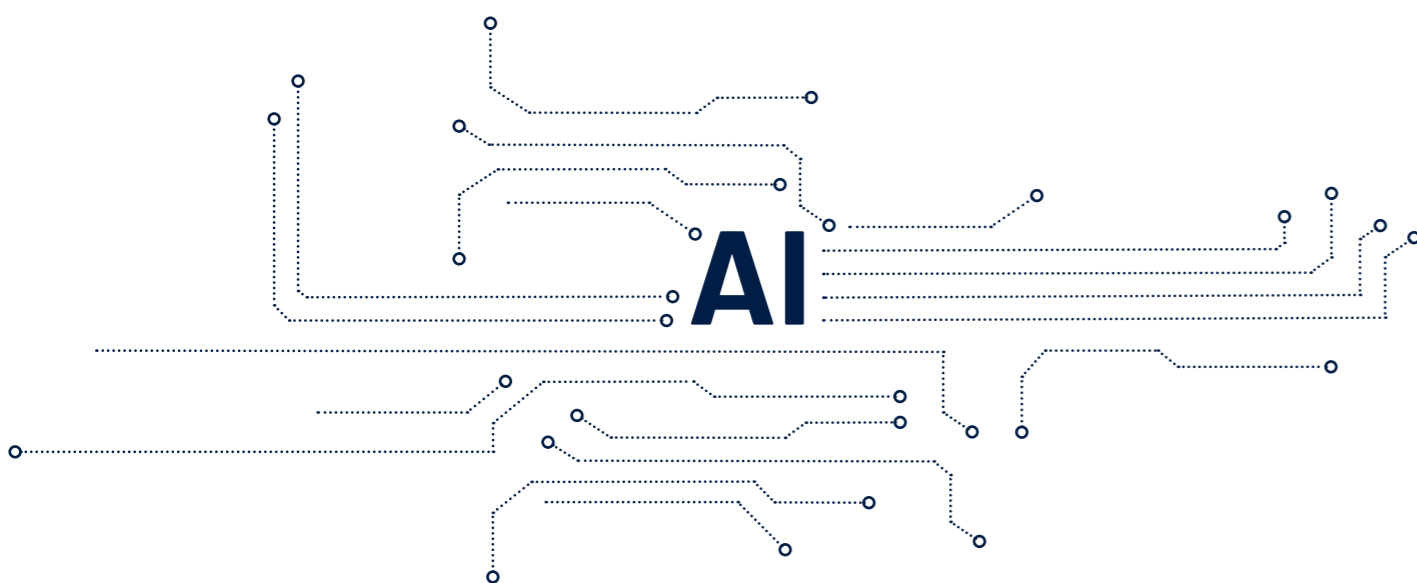


FIGURE 2: GECCO is a machine learning approach, right panel an evaluation based method for the identification of biosynthetic gene clusters (BGCs) in (meta-)genomic sequences. Left panel shows the sequences segmentation of its accuracy (in comparison to two similar methods called ClusterFinder, which is based on HMMs, and DeepBGC, which is based on Deep Learning).



REFERENCES: [1] Wirbel J et. al., 2019, Nature Medicine, 25(4):679-689. DOI: 10.1038/s41591-019-0406-6. [2] Wirbel J et. al., 2021, Genome Biology, 22(1):93. DOI: 10.1186/s13059-021-02306-1. [3] Carroll LM et. al., 2021, BioRxiv 2021.05.03.442509. DOI: 10.1101/2021.05.03.442509.

AUTHORS: Jakob Wirbel¹, Sina Barysch¹, Peer Bork¹ and Georg Zeller¹
¹ Structural and Computational Biology Unit, Meyerhofstr. 1, European Molecular Biology Laboratory, 69117 Heidelberg

DEEP LEARNING FOR PROTEIN VARIANTS DETECTION

DeProVIDEO will facilitate the identification of protein variants in mass spectrometry-based proteomics experiments

Mass spectrometry-based proteomics has become the standard for high throughput identification of proteins in complex samples. The commonly used approach for the identification relies on protein sequences, which are forwarded to identification algorithms. Usually only canonical sequences of the proteins are used, even if variants are annotated. The common algorithms can only identify sequences though, which they know of. Hence, spectra deriving from proteins containing variants cannot be identified. DeProVIDEO will facilitate the identification of variants, using all annotated variants and implement methods to deal with statistical problems arising from the application of big sequence databases.

The commonly applied approach to analyse proteins in mass spectrometry (MS)-based proteomics is by digesting the proteins in a sample into smaller amino acid sequences (peptides) and measuring these, via a liquid chromatography, on the mass spectrometer. The resulting spectra are in a further step matched to protein sequences for the identification of the original peptides in the sample. This step is performed by so called 'peptide search engines' [1, 2]. Most of the search engines depend on protein database exports in the FASTA format as an input and only peptides, which occur in the given database, can be matched against the spectra. Usually, though, the applied protein sequences contain only the canonical forms, sometimes additionally few isoforms. Any known variants, arising e.g. from mutations or being known to occur in or cause certain diseases, are not included and thus any peptide carrying a variant cannot be identified.

In the project DeProVIDEO, two problems are addressed: first, the variants need to be added to the peptide sequences in a way the search engines can make use of them. Second, the largely increased search space leads to problems in the estimation of false positives using the commonly used *target decoy approach*. This will be addressed by utilizing a spectrum centric approach and an improved spectrum identification method using deep learning.

ADDING VARIANTS TO THE DATABASES

Currently, if variants are analyzed in MS-based proteomics, there are two methods applied: the first method depends on sequencing data, be it transcriptomic or genomic data, to create a sample specific protein database containing the expected mutated sequences. This approach, though, would still miss all variants which are not sequenced by the applied complementary omics. The second approach is to create a protein database with only selected variants 'by hand'.

But the most often used protein database for proteomics, UniProt KB [3], also holds the annotations for known amino acid variants, even though an export into a format used by search engines, e.g. the most commonly used FASTA format, is not possible yet. To allow the identification of mutated peptides by MS-based proteomics, we are handing all possible combinations of varied peptides for the proteins to the search engine. By doing this, the number of peptides increases exponentially with the number of possible variants per peptide. If e.g. for a peptide, there is only one annotated variant, it only doubles this peptide. However, for two varied residues, at least four combinations must be considered, for three varied residues eight and so forth. In total, for the protein with the most variant annotation (P53), the number of possible peptides is higher than 10^{200} . This obvi-

ously cannot be exported into a FASTA file, as would be the normal process for a spectrum analysis by a search engine.

Instead, in an ongoing project, we are using a graph structure to encode the variants on the peptide sequences. This graph can then be loaded into a server's memory and any peptides of interest can quickly be queried per spectrum. To query the regular peptides, which contain no variant information, we use a database we recently published called 'MaCPepDB – mass centric peptide database' [4]. This database contains all regular peptides of the UniProt's proteins and is specifically tuned to allow querying for masses and also modifications, which are another important setting for spectrum identifications.

In the future, we will combine these two approaches, using the MaCPepDB for peptides, which are feasible to store on current cluster database setups, and the graph approach to store proteins, which contain too many variants and would lead to combinatorial explosions. With both approaches combined, we are able to quickly get all peptides of interest – with and without variants – for the spectrum identification.

SPECTRUM-WISE FDR ESTIMATION

After adding all variant peptides for each spectrum, the estimation of the false discovery rate (FDR) using the common-

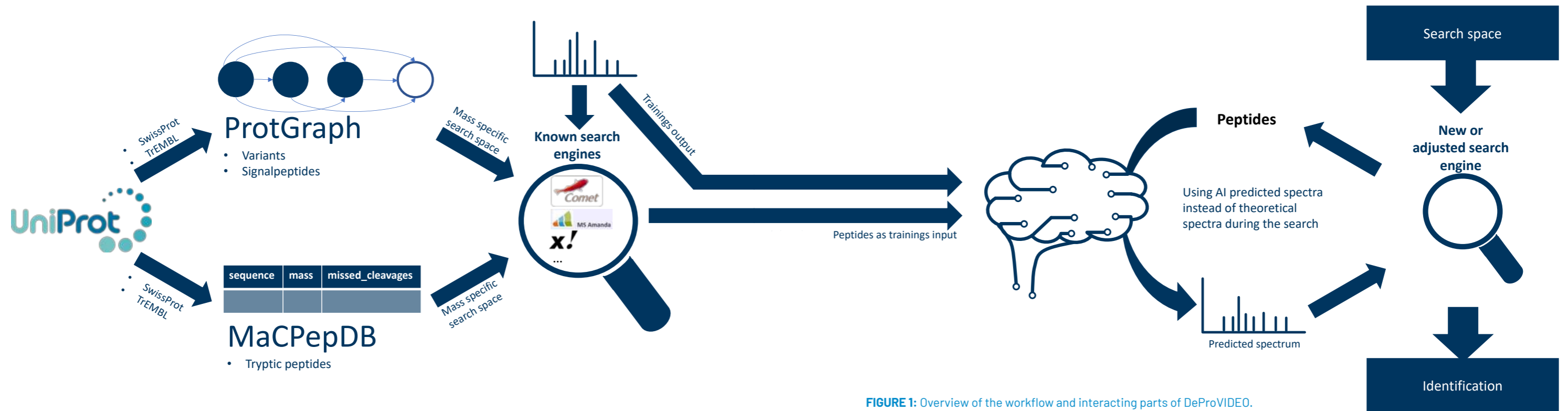


FIGURE 1: Overview of the workflow and interacting parts of DeProVIDEO.

ly applied target decoy approach (TDA) will result in an overestimation of false positives. To handle this, we will implement an approach in DeProVIDEO, which will not use a global FDR estimation, but a spectrum-wise approach. For this, an abundance of peptides fitting to the respective spectra is required, which is given by all the variants. If still more peptides are needed, so called decoy peptides (sequences, which do not exist in the original database and stand for false positive identifications) are generated matching the mass of the spectrum.

DEEP LEARNING FOR IMPROVED SPECTRUM MATCHES

Almost any database dependent peptide search engine exploits for the identification of spectra the fact, that it is relatively straightforward to create a theoretical tandem mass spectrum of a peptide. These spectra are then matched to the actual measured spectra and similarity scores as well as statistical metrics for the quality of the match can be calculated.

ed. These theoretical spectra lack the information of intensities, though, and only predict where the actual mass peaks are expected.

To calculate better metrics for the match, deep learning methods for the generation of theoretical spectra will be applied. The recently published software Prosit [5] is able to perform exactly this: given a peptide sequence and some mass spectrometer specific parameters, a spectrum including peak intensities will be returned. Theoretical spectra generated in this way will be used for the calculation and scoring of peptide spectrum matches to improve the match quality. With this approach combined with the spectrum-wise FDR estimation, we hope to gain good quality matches of peptides with and without variant information.

By reanalyzing MS experiments from the PRIDE repository [6] with the new variants containing databases, we are able to generate a more accurate training set to retrain and improve existing deep learning

approaches like Prosit or develop an entirely new approach, taking further mass spectrometric information into account.

Despite the described improvements to the spectra identification, we are expecting some spectra to be still unidentifiable by the searches that depend on the databases. These can either be due to proteins, which are not sequenced yet (especially for non-human data), but also for peptides carrying variants, which are not yet annotated in UniProt KB. These unidentifiable spectra will then be further analyzed with various state-of-the-art de novo approaches, which try to find the original peptide sequence of a spectrum without the application of sequence databases and theoretical spectra.

Moreover, the identified data that we are going to collect by reanalyzing PRIDE datasets may help to create a new deep learning based *de novo* approach to increase the number of successful identifications of former unidentifiable spectra.



CONCLUSION & OUTLOOK

With the described improvements of the database dependant spectrum identifications using variant information, including the spectrum-wise FDR estimation and a deep learning supported spectrum matching as well as de novo identification, we will be able to identify significantly more peptides carrying variants, even without sequencing

of the analysed samples. This will not only allow a highly improved coverage of identifiable spectra of a given mass spectrometry experiment, but will also increase the knowledge of the actual occurrence of variants in any samples. Finally, we are going to include these novel approaches into workflows and other services of the service center 'Bioinformatics for Proteomics' (BioInfra.Prot)[2] of de.NBI.

REFERENCES: [1] Eisenacher M, et al. 2012 Methods Mol Biol. 893:445-88. DOI: 10.1007/978-1-61779-885-6_28. [2] Turewicz M, et al. 2017 J Biotechnol. 2017 Nov 10;261:116-125. DOI: 10.1016/j.jbiotec.2017.06.005. [3] UniProt Consortium. 2021 Nucleic Acids Res. 49(D1):D480-D489. DOI: 10.1093/nar/gkaa1100. [4] Uszkoreit J et al. 2021 J Proteome Res. 20(4):2145-2150. DOI: 10.1021/acs.jproteome.0c00967. [5] Gessulat S et al. 2019 Nat Methods. 16(6):509-518. DOI: 10.1038/s41592-019-0426-7. [6] Perez-Riverol Y et al. 2019 Nucleic Acids Res. 8;47(D1):D442-D450. DOI: 10.1093/nar/gky1106.

AUTHORS: Dirk Winkelhardt^{1,2}, Dominik Lux^{1,2}, Martin Eisenacher^{1,2}, Michael Turewicz^{1,2}, Julian Uszkoreit^{1,2}
¹ Medizinisches Proteom-Center, Ruhr University Bochum, Universitätsstraße 150, 44801 Bochum
² Center for Protein Diagnostics (ProDi), Medical Proteome Analysis, Ruhr University Bochum, Universitätsstraße 150, 44801 Bochum



DeePSIVal: DEEP LEARNING PEPTIDE SPECTRUM IDENTIFICATION VALIDATOR

Deep Learning approaches for peptide-spectrum-match validation



LC-MS/MS-based proteomics is a valuable tool for the comprehensive analysis of medical and biotechnological samples, slowly entering process monitoring and routine diagnostics. Therefore, fast and reliable bioinformatic protein identification is crucial. The current bottleneck for the bioinformatic analysis time is the validation of the peptide-spectrum-matches (PSMs), carried out by false-discovery estimation (FDR) since it requires the finished LC-MS/MS measurement. To overcome this limitation, we developed and tested classification algorithms enabling the streaming PSMs validation. A convolutional neuronal network obtained the highest accuracy with values above 0.95 compared to the original FDR estimation. The new classification algorithm provides a faster PSM validation and a big step towards near-real-time processing of LC-MS/MS data.

The development of high-throughput methods such as next-generation sequencing for genomics or liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS) for proteomics revolutionized research in life science by providing comprehensive datasets for research in medicine, pharmacology, ecology, and biotechnology. In particular, proteomics enables identifying a multitude of diagnostic and prognostic biomarkers that are actually expressed under the given conditions. Metaproteomics applies proteomics methods to complex multi-species systems such as the human gut microbiome or microbial communities in the environment. Moving proteomics and metaproteomics towards routine analysis requires the fast and reliable identification and validation of the obtained protein matches [1]. To speed up the analysis, improvements to the laboratory workflow and data analysis are made continuously [2, 3]. Here, we present one such improvement to data analysis: peptide spectrum match validation using deep learning.

PEPTIDE SPECTRUM MATCHES AND VALIDATION

LC-MS/MS-based shotgun (meta)proteomics targets the identification and quantification of all proteins within a

sample. During the sample preparation, proteins are digested enzymatically into shorter peptides, which can be measured by the mass spectrometer. This means that mass spectra correspond to the shorter peptides, which later have to be mapped back to proteins. Protein database search engines compare the experimental spectra against theoretical peptide spectra derived from comprehensive protein sequence databases, revealing the best peptide-spectrum match (PSM) for a given spectrum. The PSM quality is typically expressed as a score representing the similarity of the theoretical and the experimental spectrum. Since similarity scores are continuous, a decision has to be made at which score threshold PSMs are accepted as true. The commonly used target-decoy-strategy addresses this problem by doing an additional search against a modified version of the original protein sequence database that is assumed to be false, resulting in decoy PSMs and their scores. The scores of these decoy PSMs can then be used to estimate a false discovery rate (FDR) for the search and adjust the score threshold accordingly, typically to FDR=0.01. However, the target-decoy approach suffers from several problems, including the doubling of search times – the most computing-intensive step of proteomics data analysis – and a lowered sensitivity

for true positives while searching against large sequence and decoy databases.

