Machine Learning at Genome Scale

Summary of the PhD Thesis Points:

1. Improving microbiome-based disease prediction with SuperTML and data augmentation
2. Supervised Multiple Kernel Learning approaches for multi-omics data integration
3. MINN: A Metabolic-Informed Neural Network for Integrating Omics Data into Genome-Scale Metabolic Modeling

Gabriele Tazza

Supervisor:

László Vidács, Dr. and Tibor Gyimóthy, Prof.

Doctoral School of Computer Science
University of Szeged

Department of Software Engineering

2026

Introduction

The recent advancements in high-throughput technologies, such as Next-Generation Sequencing (NGS) and mass spectrometry, have allowed researchers to generate large amounts of omics data quickly and at a reduced cost. This changed our approach to the study of biological systems. For example, in the past, the research point of view in biology was almost entirely reductionist. The aim was to investigate very specific mechanisms, designing experiments and testing hypotheses. Now, with the increasing availability of data and computational resources, the goal has moved toward a more data-driven approach. Here, the goal is to extract knowledge directly from a large amount of data, which makes it possible to focus on the system in its entirety instead of one mechanism at time. In this context, Machine Learning (ML) and, in particular, Deep Learning (DL) have become particularly useful when the mechanisms of a biological system are unknown or too complex to be approached using only reductionist approaches. However, ML/DL methods come with limitations: they need large quantities of training data to be able to generalize well. In many biology-related applications, datasets can have few samples and a high number of features, which makes it challenging to apply ML/DL models. This issue is known as the ”curse of dimensionality”, a well known problem that can limit the efficacy of ML/DL in the biological field.

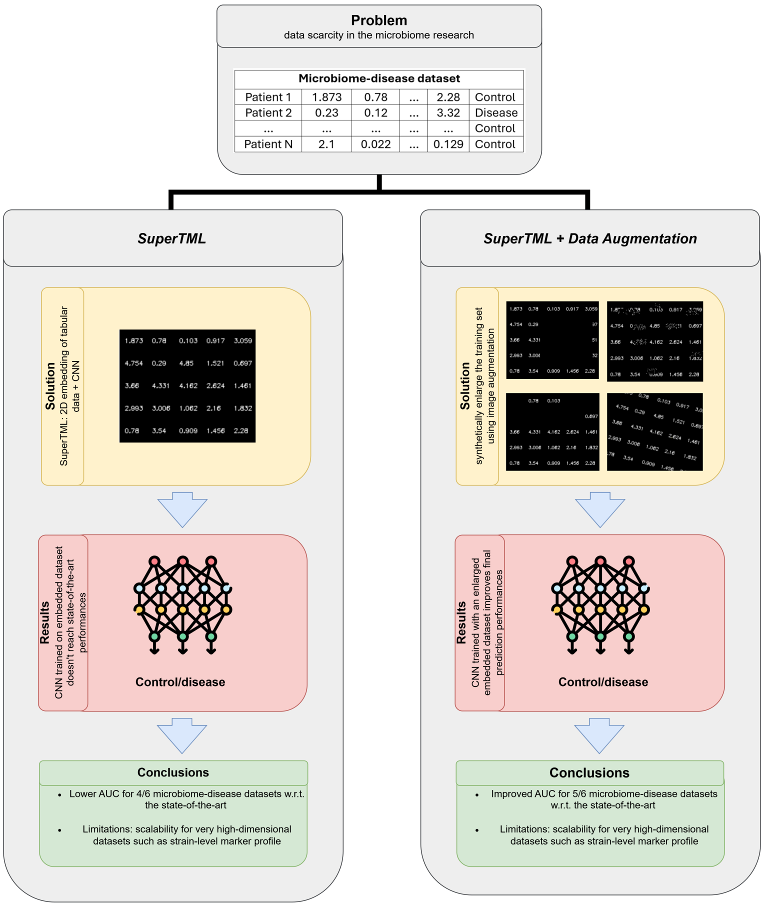
Omics data can be of several types, for example: transcriptomics, proteomics, or epigenomics. Each one represents a different aspect of the biological system, and the approach of simultaneously analyzing and integrating them is referred to as multi-omics analysis. This integration is crucial because it provides a more complete and holistic view of the biological system, which would be impossible using any single omic layers alone. However, multi-omics integration is not always straightforward, as it requires approaches that handle the heterogeneity across layers while keeping the biological relevance of each individual data type and for this reason it remains an open challenge in biology.

