

Introduction to Machine Learning in Microbiome Science

Sascha Patz

Julius Kühn Institute (JKI), Federal Research Centre for Cultivated Plants
Institute for National and International Plant Health
Brunswick, Germany

BiomeFUN 2025 | Day 5

What you will learn?

- 1) **Intro to Artificial Intelligence: Machine Learning > Deep Learning**
- 2) **Concepts and Applications in Microbiome Research**
- 3) **Machine Learning & Example**
- 4) **Deep Learning & Example**
- 5) **Pitfalls: What is Shaping Feature Importance and Downstream Predictions?**
- 6) **Recap**

Importance of Artificial Intelligence in Microbiome Research

Summer of 1956 - Dartmouth Conference by John McCarthy

- AI: simulating extensions of human intelligence
- Theoretical methods, techniques, and applied systems

1959 at Bell Labs, IBM, and Stanford Arthur Samuel

- ML: initially conceptualized as a specialized branch of AI
- Discern features from extensive, diverse datasets

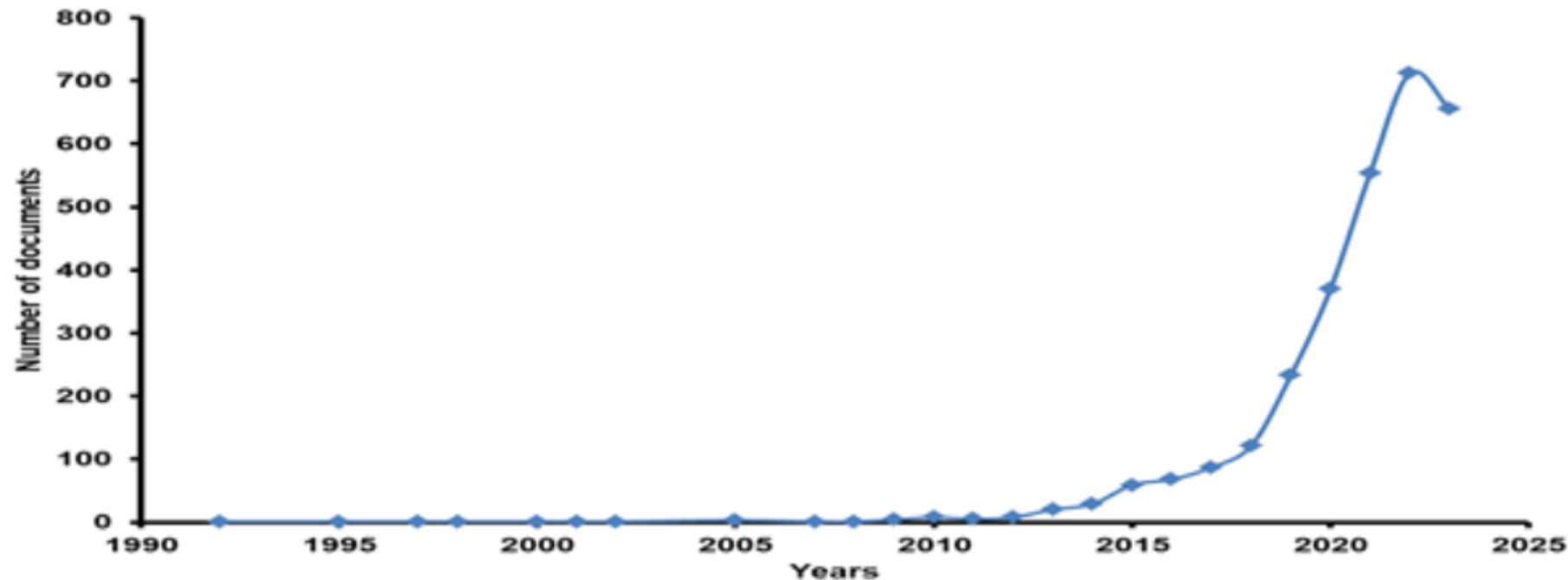
Late 1960s to 1990 Personal Computers entering the Commercial Market

- Boolean Networks for genetic regulatory systems
- AI-based biomedical expert system (MYCIN Stanford)
- Neural networks for protein sequence & structure prediction
- ML (SVMs, decision trees) for gene expression analysis & protein classification

2000 – 2022 Human Genome Project (not only 😊)

- AI/ML established in genomics, proteomics, systems biology
- Advanced sequencing techniques
- Computational Power and Storage Solutions

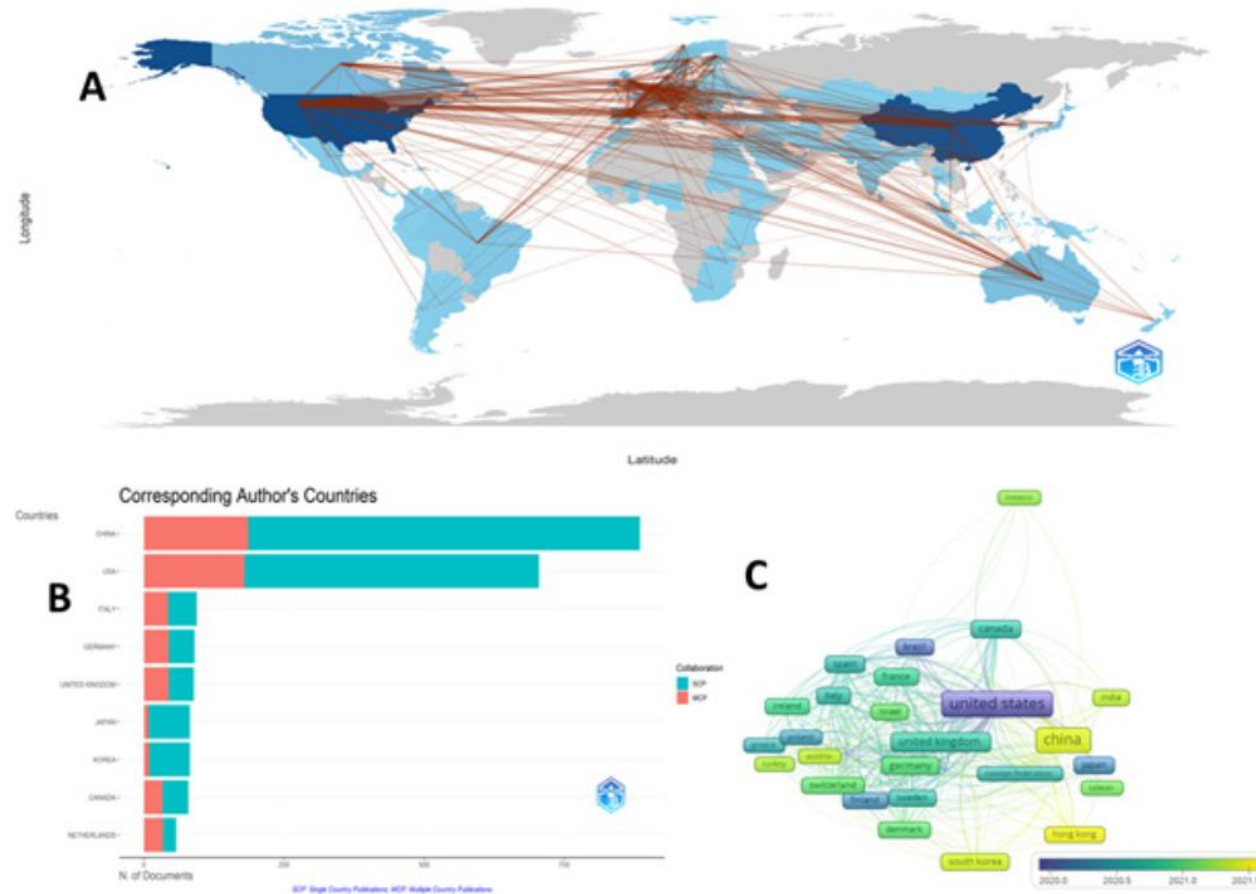
Importance of Artificial Intelligence in Microbiome Research



- (1) Rapidly collecting of substantial volumes of digital data
- (2) Exponential increase of affordable computing power (Moore's law)
- (3) Increase of data storage
- (4) Global system of interconnected computer networks (e.g. AWS, Google, ...)

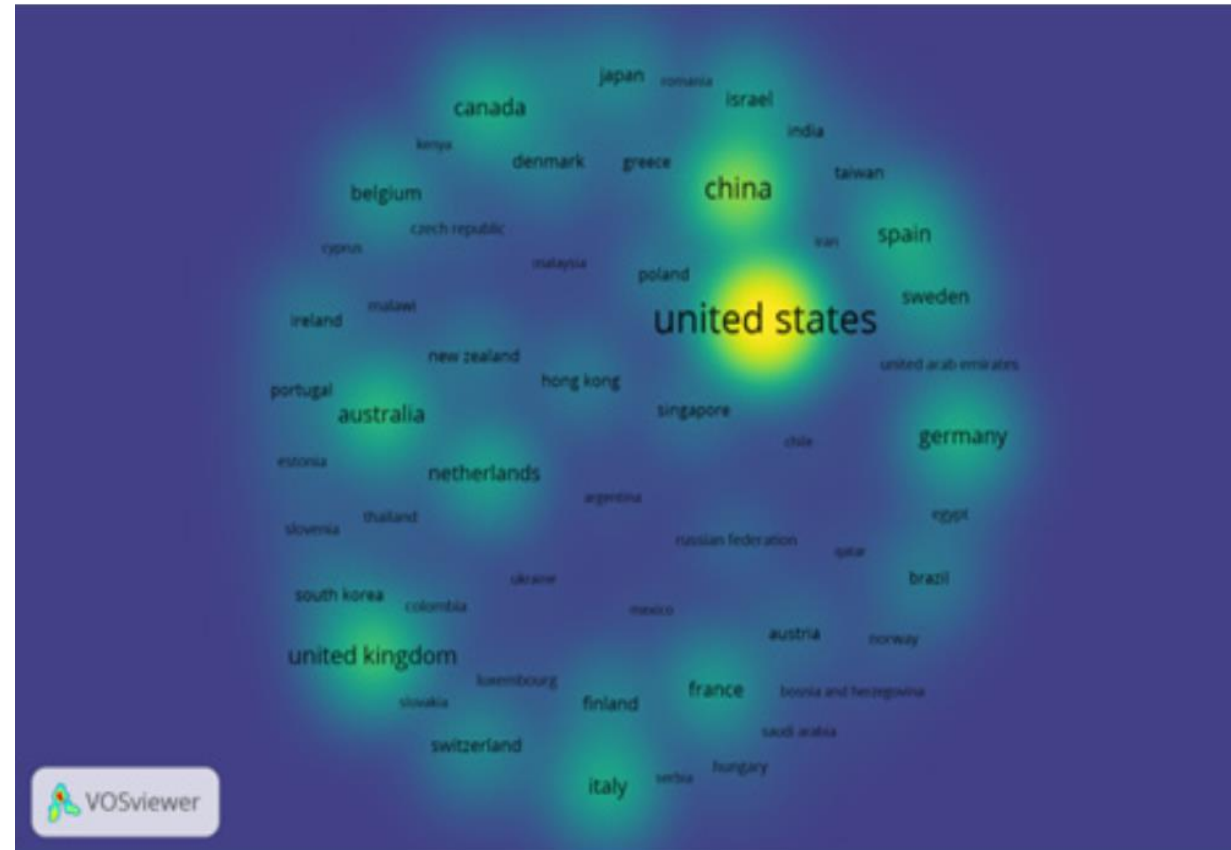
Importance of Artificial Intelligence in Microbiome Research

Country Collaboration Map



Importance of Artificial Intelligence in Microbiome Research

Country Collaboration Map



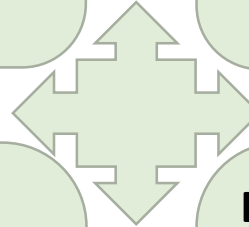
Applications of AI Approaches in Microbiome Research

Sustainable Agriculture / Environments

- Stress-mitigating microbiome marker taxa
- Soil health measures
- Microbiome-assisted plant breeding
- SynCom prediction

Microbiome Engineering & Monitoring

- Microbial interactions
- Microbial community function
- Degree of compositional & functional manipulation
- Pathogen detection and tracking (Image Classification, Object Detection)



Health & Precision Medicine

- Disease prevention and treatment
- Drug development
- Infectious diseases key taxa
- Classifying antimicrobial drug resistance
- Predicting disease outbreaks
- Precision Medicine
- Cosmetics, Pharmaceuticals

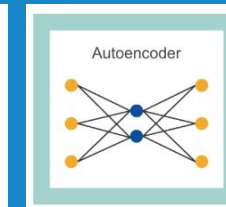
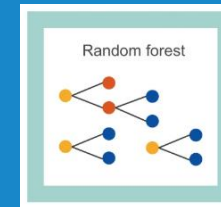
Economic & Anthropogenic Impact Evaluation

- Quality assurance
 - Compliance with health standards
- Likelihood-reduction of product recalls
 - Consumer / Food safety
- Sample disposal
 - Financial implications: Cost Reduction
 - Environmental considerations: Waste Reduction

AI > ML > DL

Artificial Intelligence (AI)**Machine Learning (ML)**

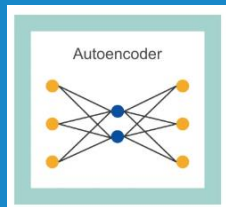
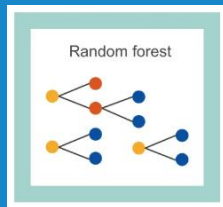
- Pattern recognition, classification, prediction
- Limited in capturing complex non-linear interactions
- Linear/logistic regression, Random Forest (RF), Support Vector Machines (SVM), Gradient Boosting Machines (XGBoost, LightGBM), k-Nearest Neighbors, Naïve Bayes, mixed effects models, ensemble methods



AI > ML > DL

Artificial Intelligence (AI)

Machine Learning (ML)



Supervised

- Learning from **labeled data** (known input/output)
- **Classification/Prediction unknown data**
- E.g. Random Forest, SVM, Gradient Boosting

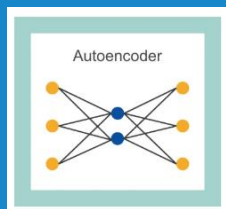
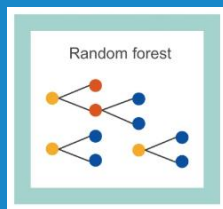
Unsupervised

- Learning from **unlabeled data** (inputs only, no outputs)
- **Clustering → Pattern / Structure Prediction**
- E.g. K-means, hierarchical clustering, t-SNE

AI > ML > DL

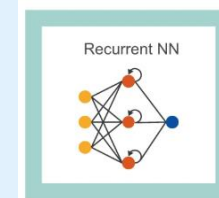
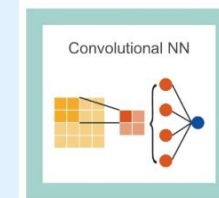
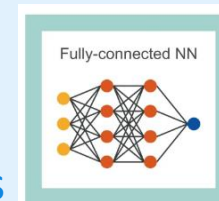
Artificial Intelligence (AI)

Machine Learning (ML)