PSM CLASSIFICATION USING MACHINE LEARNING OR DEEP LEARNING

Classification problems such as the PSM validation suggest the use of machine learning or deep learning approaches.

Percolator is a machine learning algorithm that uses a PSM feature presentation to maximize the number of identified peptides for a collection of candidate PSMs at a target false discovery rate FDR [4]. The algorithm trains a support vector machine for a fixed number of iterations with decoy peptides as false examples and high-scoring matches from the collection as positive examples. This approach uses PSM features, requires a target-decoy search to be performed, and needs to be applied to each search individually and can thus be understood as an improvement of the target-decoy approach. In contrast, Nokoi was a machine learning approach using logistic regression and PSM features to train a general model that can theoretically be applied to any PSM, replacing the target-decoy approach entirely [5]. The performance of this tool is restricted, working well for simple samples but failing in more general cases.

Deep learning offers a more complex alternative to classification problems than machine learning. The most ambitious use of deep learning in proteomics is DeepNovo, a deep learning approach to de novo peptide sequencing introduced by Tran, Ngoc Hieu et al. [6]. DeepNovo utilizes convolutional neural network (CNN) and recurrent neural network (RNN) to learn the features of tandem mass spectra, fragment ions, and produces peptide sequences. The combinations of these deep learning structures and local dynamic programming allow the identification of MS/MS spectra with known peptides and the discovery of novel peptide sequences. While this system works well for the intended use-case of de novo sequencing, it is doubtful that it can improve upon current protein database search engines and subsequent PSM validation, as the number of identified peptides will be lower.

In our work, we applied the lessons from these tools and developed a deep learning approach to PSM validation: DeepSIVAL.

DeepSIVAL CONCEPT

DeepSIVAL uses the mass spectrum and the peptide sequence of the PSM as input data. We adopted and modified the DeepNovo input processing for spectra, which first transforms the mass spectrum into a binned representation whereby 0.1 Da is the value represented by one bin. It then calculates the m/z value of possible fragment ions for the next amino acid candidate and extracts 10 bins in a window of 1 Da around the fragment ion mass from the experimental spectrum. The windows are combined with the positional amino acid information into a matrix of shape (possible Fragment-ions)*(Max sequence length)*(Spectrum Window) (12*30*10).



This approach can also be utilized for the already identified sequences. Whereby the fragment ions for the entire sequence are calculated, and only the appropriate fragment ion windows are extracted from the experimental spectrum.

For the classification of PSMs a convolutional neural network is utilized. The neural network is organized with three convolutional layers with batch normalization and rectified linear unit (ReLU) activation followed by a fully connected classifier. The first convolutional layer is organized as 128 filters with 3*6 kernel, the second layer has 64 filters with a 2*5 kernel, and the third layer has 32 filters with a 2*2 kernel. The classifier combines a fully connected layer of 32 neurons with 2 softmax output units that produce a probability distribution over the identity classes.

A notable difference that machine learning and deep learning approaches offer compared to the target-decoy approach is, that the validation is done on the level of an individual spectrum, requiring no information of other PSMs, thus enabling parallelization and streaming.

RESULTS

The accuracy indicates the number of data points (PSMs) that are classified correctly, where our evaluation uses the

classification from the target-decoy approach as baseline for PSM validation. In an earlier attempt, we implemented a machine learning tool to validate PSMs, building on the tool Nokoi [5]. Among other things, we explored different models such as logistic regression and different types of species-based training data sets. A problem emerged, where models trained using the training data set for one species suffered a drastic drop in prediction accuracy when applied to the test data of another species, with the best model showing a drop from ~96% to ~80%. While this issue is still present using DeepSIVAL, this effect is drastically reduced as shown in Table 1. In case of the protein mix, the number of PSMs used for training was much lower than for the *E. coli* or HeLa data sets.

CONCLUSION & OUTLOOK

We demonstrate that PSM validation using convolutional neuronal networks is feasible. DeepSIVAL only requires the mass spectrum and the peptide sequence of a single PSM as input and enables a fast PSM validation that represents a big step towards near-real-time processing of LC-MS/MS data.

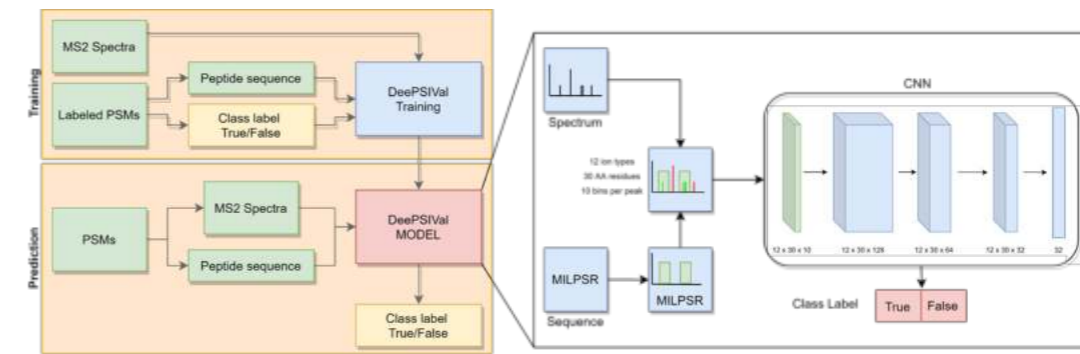


FIGURE 1: The DeePSIVAL workflow, showing the model creation by training and the prediction workflow for use in a larger pipeline. To the left the basic principle of the CNN model is illustrated.

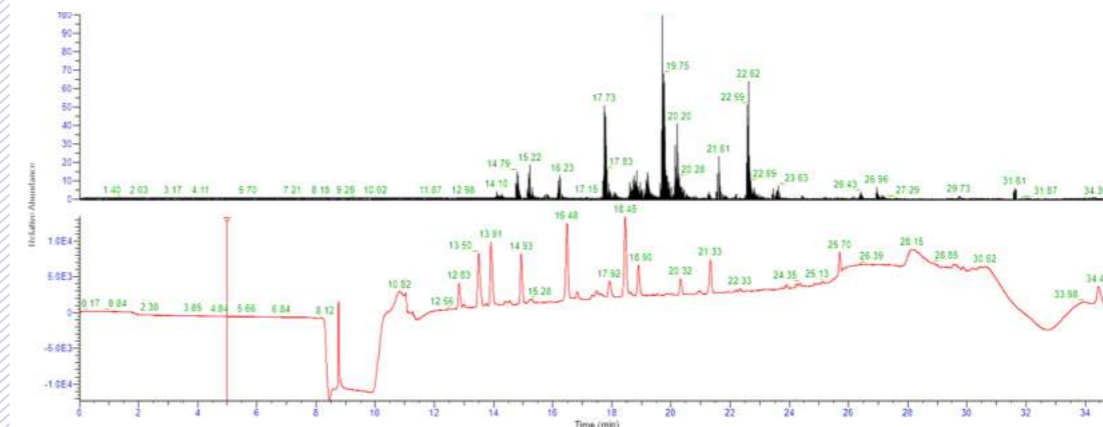


FIGURE 2: A single shotgun proteomics experiment of just forty minutes can result in tens of thousands of potential peptide spectra of greatly varying intensity and quality. Automation of data analysis is unavoidable

and only very few high intensity peptides that are visible in a chromatogram as shown in this figure are easily identified and validated. The vast majority of spectra that are recorded belong to lower abundance peptides

that require sophisticated protein database search engines and thorough validation for automatic analysis.

Accuracy Training vs Test	<i>E. coli</i>	HeLa	Protein Mix
Ecoli	0.973	0.960	0.953
HeLa	0.958	0.975	0.956
Proteinmix	0.939	0.939	0.984

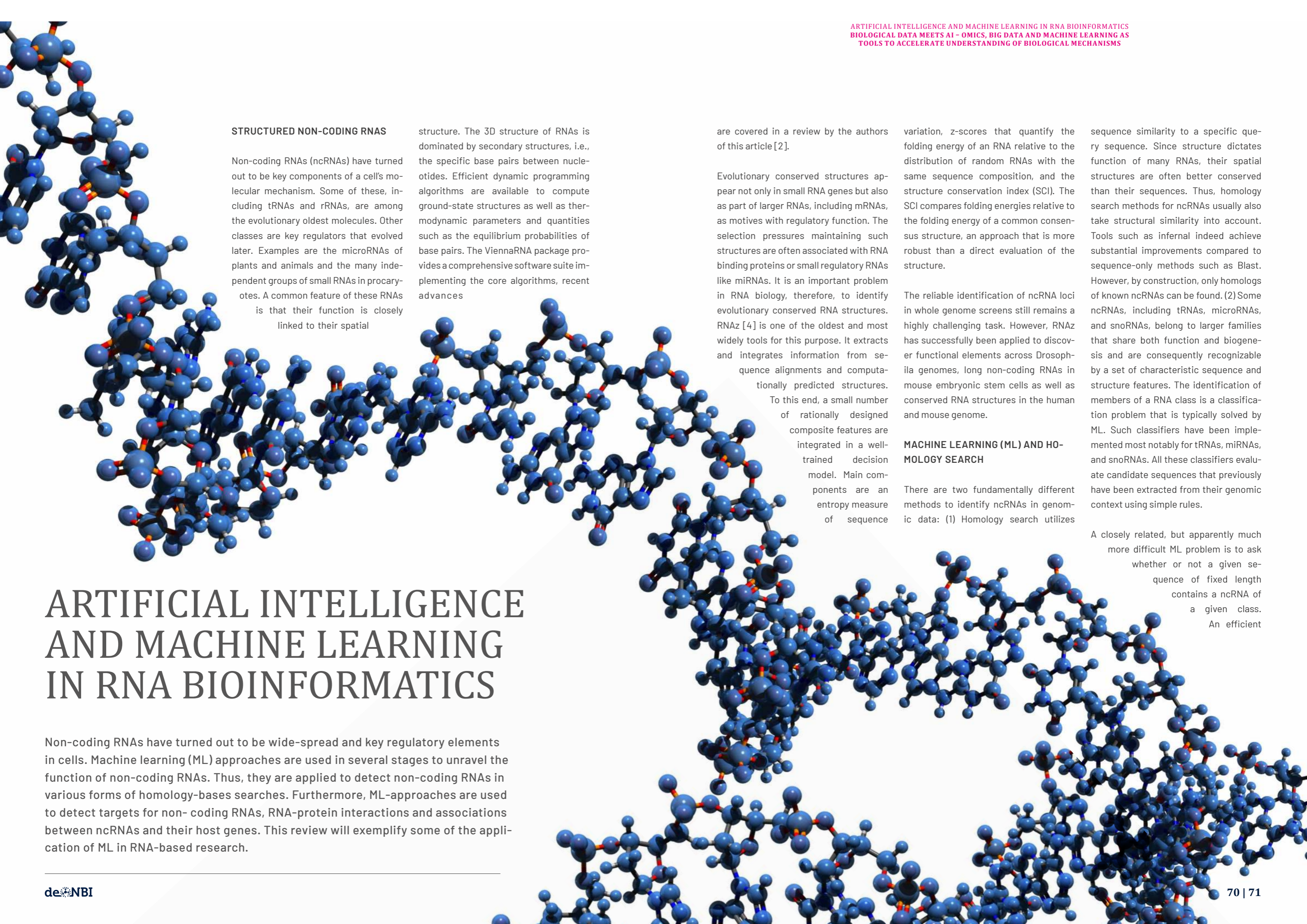
TABLE 1: Accuracy for three models trained with three datasets, tested against three datasets.

REFERENCES: [1] Heyer R. et al. 2017 J Biotechnol, 261:24-36. DOI: 10.1016/j.jbiotec.2017.06.1201. [2] Zoun R. et al. 2019 Gesellschaft für Informatik, DOI: 10.18420/btw2019-33. [3] Heyer R. et al. 2019 Front Microbiol, 10: 1883. DOI: 10.3389/fmicb.2019.01883. [4] Käll L et al. 2007 Nat Method., 4(11):923-5. DOI: 10.1038/nmeth1113. [5] Gonnelli G et al. 2015 J Proteome Res, 14, 4, 1792-1798. DOI: 10.1021/pr501164r. [6] Ngoc Hieu T et al. 2017 Proc. Natl. Acad. Sci., 114(31), 8247-8252, DOI: 10.1073/pnas.1705691114.

AUTHORS: Kay Schallert², Daniel Micheel¹, Robert Heyer^{1,2}, Gunter Saake¹, Dirk Benndorf²

¹ Otto von Guericke University, Institute for Technical and Business Information Systems, Database and Software Engineering Group, Universitätsplatz 2, G29, 39106 Magdeburg

² Otto von Guericke University, Faculty of Process- and Systems Engineering, Chair of Bioprocess Engineering, Universitätsplatz 2, G25, 39106 Magdeburg



STRUCTURED NON-CODING RNAs

Non-coding RNAs (ncRNAs) have turned out to be key components of a cell's molecular mechanism. Some of these, including tRNAs and rRNAs, are among the evolutionary oldest molecules. Other classes are key regulators that evolved later. Examples are the microRNAs of plants and animals and the many independent groups of small RNAs in prokaryotes. A common feature of these RNAs is that their function is closely linked to their spatial

structure. The 3D structure of RNAs is dominated by secondary structures, i.e., the specific base pairs between nucleotides. Efficient dynamic programming algorithms are available to compute ground-state structures as well as thermodynamic parameters and quantities such as the equilibrium probabilities of base pairs. The ViennaRNA package provides a comprehensive software suite implementing the core algorithms, recent advances

are covered in a review by the authors of this article [2].

Evolutionary conserved structures appear not only in small RNA genes but also as part of larger RNAs, including mRNAs, as motives with regulatory function. The selection pressures maintaining such structures are often associated with RNA binding proteins or small regulatory RNAs like miRNAs. It is an important problem in RNA biology, therefore, to identify evolutionary conserved RNA structures. RNAz [4] is one of the oldest and most widely used tools for this purpose. It extracts and integrates information from sequence alignments and computationally predicted structures.

To this end, a small number of rationally designed composite features are integrated in a well-trained decision model. Main components are an entropy measure of sequence

variation, z-scores that quantify the folding energy of an RNA relative to the distribution of random RNAs with the same sequence composition, and the structure conservation index (SCI). The SCI compares folding energies relative to the folding energy of a common consensus structure, an approach that is more robust than a direct evaluation of the structure.

The reliable identification of ncRNA loci in whole genome screens still remains a highly challenging task. However, RNAz has successfully been applied to discover functional elements across *Drosophila* genomes, long non-coding RNAs in mouse embryonic stem cells as well as conserved RNA structures in the human and mouse genome.

MACHINE LEARNING (ML) AND HOMOLOGY SEARCH

There are two fundamentally different methods to identify ncRNAs in genomic data: (1) Homology search utilizes

sequence similarity to a specific query sequence. Since structure dictates function of many RNAs, their spatial structures are often better conserved than their sequences. Thus, homology search methods for ncRNAs usually also take structural similarity into account. Tools such as infernal indeed achieve substantial improvements compared to sequence-only methods such as Blast. However, by construction, only homologs of known ncRNAs can be found. (2) Some ncRNAs, including tRNAs, microRNAs, and snoRNAs, belong to larger families that share both function and biogenesis and are consequently recognizable by a set of characteristic sequence and structure features. The identification of members of a RNA class is a classification problem that is typically solved by ML. Such classifiers have been implemented most notably for tRNAs, miRNAs, and snoRNAs. All these classifiers evaluate candidate sequences that previously have been extracted from their genomic context using simple rules.