The PhD thesis fits into this broader context and is framed within the European project E-MUSE: Complex microbial Ecosystems MUltiScale modElling (https://www.itn-emuse.com/https://www.itn-emuse.com). E-MUSE is a Marie Skłodowska-Curie Action Innovative Training Network focused on developing new methodologies for modeling complex biological systems. The project aims to improve our understanding of microbial ecosystems, with a particular emphasis on fermented food products such as cheese. The Work Package 2 specifically focuses on data-driven approaches. Among these, it explores the ”curse of dimensionality” in omics dataset and multi-omics integration using statistical, network-based, and ML/DL approaches. Within this framework, the PhD thesis explores the use of ML, in particular DL methods to address these challenges. The common theme of the work is the methodological development of ML approaches at genome-scale.

After this Introduction, we divided this summary into three chapters: one for each thesis point. In the first one, we introduce a DL method to analyze metagenomics data in the context of data scarcity. It is based on SuperTML, a method originally used for small tabular datasets, that embeds data into a 2D format, enabling the use of image augmentation techniques. In the second chapter, we summarize the second thesis point. We explore supervised multiple kernel learning (MKL) methods for multi-omics data integration and present a novel MKL method based on DL optimization, called DeepMKL. In addition, we present a novel feature importance method for biomarker discovery. In the third chapter, we presented the third thesis point: a hybrid framework that combines data-driven and mechanistic approaches to predict metabolic fluxes in the context of data scarcity. The proposed method addresses this problem by regularizing a neural network through the incorporation of prior knowledge from genome-scale metabolic modeling, which reduces the amount of data needed for training.
Each chapter contains an Author's Contribution section, which details the exact contribution of the author for that thesis point. These chapters are then followed by a summary in Hungarian.

Improving microbiome-based disease prediction with SuperTML and data augmentation

The first thesis point addresses a common issue in bioinformatics: the scarcity and the high dimensionality of the datasets. Specifically, it addresses this problem in the microbiome research domain. The human microbiome is the complex of all the microorganisms that live in our body, especially in the gut. And it is linked with several health aspects. Alterations in its composition appear in conditions such as diabetes, inflammatory bowel disease, obesity, liver cirrhosis, and colorectal cancer [1, 3]. In microbiome research, the availability of data is still limited, and this makes it challenging to adopt traditional deep learning models such as feed-forward neural networks because they tend to fall into the "curse of dimensionality" and overfit.


Problem
data scarcity in the microbiome research
SuperTML
SuperTML + Data Augmentation
Solution
SuperTML: 2D embedding of tabular data + CNN
Solution
synthetically enlarge the training set using image augmentation
Results
CNN trained on embedded dataset doesn't reach state-of-the-art performances
Control/disease
Results
CNN trained with an enlarged embedded dataset improves final prediction performances
Control/disease
Conclusions
Conclusions

Microbiome-disease dataset

Patient 1	1.873	0.78	...	2.28	Control
Patient 2	0.23	0.12	...	3.32	Disease
...	...	...	...	...	Control
Patient N	2.1	0.022	...	0.129	Control

Conclusions

- Lower AUC for 4/6 microbiome-disease datasets w.r.t. the state-of-the-art
- Limitations: scalability for very high-dimensional datasets such as strain-level marker profile

Conclusions

- Improved AUC for 5/6 microbiome-disease datasets w.r.t. the state-of-the-art
- Limitations: scalability for very high-dimensional datasets such as strain-level marker profile

Figure 1: *High-level view of SuperTML for microbiome-disease prediction: 2D embedding of tabular data and CNN for control/disease predictions. Left branch: SuperTML without augmentation, which achieves lower AUC scores for 4 out of 6 datasets compared to state-of-the-art methods. Right branch: the augmented SuperTML framework, where image augmentation improves performance, achieving higher AUC scores for 5 out of 6 datasets.*

In recent years, an alternative deep learning approach, SuperTML [10], showed promising performances with small tabular datasets. It is based on a 2D embedding of the feature vector into an image: the values are printed on a black background, and then a CNN is employed for the downstream classification task. Additionally, this framework enables the use of image augmentation as in a standard image processing task [?]. Image augmentation is a method to synthetically enlarge the dataset dimension and for this reason is particularly useful in our case.
In our work, we applied SuperTML and the use of image augmentation to six different microbiome datasets, where the task was to predict the absence or presence of a certain disease based on metagenomics features. We evaluated our approach against classical feedforward neural networks and the state-of-the-art DeepMicro. The datasets used in our analysis and their characteristics are shown in Table 1.