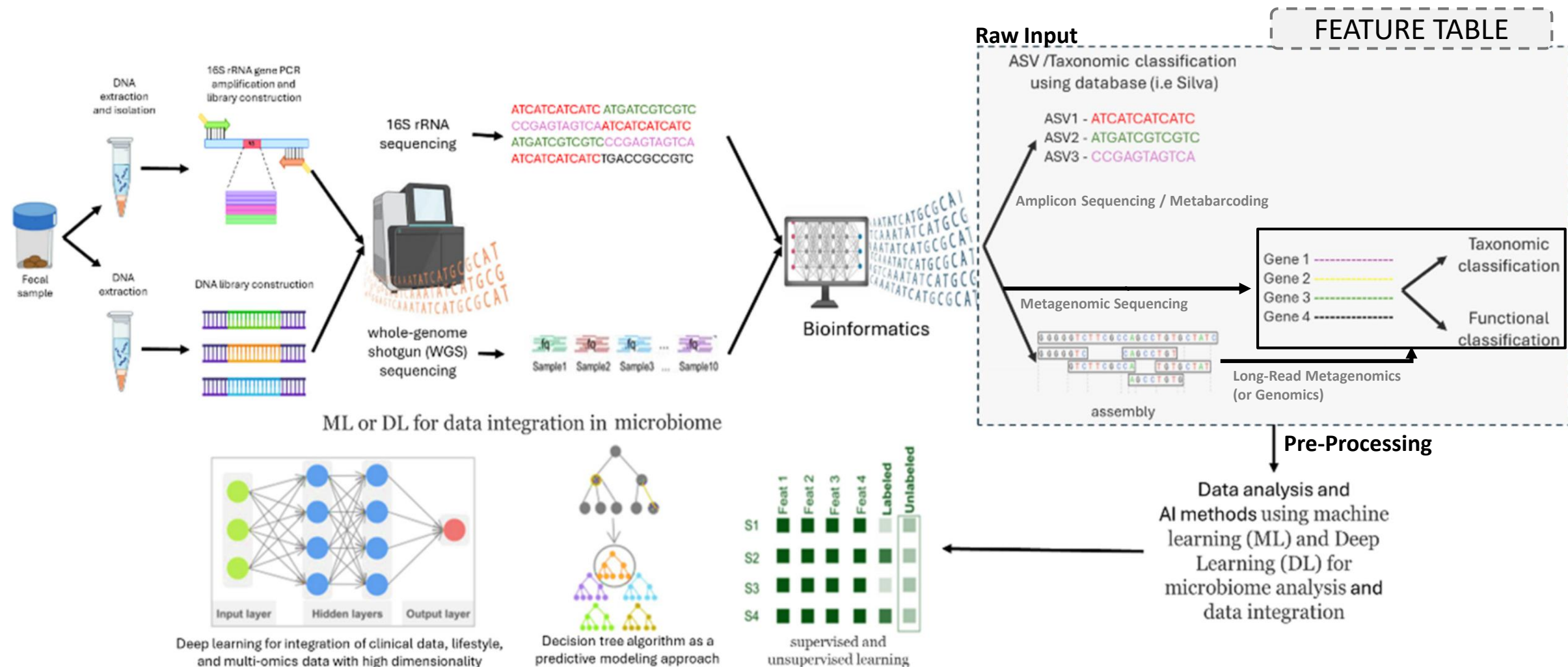
Deep Learning (DL)

- Handling with complex non-linear interactions
- E.g. Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), Autoencoders (e.g. VAE), Graph Neural Networks (GNNs)



... Natural Language Processing (NLP): e.g. LLMs

Embedded Microbiome Analysis Workflow – Recap last days 😊



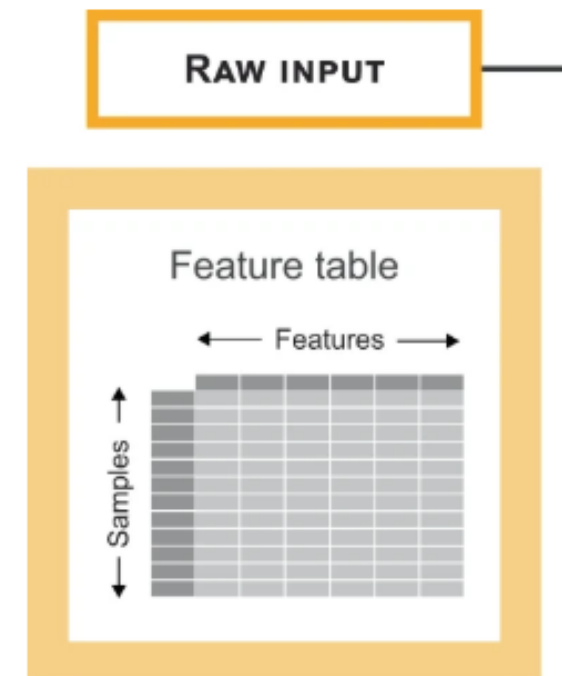
Downstream AI-assisted Microbiome Analysis



Downstream AI-assisted Microbiome Analysis

RAW INPUT: Feature Tables

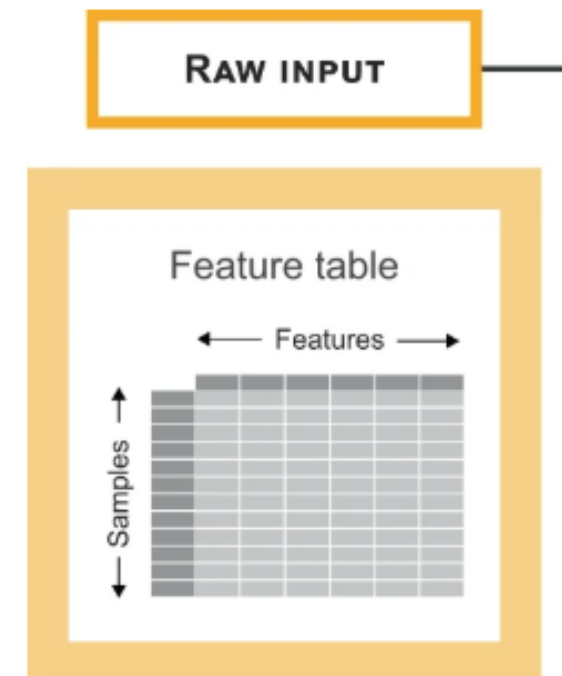
- Raw sequencing reads as **quantification (feature) tables**
- High **variability in read depths** per sample
- **Data heterogeneity** (sequencing methods, pipelines)
- **Sparsity** due to zero counts
- **Non-Gaussian distribution**



Downstream AI-assisted Microbiome Analysis

RAW INPUT: Feature Tables

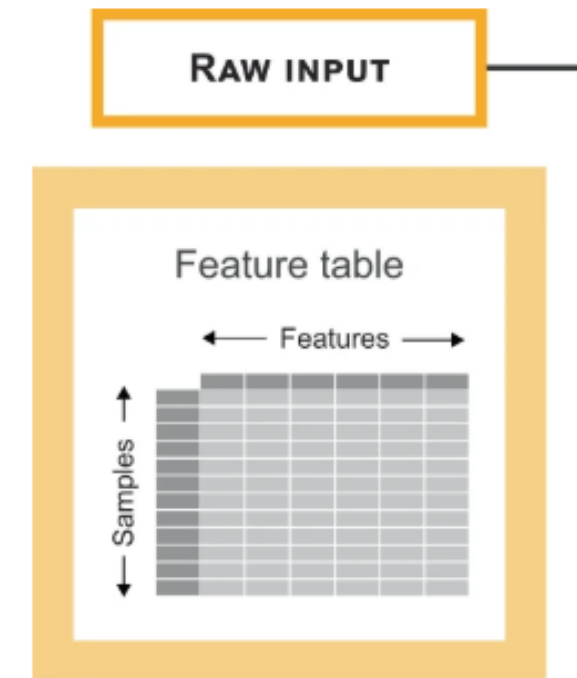
- Raw sequencing reads as quantification (feature) tables
- High **variability in read depths** per sample
- **Data heterogeneity** (sequencing methods, pipelines)
- **Sparsity** due to zero counts
- **Non-gaussian distribution**
- Compositional data
 - relationships between its components
 - **not independent**
 - **sum is arbitrary**



Downstream AI-assisted Microbiome Analysis

RAW INPUT: Feature Tables

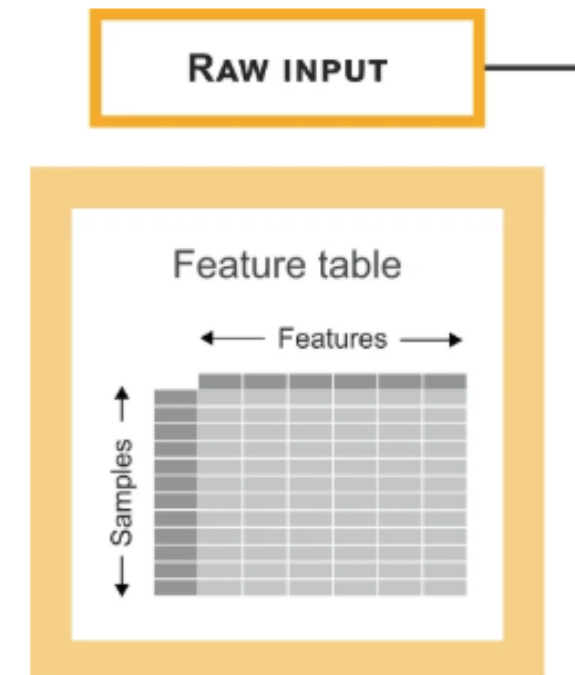
- Raw sequencing reads as quantification (feature) tables
- High **variability in read depths** per sample
- **Data heterogeneity** (sequencing methods, pipelines)
- **Sparsity** due to zero counts
- **Non-gaussian distribution**
- **Compositional data**
- **High Dimensionality**
 - More features (microbial genes or taxa) than samples
 - Computational cost
 - Overfitting issue
 - Poor generalization



Downstream AI-assisted Microbiome Analysis

RAW INPUT: Feature Tables

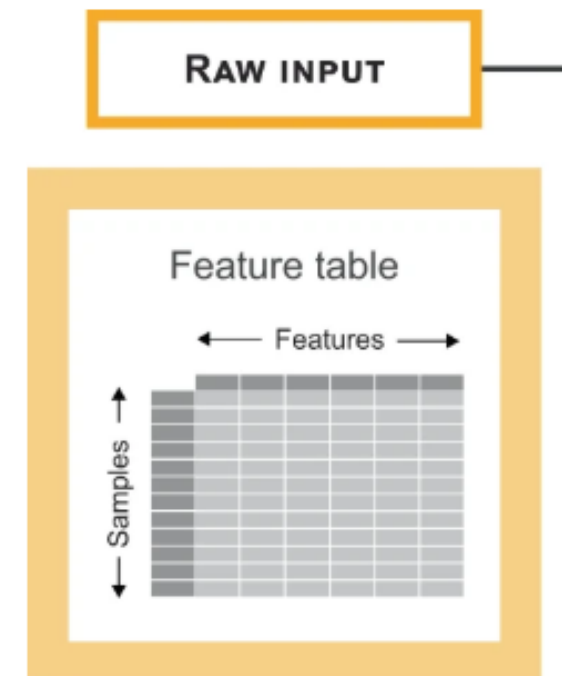
- Raw sequencing reads as quantification (feature) tables
- High **variability in read depths** per sample
- **Data heterogeneity** (sequencing methods, pipelines)
- **Sparsity** due to zero counts
- **Non-gaussian distribution**
- **Compositional data**
- **High Dimensionality**
- Relatively low number of samples
 - Impoverishes generalization to other datasets



Downstream AI-assisted Microbiome Analysis

RAW INPUT: Feature Tables

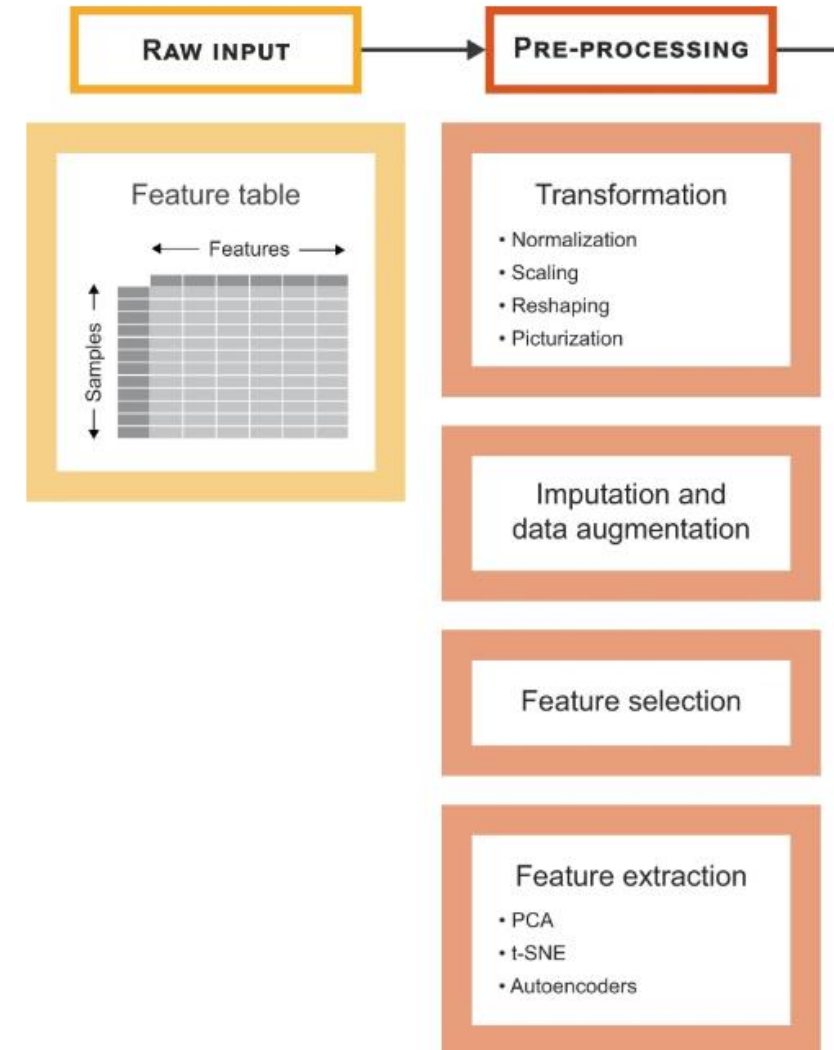
- Raw sequencing reads as quantification (feature) tables
- High **variability in read depths** per sample
- **Data heterogeneity** (sequencing methods, pipelines)
- **Sparsity** due to zero counts
- **Non-gaussian distribution**
- **Compositional data**
- **High Dimensionality**
- **Relatively low number of samples (generalization)**
- **Lacking or inconsistent Metadata**



Downstream AI-assisted Microbiome Analysis

PRE-PROCESSING

- Common distance and association measures:
 - **invalid for compositional data**
- **Non-linearity**
 - among and between features and the target/phenotype
 - Increasing Feature 1 → decreasing Feature 2



Downstream AI-assisted Microbiome Analysis

PRE-PROCESSING

- Common distance and association measures are **invalid for compositional data**
- **Non-linearity**: among and between features and the target/phenotype
- **Statistical methods**
 - **Outlier detection**
 - **Normalization**
 - **Batch effect** correction
 - E.g. platform, library prep, batch processing
 - **Log-ratio transformations**
 - Compresses large values and spreads out small ones, stabilizing the variance
 - Cannot deal with **sparsity**
 - Data is often **imputed**
 - Commonly **zeros** are replaced with **pseudo-counts**
 - Other Methods:
 - **CLR transformation (log-ratio)**
 - Values can be negative or positive (centered around 0 from -5 to +5)
 - Each feature expressed relative to the geometric mean of the sample