A closely related, but apparently much more difficult ML problem is to ask whether or not a given sequence of fixed length contains a ncRNA of a given class. An efficient

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING IN RNA BIOINFORMATICS

Non-coding RNAs have turned out to be wide-spread and key regulatory elements in cells. Machine learning (ML) approaches are used in several stages to unravel the function of non-coding RNAs. Thus, they are applied to detect non-coding RNAs in various forms of homology-based searches. Furthermore, ML-approaches are used to detect targets for non-coding RNAs, RNA-protein interactions and associations between ncRNAs and their host genes. This review will exemplify some of the application of ML in RNA-based research.

solution to this version of the ncRNA classification problem would provide an alternative to homology search for large evolutionary distance, where sequence similarity comes close to or even falls below the detection limit. A first pilot study [1] has been encouraging, but also identifies this as a difficult problem for future research.

CLASSIFICATION OF ncRNAs FROM READ PROFILE

An alternative to identify ncRNAs in genomic context is the identification from transcriptomics data, inspired by the observation that microRNAs exhibit a characteristic read profile with two blocks, corresponding to the locations of the microRNA and the complementary microRNA*. The tool 'BlockClust' [7] uses 'blockbuster' to split read profiles into blocks representing certain read patterns. This read patterns are then encoded using a graph representation, which allows for a better representation of different combinations of properties. These are then processed using a graph-kernel technique, obtaining new features that represent pairs, triplets and higher order combinations of the original features, while still being able to be classified using efficient linear models such as linear support vector machines. In this way, the resulting classifier is in fact non-linear.

ASSOCIATION OF HOST GENE FUNCTION FOR LONG NON-CODING RNAs

A wide variety of molecular and biological functions have been reported for long non-coding RNAs (lncRNAs). Specific lncRNAs regulate chromosome architecture and chromatin remodeling, modulate inter- and intrachromosomal interactions and recruit or prevent the recruitment of chromatin modifiers. Other lncRNAs regulate transcription by forming R-loops thus recruiting transcription factors and

interfere with the Pol II machinery to inhibit transcription. There is, however, no clear-cut correspondence between sequence or secondary structure feature and lncRNA function. In contrast to protein-coding genes, where function is closely tied to protein families and specific sequence motifs, sequence similarity appears to be a poor predictor of functional similarity in lncRNAs. In stark contrast to small structured RNAs, it has remained impossible to predict the biological function or molecular mechanism of a lncRNA from its sequence alone.

Machine learning is a step forward towards the prediction of the biological function or molecular mechanism of ncRNA.

Unsupervised clustering of normalized k-mer abundances revealed an association of k-mer profiles with lncRNA function, in particular with protein binding properties and sub-cellular localization. Still, it remains an open question whether there are distinct, well-separated classes of lncRNAs or whether the universe of lncRNAs is organized as a continuum of functions and associated molecular features. In contrast to their highly conserved and heavily structured payload, the non-coding host genes of both miRNAs and snoRNAs feature poorly conserved sequences. So far, no connections between the function of the host genes and the function of their payloads have been reported, however.

In [6], we investigated whether there is evidence for an association of host gene function or mechanisms with the type of payload. To assess this hypothesis, we tested whether the miRNA host genes (MIRHG), snoRNA host genes (SNHG), and other lncRNA genes can be distinguished based on sequence and/or struc-

ture features unrelated to their payload. A positive answer would imply a functional and mechanistic correlation between host genes and their payload. We obtained a negative answer, however, indicating that the functions of host genes are not strongly constrained by the prior, primary function of the payload. While ML classifiers readily distinguish the three classes if presented with the payload sequences, they become virtually indistinguishable as soon as only sequence and structure of parts of the host gene distal from the snoRNAs or miRNA payload are used for classification. The functions of MIRHG and SNHG thus are largely independent of the functions of their payloads. Furthermore, there is no evidence that the MIRHG and SNHG form coherent classes of lncRNAs distinguished by features other than their payloads. The study [6] shows that ML approaches can also be employed to provide evidence for the independence of features by observing that for certain problems efficient classifiers are unattainable.

MACHINE LEARNING FOR (nc)RNA TARGET PREDICTION

A fundamentally important feature of ncRNAs and RNAs in general is their vast potential for interaction, intramolecular, as well as intermolecular with all sorts of other (nc)RNAs and proteins. This plethora of possible interactions makes them key regulators of many biological processes. So far, however, there exists no universal approach to unravel the systemic effects of such interactions. For that reason, it is important to get a genome-wide overview of likely interaction partners for a ncRNA. For RNA-RNA interactions, there are several tools for predicting a joint structure between two RNAs, with IntaRNA being one of the most popular ones. For the prediction of microRNA targets, established tools did not rely on the thermodynamic predic-

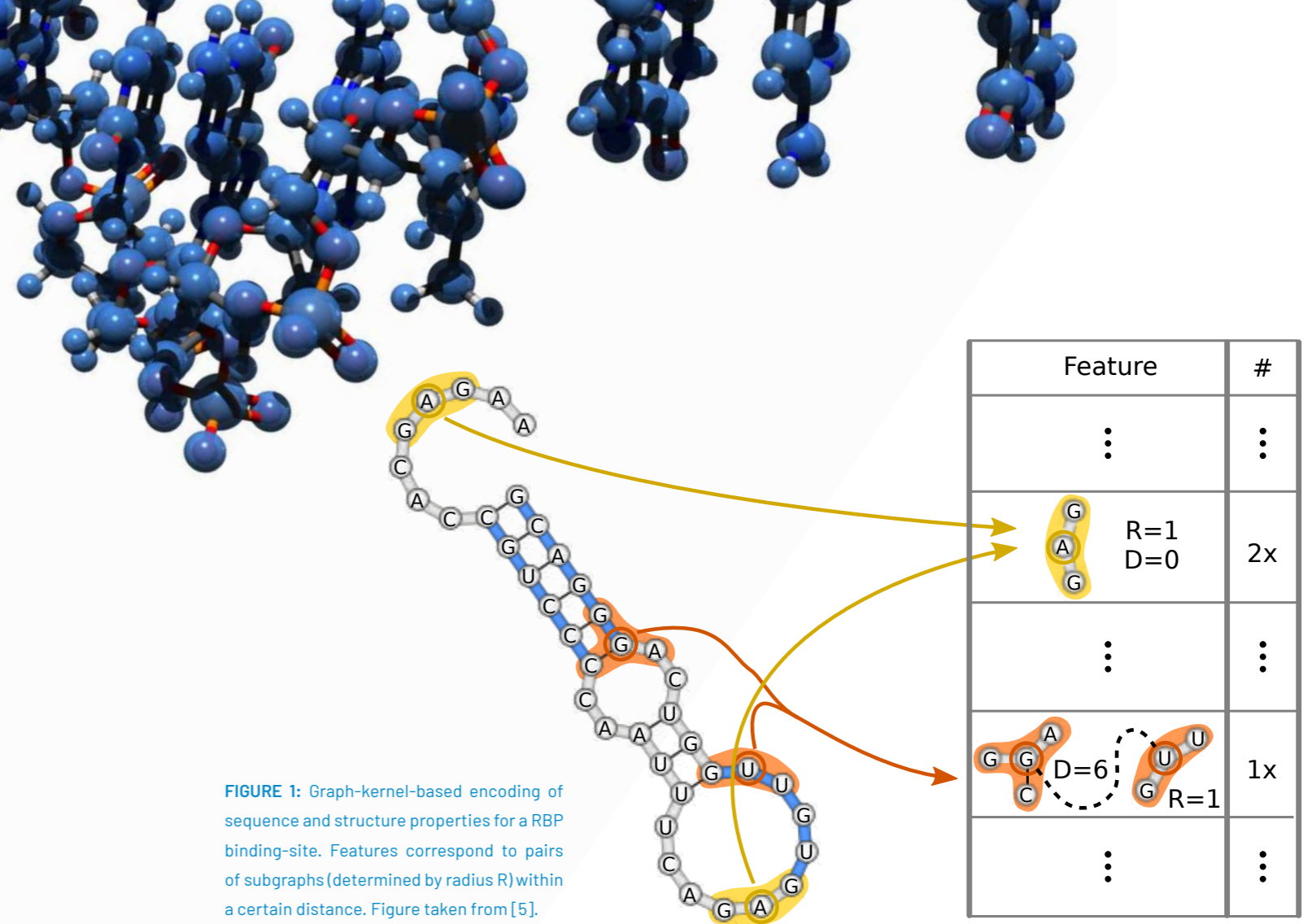


FIGURE 1: Graph-kernel-based encoding of sequence and structure properties for a RBP binding-site. Features correspond to pairs of subgraphs (determined by radius R) within a certain distance. Figure taken from [5].

tion of RNA-RNA interactions alone as the structure of an RNA is influenced by RNA-protein interactions. Instead, several additional features are combined into a score, either by a hand-crafted scoring approach, or via ML as in e.g., mirSVR or Miranda2.

For RNA-protein interactions, a purely sequence-based approach is impossible due to the complexity of the inter-

action. Here, the method of choice is to predict possible binding sites by learning sequence-structure properties from known interactions. This type of data is available on a genome-wide scale via specific sequencing protocols (such as CLIP-seq) that enrich RNA bound by a specific RNA-binding protein (RBP). As this data inherently is cell type specific, one needs prediction tools that determine likely bound RNAs that are not

expressed in the cell-type used for the CLIP-seq experiments. Here, GraphProt [5] used a graph-kernel approach to encode both sequence and structure properties of RBP binding sites (see Figure 1). MechRNA [3], constitutes a pipeline which integrates RNA-RNA interaction prediction with GraphProt to identifying possible mechanisms of lncRNA regulation.

REFERENCES: [1] Costa M C S F et al. 2021 *Bioinformatics*, vol. 3. 2021 176–183. DOI: 10.5220/0010346001760183. [2] Fallmann et al. 2017 *J Biotechnol* 2017261 97–104. DOI:10.1016/j.jbiotec.2017.07.007. [3] Gawronski A R et al. 2018 *Bioinformatics* 201834 3101–3110. DOI:10.1093/bioinformatics/bty208. [4] Gruber A R et al. 2010 *Pac. Symp. Biocomput.* 15: 69–79. DOI:10.1142/9789814295291_0009. [5] Maticzka D et al. 2014 *Genome Biol* 201415 R17. DOI:10.1186/gb-2014-15-1-r17. [6] Sen R et al. 2020 *Th. Biosci.* 2020139 349–359. DOI:10.1007/s12064-020-00330-6. [7] Videm P et al. 2014 *Bioinformatics* 201430 i274–i282. DOI:10.1093/bioinformatics/btu270.

AUTHORS: Jörg Fallmann¹, Peter F. Stadler¹, Rolf Backofen²,
¹Institute for Informatics, University Leipzig, Härtelstr. 16-18, 04107 Leipzig
²Albert-Ludwigs-University Freiburg, Georges-Koehler-Allee, Geb 106, 79110 Freiburg

EXPLAINABLE TRANSCRIPTOME (?) ANALYSES

The route of bulk, single-cell, and spatial transcriptomics analyses taken by explainable AI algorithms

RNA-sequencing as one of the most important and versatile high-throughput technologies continues its success story with novel developments within the area of single-cell and spatial transcriptomics. Likewise, molecular resolution and data interpretation complexity are increasing, which is why Machine learning algorithms are already essential for fast and accurate analyses. Here, we discuss possibilities of specific explainability algorithms around common bulk experiments, single-cell, as well as spatial transcriptomic investigations that are used to facilitate the overall data interpretation. The reader will get to know a broad variety of current AI applications for RNA-Seq data.

WHY DO TRANSCRIPTOMES NEED AI FOR AN EXPLANATION?

Thanks to the sequencing technology that enabled a fast and reliable measurement of the active state of the DNA, namely the RNA. Nowadays, transcriptomics analyses are of major interest for high-throughput investigations because of the ease of use and broad variety of applications available. Current applications like single-cell, single-nuclei, and spatial RNA-Sequencing (RNA-Seq) are adding great value and unprecedented molecular resolution to standard bulk RNA-Seq (Figure 1). The increasing complexity of experimental applications also resulted in more advanced computational data analysis procedures, in which Artificial Intelligence (AI) algorithms have an essential role. The application areas of ML supporting the interpretation and explanation of data include the general areas of data clustering, classification, and annotation, but also more specific ones, such as oversampling, trajectory prediction, or Deep fusion models to learn specific cell types. Here, we give a brief excerpt about current AI applications utilized in de.NBI for the transcriptomic domain (as part of the RBC – de.STAIR partner project) that facilitate the interpretation of RNA-Seq data.

INDEPENDENT BULK RNA-SEQ DATA ANALYSIS VIA AI

The advancements in RNA-Seq technology made it more feasible for researchers to sequence larger cohorts; thus it is not uncommon anymore to have several hundred samples per experimental group. Since RNA-Seq data is already high-di-

mensional by nature, featuring more than 200.000 coding and non-coding transcripts respectively in humans, the pairing with large amounts of samples highly attracts AI-based approaches for its downstream analysis.

Dimensionality reduction approaches, such as Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE), or Uniform Manifold Approximation and Projection (UMAP), are used for preliminary clustering of the data. The unsupervised abstraction of high-dimensional data into the two-dimensional space with retention of the variability contained within the data on the one hand enables straightforward detection of batch effects, and on the other hand yields insight in the composition of the cohort, with possible indications towards subpopulations.

Feature selection approaches using algorithms, such as Elastic Net Regression, Random Forest, or Gradient-Boosting methods, can be used to great effect in tandem with conventional Differential Expression approaches [1]. As an example, conservative fold change thresholds can be used to first trim a larger gene (=feature) list, and subsequently ML-based algorithms with their ability to identify inter-variable relationships narrow down this list to a handful of genes of interest [2].

SINGLE-CELL TRANSCRIPTOMICS FOR CELL SUBPOPULATION CHARACTERIZATION

The single-cell and single-nuclei RNA-Seq technologies (scRNA-Seq) provid-



ed us with an unprecedented resolution of molecular detail. Starting with initial tissue and cell preparation, cell capturing and library preparation, sequencing and raw data processing, the single-cell technology ultimately helps us in visualization and characterization of molecular profiles at a cell or nuclei level. Several AI-based techniques have found their application in visualization of the high-dimensional data and several other downstream analyses [3].

A popular application of AI in this area is to identify, quantify, and characterize cell populations in heterogeneous samples and tissues from scRNA-Seq. Intelligent feature selection and use of unsupervised learning approaches are the driving force in this area of research [4]. Advanced meta-learning based tools, such as MARS, can automatically identify and annotate known and novel cell types. A related challenge addressed by AI in this domain is to distinguish between closely related cell populations, potentially revealing functionally distinct groups with

complex relationships. This is realized by investigation of cell transitions through temporal cell-states to observe gradual transcriptional changes occurring in cells. Popular tools in these domains are Monocle and Slingshot both of which leverage on several unsupervised learning and dimension reduction algorithms. Instead of discrete characterization of cells, the space of cells can be realized as a continuum via interesting concepts, such as cell-trajectory prediction utilized by tools like sc-Velo. These types of analyses can be used to explain differentiation directions and the specific velocity of individual cells.

We have recently developed an AI-based tool, namely single-cell Synthetic Oversampling (scSynO), that uses gene expression counts of already identified rare cells as an input to generate synthetic cells. Afterwards, the newly generated cells are used to identify similar (rare) cells of the same kind in other publicly available experiments. For this reason, we applied the Localized Random Affine

Oversampling (LoRAS) algorithm to generate synthetic samples from rare-cell populations [5]. Training an AI-based classifier on such data enhances its likelihood for detecting these rare-cell types from a vast distribution of cells. scSynO can be integrated with existing workflows for further downstream analysis.