Table 1: *Summary of disease datasets.*

Disease	**Dataset**	**# Samples**	**# Controls**	**# Patients**	**# Features**
Inflammatory Bowel Disease	IBD	110	85	25	443
Type 2 Diabetes	EW-T2D	96	43	53	381
Type 2 Diabetes	C-T2D	344	174	170	572
Obesity	Obesity	253	89	164	465
Liver Cirrhosis	Cirrhosis	232	114	118	542
Colorectal Cancer	Colorectal	121	73	48	503

In our experiments, we tried several Image Augmentation transformations and also implemented a custom one, called CellDropout, designed specifically for the kind of image generated by SuperTML.
The results showed that SuperTML, even when used alone, consistently outperformed FNNs across five out of six microbiome datasets. Additionally, when enhanced with augmentation, SuperTML achieved the highest AUC scores in five out of six datasets, demonstrating its competitive performance. Despite the clear efficacy of image augmentation, we could not observe a single transformation that consistently worked better in the majority of the datasets. This leaves the open question about which kind of transformations are more appropriate for this kind of representation.
In this thesis point, we also analyze the limitations of SuperTML and disscuss possible future directions. First, the dimension of the image grows with the number of features. This poses an important computational limit for high dimensional datasets as multi-omics ones. In order to mitigate this, a potential first approach could be to add a dimensionality reduction step, such as an autoencoder, on top of our framework. Finally, another limitation is the black-box nature of this approach. The 2D embeddings make it harder to retrieve the feature importance. This issue could be mitigated using pixel-level attribution methods to identify the most relevant pixels and relate them to features, which would be useful for biomarker discovery and future applications in precision medicine.

Author’s contributions

For the first thesis point, the author is responsible for contributing to the conceptualization and design of the work, the idea of transforming microbiome–disease classification tasks into image classification using SuperTML and comparing the results with DeepMicro. The author conducted the full literature survey related to SuperTML and data augmentation, and implemented all experiments presented in the chapter. In addition, the author designed and developed the novel CellDropout transformation entirely from scratch.

Supervised Multiple Kernel Learning approaches for multi-omics data integration

In this second thesis point, we address the challenge of multi-omics integration. Omics datasets are high dimensional and heterogeneous, and this creates several challenges when trying to combine them into a single predictive framework. Kernel methods, however, offer a natural and elegant way to deal with this scenario, because they represent each dataset in terms of a matrix of pairwise similarities and provide a nonlinear version of many algorithms. Despite this, kernel approaches remain an underused tool in genomic data mining [14]. In our work, we explore different Multiple Kernel Learning (MKL) strategies, starting from classical convex linear combinations of kernels. We consider simple approaches such as assigning equal weights to each kernel, and also supervised MKL methods where the weights are optimized to improve classification performance. At the same time, we adapt an unsupervised fusion method, STATIS-UMKL [8], to a supervised setting by using the fused meta-kernel as input for an SVM classifier. This method constructs a consensus kernel by maximizing the average similarity between kernels, and we test whether this unsupervised fusion can still provide benefits when used in a supervised classification task.

Additionally, we investigate the possibility of integrating kernels using deep learning. Recent deep learning architectures have shown the ability to learn homogeneous representations from heterogeneous sources, making them suitable for multi-omics integration. With this motivation, we introduce DeepMKL, a framework that first ap-