Downstream AI-assisted Microbiome Analysis

PRE-PROCESSING

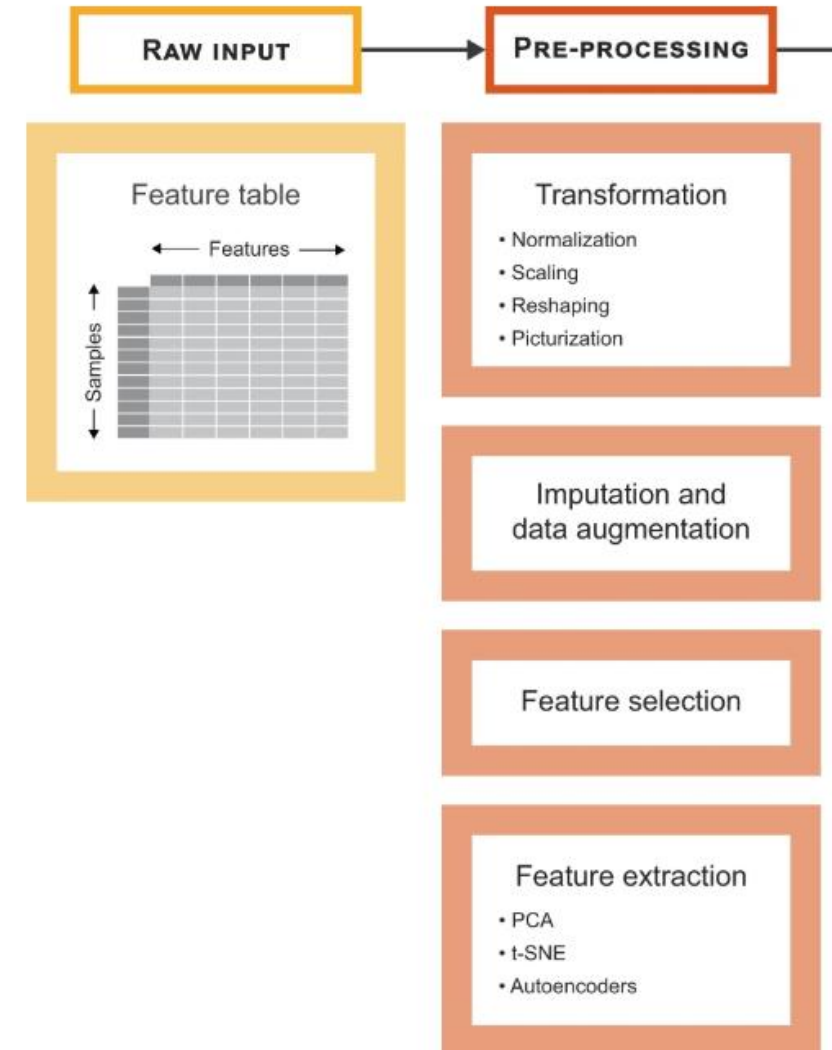
- Common distance and association measures are **invalid for compositional data**
- **Non-linearity**: among and between features and the target/phenotype
- Statistical methods (**Normalization, Batch Effects, Log-ratios, pseudo-counts, ...**)
- **Dimensionality Reduction**:
 - **Feature cleaning/filtering**
 - E.g. low-abundance filtering, variance-/correlation-based
 - **Feature selection**
 - Selecting optimal subspace of relevant, non-redundant features
 - Supervised: e.g. Correlations, ANOVA
 - Wrapper: Recursive feature elimination (RFE), Statistically equivalent signatures (SES)
 - **Embedded in Models: LASSO regression, Random Forest importance**
 - **Feature extraction**
 - Compressed representation of the input features
 - E.g. PCA (linear transformation), tSNE (non-linear), Autoencoders (non-linear)



Downstream AI-assisted Microbiome Analysis

PRE-PROCESSING

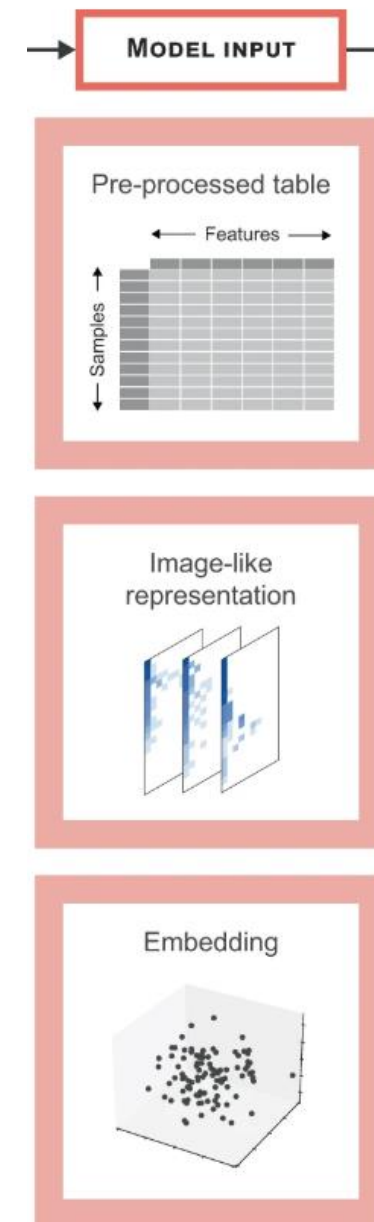
- Common distance and association measures are **invalid for compositional data**
- **Non-linearity**: among and between features and the target/phenotype
- Statistical methods (**Normalization, Batch Effects, Log-ratios, pseudo-counts, ...**)
- Dimensionality Reduction
- **BUT: strongly affects the performance** of machine learning methods



Downstream AI-assisted Microbiome Analysis

MODEL INPUT

- Dependent on model selection
 - Preprocessed Table
 - Image-like Representation for e.g. Abundance or Absence/Presence Tables
 - Embeddings



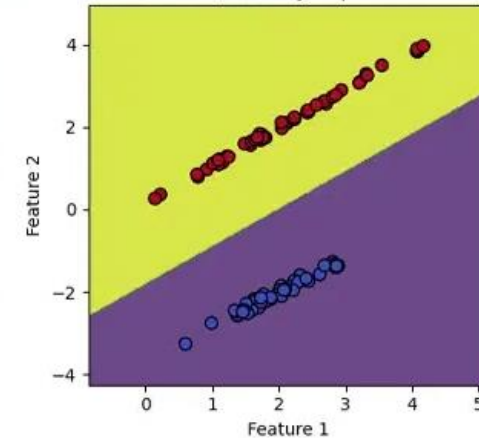
Downstream AI-assisted Microbiome Analysis

MODEL TRAINING & TUNING & SELECTION

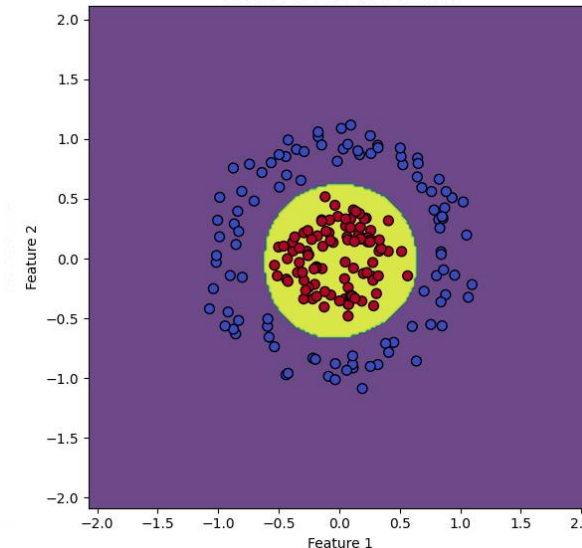
- Classification & Model Selection

- **Linear Classifiers:** Linear SVMs, Logistic regression, linear discriminant analysis, partial least squares discriminant analysis (PLSDA)
- **Non-linear classifiers:** SVMs, decision trees, random forests, artificial neural networks, gradient boosting, kernel PLS-DA.
„Kernel Trick“

Linear SVM (Linearly Separable Data)

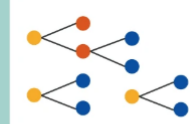


Non-linear SVM (RBF Kernel)



MODEL TRAINING AND TUNING

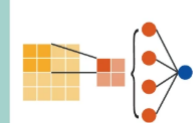
Random forest



Fully-connected NN



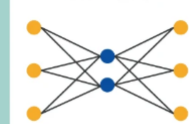
Convolutional NN



Recurrent NN



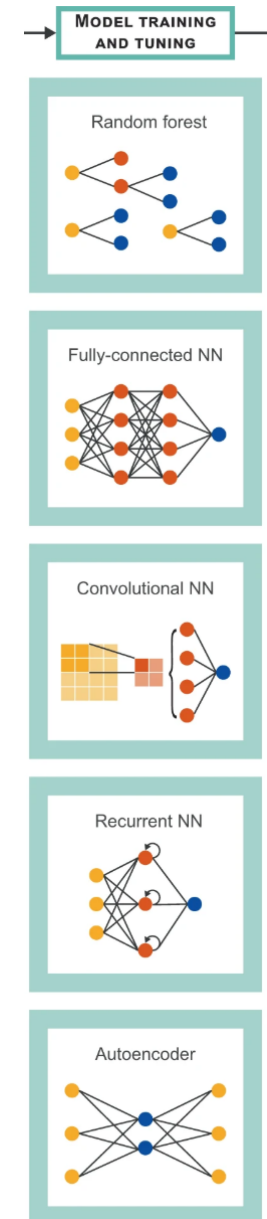
Autoencoder



Downstream AI-assisted Microbiome Analysis

MODEL TRAINING & TUNING & SELECTION

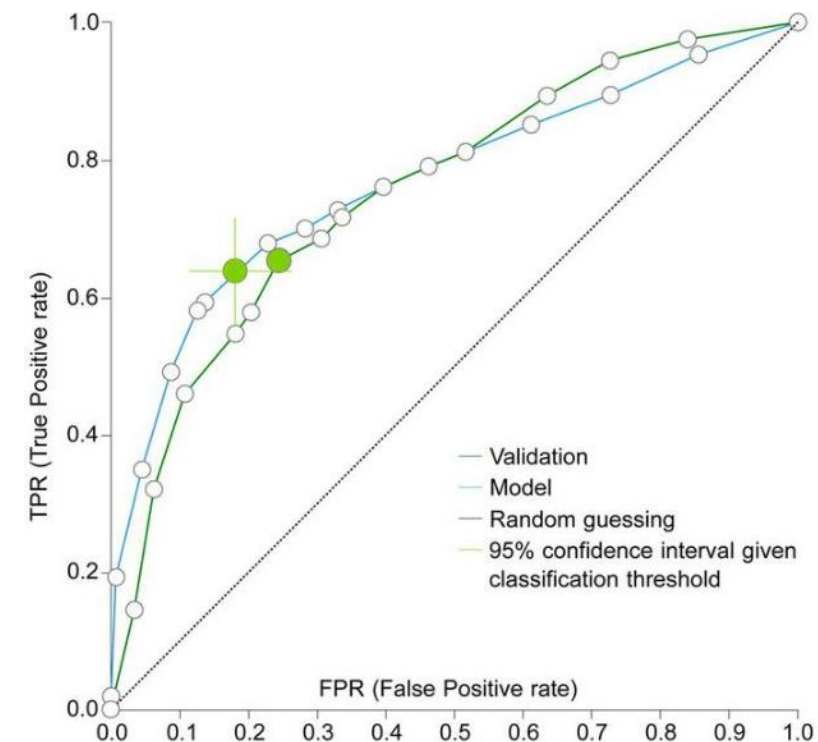
- Performance estimation
 - Quality Evaluation of a predictive model
 - **Holdout method**: typically 70/30 split into training and test data
 - **K-fold (nested) Cross Validation**: train k-1 fold (test on 1 fold) repeat ...
 - Monte Carlo cross validation (random partitioning many times)



Downstream AI-assisted Microbiome Analysis

MODEL TRAINING & TUNING & SELECTION

- Optimization metrics
 - **Threshold-independent measures:**
 - **AUROC:** Area under the receiver operating characteristic curve



Downstream AI-assisted Microbiome Analysis

MODEL TRAINING & TUNING & SELECTION

- Optimization metrics

- **Threshold-independent measures:**
 - AUROC
 - ...
- **Threshold-dependent measures:**
 - Accuracy
 - Precision
 - Recall (“Sensitivity”)
 - F1 score
 - ...

The proportion of all predictions the model got correct.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Of all predicted positives, how many are actually positive?

$$\text{Precision} = \frac{TP}{TP + FP}$$

Of all actual positives, how many did the model correctly identify?

$$\text{Recall} = \frac{TP}{TP + FN}$$

Harmonic mean of precision and recall

$$\text{F1} = 2 \times \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Downstream AI-assisted Microbiome Analysis

MODEL TRAINING & TUNING & SELECTION

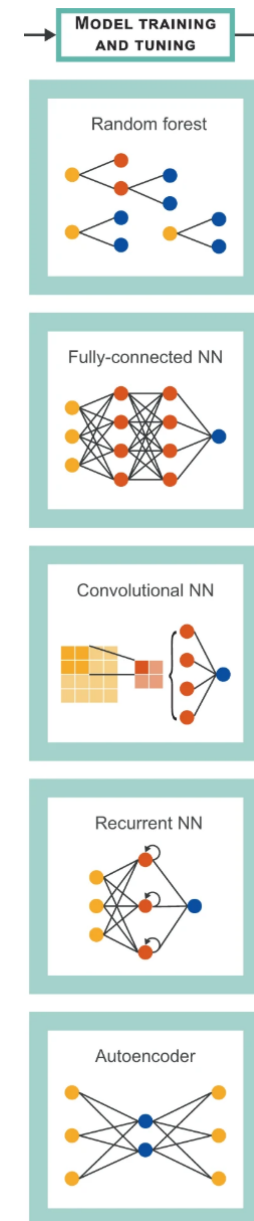
- Model Tuning

- **Hyper-Parameter Optimization (HPO)**

- Random Forest: number of trees, max depth
- SVM: kernel type
- Neural Network: learning rate, number of layers, number of neurons

→ **Finding the best combination** of hyperparameters to **maximize performance**

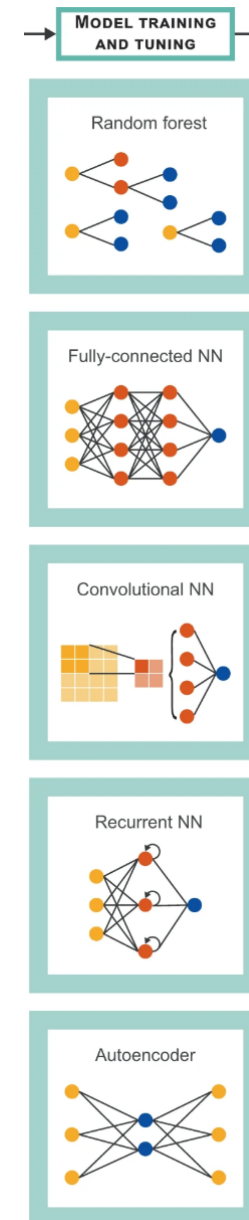
- **Techniques:** grid search, random search, Bayesian optimization (to prior)
- Combined Algorithm Selection and HPO (CASH)
- Early stopping, model checkpoints



Downstream AI-assisted Microbiome Analysis

MODEL TRAINING & TUNING

- Model Interpretability
 - E.g. Feature importance
 - Percentage drop in predictive performance when the feature is removed from the model

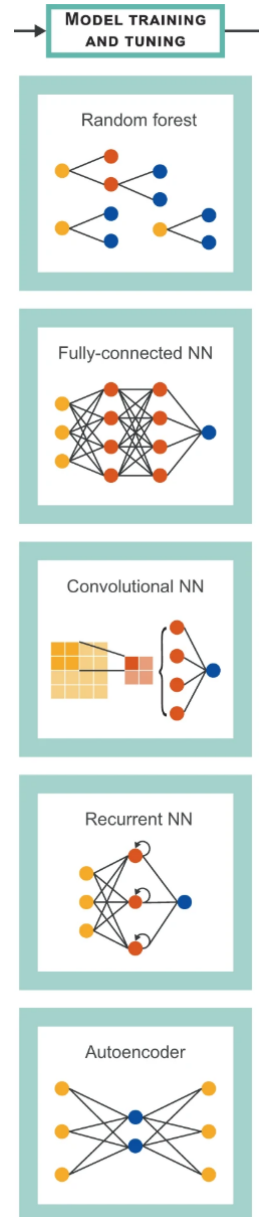


Downstream AI-assisted Microbiome Analysis

MODEL TRAINING & TUNING

- Model Interpretability

- E.g. Feature importance
 - percentage drop in predictive performance when the feature is removed from the model
- **Gini Importance** („Mean Decrease in Impurity“)
 - Every split in a tree reduces impurity
 - Reduction assigned to the split-causing feature
 - Normalized Sum of all reductions across all trees