AI ALLOWS FOR SINGLE-CELL RESOLUTION OF SPATIAL RNA-SEQ DATA

The spatial RNA-Seq technology joins the two worlds of sequencing data analysis with image processing, which greatly expanded the knowledge of native multicellular biological systems. While the field is moving forward at a rapid pace, there are still multiple challenges, including sensitivity, labor extensiveness, tissue-type dependence, and limited capacity to obtain detailed single-cell information [6]. Some limitations might be circumvented by integrating single-cell RNA-Seq data but the technical limitation to achieve single-cell resolution might only be over-



FIGURE 1: Overview of current RNA-Seq technologies that await developments for AI assistance. Images were taken from Pixabay (Acknowledgements to Rubén Calvo, ElasticComputeFarm, and Devon Breen).

come by computational approaches. Current AI-based methodologies, such as XFuse, utilize a deep generative model to spatial expression data. The underlying model fuses low-sensitivity, low-resolution expression data with high-resolution histological image data to infer denoised full-transcriptome spatial gene expression at the same resolution as the image data (Figure 2).

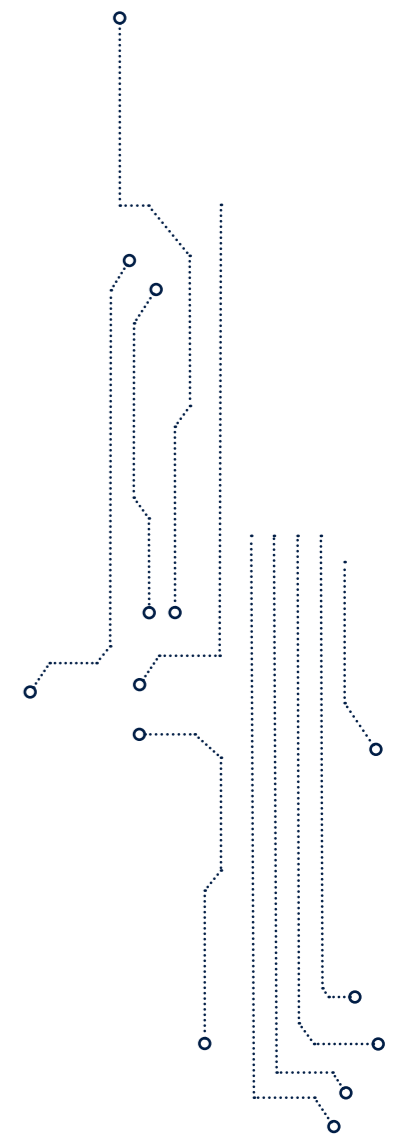
Additional software tools, such as STUtility, can add multiple features for spatial analysis, image processing, and visualization. Thus, aligned images can be stacked to create a turntable 3D model of the tissue, which e.g., can be used to visualize gradually shifting changes in gene expression or allows for an actual 3D tissue reconstruction.

However, taking into consideration the high costs per experiment, a first step for a broader applicability and higher awareness would encompass large collaborative efforts to create comprehensive and publicly available cell atlases. Based upon

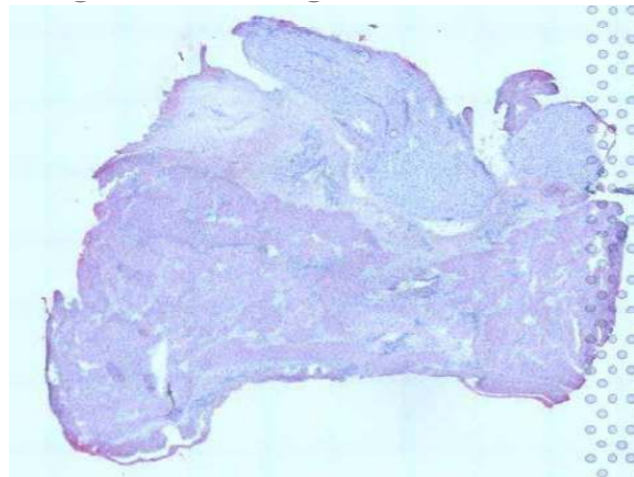
these new data analysis tools and benchmarks can be provided to set standardized procedures for this new technology.

POTENTIAL USE OF AI-BASED EXPLAINABILITY ALGORITHMS

AI algorithms, especially ones utilizing Deep Learning (DL), are often described as 'Black Box' models. They are far too complex to simply grasp their inner workings and mechanisms behind the decision-making. Luckily, algorithms are being developed to shed light into these models and help the user to understand their reasoning. Methods like Grad-CAM and Integrated Gradients are able to highlight the important parts of an input according to the model they are used on. We applied this on a network, which was trained to distinguish between maturity states of differentiating cardiomyocytes based on fluorescence stained images. These algorithms can be applied on transcriptomics data sets and might be able to extract important gene combinations in the future [7].



Bright-field image



Enriched with single-cell data



Clustered by Seurat



Single-cell resolution via XFuse

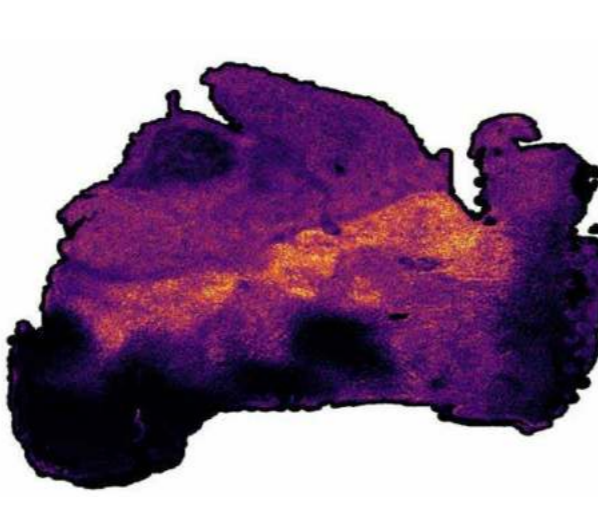


FIGURE 2: Use of AI-based algorithms for spatial RNA-Seq data to integrate, cluster, and enhance the interpretation. Data was generated by the ESF funded iRhythmic project (<https://irhythmic.med.uni-rostock.de/>).



CONCLUSION & OUTLOOK

Taken together, AI fosters explaining complex RNA regulation and is highly utilized in novel technologies, e.g., single cell and spatial transcriptomics, in which the data structure and general amount is even more complex and larger in size than in traditional bulk sequencing. AI is an inevitable tool for the analysis of these data types.

REFERENCES: [1] Wolfien M et al. 2020 EBioMedicine. 57 102862. DOI:10.1016/j.ebiom.2020.102862. [2] Wenric S et al. 2018 Front. Genet. 9 297. DOI:10.3389/fgene.2018.00297. [3] Wolfien, M et al. 2021 Bioinformatics, pp. 19–35. DOI:10.36255/exonpublications.bioinformatics.2021.ch2. [4] Teschendorff AE and Feinberg AP 2021 Nat. Rev. Genet. 1–18. DOI:10.1038/s41576-021-00341-z. [5] Bej S et al. 2021 Mach. Learn. 110 279–301. DOI:10.1007/s10994-020-05913-4. [6] Asp M et al. 2020 BioEssays 1900221. doi:10.1002/bies.201900221. [7] Yap, M. et al. 2021 Sci. Rep. 11 2641. DOI:10.1038/s41598-021-81773-9.

AUTHORS: Maximilian Hillemanns¹, Saptarshi Bej¹, David Brauer¹, Markus Wolfien¹, and Olaf Wolkenhauer^{1,2,3}

¹ Department of Systems Biology and Bioinformatics, Universitätsplatz 1, University of Rostock, Rostock,

² Stellenbosch Institute of Advanced Study, Wallenberg Research Centre, Stellenbosch University, Stellenbosch (South Africa)

³ Leibniz-Institute for Food Systems Biology, Technical University of Munich, Lise-Meitner-Straße 34, Freising

BEHIND THE SCENES – AI AS AN ENABLER OF SCIENTIFIC DISCOVERY IN THE LIFE SCIENCES

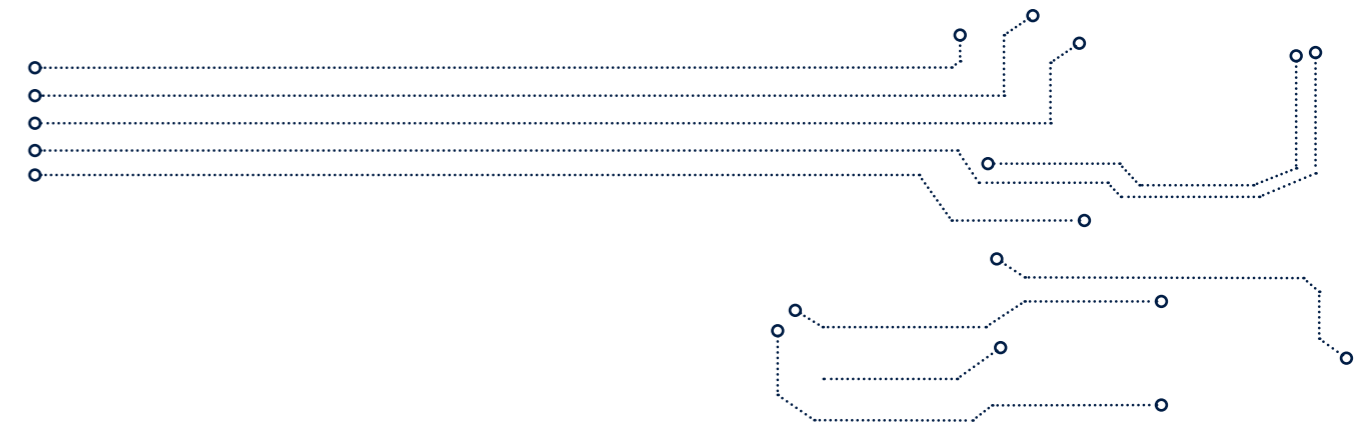
The development and application of the methods of AI to decision support and knowledge acquisition has been increased just recently. The aim is to go from data to insight faster and more efficiently. These efforts are indispensable, for example, to improve the evaluation of complex data sets, thus contributing to a better but also more complete picture of a disease or a certain condition.



FROM GENOMES TO PHENOTYPES

How AI helps mobilizing and analyzing big data and predicting properties for the many uncultured bacteria

In bacterial research, there is a great imbalance of available information per species. While a few species are well-studied, only sequence information is at best available for the vast majority of bacteria. We are applying AI-driven approaches to reduce this gap. The starting point for this approach is the database *BacDive*, which is the largest source for standardized phenotypic data on bacteria. The data basis is further extended with AI-assisted text mining to extract phenotypic data from the literature and extract genomic phenotypes. This builds the starting point for AI-driven analysis to predict the physiology and the appropriate cultivation conditions for not yet cultured bacteria.



There is a great imbalance in microbiology regarding the abundance of data for different species. Bacteria can be roughly divided into three groups, (1) a few well-studied model organisms, (2) thousands of species that can be cultured but are not yet examined in detail, and (3) presumably millions of organisms that appear in sequencing datasets at best and that are neither cultivated nor studied. The collaborative project DiASpora addresses the latter two cases by extracting and integrating available information from the literature and drawing conclusions from well-studied organisms. This is done by applying artificial intelligence (AI) techniques to comprehensive datasets from *BacDive*. Researchers from DSMZ, ZB MED and TIB are sharing their expertise in this joint project.

The Bacterial Diversity Metadatabase *BacDive* is the largest database for standardized phenotypic information. *BacDive* comprises data for over 14,000 bacterial species and over 80,000 strains. The data cover over 600 different data fields including taxonomy, morphology, physiology, and cultivation conditions. Due to the standardized data fields *BacDive* enables systematic analysis, like comparisons over a wide range of bac-

terial species, as well as finding strains based on certain attributes. Still, the coverage of the database shows gaps that need to be filled to allow comprehensive analyses over all known species.

WHERE THE DATA HIDE: USING AI TO SUPPORT TEXT MINING OF PUBLISHED KNOWLEDGE

The process of describing organisms is a basic but essential component for studying diversity, taxonomy and evolution in biological sciences [1]. Scientists have amassed huge amounts of taxonomic literature over centuries that provides comprehensive phenotypic information for each species hidden in research papers.

BacDive integrates systematically extracted data and already provides access to standardized data for over 6500 species descriptions of bacteria and archaea. So far, extracting the data and transforming them into standardized *BacDive* datasets needs a significant amount of manual work. Data in text format is mostly unstructured and the natural language used is highly variable.

To mobilize these hidden data from published literature, we apply a combination

of rule-based and AI-based models for information extraction. We use classical Natural Language Processing (NLP) approaches including Part-of-Speech (POS) tagging for entity recognition and build relationships between various entities and other parts of sentences. So far, we have used a comprehensive list of keywords from the *BacDive* database to build classifiers to identify relevant sentences. We are currently using a rule-based method to extract phenotypic characters from text using syntactic patterns. In parallel, we are also working towards an unsupervised information extraction system using word embedding and deep learning approaches including Long Short-Term memory (LSTM) models.

The aim is to work towards an automated approach that is robust by including feedback cycles with domain experts to train the AI. In this way extracting information can be accelerated significantly and performed with a high degree of confidence.

JUMPING TO CONCLUSIONS: USING AI TO INFER PHENOTYPIC INFORMATION FROM GENOMES

In contrast to phenotypic data that are sophisticated and hard to determine, genome sequences are becoming in-

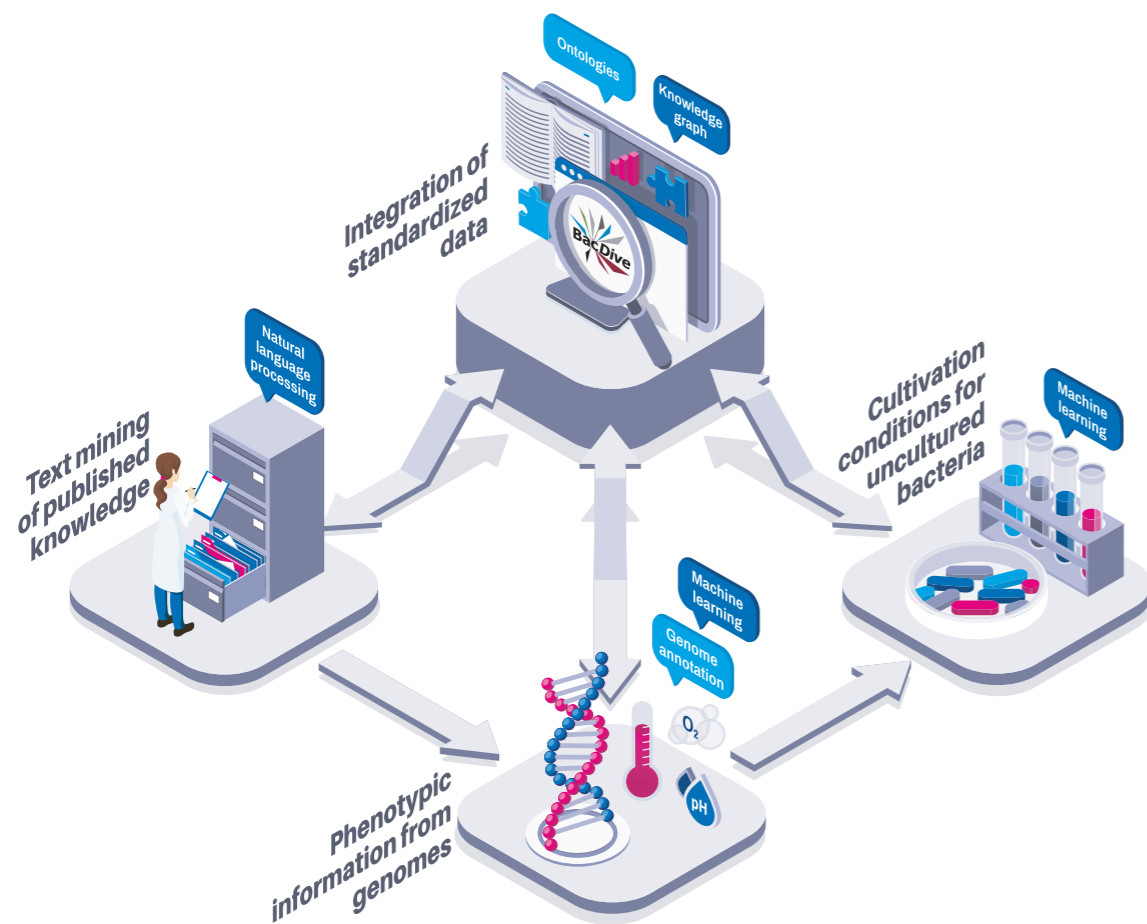


FIGURE 1: The different aspects of the DiASPora project. Platforms represent the major working packages and speech bubbles represent the main techniques used.