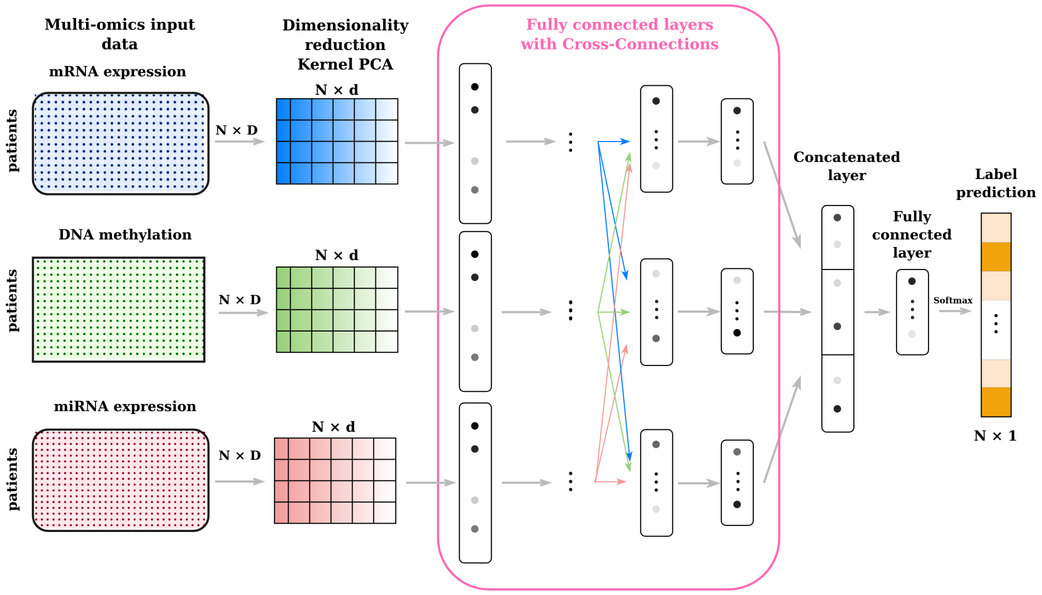


Figure 2: *Cross-modal Deep MKL (concat) takes in input the Kernel PCA dense embeddings of different omics datasets. It extracts the features using different feedforward sub-networks that are linked by cross-connections, then fuses the learnt representations by concatenating them for the final classification*

plies Kernel PCA to each omic dataset to obtain dense embeddings, and then uses a multi-modal neural network to learn separate representations and integrate them for classification. We also propose a second architecture, Cross-modal DeepMKL (Figure 2), where cross-connections between modalities allow information exchange before the final fusion, potentially improving the representation learning. To evaluate all these methods, we tested them on four publicly available multi-omics datasets widely used in biomedical machine learning: ROSMAP, BRCA, LGG, and KIPAN. Each dataset contains three omics layers (mRNA, DNA methylation, and miRNA). To ensure a fair comparison with state-of-the-art methods such as MOGONET and Dynamics[6, 13]. we applied the same evaluation pipeline and the same performance metrics. The results show that kernel-based methods are consistently comparable and often outperform the state-of-the-art deep learning approaches. Methods such as SimpleMKL [9], SEMKL [15], and STATIS-UMKL combined with SVM achieved strong and stable results across the datasets, in some cases surpassing both MOGONET and Dynamics. Even SVM applied to early concatenation performed better than previously reported once proper tuning and kernel choices were applied. DeepMKL and Cross-modal DeepMKL delivered competitive performance on larger datasets, while, as expected, their effectiveness was reduced on smaller ones, confirming the known limitations of deep architectures in low-sample scenarios. Additionally, we introduce a two-step framework for biomarker discovery based on DeepMKL. In the first step, we apply Integrated Gradients [11] to identify the most influential kernel principal components for the classification task. In the second step, we use KPCA-IG [2] to trace back the contribution of the original variables, making it possible to recover biologically relevant features from a transformed representation that normally hides the input structure. This combination allows us to derive consistent and meaningful biomarker candidates, as shown in our experiments on BRCA and ROSMAP. In conclusion, this thesis point highlights how MKL represents a fast and reliable solution that can compete with more complex deep learning architectures. The integration of deep learning with MKL and its interpretability pipeline offers a promising direction for future work in multi-omics data analysis, biomarker discovery, and precision medicine.

Author’s contributions

For the second thesis point, the author is responsible for contributing to the conceptualization and design of the work, specifically comparing deep learning state-of-the-art approaches to multiple kernel learning ones, based on SVM and deep learning, on biomedical multi-omics datasets. The author conducted the literature survey regarding the deep learning approaches in the Related Work section and implemented all data preprocessing steps. Furthermore, the author carried out the conceptualization and implementation of all DeepMKL architectures together with the relative evaluation pipeline and experiments. Finally, the author conceptualized and implemented the novel two-step interpretability method used for biomarker discovery.

MINN: A Metabolic-Informed Neural Network for Integrating Omics Data into Genome-Scale Metabolic Modeling

In this chapter, we address the challenge of hybrid modeling in the context of genome-scale metabolic modeling. Mechanistic models, such as GEMs and FBA, offer a structured way to describe metabolism and simulate flux distributions under stoichiometric constraints [12]. However, their predictive power is limited by incomplete biological knowledge and by the presence of many feasible solutions. On the other hand, data-driven models can extract complex patterns from multi-omics data, but they require large datasets and often lack interpretability. For this reason, hybrid models have recently gained attention. They combine the strengths of mechanistic models with the prediction power of pure data-driven models. Recently, an approach that leveraged GEM structures and FBA constraints within neural networks to predict growth rates from media compositions came out, opening the way for the implementation of new hybrid models [4] .