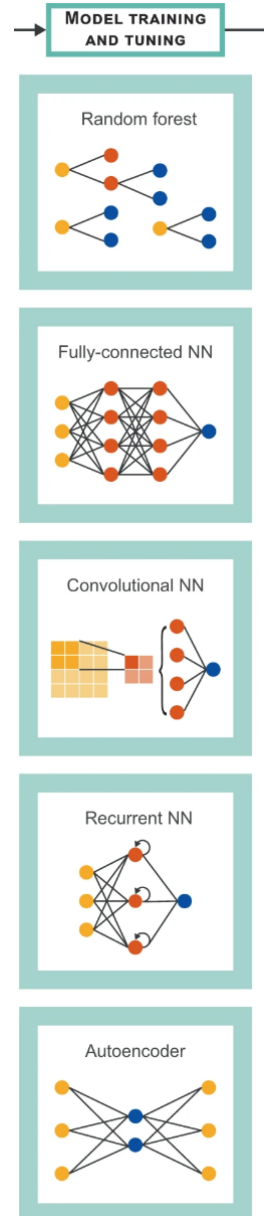


Downstream AI-assisted Microbiome Analysis

MODEL TRAINING & TUNING

- Model Interpretability

- E.g. Feature importance
 - percentage drop in predictive performance when the feature is removed from the model
 - **Gini Importance** („Mean Decrease in Impurity“)
 - Every split in a tree reduces impurity
 - Reduction assigned to the split-causing feature
 - Normalized Sum of all reductions across all trees
 - **SHapley Additive exPlanations (SHAP)**
 - Directionality
 - Game theory: each feature as player
 - Feature contribution to the prediction (all Combinations)



Machine Learning (ML)

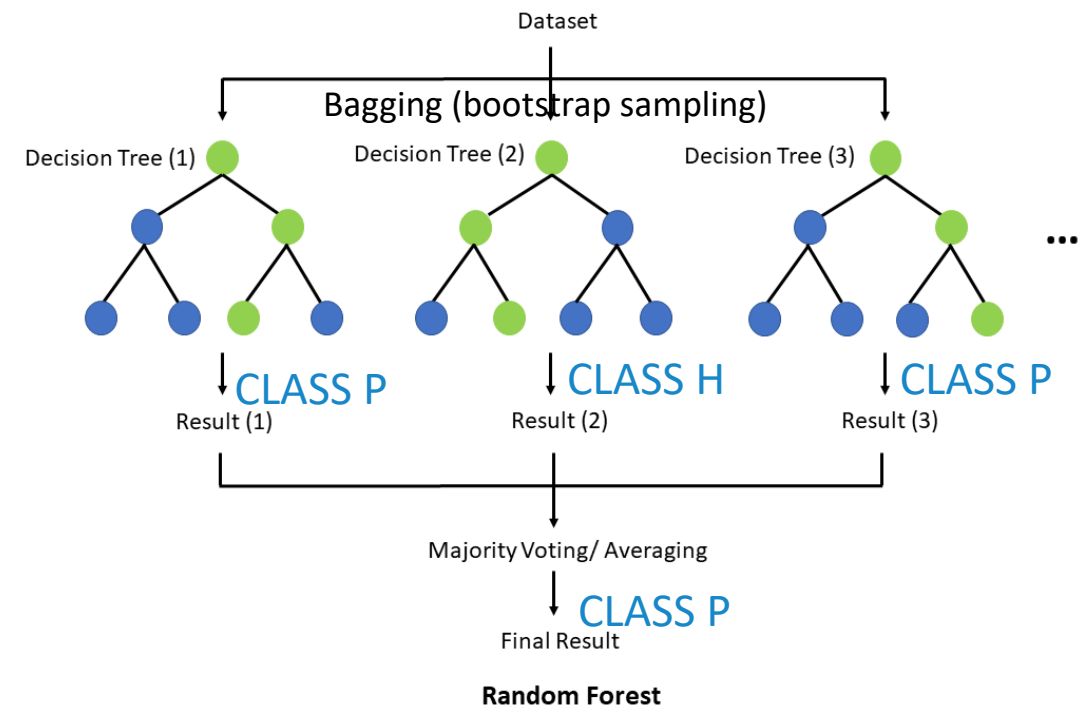
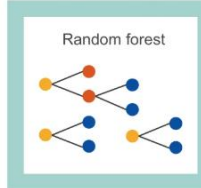
Logistic Regression

- **Linear classifier**
- **Supervised algorithm**
- **Categorical/binary** outcome (e.g. disease vs. no disease)
- Employs a **logistic (sigmoid) function** to model independent variables
- Probability of a sample belonging to a class
- Tending to **overfitting for small sample sizes**
 - ALTERNATIVE: Partial Least Squares Discriminant Analysis PLS-DA

Machine Learning (ML)

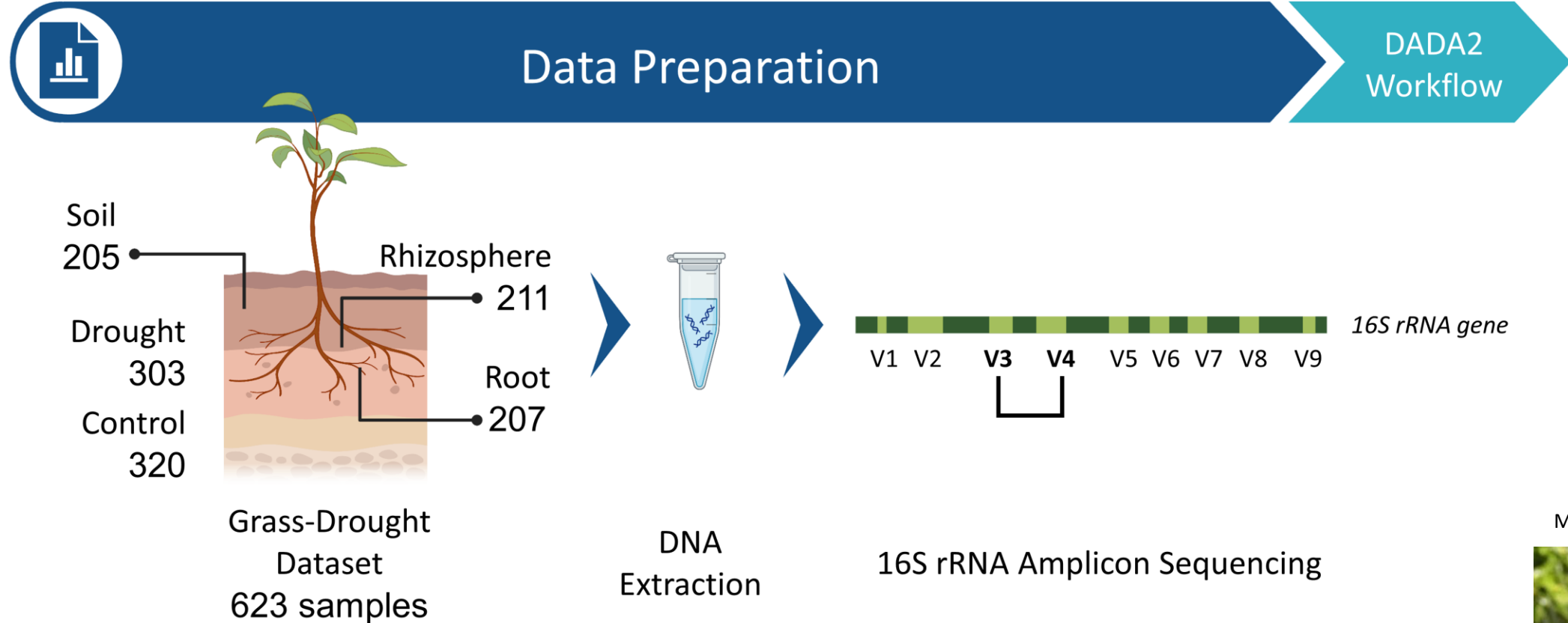
Random Forests (RFs)

- **“Ensemble learning”**: Aggregates decision trees as a forest
- **Bagging** (bootstrap sampling)
- **Populating a tree** (until limit set)
- **At each split a feature subset** evaluation
- Strongest features defines split for target prediction (YES/NO)
- **Majority Voting** over all Trees
- **Overfitting and instability in case trees are too deep or noisy data**
- Alternative: Gradient Boosting (e.g. XGBoost)



Example RF/LG

Drought Stress Marker Taxa and Classification

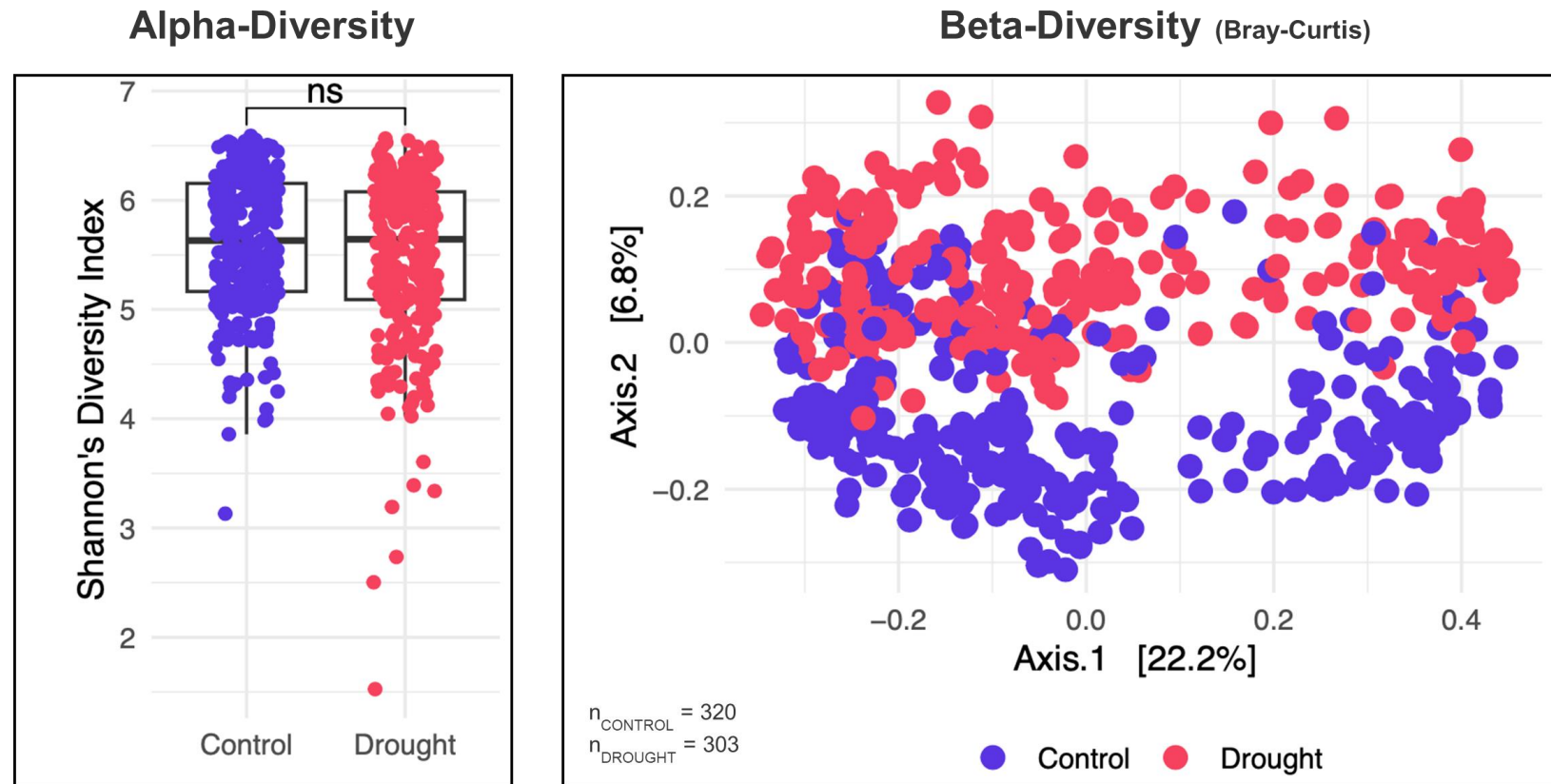


Michelle Hagen



Example RF/LG

Drought Stress Marker Taxa and Classification



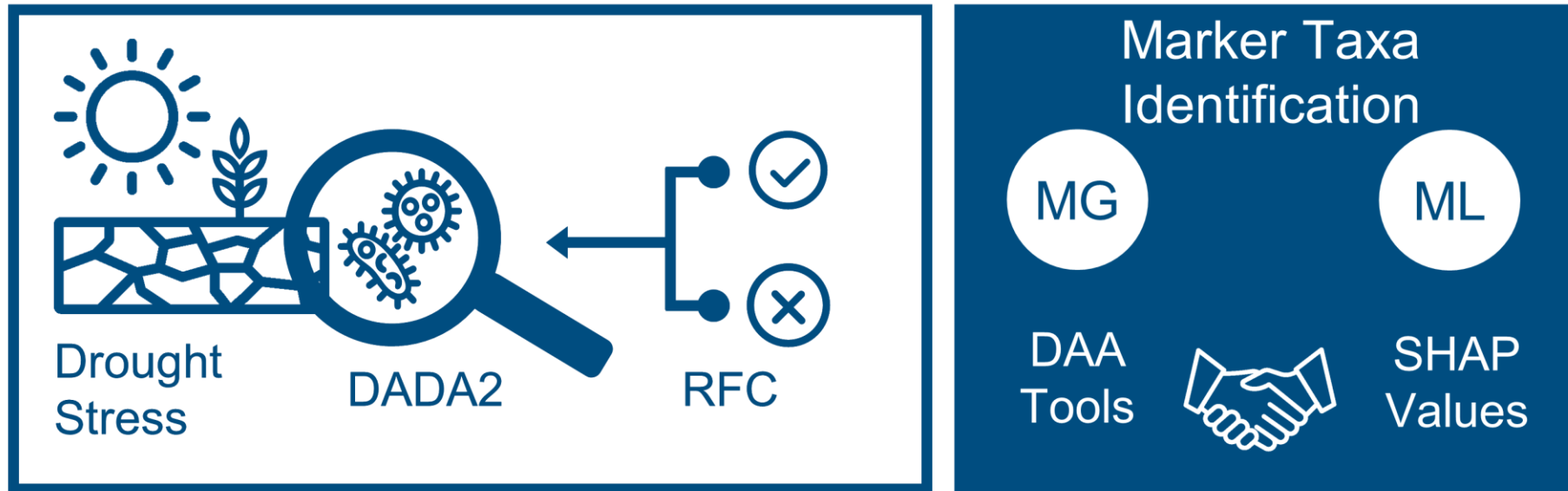
Michelle Hagen



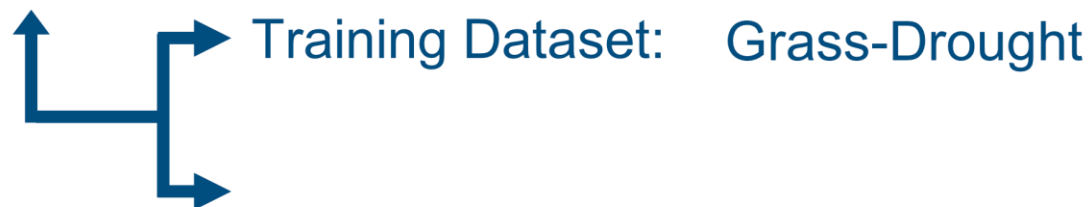
Example RF/LG

Drought Stress Marker Taxa and Classification

Training a Random Forest Classifier using 5-fold nested cross validation and 1 hold out data set



Classifier Performance



560 Samples
(5f NCV: 4:1)

Michelle Hagen



Example RF/LG

Machine Learning

Performing a **nested cross-validation** with feature tables for the prediction of drought stress in the soil metagenome from **phylum** to **genus** rank using a **Random Forest Classifier**. The feature tables can be found in `data/feature_tbl_{rank}.csv` and have the following format:

Samples	<-Columns with Taxa of corresponding Rank->	Target: Watering_Regm
Sample Name	Relative Abundances	0/1 (for Control/Drought)

For each rank, the nested cv is performed with hyperparameter tuning and model selection (the five best models are saved in `data/ RFC_{fold_no}_{rank}.sav`). Mean performance metrics (accuracy, F1 score, precision, recall, ROC with AUC) are calculated. The script creates tables containing the mean SHAP feature importance `SHAP_feature_importance_{rank}.csv` and enrichment information per fold `SHAP_enriched_{rank}.csv` for each taxon of the corresponding rank.

```
In [4]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
from sklearn.model_selection import KFold
from sklearn.model_selection import GridSearchCV
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score
from sklearn.metrics import f1_score
from sklearn.metrics import precision_score
from sklearn.metrics import recall_score
from sklearn.metrics import auc
from sklearn.metrics import RocCurveDisplay
import pickle
import shap
```

https://github.com/Computomics/SoilMicrobiomeDroughtML/blob/main/Machine_Learning.ipynb

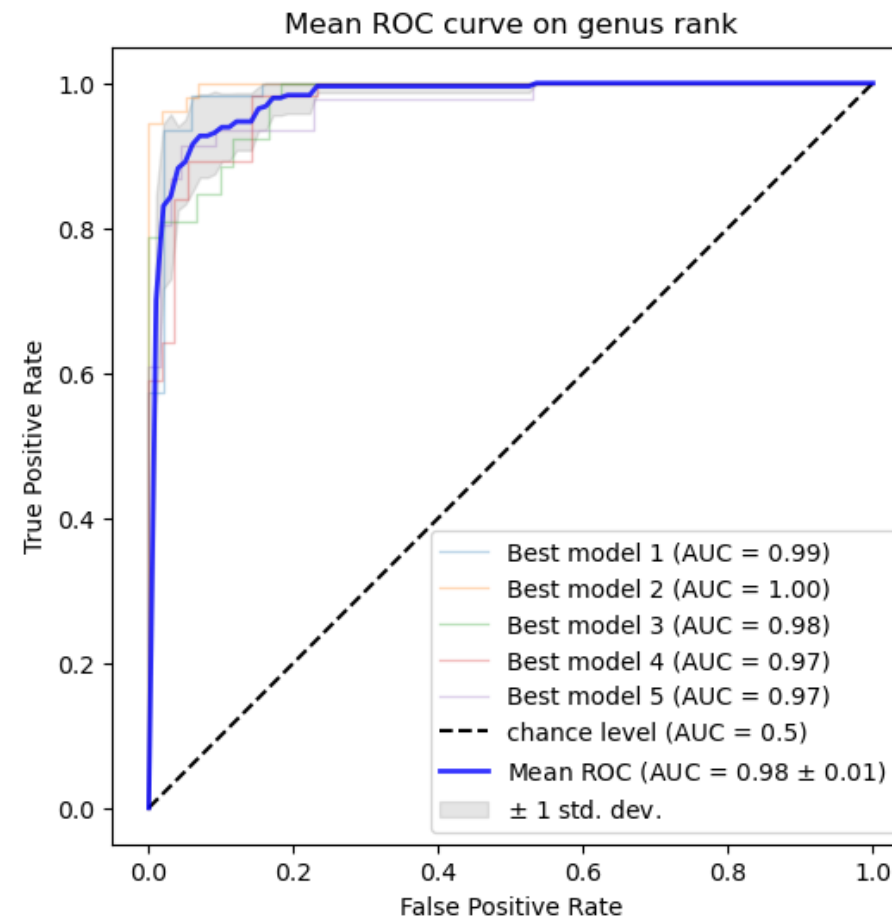
Michelle Hagen



Example RF/LG

Drought Stress Marker Taxa and Classification

Training a Random Forest Classifier using 5-fold nested cross validation and 1 hold out data set → AUROC



Very High!

Michelle Hagen



Example RF/LG

Drought Stress Marker Taxa and Classification

Training a Random Forest Classifier using 5-fold nested cross validation and 1 hold out data set → AUROC

Performance

Tab. 1 Random Forest Classifier Performance. Table displaying the mean accuracy, F1 score, precision, recall, and AUC of the classifier on different taxonomic ranks, with the best-performing rank for each metric marked in bold.

Metric	Phylum	Class	Order	Family	Genus
Accuracy	0.900 ± 0.025	0.911 ± 0.020	0.914 ± 0.023	0.921 ± 0.017	0.923 ± 0.029
F1 score	0.895 ± 0.029	0.906 ± 0.023	0.912 ± 0.024	0.919 ± 0.017	0.921 ± 0.030
Precision	0.891 ± 0.041	0.902 ± 0.032	0.890 ± 0.038	0.902 ± 0.038	0.892 ± 0.041
Recall	0.899 ± 0.024	0.911 ± 0.022	0.936 ± 0.022	0.939 ± 0.031	0.954 ± 0.029
AUC	0.960 ± 0.020	0.960 ± 0.010	0.970 ± 0.010	0.970 ± 0.010	0.980 ± 0.010

Tab. S2: Logistic Regression Performance of the Grass-Drought Dataset. Table displaying the mean accuracy, F1 score, precision, recall, and AUC of the classifier on different taxonomic ranks of the Grass-Drought dataset, with the best-performing rank for each metric marked in bold. The Grass-Drought dataset underwent preprocessing, aligning with the approach utilized by the Random Forest Classifier, followed by a five-fold nested cross-validation.

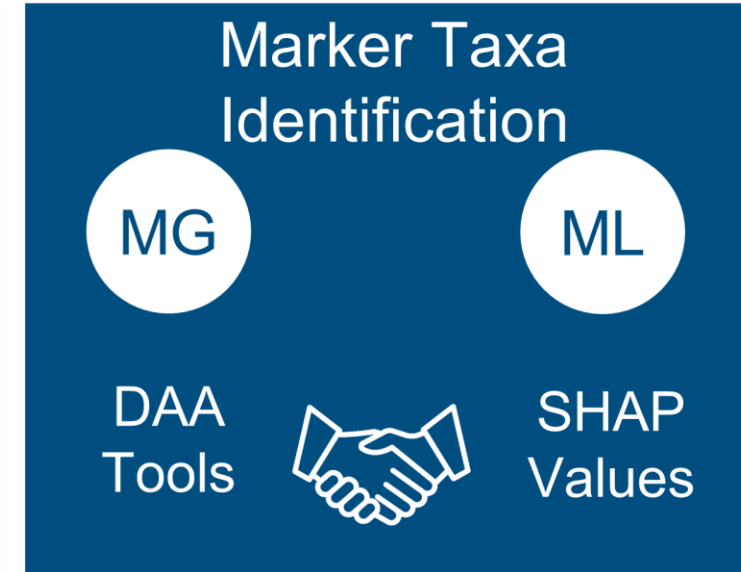
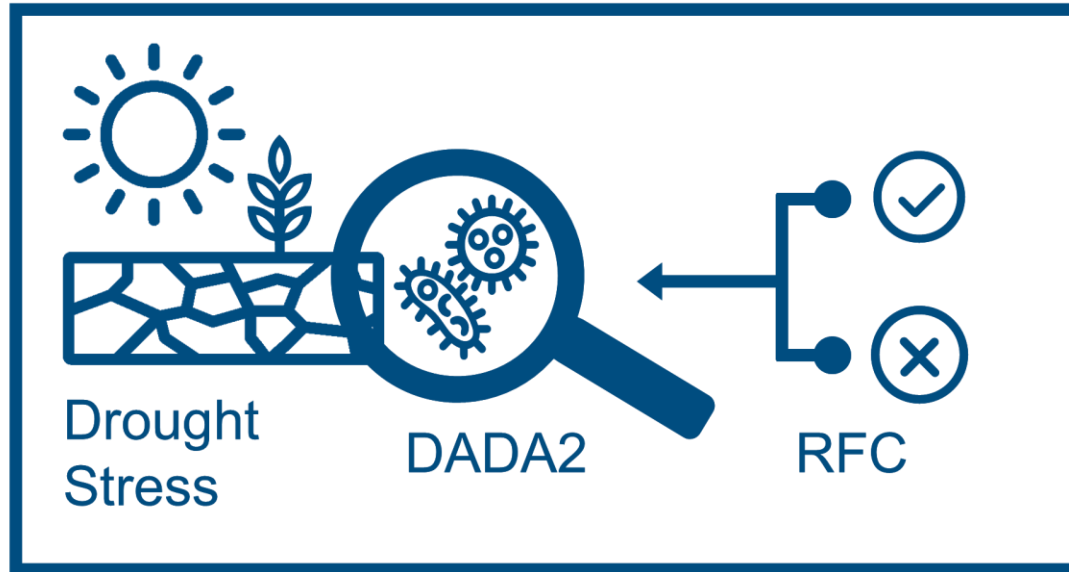
Metric	Phylum	Class	Order	Family	Genus
Accuracy	0.836 ± 0.026	0.829 ± 0.028	0.900 ± 0.043	0.909 ± 0.035	0.916 ± 0.028
F1 score	0.821 ± 0.034	0.816 ± 0.040	0.897 ± 0.044	0.907 ± 0.033	0.914 ± 0.030
Precision	0.853 ± 0.066	0.831 ± 0.059	0.875 ± 0.050	0.887 ± 0.042	0.889 ± 0.036
Recall	0.796 ± 0.045	0.803 ± 0.044	0.920 ± 0.043	0.929 ± 0.028	0.941 ± 0.036
AUC	0.909 ± 0.030	0.914 ± 0.027	0.964 ± 0.019	0.964 ± 0.023	0.968 ± 0.019

Michelle Hagen

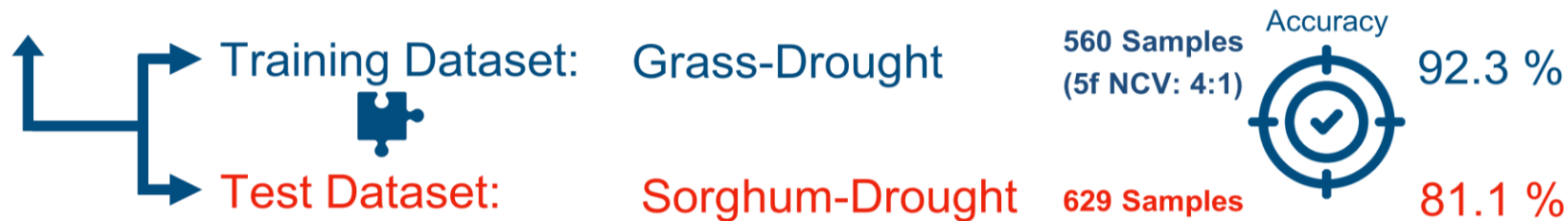


Example RF/LG

Drought Stress Marker Taxa and Classification



Classifier Performance



Michelle Hagen



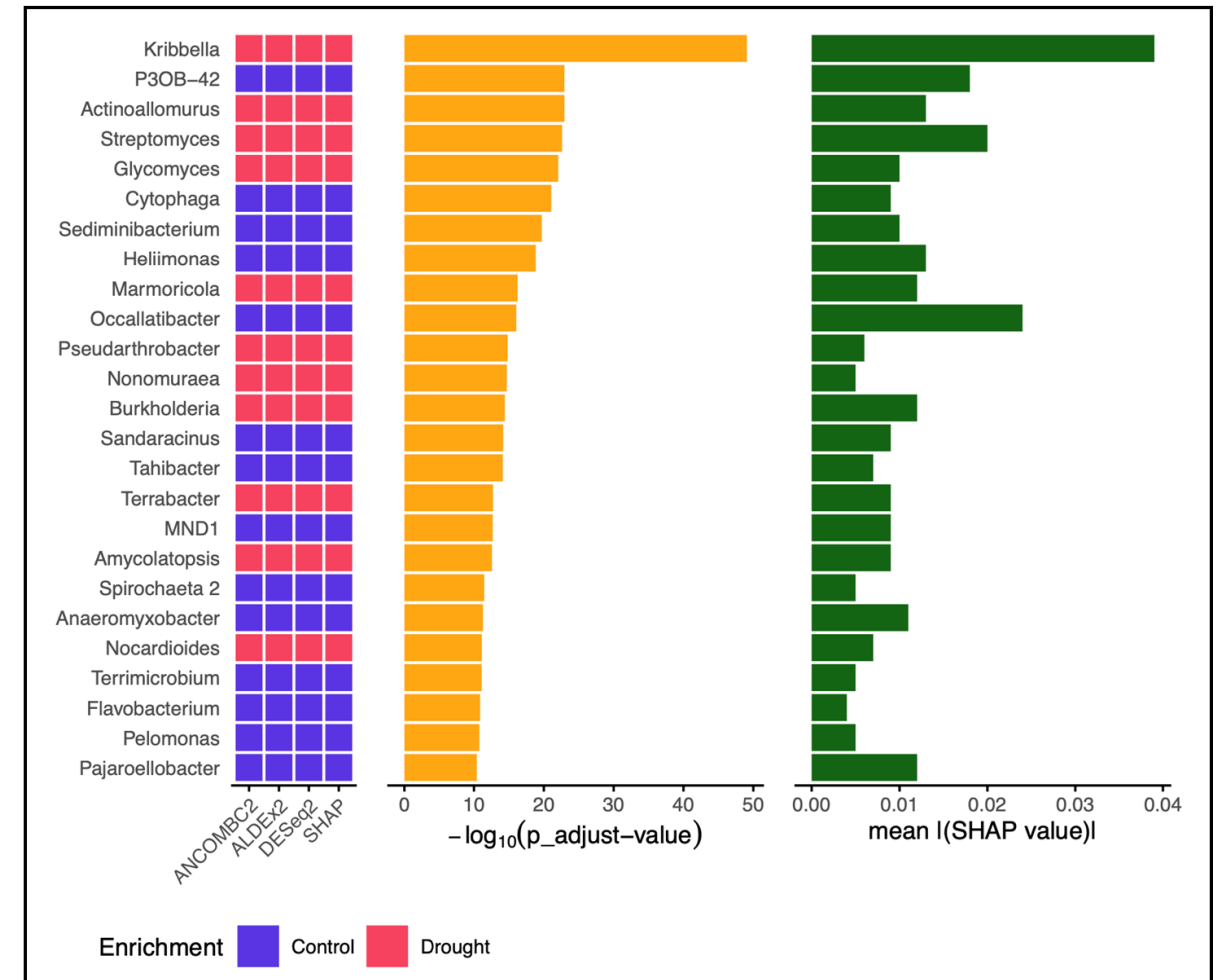
Example RF/LG

Drought Stress Marker Taxa and Classification

RESULTS:

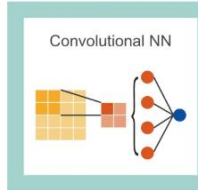
- Trained RF classifier with an accuracy of ~92%
- 81% accuracy on an independent dataset
- **Feature Importance** reveals ~2,600 Drought Stress related ASVs
- ASVs were associated to Isolated Strains and high-performing SynComs derived
- Concept was then applied to Halotolerant Plant-associated Microbial Communities
 - See Abdelfadil, M.R., Patz, S., Kolb, S. *et al.* Unveiling the influence of salinity on bacterial microbiome assembly of halophytes and crops. *Environmental Microbiome* **19**, 49 (2024).
<https://doi.org/10.1186/s40793-024-00592-3>