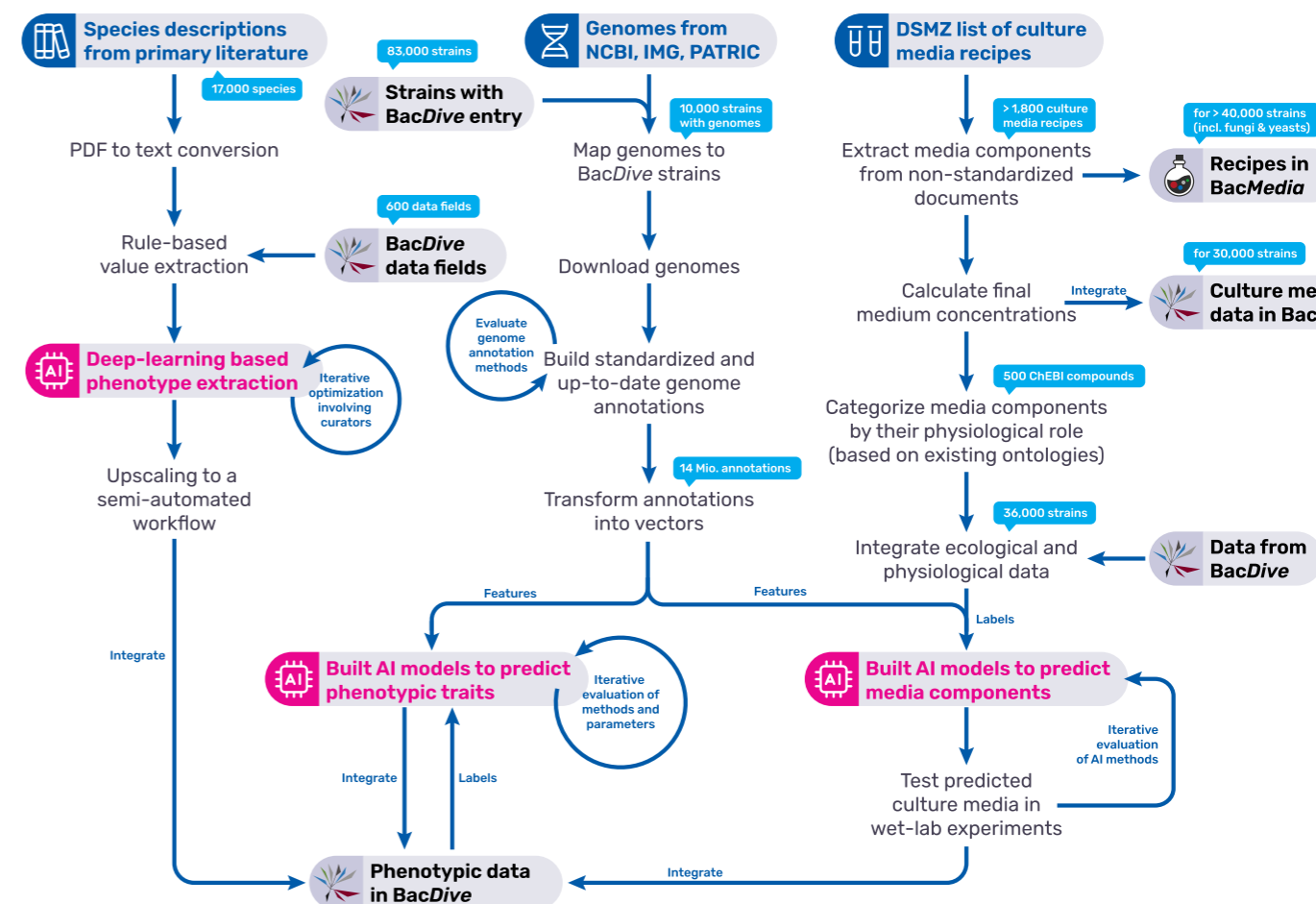


FIGURE 2: Schematic workflow of the DiASPora project. External data resources, BacDive data and AI methods are highlighted in dark blue, black and pink, respectively.

creasingly available for a large number of bacterial species [2], providing the opportunity to infer phenotypic information from genome sequences. Previous approaches already demonstrated the possibilities of this procedure, although they were limited regarding their training data. Traitair for example was able to predict a set of 67 different traits with up to 73% accuracy, although it was trained on only 234 bacterial species and the algorithm

discarded strain specific phenotypes [3]. The PICA algorithm was used to build the PhenDB database [4] that currently holds 39 phenotype models. The models were trained on up to 427 strains and have an accuracy between 63% (psychrophilic) and 99% (obligate intracellular).

We extend these approaches by using much larger sets of manually curated phenotypic data from BacDive. To this

end, we linked genome information from other databases to BacDive. An extensive data mapping approach resulted in almost 10,000 strains that have sequenced genomes available. We performed state-of-the-art genome annotations using various methods. After thoroughly reviewing the data, we used Pfam classes for training our AI models.

The goal is to enrich phenotypic data of these 10,000 strains in BacDive by an AI-guided approach. The following six phenotypic categories were best suited for a proof-of-concept: growth temperature, salinity, gram staining, oxygen requirement, motility, spore formation. We trained Support Vector Machines (SVM) and Random Forest (RF) Classifiers on one-hot encoded Pfam annotations to predict phenotypic traits. Besides the

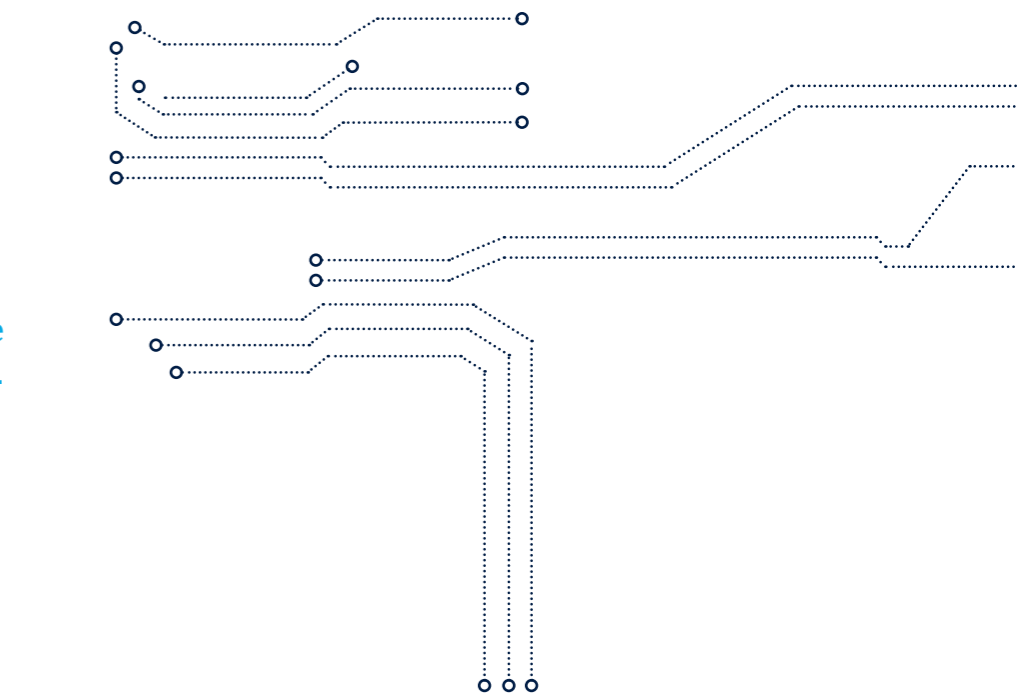
great potential of this approach, it is limited by the availability and distribution of data. While we are improving the availability of data in the first part of the project, uneven or biased data may be challenging. However, first results showed an accuracy between 78 and 97%, depending on the trait category. These preliminary results highlight the importance of data quality and quantity for successful AI application.

CULTIVATING THE UNKNOWN: USING AI TO PREDICT CULTIVATION CONDITIONS FOR SO FAR UNCULTURED BACTERIA

The enhanced data basis in BacDive is subsequently used to predict cultivation conditions for so-far-uncultivated bacteria. This last part of the project combines datasets from the DSMZ list of media, the previously applied AI tech-



The newly created culture media database *BacMedia* offers instructions to cultivate over 40,000 microbial strains.



CONCLUSION & OUTLOOK:
Standardized and machine-interpretable data ready to explore

The DiASPora project aims for enhancing biodiversity information on poorly studied organisms. Efforts in text mining allow the extraction of published knowledge while AI-assisted prediction of phenotypic traits extrapolate this knowledge. This huge increase

in knowledge will improve our understanding of the poorly studied majority of species in bacterial research. Transforming all data into a machine-interpretable knowledge graph will allow innovative search options for the discovery of hidden data relationships. All this data will extend the *BacDive* database and will therefore be accessible by the microbial research community according to the FAIR principles.

niques and the predicted phenotypic data. In this way, we extend the efforts of the KOMODO approach [5] that solely used taxonomic relationships to predict media for uncultivated bacteria.

Since the KOMODO database is outdated, we decided to build a new database. Therefore, over 1,800 culture media had to be extracted from non-standardized documents and transferred into a machine-interpretable data format. We transformed the data into a relational

database and implemented a user interface for this newly created culture media database. The database is freely available under the name *BacMedia* (bacmedia.dsmz.de) and offers instructions to cultivate over 40,000 microbial strains. The data were enhanced using phenotypic information from *BacDive* and the previous studies.

First AI models demonstrated the feasibility of this approach: as an example, the supplementation of biotin could be pre-

dicted with an accuracy of almost 90 %. This approach has yet to be applied to a larger amount of molecular components, as well as tested in the laboratory. However, this approach is also limited by the amount of available data. For instance, the majority of media are complex media, meaning their exact composition can hardly be determined. We address this by integrating ecological and physiological data in our analyses.

REFERENCES: [1] Overmann J 2013 *The Prokaryotes*, 4th edition, Prokaryotic Biology and Symbiotic Associations. 149-207. DOI: 10.1007/978-3-642-30194-0_7. [2] Mukherjee S et al. 2017 *Nat Biotechnol.*;35(7):676-683. DOI:10.1038/nbt.3886. [3] Weimann A et al. 2016 *mSystems*; 27;1(6):e00101-16. DOI: 10.1128/mSystems.00101-16. [4] Feldbauer R et al. 2015 *BMC Bioinformatics*. 16 Suppl 14(Suppl 14):S1. DOI: 10.1186/1471-2105-16-S14-S1. [5] Oberhardt MA et al. 2015 *Nat Commun*. 13;6:8493. DOI: 10.1038/ncomms9493.

AUTHORS: Julia Koblit¹, Arindam Halder², Lorenz C. Reimer¹, Konrad U. Förstner² and Jörg Overmann¹
¹ Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures GmbH, Inhoffenstraße 7b, 38124 Braunschweig
² ZB MED - Information Centre for Life Sciences, Gleueler Straße 60, 50931 Cologne

TOOL RECOMMENDER SYSTEM IN GALAXY USING DEEP LEARNING

Researchers in Galaxy, a web-based and open-source platform for scientific data-processing, create data analysis pipelines known as workflows. They can be complicated to create, especially for new researchers, from thousands of accessible tools in Galaxy. In order to assist them in forming workflows, a recommender system is devised which suggests tools at each step of forming a workflow. These workflows consist of sequences of tools and their higher-order dependencies are learned by a variant of a recurrent neural network. A Galaxy API is created to fetch recommendations from the model created by the neural network and they are shown using two user interface integrations in the European Galaxy Sever. An accuracy of 98 % is achieved when using this recommender system for the top-1 metric.





seq, variant-calling, Hi-C, assembly, single-cell, proteomics, and many others. Workflows, created by researchers in Galaxy, are decomposed into multiple tool sequences (Figure 1). Tools are connected one after another in these tool sequences and have similar nature as other sequential data such as text and speech. There are multiple studies in the fields of natural language processing, and speech recognition that apply deep learning techniques on sequential data to obtain good accuracy in predicting future items. Therefore, in our work, a variant of recurrent neural network (RNN)—gated recurrent units (GRU)—is used to create the tool recommender system in Galaxy.

RESULTS

Three different neural network architectures — dense neural network (DNN), convolutional neural network (CNN), and gated recurrent units neural network (GRU) — are compared on their performances in recommending tools (Figure 2). The models, obtained after training these three architectures, are used to predict tools for the tool sequences in the test data after every training iteration. Top-k precision (precision@k) is a popular metric for evaluating a recommender system. Precision@k implies how many of the k predicted tools are compatible. For example, k = 2 implies that there are 2 predicted tools with the highest predicted scores. If only 1 of them is correct, then the precision@2 is $1/2 = 0.5$. Precision@1 and precision@2 metrics are used in this approach to evaluate the quality of the tool recommender system. Overall, the GRU neural network shows a superior performance compared to DNN and CNN by achieving 98% top-1 precision.

To ensure a reproducible data analysis, many workflow systems such as Bcbio-nextgen, Omics Pipe, and many others have been developed [1, 2]. A workflow is a stepwise data processing pipeline, for example, quality control, preprocessing, quantification, and statistical analysis. These steps together transform any raw data into meaningful scientific outcomes. A workflow represents one unit of software which can be shared, saved, and reused enabling reproducible research. But, creating a meaningful workflow is a difficult task. An important question would be how to ascertain that a given workflow generates a valid output. Therefore, it becomes necessary to use correct tools at each step of creating a workflow.

create meaningful workflows, a recommender system is created which has several benefits. First, it will make researchers more efficient by saving their time wasted in creating erroneous workflows. Second, it will help them avoid the step of searching for tools separately, which will shorten the time spent in creating workflows. Third, it will promote high-quality tools that have been used more often in the past to the top of the recommendations. Finally, it can be extended to promote the newly added tools in Galaxy by showing them alongside the recommended tools predicted using the neural network approach.