In this work, we present a Metabolic-Informed Neural Network (MINN), a method inspired by [4] , designed to use multi-omics data to predict metabolic fluxes under different growth rates and single-gene knockouts. The architecture is composed of a

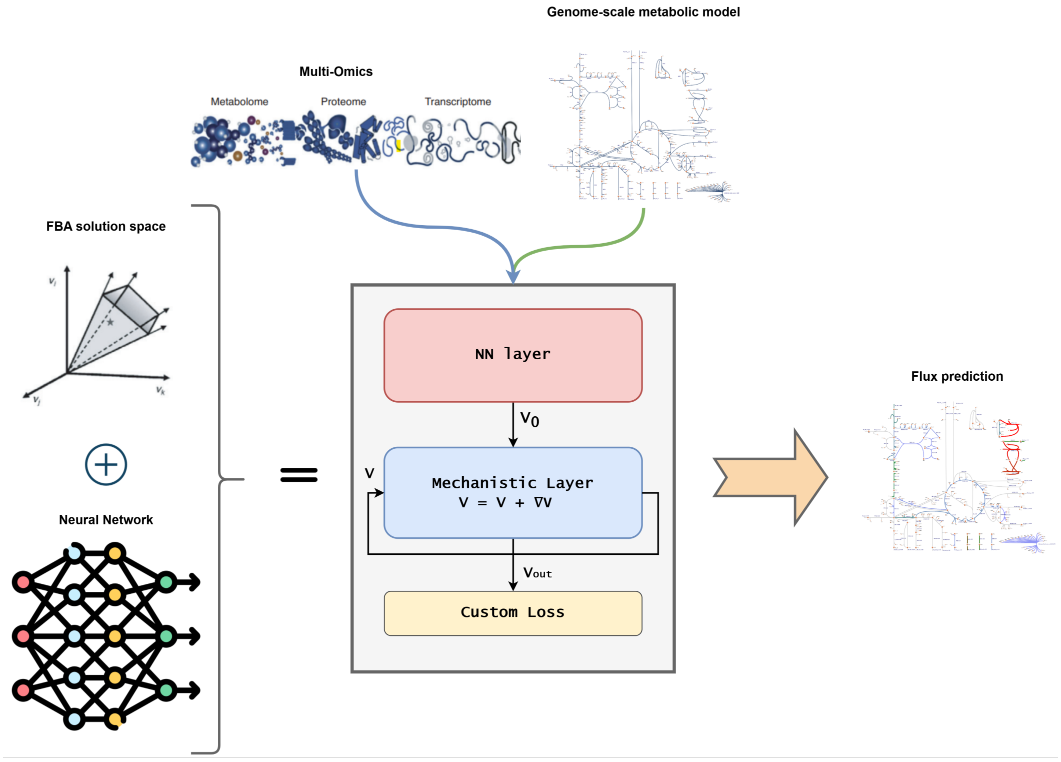


Figure 3: *High-level scheme of the MINN framework*

first, fully data-driven neural network that predicts an initial flux distribution, and a mechanistic layer that refines this prediction by enforcing the FBA constraints. A key issue that emerged in our analysis is the conflict between the data-driven objective and the mechanistic one. To mitigate this, we introduce different strategies. One of them is a balancing coefficient that controls the weight of the data-driven component. Additionally, we also explore other hybrid optimization strategies, including a bound on the mechanistic loss, an adaptive loss balancing mechanism, and a dynamic loss weight scheduler. In our experiments, we evaluate MINN on the ISHII dataset[7], which contains transcriptomics, proteomics and fluxomics measurements for E. coli across different growth conditions and knockout strains. We compare the MINN with pure machine learning models and with pFBA [5]. The results show that the MINN achieves comparable or better predictive performance than the machine learning baselines, and also outperforms pFBA. The configuration with the balanced loss shows the most stable and robust predictions. We also introduce MINN-reservoir, a variation trained in two steps. First, a MINN without omics inputs is trained to approximate an FBA solver, producing a pretrained block that predicts flux distributions directly from exchange fluxes. Then, this block is embedded into a new architecture that uses multi-omics data to generate additional constraints for pFBA. This approach allows the model to generate constraints in a data-driven way, enriching pFBA with information derived from omics data. In our experiments, the MINN-reservoir slightly improves the predictive metrics and significantly reduces the variance across the splits, showing more stable behavior. To conclude, this chapter shows that MINN and its variants improve predictive performance, stabilize the learning process, and offer a flexible framework to integrate data-driven and mechanistic components.

