Reliable Machine Learning for Omics Data: Evaluation Protocols, Hybrid Models, and Applications in Foodomics

PhD Thesis

Dario Ruggeri
Supervisor: László Vidács, PhD

Doctoral School of Computer Science
Department of Software Engineering
Faculty of Science and Informatics
University of Szeged

Szeged
2026

1 Introduction

Machine Learning (ML) and Deep Learning (DL) methods are increasingly applied to omics data to model complex relationships between molecular features and phenotypic or functional outcomes. Their flexibility and capacity to capture non-linear interactions make them attractive tools in domains such as genomics, transcriptomics, proteomics, metabolomics, and foodomics. As a result, ML-based approaches are now widely used for predictive modeling, pattern discovery, and hypothesis generation in biological and agricultural research [12, 21].

At the same time, the application of ML and DL to omics data introduces substantial methodological and practical challenges. Omics datasets are typically characterized by very high dimensionality and limited sample sizes. In these settings, models are particularly prone to overfitting, evaluation results can be highly sensitive to experimental design choices, and the interpretability of learned representations is often limited [18, 25]. Consequently, the reliability of ML-based conclusions in omics depends not only on model architecture, but also on evaluation protocols, training strategies, and the overall organization of experiments.

The central motivation of this PhD thesis is to investigate how ML methods can be applied more reliably and transparently to omics data analysis. Rather than focusing solely on improving predictive performance, the thesis emphasizes methodological soundness, interpretability, and reproducibility. The work addresses questions such as how model evaluation should be performed in data-scarce, high-dimensional regimes; how training strategies influence generalization and stability; and how domain knowledge can be incorporated into learning algorithms through hybrid modeling approaches [16, 30].

The thesis focuses on two complementary aspects. The first concerns methodological choices that affect model generalization, evaluation reliability, and interpretability, with particular emphasis on DL models trained on omics data. This includes the analysis of evaluation strategies, training procedures, and objective formulations, as well as the study of hybrid models that combine data-driven learning with mechanistic constraints. The second aspect addresses applied and technical challenges that arise in real-world omics studies, including robust model design, explainable learning approaches, and the management of increasingly complex experimental workflows.

The contributions of the thesis are developed through a combination of methodological analyses and applied case studies, primarily in the domains of foodomics and agronomy. These studies illustrate how careful experimental design, appropriate evaluation practices, and structured tooling can improve the reliability and practical usefulness of ML models in applied omics research. Together, they aim to support results that are not only predictive, but also interpretable, scientifically meaningful, and reusable across studies.

The dissertation is organized into two major parts. The first part focuses on methodological aspects of ML for omics data, including evaluation strategies, training practices for DL models in high-dimensional, low-sample-size settings, and hybrid modeling approaches that integrate data-driven learning with mechanistic constraints. The second part addresses applied ML problems in foodomics and agronomy, with particular emphasis on predictive performance, model robustness, explainability, and the organization and reproducibility of complex experimental workflows.

2 Evaluation Strategies and Hybrid Models for Omics Machine Learning

This thesis group focuses on methodological aspects that critically influence the reliability of ML models applied to omics data. In high-dimensional, low-sample-size settings, seemingly standard choices related to model evaluation, training procedures, and objective formulation can substantially affect performance estimates, generalization behavior, and interpretability [14, 19]. As a result, methodological rigor is a prerequisite for drawing meaningful conclusions from applied ML studies in omics domains.

A central challenge addressed in this thesis group concerns the design of sound evaluation and training strategies for DL models. Techniques such as Cross Validation (CV) and Early Stopping (ES) are widely adopted to control overfitting and estimate generalization performance [17, 29], yet their interaction is often handled inconsistently in practice. Inappropriate combinations of these techniques may lead to biased performance estimates, unintended data leakage, or overly optimistic conclusions, particularly in data-scarce regimes that are common in omics research.

In addition to evaluation-related issues, this thesis group investigates hybrid modeling approaches that combine data-driven learning with mechanistic constraints. In several scientific domains, including systems biology and foodomics, prior knowledge in the form of mechanistic models can complement flexible ML methods [23, 30]. Integrating such knowledge into Neural Network (NN) models introduces multi-objective optimization problems, where predictive accuracy must be balanced against consistency with known system constraints. Understanding how these competing objectives interact during training is essential for achieving stable, interpretable, and biologically meaningful models [13, 26].