SEQUENTIAL LEARNING ON WORKFLOWS

More than 18,000 workflows have been collected from the European Galaxy Server to create the recommender system. These workflows come from different scientific analyses such as RNA-

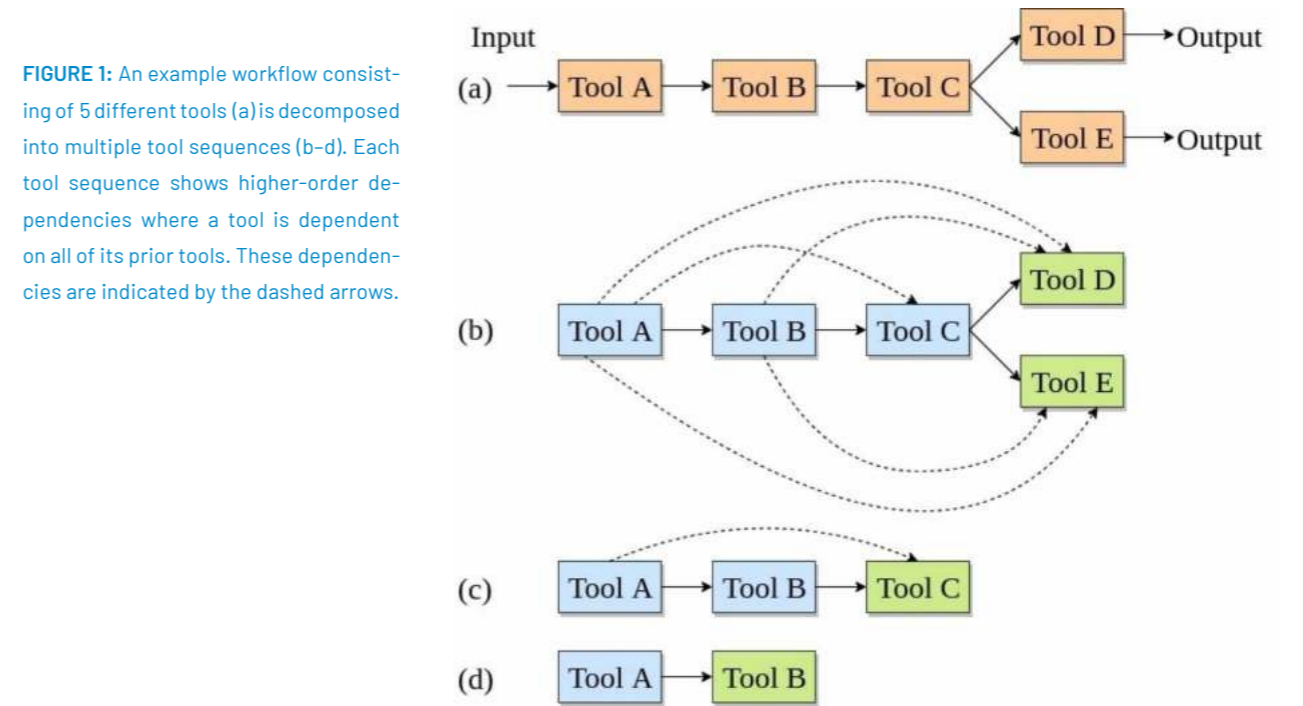


FIGURE 1: An example workflow consisting of 5 different tools (a) is decomposed into multiple tool sequences (b-d). Each tool sequence shows higher-order dependencies where a tool is dependent on all of its prior tools. These dependencies are indicated by the dashed arrows.

FIGURE 2: Top-k (precision@k) shared precision for DNN, CNN, and GRU neural networks with cross-entropy loss function in (a), (c), and (e), respectively. Topk (precision@k) shared precision for DNN, CNN, and GRU neural networks with weighted cross-entropy loss function in (b), (d), and (f), respectively.

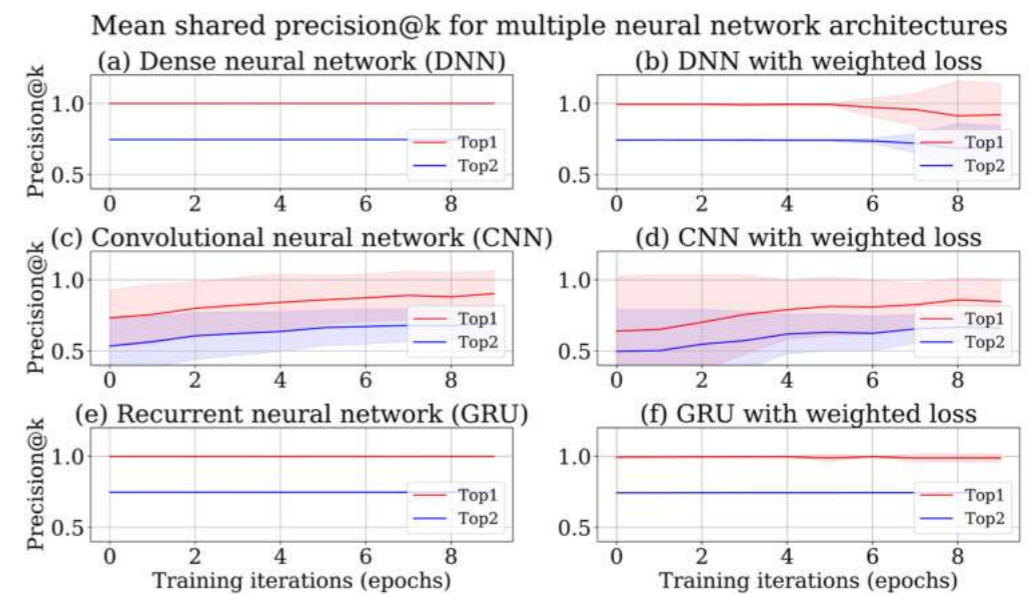
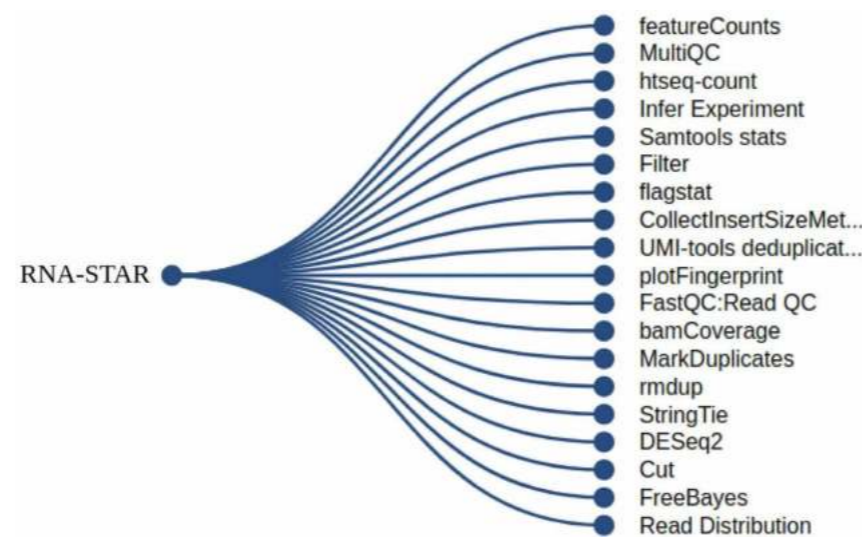


FIGURE 4: The figure shows recommended tools as leaves (on the right) of the tree after executing the RNA-STAR tool. Clicking on any recommended tool opens its definition in Galaxy and can be used for further analysis with the data files produced by the previous tool (RNA-STAR).



CONCLUSION & OUTLOOK

A recommender system to suggest tools in Galaxy is created by learning on workflows coming from different scientific analyses using a variant of RNN (GRU). The tools recommended by the system are highly relevant as confirmed by their similarities with the tools used in Galaxy Training Network [7] tutorials for multiple scientific analyses. These recommended tools can be easily accessed using simple user interface interactions in the European Galaxy Server (Figure 3 and 4). Together, they offer a good user experience for researchers to choose

high quality tools for their data analysis. The approach used to create the recommender system avoids storing any metadata of tools or workflows and uses only patterns of tool connections from workflows to suggest tools at each step of creating a workflow. The neural network creates the model to recommend tools. An API, residing with other Galaxy APIs, accesses this model and makes use of the input tool or a tool sequence provided by a researcher to recommend tools in real time. It is expected that this system is extremely useful for researchers

new to Galaxy who are not aware of all the tools in Galaxy. This system shows them only a handful of tools from a large collection (>3000) of tools which helps in exploratory data analysis.

Many tools come with annotations which can be used to improve the recommendations by adding more importance to those which carry annotations in comparison to those which do not in the list of recommendations. Tools containing similar annotations may have similar functionalities, and using these similarities, recommendations can be further improved by showing similar tools for each recommended tool. On any Galaxy server, tools and workflows are created and updated regularly. Therefore, it becomes necessary to learn the latest tools and workflows using the GRU neural network. This task should be periodic to keep the tool recommendation model updated with the latest tools and workflows. Galaxy administrators can overwrite the recommended tools predicted using the trained model by a different set of tools using the parameters specified in Galaxy configuration.

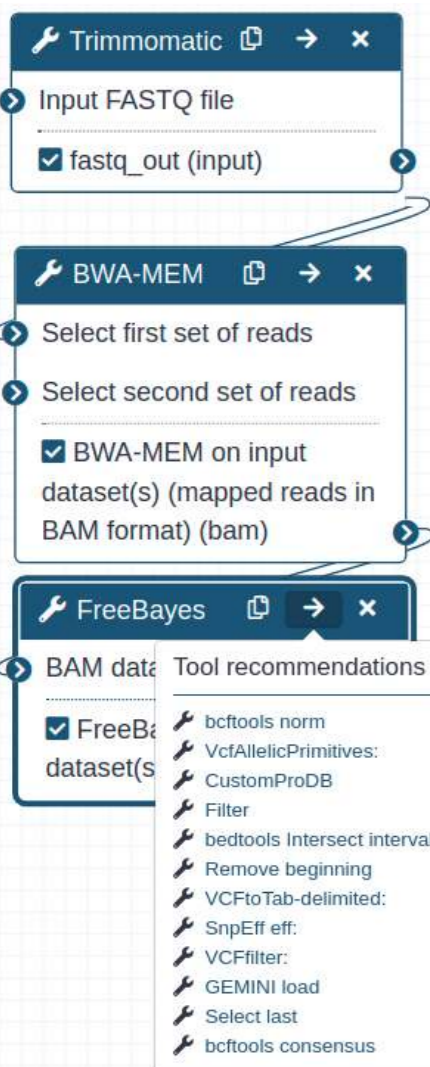
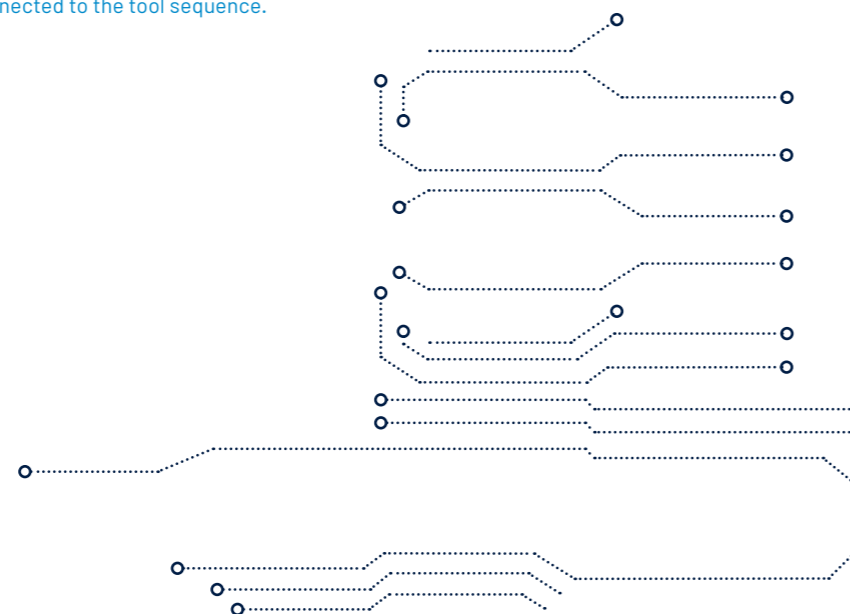


FIGURE 3: Recommended tools, listed in the 'Tool recommendations' dropdown, in the workflow editor of the European Galaxy server for the Trimmomatic > BWA-MEM > FreeBayes tool sequence. The recommended tools for the tool sequence can be seen in a dropdown while hovering on the right arrow button visible in the top right corner of the 'FreeBayes' tool. Clicking on any recommended tool such as 'bcftools norm' in the dropdown opens a new block for the chosen tool that can be connected to the tool sequence.

ILLUSTRATIONS

Two examples are provided to show the real-time use of the recommender system in the European Galaxy Server. The first shows recommended tools for a tool sequence with 3 tools, Trimmomatic [3] > BWA-MEM [4] > FreeBayes [5], in the workflow editor (Figure 3). Another example of tool recommendations after using RNA-STAR is shown in Figure 4. It shows follow-up tools such as bamCoverage [6], MultiQC, and a few others.



ACKNOWLEDGEMENT

Tool recommender system in Galaxy using deep learning project has been supported by the German Research Foundation (DFG) under Germany's Excellence Strategy (CIBSS - EXC-2189 - Project ID 390939984)

and German Federal Ministry of Education and Research (BMBF grant 031A538A de.NBI). The article processing charge was funded by the University of Freiburg in the funding programme Open Access Publishing.

Links:
 ORCID <http://orcid.org/0000-0002-2068-4695>
 Code repository https://github.com/anupruezh/galaxy_tool_recommendation
 Model https://github.com/galaxyproject/galaxy-test-data/blob/master/tool_recommendation_model.hdf5
 API: https://github.com/usegalaxy-eu/galaxy/blob/release_20.05_europe/lib/galaxy/webapps/galaxy/api/workflows.py#L638
 GigaDB <http://dx.doi.org/10.5524/100838>
 RRID https://scicrunch.org/resolver/RRID:SCR_018491
 Biotools ID [tool_recommendation_system_in_galaxy](https://biotools.org/tool_recommendation_system_in_galaxy)
 European Galaxy Server: <https://usegalaxy.eu>

REFERENCES: [1] Ewels P et al. 2017 F1000Res 5:2824. DOI: 10.12688/f1000research.10335.2. [2] Leipzig J 2017 Brief Bioinform;18(3):530-6. DOI: 10.1093/bib/bbw020. [3] Bolger AM et al. 2014 Bioinformatics 30:2114-20. DOI: 10.1093/bioinformatics/btu170. [4] Li H 2013, arXiv:1303.3997. DOI:10.6084/M9.FIGSHARE.963153.V1. [5] Garrison E and Marth G. 2012, arXiv:1207.3907. [6] Ramirez F et al. 2016 Nucleic Acids Res;44(W1):W160-5. DOI: 10.1093/nar/gkw257. [7] Batut B et al. 2018 Cell Syst;6:752-8. DOI: 10.1016/j.cels.2018.05.012.

AUTHORS: Anup Kumar¹, Helena Rasche¹, Bjoern Gruening¹ and Rolf Backofen^{1,2}
¹ Bioinformatics Group, Department of Computer Science, University of Freiburg, Georges-Koehler-Allee 106, 79110 Freiburg
² Signalling Research Centres BIOSS and CIBSS, University of Freiburg, SchaezenlestraÙe 18, 79104 Freiburg

TOWARDS SMART WAYS TO HELP DATA CURATION

Natural language processing for the life sciences

Anecdotically, most discussions about data center around 'big data', datasets that are huge in size and huge in the number of dimensions. The FAIR data movement stresses that data needs to fulfill certain requirements in order to be useful in the long run. Rendering data FAIR is a challenge that is beyond most experimentalists, creating the need for data stewards and curators. Also, condensed insight still is mainly to be found in scientific papers. For some domains, there are scientific databases that further condense insight from papers into standardized datasets, again using the services of data stewards and curators. In any case, there is a need for humans to look at data in order to provide quality control. And while in some domains, like e.g. object recognition in street photos nearly every human in the world can provide simple answers (we all know the 'Please mark all traffic lights in this image!' captchas), finding human 'ground truth' becomes much harder when looking at the results of biochemical, biological, and biomedical experimentation. The challenge of finding ground truth is not limited to builders of research data bases, but they are shared by anyone who wants to enhance the value of existing, specialized data. Within this article we describe two ways of making use of data in order to help data curation, ChemHITS, and the DeepCurate project.