Author’s contributions

For the third thesis point, the author is responsible for contributing to the conceptualizing and designing the MINN architecture for integrating multi-omics into genome-scale metabolic modeling. The author conducted the literature survey regarding the hybrid modeling methods in the Related Work section. In addition, the author implemented the MINN and MINN-reservoir experiments and the related code used in this analysis, except for the hybrid optimization strategies and the code strictly related to handling the mechanistic aspects.

Összefoglalás

A doktori disszertáció genomléptékű gépi tanulási megközelítéseket mutat be. Különösen olyan módszerekre összpontosít, amelyek a bioinformatikában gyakori adatritkaság és multi-omikai integráció problémáját kezelik. A dolgozat három fő tézispontból áll. Az első részben a mikrobiom adatokból történő betegségpredikció kihívásait vizsgáltuk. Ezekben az adatokban a kis mintaszám és a nagy dimenziószám gyakran korlátozza a hagyományos neurális hálózatok teljesítményét. Ennek kezelésére a SuperTML keretrendszert teszteltük, amelyet eredetileg kis táblázatos adathalmazokhoz fejlesztettek ki. Eredményeink azt mutatják, hogy a SuperTML életképes alternatívát jelent a legkorszerűbb módszerekkel szemben, és teljesítménye tovább javul, ha adataugmentációs technikákkal kombináljuk. A vizsgált adathalmazok többségében ez a megközelítés felülmúlta a hagyományos modelleket, ami rámutat az augmentáció regularizációs szerepének fontosságára. A második részben a multi-omikai adatok integrációjának kihívását elemeztük. Ezen adatforrások sokfélesége és összetettsége gyakran csökkenti a hagyományos bioinformatikai módszerek hatékonyságát. Ennek megoldására két új, Multiple Kernel Learning (MKL) alapú megközelítést vezettünk be. Az MKL viszonylag ritkán használt módszer, ugyanakkor nagy potenciállal rendelkező keretrendszert kínál erre a feladatra. Olyan megoldást dolgoztunk ki, amelyben Support Vector Machine módszer segítségével felügyelet nélküli tanulási módszereket alakítunk át felügyelt tanulási feladattá. Emellett bemutattuk a DeepMKL-t, egy mélytanulás-alapú keretrendszert, amely a kernel függvények integrációját konvex lineáris optimalizálás nélkül valósítja meg. Négy nyilvánosan elérhető orvosbiológiai adathalmazon végzett kísérletek azt mutatták, hogy mindkét megközelítés megbízható és versenyképes megoldást jelent. Teljesítményük összevethető, sőt esetenként jobb, mint a bonyolultabb, legkorszerűbb módszereké. Ezen felül egy kétlépéses biomarker-azonosítási stratégiát is javasoltunk, amely a DeepMKL-t egy új magyarázhatósági eljárással kombinálja. Kísérleteink alapján az MKL módszer robusztus megoldást nyújt a multi-omikai integrációra, és versenyképes alternatívát kínál a fejlettebb architektúrákkal szemben. A harmadik részben a Metabolic-Informed Neural Network (MINN) módszert mutatjuk be, amely egy hibrid keretrendszer. Célja, hogy multi-omikai adatokat genomléptékű metabolikus modellekkel integráljon fluxuspredikcióhoz. A kizárólag adatalapú és a kizárólag mechanisztikus megközelítésekkel szemben a MINN egyesíti a neurális hálózatok rugalmasságát a metabolikus modellek strukturált korlátaival. Az architektúra több változatát teszteltük, hogy kezeljük az előrejelzési pontosság és a biológiai konzisztencia közötti kompromisszumot. Emellett stratégiát javasoltunk a MINN parsimonious Flux Balance Analysis (pFBA) módszerrel való összekapcsolására az értelmezhetőség növelése érdekében. Egy kisméretű, E. coli multi-omikai adathalmazon, amely egyszeres génkiütéses törzseket tartalmazott, a MINN felülmúlta mind a klasszikus gépi tanulási módszereket, mind a mechanisztikus modelleket. Eredményeink azt mutatják, hogy a biológiai korlátok bevezetése stabilizálja a tanulási folyamatot és csökkenti a túlillesztést.