Feature Importance Comparison

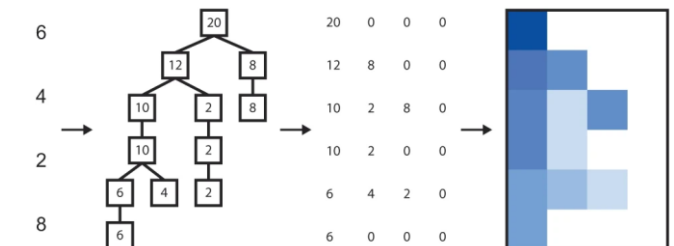
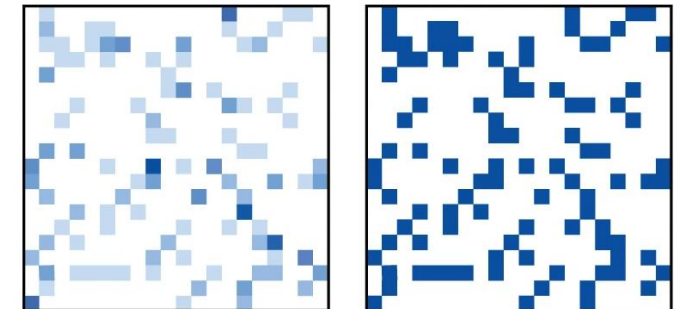


Deep Learning (DL)

Convolutional Neural Network (CNN)

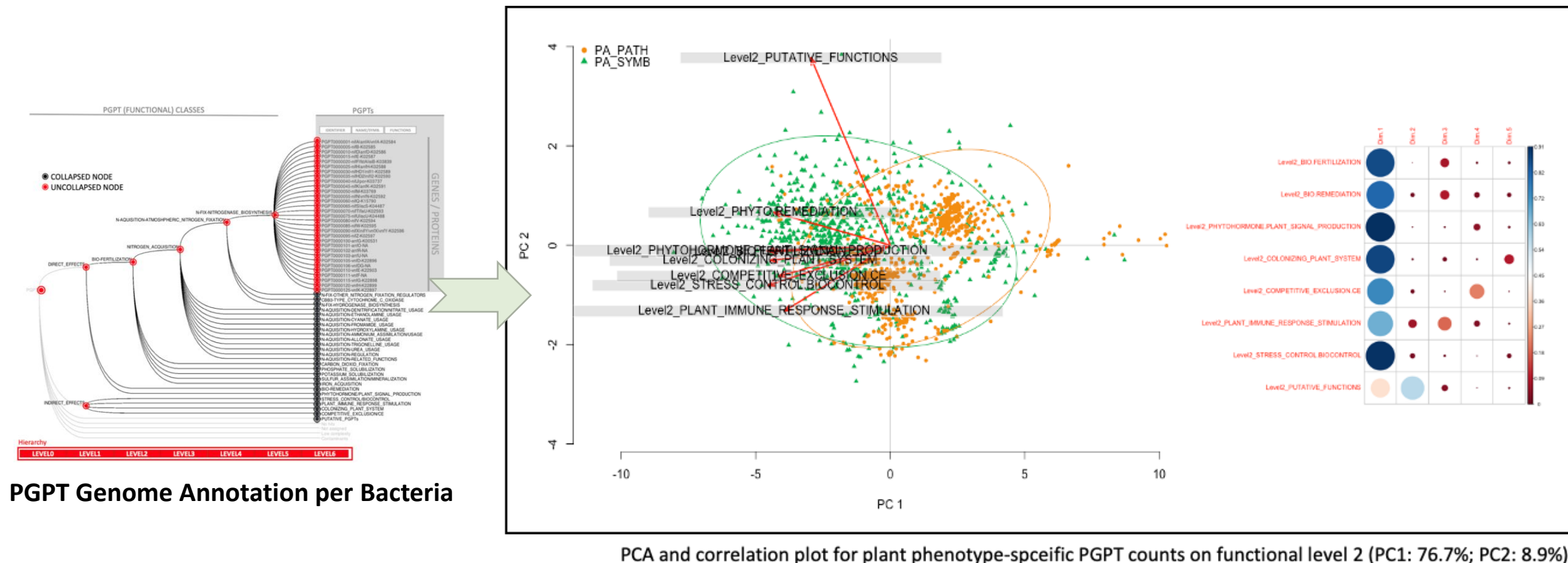


- **Spatial information**, such as **images**
- Inductive capabilities **by summarizing local structure**
- Learns features from input data
- Nguyen et al. rendered an **OTU table into an image**
 - by reshaping **each sample into a square**
 - where **each pixel was colored based on the abundance or presence** of microbial taxa
- Using **inherent phylogenetic Information**
 - taxoNN **rearranges an OTU table based on its inherent phylogenetic** information
 - PopPhy-CNN populates a **phylogenetic tree with OTU abundances** transformed into 2-D matrix
- Generally, these approaches outperformed their benchmarks (traditional ML or FCNNs) in host phenotype prediction



Example CNNs

Prediction of Symbiotic and Pathogenic Bacteria by Targeted Gene Annotation

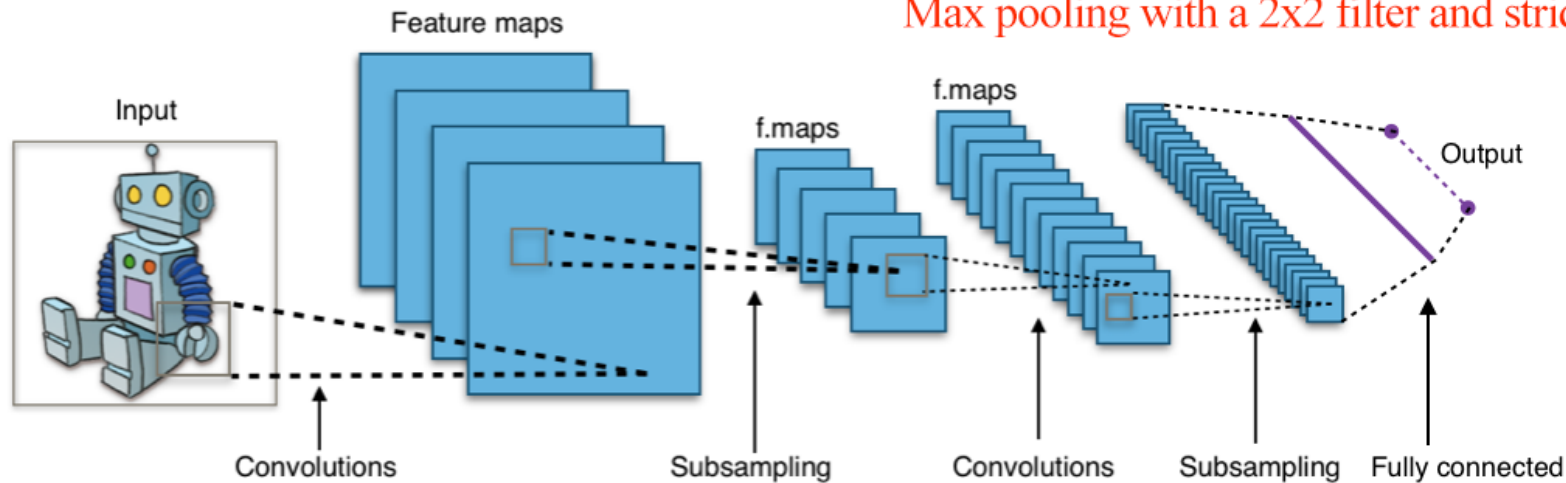
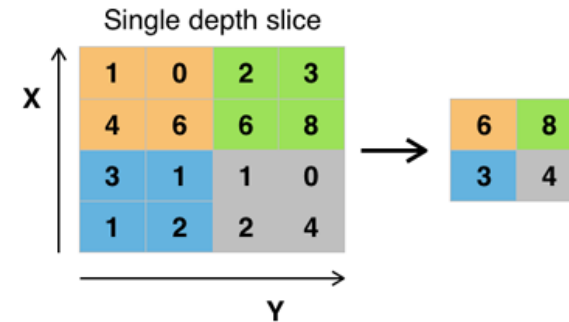


Example CNNs

Prediction of Symbiotic and Pathogenic Bacteria by Targeted Gene Annotation

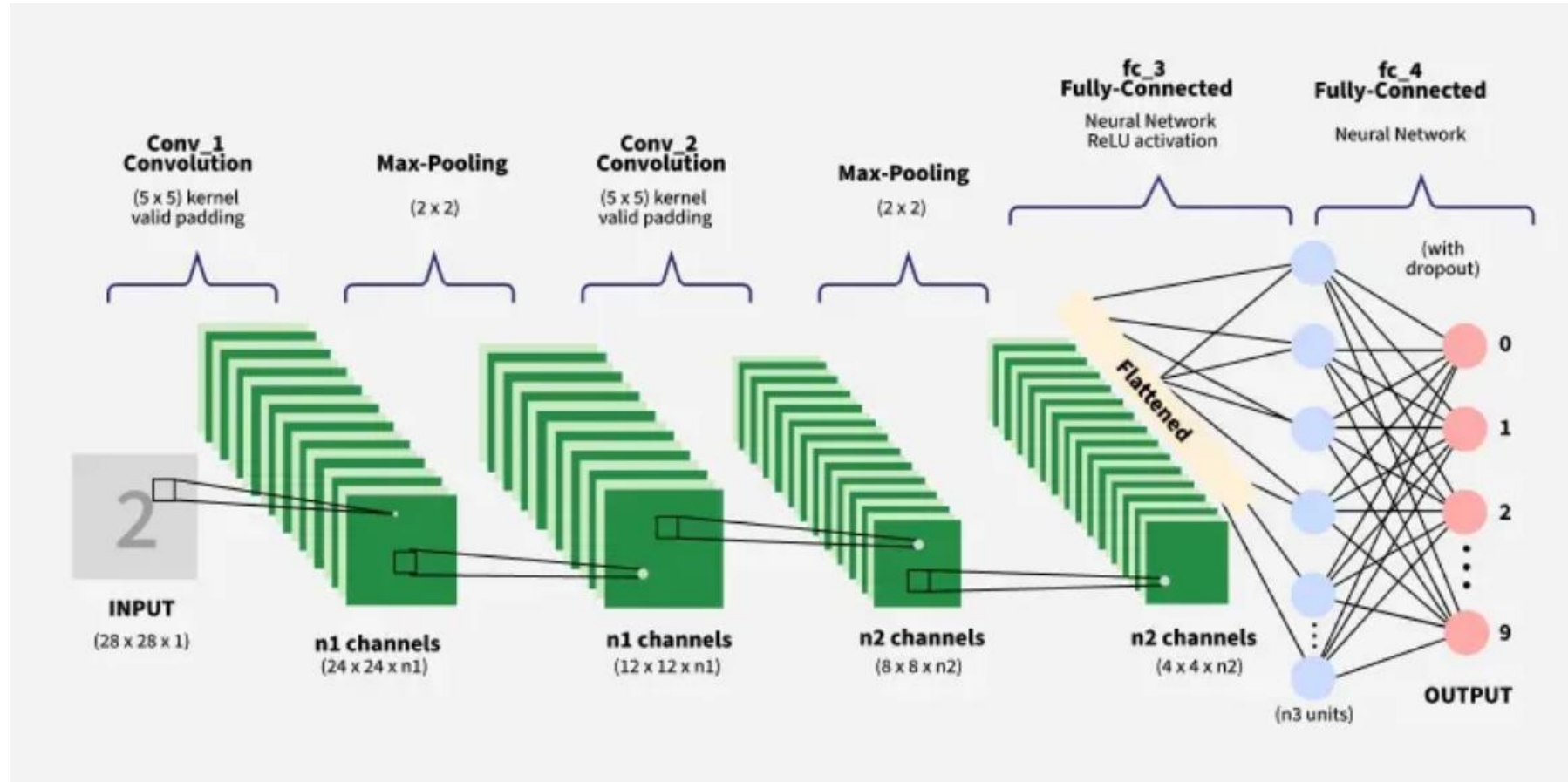
PyTorch

CNN



Example CNNs

Prediction of Symbiotic and Pathogenic Bacteria by Targeted Gene Annotation

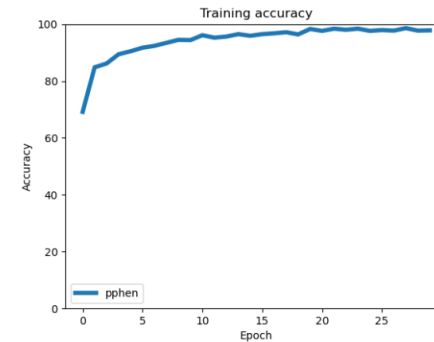


Example CNNs

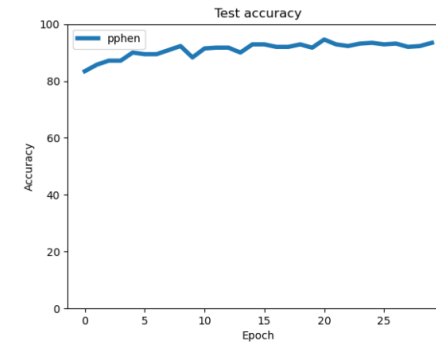
Prediction of Symbiotic and Pathogenic Bacteria by Targeted Gene Annotation



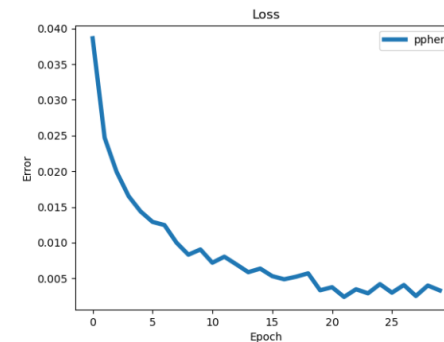
Pickle File Genome
Gene Count 2D



(a) Training accuracy of only pphen



(b) Test accuracy of only pphen



(c) Loss error of only pphen



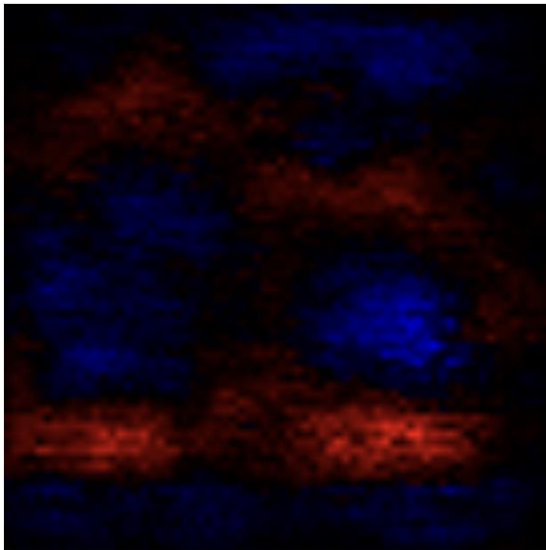
(d) Accuracy of pphen classes in the test dataset