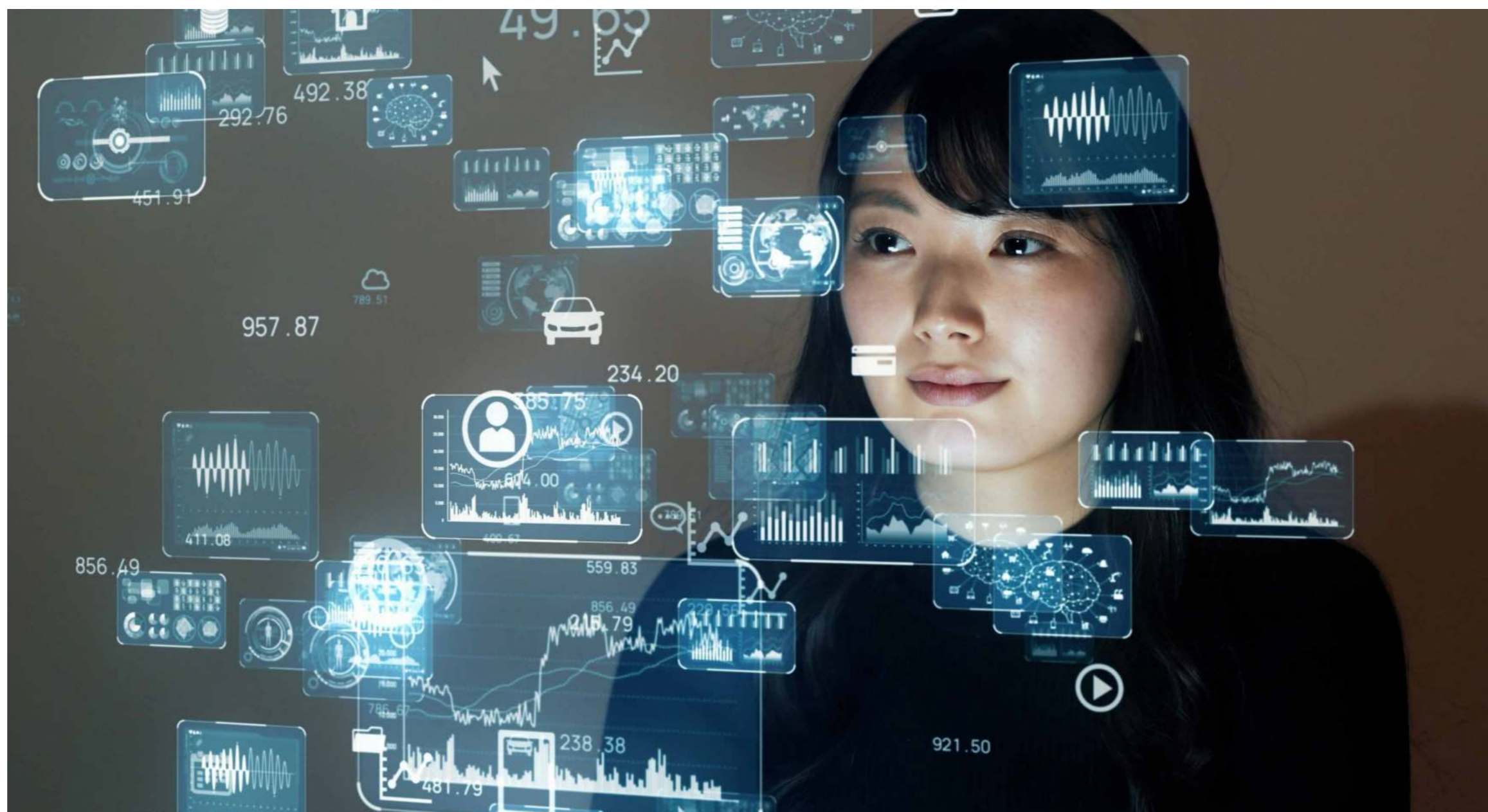
Sometimes it is said that 'data is the new gold' or the 'new oil' This underlines the value of data, especially in the context of training smart systems. As training data is so valuable, it must be put to use in a way that minimizes the need for human input and makes the maximum use of human input. In the following, we outline two methods that use human-curated public biological databases in order to either directly

use them for a new purpose (ChemHITS) or in order to repurpose them as training data for deep learning in neural networks. In both cases it is our goal to bridge the gap between unstructured textual data and structured, ontology-annotated graph and tabular data.

An ontology is a formal representation of knowledge. Ontologies are used to disambiguate concepts. An ontology

identifier can resolve if 'plasma' pertains to 'blood plasma' or 'cytoplasma', for example.

ChemHITS (next section) is rule-based and uses a dictionary of compounds to map compound names to ontology terms. DeepCurate (following section) researches the use of database data as training data for deep learning networks.



ChemHITS

Chemical compounds can be found having many different names – trivial, as well as systematic names. Hence, the unambiguous identification of a chemical compound solely based on its name requires comprehensive chemical knowledge and often searches in chemical databases. As many publications exclusively describe a chemical

compound by its name (as opposed to name and standardized identifier or name and structure formula) the matching of these diverging notations can be tedious. However, this identification is crucial for the integration of biochemical data in databases or for the setup of biochemical models based on published data found in the literature.

We have developed ChemHits, an application which detects and matches synonymic names of chemical compounds and thereby facilitates merging of corresponding data referring to the same compound, but described with different names. It applies transformation rules to systematically process chemical compound names to a unique generic normalized name form. It is capable of normalizing a given name of a chemical compound and matching it against names in (bio-) chemical databases, like KEGG COMPOUND, SABIO-RK or ChEBI, even when there is no exact name-to-name match. The tool is also able to match a complete list of compound names against these databases which makes it useful for the automatic cross-annotation of chemical data in databases.

The key method driving ChemHITS is to become invariant to compound name modifications that do not change the meaning of the compound name. For example, parts of the compound name can differ by replacing a systematic chemical description by a trivial name. For example 2-propylpentanoic acid could become valproic acid. Similarly, there can be dissent about the lead structure: acetylphenol or phenylacetate? These and similar modifications are covered by ChemHITS based on rules and dictionaries that are employed to map compound names onto a normalized form. Two different compound names with the same normalized form are very likely to have the same chemical meaning.

General information							
Organism	Phlebiopsis gigantea						
Strain	DSM 13218						
Tissue	mycelium						
EC Class	1.1.3.10						
SABIO reaction id	7933						
Variant	mutant P2OxB1H (E540K-His6)						
Recombinant	expressed in Escherichia coli BL21(DE3)						
Experiment Type	in vitro						
Event Description	-						
Substrates							
name	location	comment					
beta-Methyl-D-glucopyranoside	-	-					
Q2	-	-					
Products							
name	location	comment					
H2O2	-	-					
2-Dehydro-beta-D-methylglucoside	-	-					
Modifiers							
name	location	effect	comment	protein complex			
pyranose oxidase(Enzyme)	-	Modifier-Catalyst	-	(Q6UG02)*4;			
Enzyme (protein data)							
	UniProtKB_AC	name	mol. weight (kDa)	deviation (kDa)			
subunit	Q6UG02	-	66.7	-			
complex	-	-	266.8	-			
Kinetic Law							
type	formula		annotation				
-	-		-				
Parameter							
name	type	species	start val.	end val.	deviat.	unit	comment
kcat_Km	kcat/Km	beta-Methyl-D-glucopyranoside	13.0	-	-	M ⁻¹ s ⁻¹	-
Km	Km	beta-Methyl-D-glucopyranoside	261.0	-	12.0	mM	-
Vmax	Vmax	-	0.8	-	-	μmol/(min*mg)	-
Experimental conditions							
	start value	end value	unit				
pH	6.5	-	-				
temperature	30.0	-	°C				
buffer	100 mM potassium phosphate, 1 mM ABTS, 2 U/ml peroxidase, 5 -20 mU/ml pyranose oxidase						
comment	-						
Reference							
title	author	year	journal	volume	pages	PubMed	
Engineering of pyranose 2-oxidase from Peniophora gigantea towards improved thermostability and catalytic efficiency	Bastian S, Rekowski MJ, Witte K, Heckmann-Pohl DM, Giffhorn F	2005	Appl Microbiol Biotechnol	67	654-63	15660220	

FIGURE 1: SABIO-RK provides structured data to its users. The data shown was assembled from a variety of locations in a scientific paper.

DeepCurate

While the ChemHITS project takes a rule- and dictionary-based approach, the DeepCurate project seeks to merge different knowledge sources in order to produce deep-learning based improved Named Entity Recognition in texts with the ultimate goal of improving data extraction and curation in full paper texts. DeepCurate is a collaborative project between the NLP and SDBV groups at HITS and is funded by BMBF for three years that started in January 2020.

The motivating setting of DeepCurate is the SABIO-RK curation process. SABIO-RK is our database currently offered in de.NBI. We use SABIO-RK both as target for improved curation support and as a tool to provide this. The key idea is here: *Why not use the existing SABIO-RK data to learn more about its future data?*

For SABIO-RK, human data extractors read papers, extract data, enter it into a data input interface. Here, human curators read the data entered, verify the data in the paper, assign entities (reac-

tants, enzymes, tissues) and reactions to ontologies, making the data ready for publication in the SABIO-RK database. This two-step manual curation pipeline is the key to SABIO-RK's data quality. And, along with the SABIO-RK data itself, SABIO-RK curators meticulously gathered the paper trail leading up to the curation results.

To our knowledge, many curators still prefer using paper in the process. This is preferred, as it reduces screen time and gives more liberty in terms of posture.

DATABASE-TO-DOCUMENT BACK-PROJECTION: FROM DATABASE TO TRAINING DATA

One key challenge in performing curation is the recognition of named entities. Named entities are parts of speech that are not normal nouns (e.g. reactant) but rather names (e.g. oxygen). Obviously, for recognizing a reaction you first have to find which reactants are involved.

Named entity recognition becomes challenging because context matters. Wikipedia lists more than 20 meanings for the three letter acronym ATP which designates *adenosine triphosphate* in biological context. While most of these meanings are non-biological, the word *plasma*, for example, can designate *blood plasma* or *cytoplasm*, the plasma inside the cell. Context helps readers in making the difference between the two.

Database-to-document backprojection is about putting the database content back into its original context.

Databases extract data from its original context and then put it into a new, table based context. A typical database entry standardizes data, puts semantically same data in the same location in the same form (see Figure 1). The original context gets lost in the process, it is at least partly remaining in the paper trail, but not electronically readable.

However, deep learning training data is about *presenting the context* to deep neural networks with the goal of enabling them to learn this context and subsequently reduce the number of errors when performing named entity recognition.

With backprojection, DeepCurate puts database content into its original context, thus creating training data for deep learning. Within the project we have used this for SABIO-RK data, however the methods are largely database agnostic.

At the same time, the backprojection has another useful purpose, this is the purpose of quality control. It facilitates context checking by the curators. Experiments that show the superiority of the derived training data are still ongoing.

MANUAL ANNOTATION EXTRACTION: SEEING WHAT MATTERS TO HUMAN CURATORS

When curating papers, SABIO-RK collaborators print the paper and then *mark* it using marker pens. They then archive the marked printout. As part of the DeepCurate project, we integrate this information into the training data. We scanned all printouts of papers contained in SABIO-RK including their markings and recognized the markings within the papers.

We will use these data as training data, and at the same time, these data shows the use of markings. The *content* ending up in the database often is not marked. However, the *context*, parts that are important for the understanding of the paper is marked by the curators. We are looking forward to using this information for training, learning and named entity recognition.



CONCLUSION & OUTLOOK

Within this article we have described efforts to enrich context-less database data together with curation traces to context-rich training data for the use in deep learning methods. The methods are generally applicable, collaborations are welcome. Stay tuned for upcoming results. Please briefly state the most important key data of your research project (name of the project, supporting institution(s), partners involved, etc.)

REFERENCES: [1] M-C. Müller et al., 2020 Proceedings of the First Workshop on Scholarly Document Processing, DOI: 10.18653/v1/2020.sdp-1.9.

AUTHORS: Wolfgang Müller for the DeepCurate and ChemHITS teams¹

¹Scientific Databases and Visualization Group, Heidelberg Institute for Theoretical Studies (HITS gGmbH), Schloss-Wolfsbrunnengasse 35, 69118 Heidelberg

IMPROVING THE SEARCH FOR NEEDLES IN A HAYSTACK

Classifier-independent oversampling for imbalanced data

Many real world problems are a search for the needle in a haystack - we are trying to identify the rare, exceptional event, which may correspond to a disease in medicine or fraud in a business context. Learning from such data creates imbalanced data sets - some of the classes that we wish to identify have only a relatively small number of datapoints compared to the other classes. Imbalanced data are also common in numerous research domains including biomedicine and bioinformatics. Using oversampling with classical machine learning algorithms is a popular approach to handle classification tasks for tabular imbalanced datasets. However, due to the existence of numerous oversampling algorithms and multiple classification models, towards only a given single dataset, it is difficult to properly decide on a classifier and oversampling algorithm. We developed a multi-schematic oversampling approach that produces high classification performance, irrespective of the classifier used, by utilizing rigorous modelling of the convex data space.

IMBALANCED DATASETS AND OVERSAMPLING

Imbalanced classification problems arising in research domains including biomedicine and bioinformatics, are characterized by an unequal distribution of data instances over different classes. Those with more instances are called majority classes, while classes with fewer instances are called minority classes. A common approach is to re-balance training data through oversampling, generating synthetic examples for the minority class.

Along with several approaches, such as cost-sensitive learning, undersampling, generative networks etc., oversampling is a popular approach to improve imbalanced classification. Oversampling approaches generate synthetic minority class instances to balance the minority and majority class and facilitate a balanced learning experience for the subsequently applied classification algorithms. In particular, when tabular imbalanced datasets have fewer instances, predictive models rely on classical machine learning algorithms. In contrast, deep learning models rely on a higher volume of training data.

Among numerous oversampling approaches, an observable trend is to generate synthetic samples from the convex space of the minority class. The SMOTE algorithm, developed in 2002, is the pioneer of such algorithms [1]. The algorithm creates a synthetic sample as a convex combination of two close-enough minority class data points. A major criticism of this approach is rooted on its tendency to over-generalize the minority class resulting in an improved classification of the minority class instances at the cost of a relatively high number of misclassifications of majority class instances.

Over the past two decades more than 85 extensions of the SMOTE algorithm have been developed to overcome this problem [2]. These extensions of SMOTE have implemented multiple strategies to solve the limitations of SMOTE. For example, the ADASYN algorithm uses a weighted distribution of minority class samples to decide upon minority class samples that are more important for synthetic sample generation [3]. Borderline-SMOTE detects borderline regions in the data and aims to generate synthetic samples from these borderline regions because such regions are important for classifiers to form a decision boundary [4]. Algorithms like CURE-SMOTE, ProWSYN, and MOT2LD use several clustering and dimension reduction strategies to model the latent minority class manifold more precisely [5,6,7]. However, till now there is no conclusive evidence that a single algorithm can be considered significantly better than all others leading to the introduction of more extensions of SMOTE with passing time [2]. With many available oversampling algorithms and classification models it is getting increasingly difficult for researchers to choose an appropriate oversampling algorithm and a well-suited classification model, given a single imbalanced dataset.

RIGOROUS MODELLING OF THE CONVEX SPACE

The Localized Random Affine Shadowsampling (LoRAS) is a recently proposed oversampling approach that relies on rigorous modelling of the convex space [8]. The rationale behind the approach rests on the assumption that considers the stochastically generated synthetic samples as random variables. This overcomes the problem of overgeneralization of the minority class by SMOTE, which is due to the inability to model the variance of the synthetic samples. Thus, the resulting synthetic samples interfere with the majority class samples and thereby, hamper classifiers to create an effective decision boundary during the training process.

The LoRAS algorithm first creates shadowsamples in a minority class data neighbourhood, which are Gaussian noise added to the minority class data points. Assuming that the synthetic samples are random variables following a t-distribution in a data neighbourhood, the variance of the synthetic samples is inversely proportional to the number of shadowsamples considered for a convex combination to generate the synthetic sample. Thus, instead of generating synthetic data instances with convex combinations of only two minority samples as done by the SMOTE algorithm, the LoRAS algorithm generates synthetic samples as random convex combinations of multiple shadowsamples. Moreover, the LoRAS algorithm also applies a manifold-learning step prior to the synthetic sample generation. For this step state-of-the-art manifold learning techniques, such as t-SNE and UMAP have been used. This enables the algorithm to detect data neighbourhoods that are consistent with the latent data manifold.