List of author’s publications

Journal publications

[J1] **Gabriele Tazza**, Francesco Moro, Dario Ruggeri, Bas Teusink, and László Vidács MINN: A metabolic-informed neural network for integrating omics data into genome-scale metabolic modeling. In *Computational and Structural Biotechnology Journal*, 27, 3609–3617, 2025.

[J2] **Gabriele Tazza**, Dario Ruggeri and László Vidács Improving Microbiome-Based Disease Prediction With SuperTML and Data Augmentation . In *IEEE Access*, 13, 144505-144515, 2025.

[J3] Mitja Briscik, **Gabriele Tazza**, László Vidács, Marie-Agnès Dillies and Sébastien Déjean Supervised multiple kernel learning approaches for multi-omics data integration . In *BioData Mining* , 17, 53, 2024.

[J4] Dario Ruggeri, **Gabriele Tazza** and László Vidács Introducing MLOps to Facilitate the Development of Machine Learning Models in Agronomy: A Case Study. In *IEEE Access*, 13, 122059-122070, 2025.

Full papers in conference proceedings

[C1] **Gabriele Tazza**, Francesco Moro, Bas Teusink, and László Vidács Metabolic-Informed Neural Network for Multi-Omics Data Integration. In *Proceedings of the 13th International Conference on Simulation and Modelling in the Food and Bio-Industry, FOODSIM 2024*, Eurosis-ETI, 193-197, 2024.

Journal/Conference	Rank	**chapter 3**	**chapter 4**	**chapter 5**
[C1]	NA			x
[J3]	Q1		x	
[J2]	Q1	x		
[J1]	Q1			x

Table 2: *The connection between the thesis chapters and publications.*

Bibliography

[1] Nasrullah Abbasi, Nizamullah FNU, Shah Zeb, Muhammad Fahad, and Muhammad Umer Qayyum. Machine learning models for predicting susceptibility to infectious diseases based on microbiome profiles. *Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online)*, 3(4):35–47, December 2024.

[2] Mitja Briscik, Marie-Agnès Dillies, and Sébastien Déjean. Improvement of variables interpretability in kernel PCA. *BMC Bioinformatics*, 24(1), July 2023.

[3] Ilseung Cho and Martin J. Blaser. The human microbiome: at the interface of health and disease. *Nature Reviews Genetics*, 13(4):260–270, March 2012.

[4] Léon Faure, Bastien Mollet, Wolfram Liebermeister, and Jean-Loup Faulon. A neural-mechanistic hybrid approach improving the predictive power of genome-scale metabolic models. *Nature Communications*, August 2023.

[5] Daniel M. Gonçalves, Rui Henriques, and Rafael S. Costa. Predicting metabolic fluxes from omics data via machine learning: Moving from knowledge-driven towards data-driven approaches. *Computational and Structural Biotechnology Journal*, pages 4960–4973, January 2023.

[6] Zongbo Han, Fan Yang, Junzhou Huang, Changqing Zhang, and Jianhua Yao. Multimodal dynamics: Dynamical fusion for trustworthy multimodal classification. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20675–20685, 2022.

[7] Nobuyoshi Ishii, Kenji Nakahigashi, Tomoya Baba, Martin Robert, Tomoyoshi Soga, Akio Kanai, Takashi Hirasawa, Miki Naba, Kenta Hirai, Aminul Hoque, Pei Yee Ho, Yuji Kakazu, Kaori Sugawara, Saori Igarashi, Satoshi Harada, Takeshi Masuda, Naoyuki Sugiyama, Takashi Togashi, Miki Hasegawa, Yuki Takai, Katsuyuki Yugi, Kazuharu Arakawa, Nayuta Iwata, Yoshihiro Toya, Yoichi Nakayama, Takaaki Nishioka, Kazuyuki Shimizu, Hirotada Mori, and Masaru Tomita. Multiple high-throughput analyses monitor the response of e. coli to perturbations. *Science*, 316(5824):593–597, April 2007.

[8] Jérôme Mariette and Nathalie Villa-Vialaneix. Unsupervised multiple kernel learning for heterogeneous data integration. *Bioinformatics*, 34(6):1009–1015, October 2017.

[9] Alain Rakotomamonjy, Francis Bach, Stephane Canu, and Yves Grandvalet. SimpleMKL. *Journal of Machine Learning Research*, 9:2491–2521, 2008.