Figure S.1: Graphs for the experiment with only pphen being trained

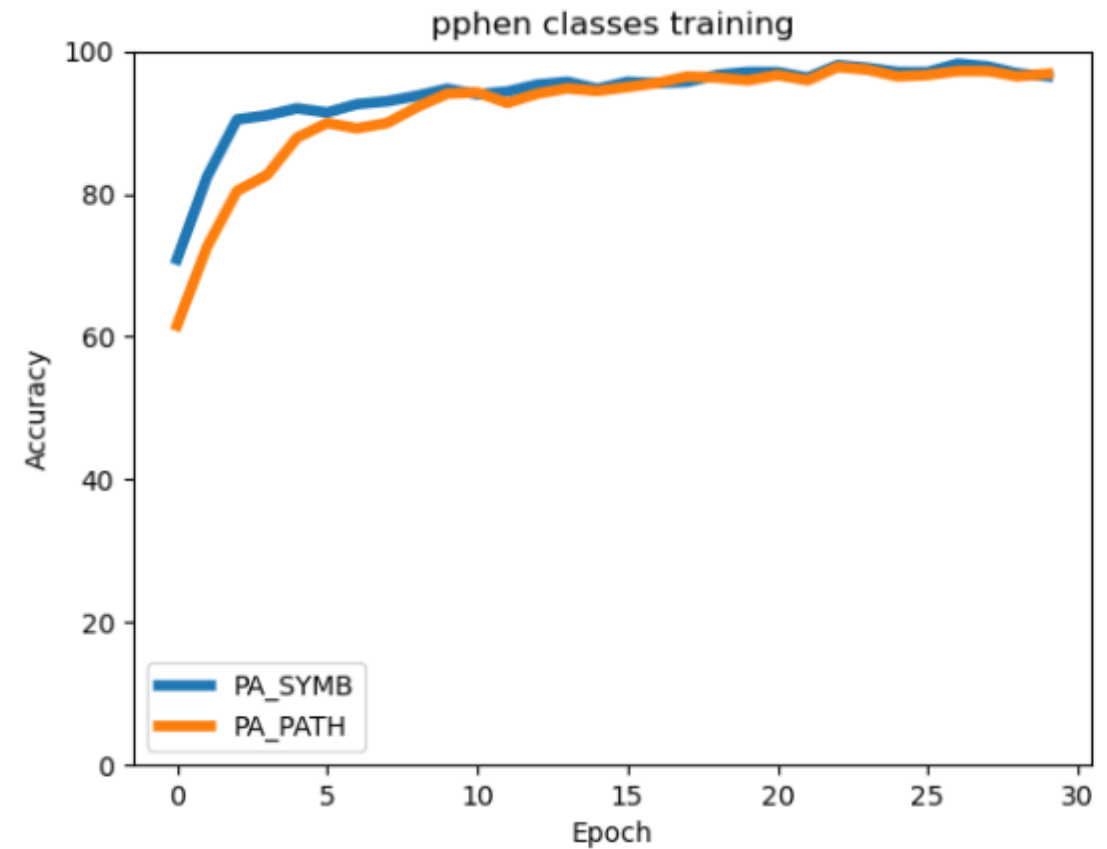
Strain-Level – Taxonomic Independent

Example CNNs

Prediction of Symbiotic and Pathogenic Bacteria by Targeted Gene Annotation

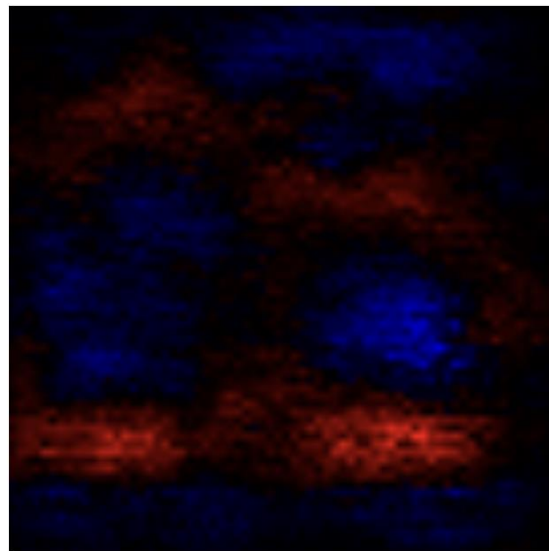


(b) Saliency map of PA_SYMB with z-score normalised abundance as input

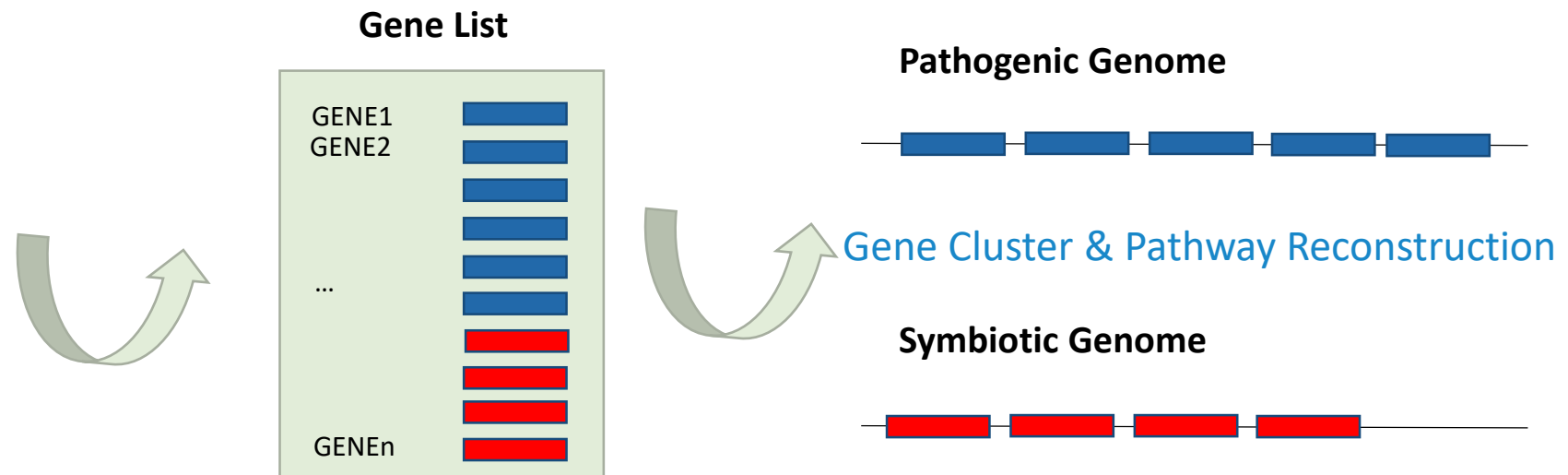


Example CNNs

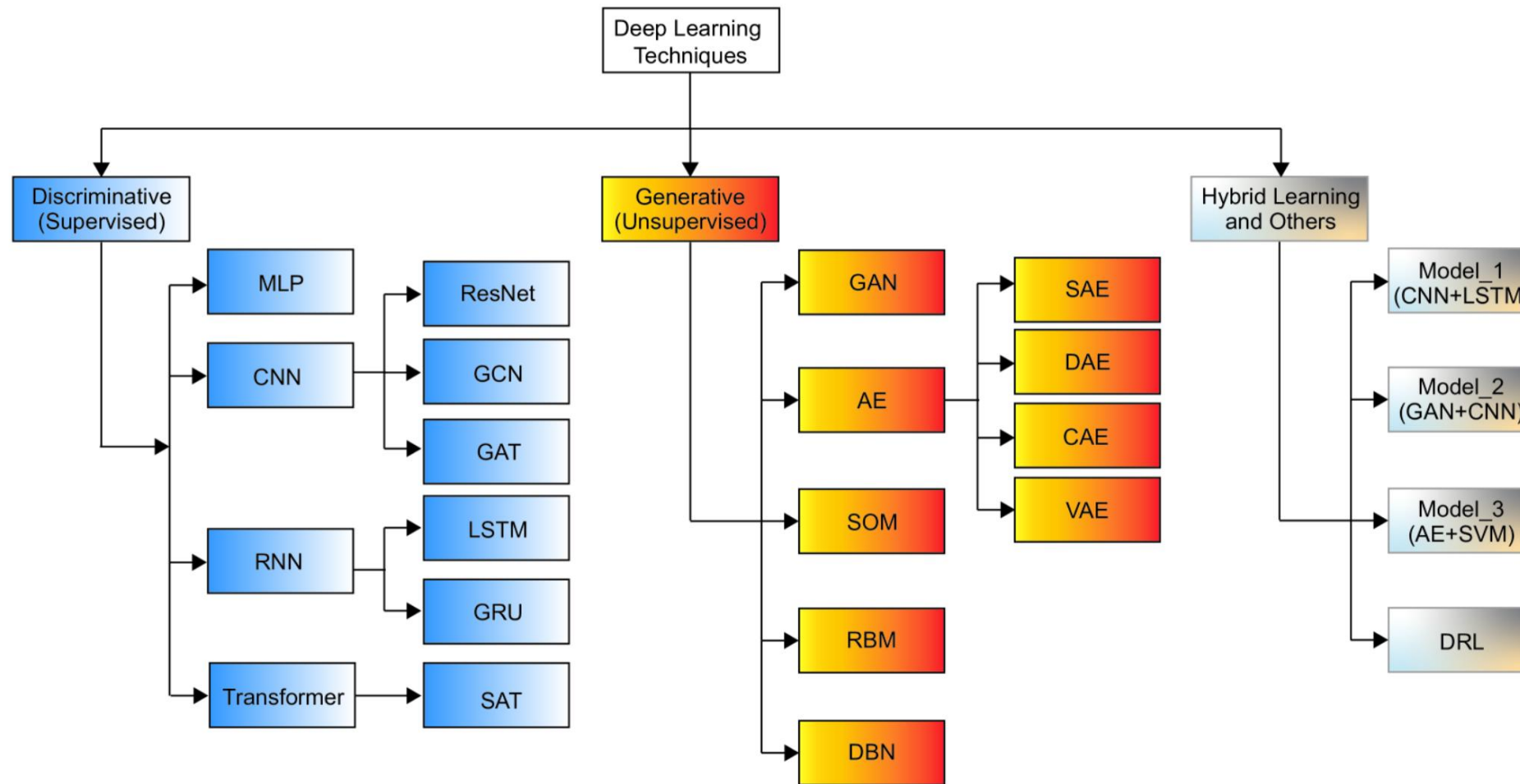
Prediction of Symbiotic and Pathogenic Bacteria by Targeted Gene Annotation



(b) Saliency map of PA_SYMB with z-score normalised abundance as input



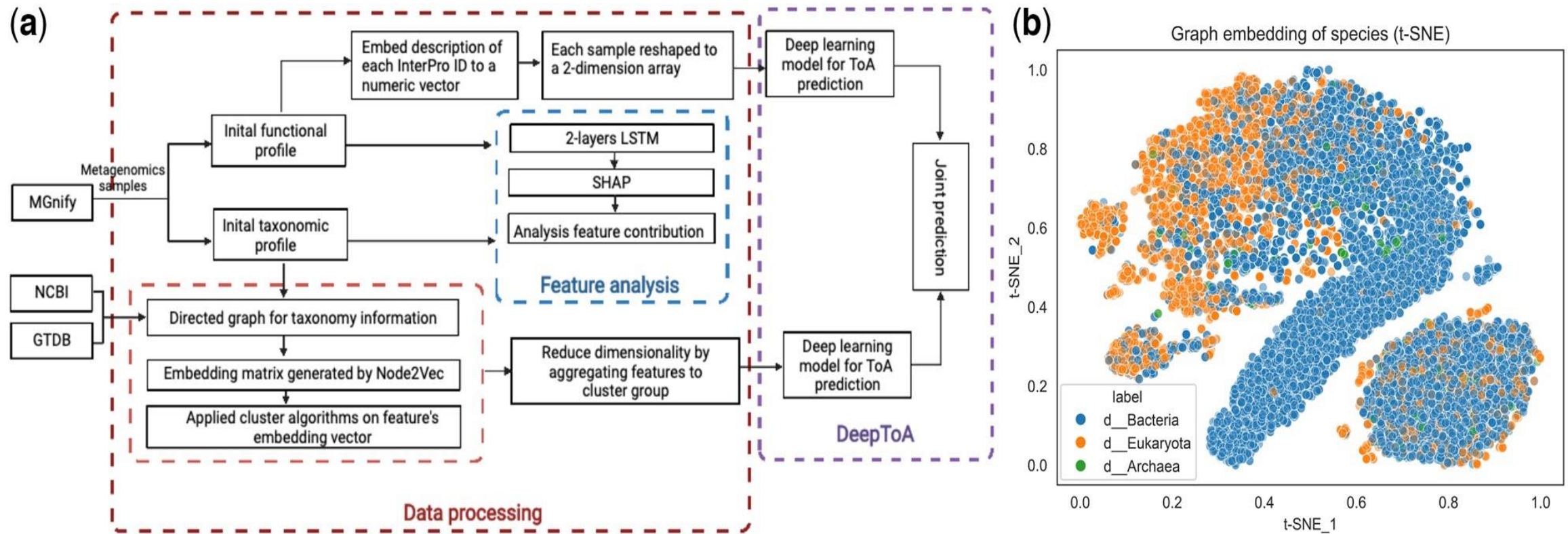
Mixed Model Frameworks



MLP: Multi-Layer Perceptron; CNN: Convolutional Neural Network; ResNet: Residual Neural Network; GCN: Graph Convolutional Network; GAT: Graph Attention Network; RNN: Recurrent Neural Network; LSTM: Long Short-Term Memory; GRU: Gated Recurrent Unit; SAT: Structure-Aware Transformer; GAN: Generative Adversarial Network; AE: Auto-Encoder; SAE: Sparse Autoencoder; DAE: Denoising Autoencoder; CAE: Contractive Autoencoder; VAE: Variational Autoencoder; SOM: Self-Organizing Map; RBM: Restricted Boltzmann Machine; DBN: Deep Belief Network; DRL: Deep Reinforcement Learning.

Example Mixed Models: CNN-LSTM

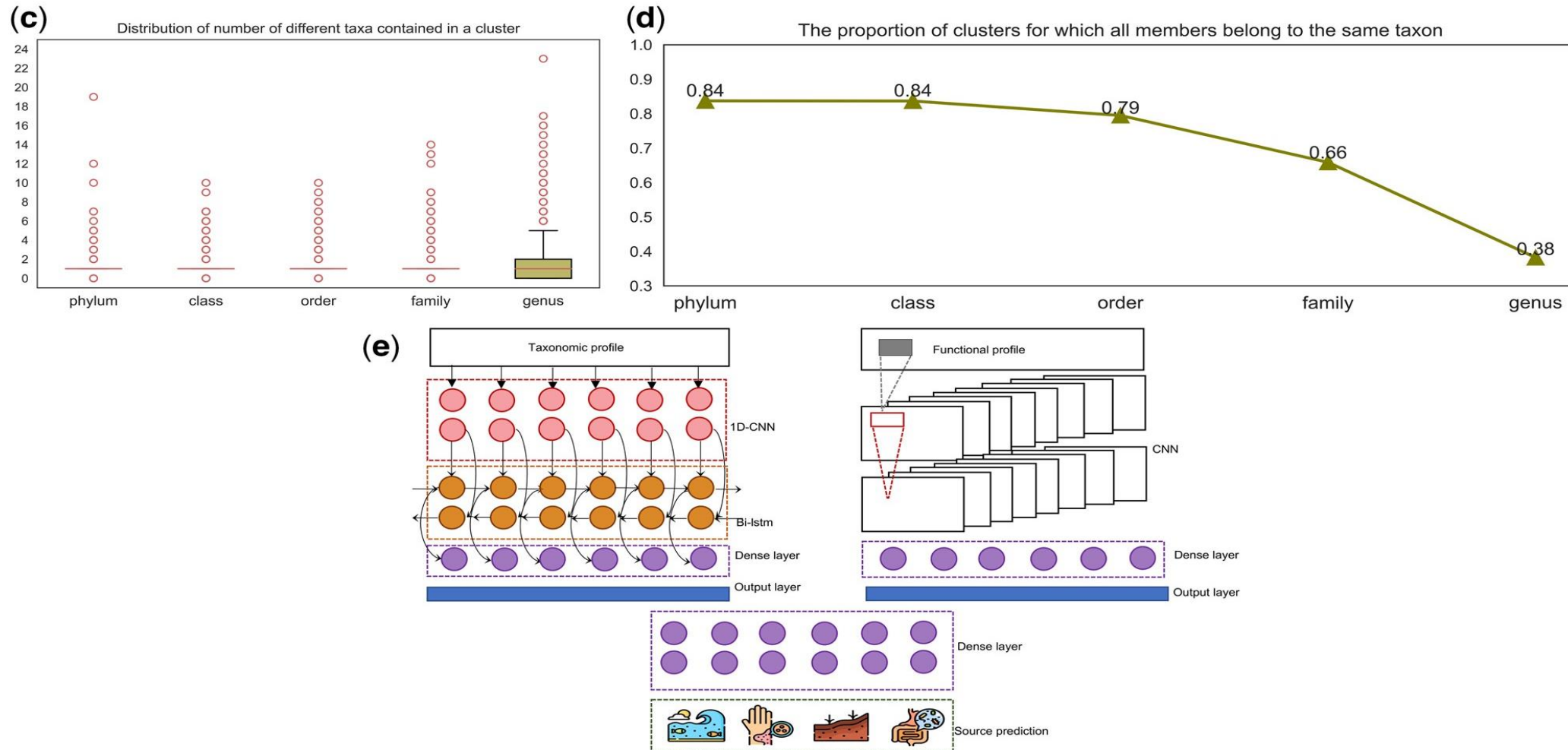
DeepToA: an ensemble deep-learning approach to predicting the theater of activity of a microbiome