The algorithm was tested on several publicly available imbalanced datasets arising from a variety of research domains. The algorithm proved to be more effective as per popular performance measures, such as F1-Score and Balanced accuracy, compared to some popular oversampling algorithms for some classifiers. It was also observed, that as opposed to many other oversampling approaches, the LoRAS algorithm avoids the generation of synthetic samples near outliers and, thereby, has an inherent outlier detection mechanism. However, for some classifiers the improvement induced by LoRAS over other algorithms was not significant. Figure 1 demonstrates the geometric idea of controlling the variance of the synthetic samples of the minority class and provides a depiction of the LoRAS algorithm in comparison to SMOTE.

CLASSIFIER INDEPENDENT OVERSAMPLING

A recent comparative study of 85 oversampling algorithms shows that the classifier-specific effectiveness of oversampling algorithms persists for all methods [2].

To develop a classifier independent oversampling approach, integration of philosophies from multiple oversampling algorithms is necessary.

The idea of controlling the variance of the synthetic samples effectively for improved modelling of the convex space is thus extended to develop the Proximity Weighted Random Affine Shad-owsampling(ProWRAS)algorithm.

The ProWRAS algorithm uses an elaborate protocol to generate synthetic

samples. First, ProWRAS partitions the minority class data as per their proximity to the majority class. This is performed to detect regions in the minority class latent manifold that is relatively closer to the majority class. This kind of partitioning or clustering is used instead of a manifold learning step in LoRAS because it clusters the minority class data relative to the majority class. Once the clusters are decided ProWRAS decides upon the number of synthetic samples to be generated from each cluster such that more synthetic samples are generated from clusters closer to the majority class. Moreover, the algorithm ensures that the synthetic samples generated from the clusters near the majority class have relatively less variance. This prevents the synthetic samples from interfering with the latent majority class data manifold. The algorithm has four different variance schemes for generation of synthetic samples: High Global Variance (HGV), Low Global Variance (LGV), High Local Variance (HLV), and Low Local Variance (LLV). The global variance schemes consider entire clusters detected by the proximity based clustering as sampling neighbourhoods, while the local variance schemes draw synthetic samples from small neighbourhoods within the clusters. The high variance schemes generate synthetic samples using convex combinations of only two shadowsamples, while the low variance schemes generate synthetic samples using convex combinations of more than two synthetic samples for clusters that are relatively near to the majority class.

The algorithm has been tested on multiple publicly available datasets against state-of-the-art oversampling algorithms using four different classifiers. The results of the study show that given an imbalanced dataset and a corresponding classifier, a proper choice of a

specific oversampling scheme from the four proposed schemes can significantly improve the classification performance, irrespective of the classifier used.

POTENTIAL APPLICATION IN BIOMEDICINE AND BIOINFORMATICS

Algorithms such as LoRAS and ProWRAS are applicable to imbalanced datasets independent of the research domain. However, in biomedical research and bioinformatics, imbalanced classification problems are common. The research direction of personalized medicine is based on customization of treatment at a personal level. Imbalanced classification problems are significantly relevant in achieving this, since customization at a personal level can only be obtained by detecting patterns in smaller populations in contrast to relatively larger, heterogeneous populations. The benchmarking studies of the discussed algorithms also feature some datasets from biomedicine and bioinformatics, such as microcalcification, detection in mammograms, detection of thyroid patients, and protein localization prediction in yeast.

Another instance of a real world application of the LoRAS algorithm is automated annotation of rare-cell types from single-cell expression data using the Single-Cell Synthetic Oversampling (sc-Syn0) tool. This tool has been tested on single-cell expression data from cardiac tissues for automated annotation of cardiac glial cells and proliferative cardiomyocytes [9].

In summary, oversampling techniques are still under frequent development and current ensemble approaches can significantly enhance ML-based classification results in almost all biomedical disciplines.

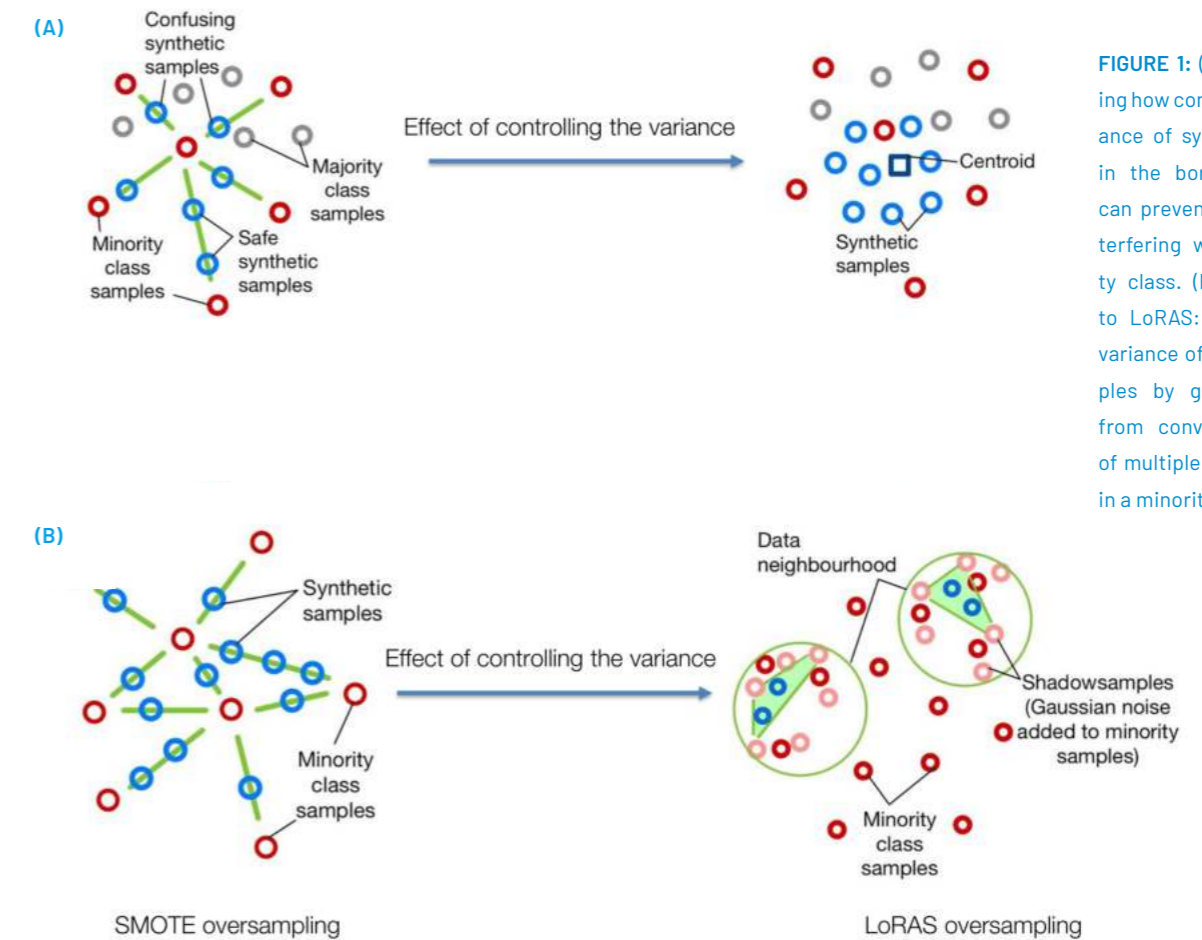


FIGURE 1: (A) Figure showing how controlling the variance of synthetic samples in the borderline regions can prevent them from interfering with the majority class. (B) From SMOTE to LoRAS: controlling the variance of synthetic samples by generating them from convex combination of multiple shadowsamples in a minority neighborhood

REFERENCES: [1] Chawla N et al. 2002 J. Artif. Intell. Res., 16, 321-357. DOI: 10.1613/jair.953. [2] Kovács G 2019 Applied Soft Computing, 83, 105662. DOI: 10.1016/j.asoc.2019.105662. [3] He H et al. 2008 IEEE international joint conference on neural networks. DOI: 10.1109/IJCNN.2008.4633969. [4] Han H et al. 2005 Advances in intelligent computing. 3644, 878-887. DOI: 10.1007/1153805_91. [5] Ma L and Fan S 2017 BMC Bioinformatics, 18, 169. DOI: 10.1186/s12859-017-1578-z. [6] Xie Z et al. 2015 DASFAA, Lecture Notes in Computer Science. DOI: 10.1007/978-3-319-18123-31. [7] Barua S et al. 2013 PAKDD, Lecture Notes in Computer Science. DOI: 10.1007/978-3-642-37456-227. [8] Bej, S et al. 2021 Mach Learn 110, 279-301. DOI: 10.1007/s10994-020-05913-4. [9] Bej S et al. 2021 bioRxiv. DOI: 10.1101/2021.01.20.427486.

AUTHORS: Saptarshi Bej¹, Prashant Srivastava¹, Kristian Schulz², Markus Wolfien¹, and Olaf Wolkenhauer^{1,2,3}

¹ Department of Systems Biology and Bioinformatics, Universitätsplatz 1, University of Rostock, Rostock

² Stellenbosch Institute of Advanced Study, Wallenberg Research Centre, Stellenbosch University, Stellenbosch, South Africa

³ Leibniz-Institute for Food Systems Biology, Technical University of Munich, Lise-Meitner-Strabe 34, Freising

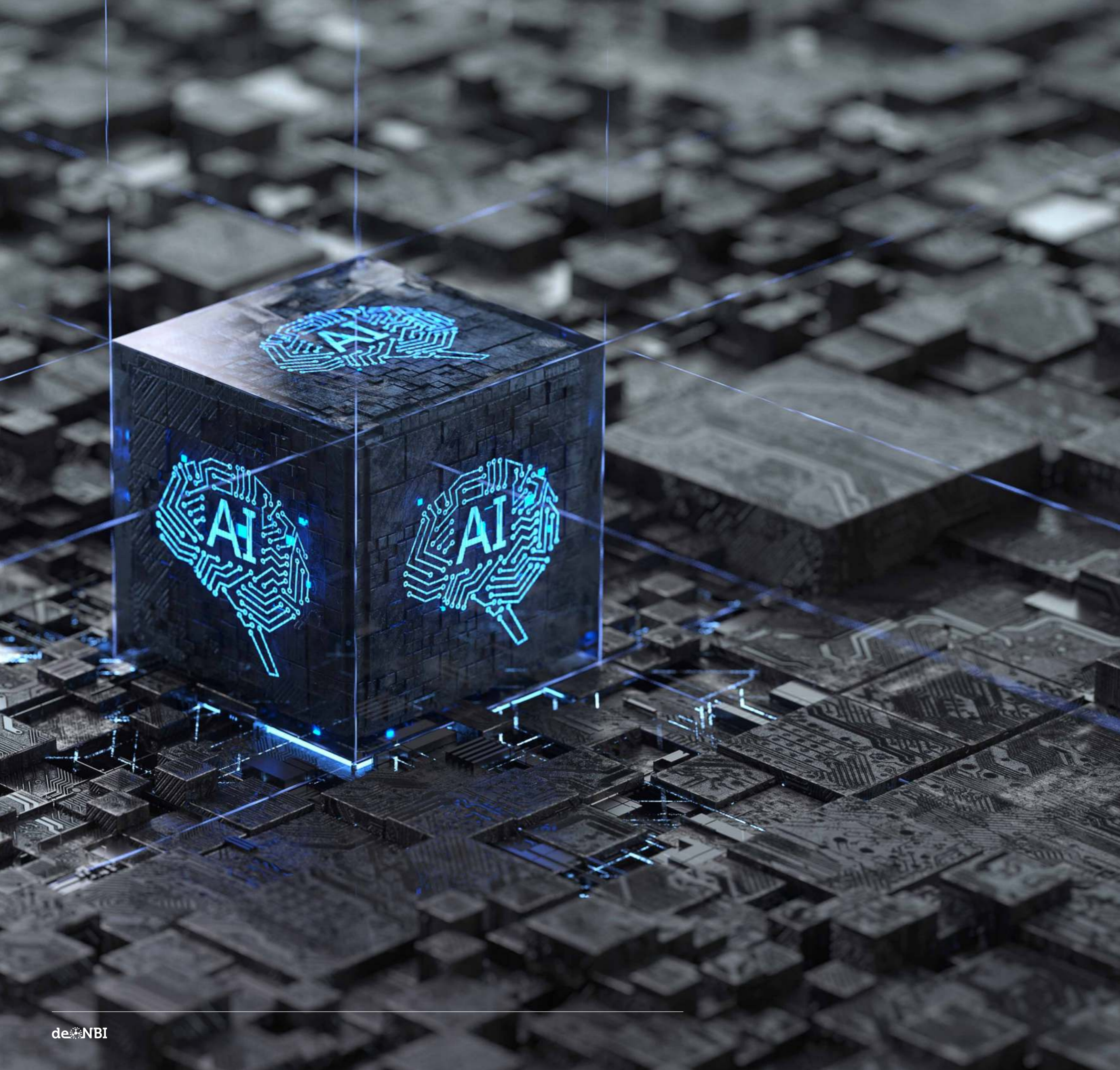


PHOTO CREDITS:


de.NBI Administration Office
Andreas Kühlken (p. 3)
© stock.adobe.com/Putilov_denis (Cover)
© stock.adobe.com/Sdecoret (p. 4,16)
© iStock.com/gorodenkoff (p. 4, 80, 86)
© iStock.com/inkoly (p. 18)
© stock.adobe.com/Kras99 (p. 22)
© iStock.com/vshivkova (p. 24)
© stock.adobe.com/Anttoniart (p. 29)
© iStock.com/vshivkova (p. 30)
© iStock.com/koto_feja (p. 34)
© iStock.com/DKosig (p. 38)
© iStock.com/metamorworks (p. 40, 94)
© stock.adobe.com/Chan2545 (p. 48)
© iStock.com/Eoneren (p. 5, 50)
© iStock.com/Grapelimages (p. 52)
© stock.adobe.com/LuckyStep (p. 56)
© stock.adobe.com/ChrisChrisW (p. 58)
© stock.adobe.com/Juan Gärtner (p. 62)
© iStock.com/jxfzsy (p. 66)
© Max-Planck-Institut Magdeburg/Stefan Deutsch (p. 68)
© stock.adobe.com/kseniyaomega (p. 70)
© iStock.com/CIPhotos (p. 74)
© iStock.com/eranicle (p. 76)
© stock.adobe.com/Tim (p. 78)
© iStock.com/ClaudioVentrella (p. 82)
© stock.adobe.com/oatawa (p. 88)
© iStock.com/Traitov (p. 90)
© iStock.com/isak55 (p. 96)
© iStock.com/MF3d (p. 98)
© iStock.com/in-future (p. 100)
© iStock.com/Andy (p. 104)

IMPRINT

Prof. Dr. Alfred Pühler
German Network for Bioinformatics Infrastructure (de.NBI)
de.NBI Administration Office
Bielefeld University
Center for Biotechnology (CeBiTec)
Universitätsstraße 27
33615 Bielefeld

Tel: +49 (0)521 106 8750
Fax: +49 (0)521 106 89046
E-Mail: contact@denbi.de

Editors-in-Chief: Prof. Dr. Alfred Pühler (Bielefeld University, CeBiTec),
Prof. Dr. Andreas Tauch (Bielefeld University, CeBiTec)
Editorial Team: Dr. Tanja Dammann-Kalionowski (Bielefeld University, CeBiTec),
Dr. Irena Maus (Bielefeld University, CeBiTec),
Dr. Vera Ortseifen (Bielefeld University, CeBiTec)

www.denbi.de
 @denbiOffice
 [linkedin.com/company/de-nbi](https://www.linkedin.com/company/de-nbi)

Date: October 2021

Design and Layout:
MEDIUM Werbeagentur GmbH, Bielefeld

DOI: <https://doi.org/10.4119/unibi/2958276>

Unless otherwise noted, this publication is licensed under Creative Commons
Attribution – Non Commercial – NoDerivatives4.0 International (CC BY NC ND).

For more information see:

<https://creativecommons.org/licenses/by-nc-nd/4.0/deed.en>
<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>

SPONSORED BY



Fkz 031A532B
(de.NBI Administration Office)