[10] B. Sun, L. Yang, W. Zhang, M. Lin, P. Dong, C. Young, and J. Dong. Supertml: Two-dimensional word embedding for the precognition on structured tabular data. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2973–2981, Los Alamitos, CA, USA, jun 2019. IEEE Computer Society.

[11] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org, 2017.

[12] Ines Thiele and Bernhard Ø Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protocols*, 5(1):93–121, January 2010. Publisher: Nature Publishing Group.

[13] Tongxin Wang, Wei Shao, Zhi Huang, Haixu Tang, Jie Zhang, Zhengming Ding, and Kun Huang. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nature Communications*, 12(1), June 2021.

[14] Christopher M. Wilson, Kaiqiao Li, Xiaoqing Yu, Pei-Fen Kuan, and Xuefeng Wang. Multiple-kernel learning for genomic data mining and prediction. *BMC Bioinformatics*, 20(1), August 2019.

[15] Zenglin Xu, Rong Jin, Haiqin Yang, Irwin King, and Michael R. Lyu. Simple and efficient multiple kernel learning by group lasso. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, page 1175–1182, Madison, WI, USA, 2010. Omnipress.

Declaration

In the PhD dissertation of Gabriele Tazza entitled "**Machine Learning at Genome Scale**", with the list of publications:

[J1] Gabriele Tazza, Francesco Moro, Dario Ruggeri, Bas Teusink, and László Vidács. *MINN: A metabolic-informed neural network for integrating omics data into genome-scale metabolic modeling*. In Computational and Structural Biotechnology Journal, 27, 3609–3617, 2025.

[J2] Gabriele Tazza, Dario Ruggeri and László Vidács. *Improving Microbiome-Based Disease Prediction with SuperTML and Data Augmentation*. In IEEE Access, 13, 144505-144515, 2025.

[J3] Mitja Briscik, Gabriele Tazza, László Vidács, Marie-Agnés Dillies and Sébastien Déjean. *Supervised multiple kernel learning approaches for multi-omics data integration*. In BioData Mining, 17, 53, 2024.

[C1] Gabriele Tazza, Francesco Moro, Bas Teusink, and László Vidács. *Metabolic- Informed Neural Network for Multi-Omics Data Integration*. In Proceedings of the 13th International Conference on Simulation and Modelling in the Food and Bio-Industry, FOODSIM 2024, Eurosis-ETI, 193-197, 2024

Gabriele Tazza 's contribution was decisive in the following results:

- In the first thesis point, "**Improving microbiome-based disease prediction with SuperTML and data augmentation**", the author contributed to the conceptualization and design of the comparative analysis based on SuperTML and image augmentation techniques in the context of microbiome-based disease prediction. The author conducted the literature review presented in the Related Work section, contributed to implementing the code for all analyses performed, and designed and implemented the novel custom transformation CellDropout. [J2]

- In the second thesis point, "**Supervised Multiple Kernel Learning approaches for multi-omics data integration**", the author contributed to the conceptualization and design of the comparison between state-of-the-art deep learning methods and multiple kernel learning approaches, based on deep learning optimization, on biomedical multi-omics datasets. The author is responsible for the literature review regarding the deep learning methods presented in the Related Work section. The development of DeepMKL and Cross-Modal DeepMKL, including both their conceptualization and implementation, was carried out by the author. The author also implemented the code used to preprocess all datasets included in the analysis. Finally, the novel feature-importance method, together with its implementation, was conceptualized and implemented by the author. [J3]

- In the third thesis point: “**MINN: A Metabolic-Informed Neural Network for Integrating Omics Data into Genome-Scale Metabolic Modeling**”, the author contributed to the conceptualization and design of the MINN architecture for integrating multi-omics data into genome-scale metabolic modeling. The author conducted a literature review regarding hybrid modeling methods, presented in the Related Work section. Furthermore, the author implemented the MINN and MINN-reservoir experiments and the code used in this analysis, with the exception of the hybrid optimization strategies and the components strictly related to handling the mechanistic aspects of this approach. [J1, C1]

These results cannot be used to obtain an academic research degree, other than the submitted PhD thesis of Gabriele Tazza

Date
23/01/2026

Signature of candidate

Gabriele Tazza

Signature of supervisors

The head of the Doctoral School of Computer Science declares that the declaration above was sent to all of the coauthors and none of them raised any objections against it.

Date Szeged, 2026. 02. 02.

signature of head of Doctoral School

UNIVERSITY OF SZEGED
DOCTORAL SCHOOL OF COMPUTER SCIENCE
1
SZEGED