LSTM: Long Short-Term Memory

Example Mixed Models: CNN-LSTM

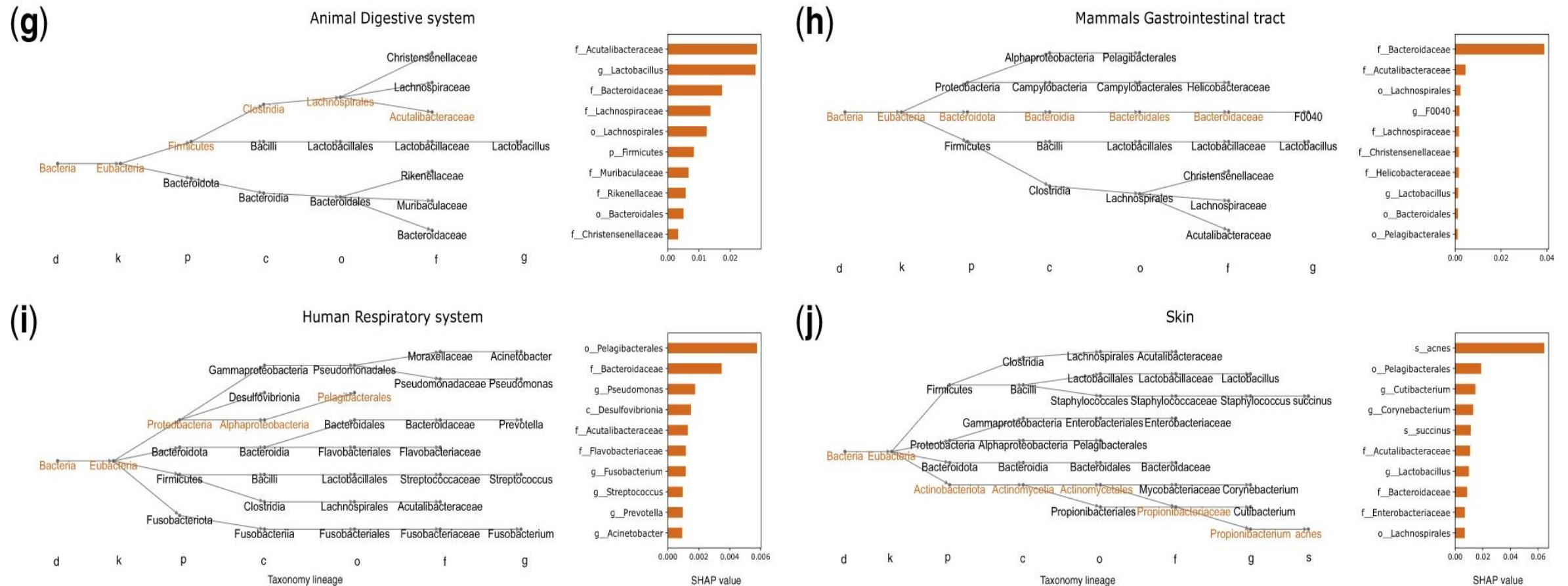
DeepToA: an ensemble deep-learning approach to predicting the theater of activity of a microbiome



LSTM: Long Short-Term Memory

Example Mixed Models: CNN-LSTM

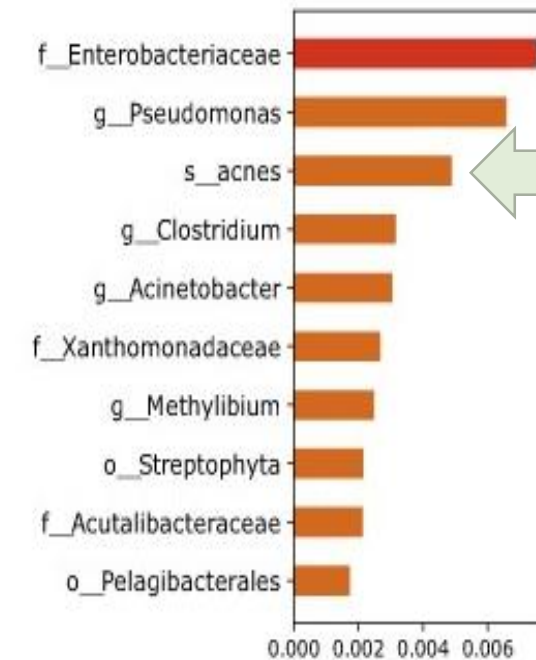
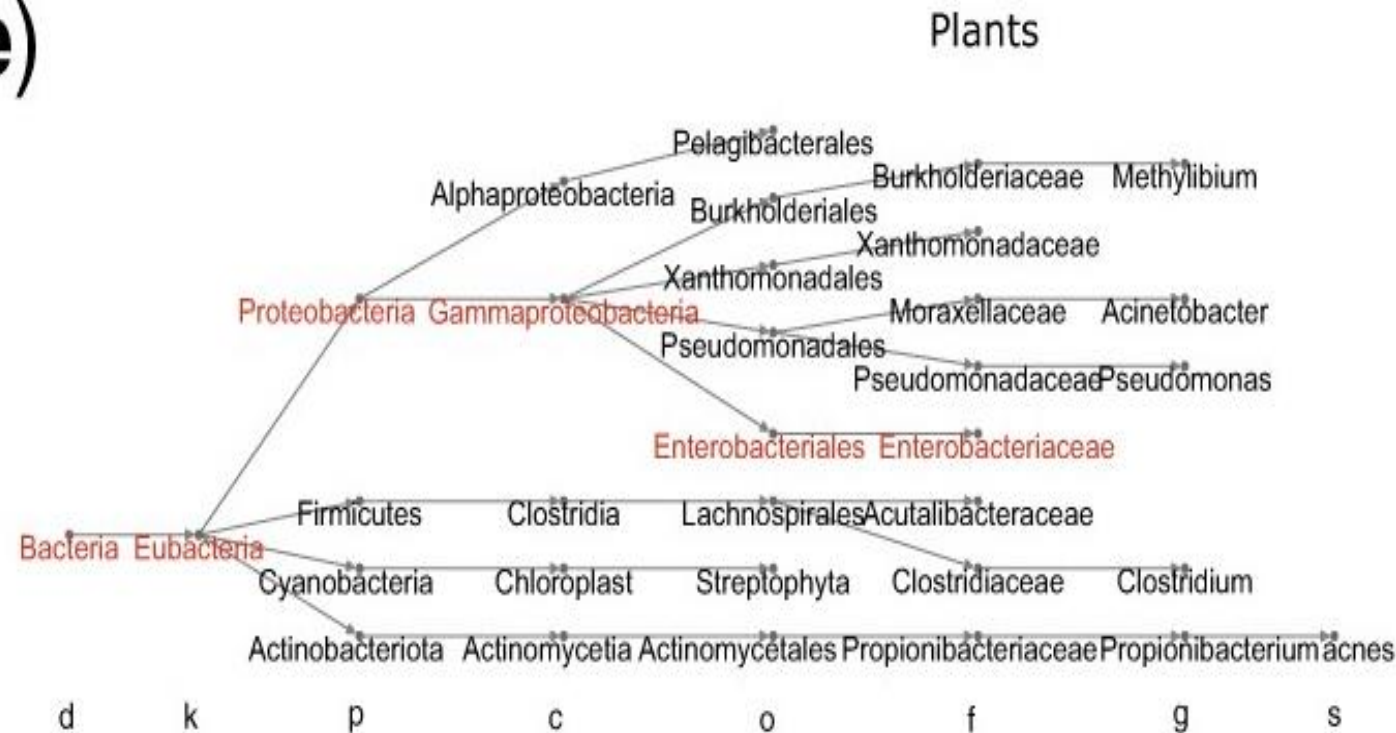
DeepToA: an ensemble deep-learning approach to predicting the theater of activity of a microbiome



Example Mixed Models: CNN-LSTM

DeepToA: an ensemble deep-learning approach to predicting the theater of activity of a microbiome

(e)



Example Mixed Models: CNN-LSTM

DeepToA: an ensemble deep-learning approach to predicting the theater of activity of a microbiome

Model	Dataset	Precision	Recall	F1-score	AUC	MCC	Acc
Bi-LSTM	Initial taxonomic profile	0.9161	0.9119	0.9131	0.9732	0.9368	0.9524
Bi-LSTM	Initial functional profile	0.9287	0.9305	0.9288	0.9735	0.9378	0.9524
Conv1D+LSTM	Processed taxonomic profile	0.9612	0.9345	0.9466	0.9798	0.9568	0.9676
Conv2D	Processed functional profile	0.9348	0.9368	0.9337	0.9813	0.9559	0.9663
Ensemble model	Processed taxonomic profile and initial functional profile	0.9622	0.9331	0.9464	0.9842	0.9628	0.9716
DeepToA	Processed taxonomic profile and processed functional profile	0.9671	0.9638	0.9622	0.9927	0.9742	0.9830

Note: For different deep-learning models and for different choices of dataset, we report the Precision, Recall, F1-score, AUC, MCC, Acc of ToA prediction. The best values are shown in bold.

Data Hungriness

- Demands **voluminous, high-quality, and correctly-labeled data** required
- **Data augmentation**: comprises a set of practices to create synthetic samples
 - Enlarging training dataset
 - Tree-based associative data augmentation (TADA) for low-sample numbers and under-represented classes
 - New OTU samples from an inferred phylogenetic tree
- **Transfer learning & hybrid models**
 - to be explored in the context of microbiome research

Data Quality and Heterogeneity

- **Data Source and deficiencies**
- **Biases of the microbiome dataset**
- **Sampling bias:** most samples from western populations/countries

- **Deduplication, class balancing**
- **Outlier removal and imputation**
- **Dimensionality Reduction:** Problem of Overfitting and poor generalization
-
- BUT: influence performance

- ML models **are tightly dependent on their training** dataset
 - Specific characteristics and scale of the dataset
 - See <https://doi.org/10.1093/bioinformatics/bty949>

 - k-mer analysis tend to be large and redundant
 - Impractical for AI
 - Conversion into a binary matrix via rapid annotation using subsystem technology (RAST)

Model Evaluation, Selection and Tuning (Rigorous validation of AI models)

- **Set of suitable hyperparameters**
 - Use Python and R libraries, such as scikit, PyTorch, Tensorflow, and mlr3
 - High-level frameworks: FastAI , PyTorch Lightning, and Keras
- **Tuning and Development**
 - Synthetic microbiome datasets like provided by the **CAMI consortium**
 - **Synthetic and pre-labeled microbiomes** guide the choice of hyperparameters and model design
 - Starting point for benchmarking and comparison
- Assess **robustness or neutral benchmarking** studies
 - Comparison across multiple datasets
- **Predictive power**
 - Attainable by marrying different data modalities: microbiome, genetic, and environmental data
 - García-Jiménez et al.:
 - **concept of multimodal embedding**
 - through **separate encoders of two modalities** (environmental variables, microbial composition)

Interpretability

- **Features importance** ranking for e.g. deep forest algorithms
- F/CNNs constraining the learning process with *a priori* knowledge, **Saliency Maps**

EXPLORE

- Familiarize / inspect dataset
- Size of the feature space
- Unbalanced classes?
- Imputation or feature Engineering?

BENCHMARKING

- Model benchmarking / tuning
- Data Splitting: Training, Validation, Test sets
- (nested) Cross-Validation
- Metrics for model comparison/ performance

PREDICTION

- Data- and task-dependent
- DL for large-scale or multi-modal data
- Autoencoder incorporating data into embeddings
- Spatial information embedding in input (e.g. phylogenetic tree) into 2D Matrix: CNNs
- Temporal dependencies: RNN framework
- Feature Importance

Acknowledgement

Prof. Dr. Daniel H. Huson
Prof. Dr. Detlef Weigel
Prof. Dr. Lars A. Angenent
Prof. Dr. Thorsten Thünen

Dr. Sebastian Schultheiss (CEO)

PD Dr. Silke Ruppel
Prof. Dr. Nabil Hegazi
Prof. Dr. Steffen Kolb
Prof. Dr. Michelle G. Giglio
Prof. Dr. Mark Wilkinson

Dr. Eva Fornefeld
Dr. Matthias Becker
Dr. Yvonne Becker

Dr. Ramadan F. Abdelfadil
Dr. Vanessa G. Tchuisseu
M.Sc. Michelle Hagen
M.Sc. Rahma
M.Sc. Hend

Your Attention Matters 😊

Thank you for the Invitation!

Prof. Dr. Ivica Dimkić

Prof. Dr. Djordje Bajić

Prof. Dr. Stéphane Compant

Dr. Nemanja Kuzmanović



Bundesministerium
für Landwirtschaft, Ernährung
und Heimat

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN



UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE
INSTITUTE FOR GENOME SCIENCES



CBGP
CENTRO DE BIOTECNOLOGÍA
Y GENÓMICA DE PLANTAS
UPM-INIA

Hernández Medina, R., Kutuzova, S., Nielsen, K.N. *et al.* Machine learning and deep learning applications in microbiome research. *ISME COMMUN.* **2**, 98 (2022).
<https://doi.org/10.1038/s43705-022-00182-9>

Fonseca D.C., da Rocha Fernandes G., Waitzberg D.L. Artificial intelligence and human microbiome: A brief narrative review. *Clin Nutr Open Sci.* 59:134–42 (2025).
<https://doi.org/10.1016/j.nutos.2024.12.009>

Mohseni P., Ghorbani A. Exploring the synergy of artificial intelligence in microbiology: Advancements, challenges, and future prospects. *Computational and Structural Biotechnology Reports*, Volume 1, 100005 (2024), <https://doi.org/10.1016/j.csbr.2024.100005>

Yelin I., Snitser O., Novich G. *et al.* Personal clinical history predicts antibiotic resistance of urinary tract infections. *Nat Med* 25, 1143–1152 (2019).
<https://doi.org/10.1038/s41591-019-0503-6>

Papoutsoglou, G., Tarazona, S., Lopes, M. B., Kammsteiner, T., Ibrahim, E., Eckenberger, J., Novelli, P., Tonda, A., Simeon, A., Shigdel, R., Béreux, S., Vitali, G., Tangaro, S., Lahti, L., Temko, A., Claesson, M. J., & Berland, M. Machine learning approaches in microbiome research: challenges and best practices. *Frontiers in microbiology*, 14, 1261889 (2023). <https://doi.org/10.3389/fmicb.2023.1261889>

Hagen, M., Dass, R., Westhues, C., Blom, J., Schultheiss, S. J., & Patz, S. (2024). Interpretable machine learning decodes soil microbiome's response to drought stress. *Environmental microbiome*, 19(1), 35. <https://doi.org/10.1186/s40793-024-00578-1>

Zeng W., Gautam A., Huson D.H. DeepToA: an ensemble deep-learning approach to predicting the theater of activity of a microbiome. *Bioinformatics*, Volume 38, Issue 20, Pages 4670–4676 (2022). <https://doi.org/10.1093/bioinformatics/btac584>

Karwowska, Z., Aasmets, O., Estonian Biobank research team. *et al.* Effects of data transformation and model selection on feature importance in microbiome classification data. *Microbiome* **13**, 2 (2025).

<https://doi.org/10.1186/s40168-024-01996-6>

Jayakrishnan, T.T., Sangwan, N., Barot, S.V. *et al.* Multi-omics machine learning to study host-microbiome interactions in early-onset colorectal cancer. *npj Precis. Onc.* **8**, 146 (2024).

<https://doi.org/10.1038/s41698-024-00647-1>

Assia Mairi, Lamia Hamza & Abdelaziz Touati. [Artificial intelligence and its application in clinical microbiology](#). *Expert Review of Anti-infective Therapy* 23:7, pages 469-490 (2025).

<https://doi.org/10.1080/13102818.2024.2349587>