The contributions presented in this thesis group are developed in Chapter 2 of the dissertation. Through methodological analyses and experimental studies, the chapter examines how evaluation protocols, training strategies, and objective formulations influence model behavior in omics applications. The results provide practical guidance for designing reliable ML workflows and for integrating mechanistic knowledge into data-driven models without compromising evaluation integrity or interpretability.

2.1 Integrating Cross-Validation and Early Stopping: Pitfalls, Bias, and Practical Guidelines for Omics Neural Networks

DL models applied to omics data are commonly trained in regimes characterized by high dimensionality and limited sample sizes, where the risk of overfitting is substantial. To mitigate this risk, practitioners frequently rely on techniques such as CV to estimate generalization performance and ES to control training dynamics [17, 31]. While both methods are well established, their combined use is often handled inconsistently in applied studies, particularly in omics research.

This work investigates how different strategies for integrating CV and ES affect model evaluation and training reliability in NNs trained on omics data. A key observation is that naïve combinations of these techniques may introduce unintended information leakage or bias performance estimates, especially when validation data used for ES monitoring overlap with data later employed for model evaluation. Such issues are particularly problematic in

small-sample settings, where even minor methodological inconsistencies can lead to overly optimistic conclusions [19].

Through a systematic analysis of commonly adopted evaluation schemas, this study highlights typical pitfalls in the joint use of CV and ES and clarifies their methodological implications. Several alternative strategies are examined, differing in how data partitions are defined and how stopping criteria are selected during training. The results demonstrate that careful separation of training, validation, and evaluation roles is essential for obtaining reliable performance estimates and for ensuring that reported results genuinely reflect model generalization.

The findings of this subsection provide practical guidelines for the design of evaluation protocols in omics ML. Figure 1 illustrates a recommended evaluation schema in which ES is monitored on an inner validation split, while CV is used exclusively for performance estimation. By clarifying how CV and ES can be combined without compromising evaluation integrity, the work contributes to more robust and transparent experimental practices in applied DL studies.

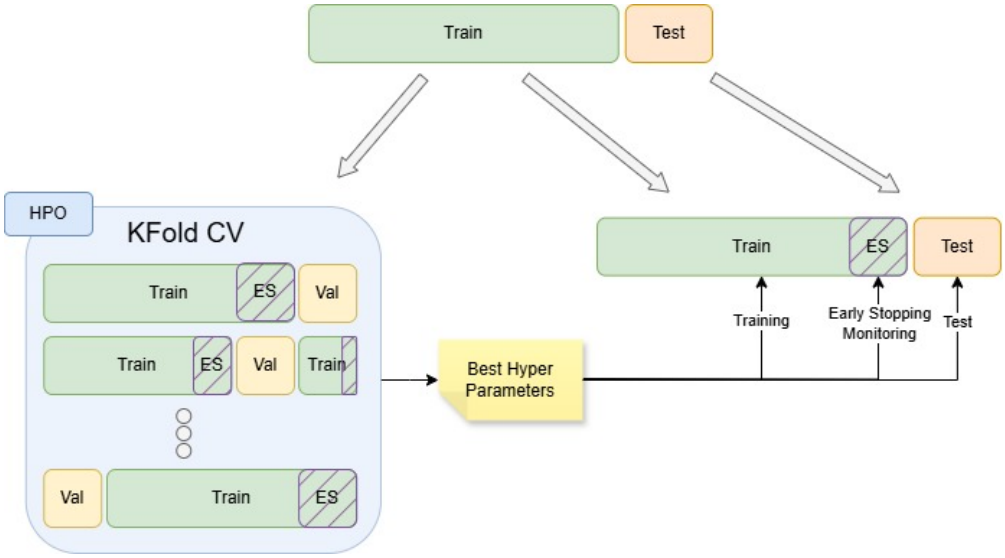

Figure 1: *Use of an inner validation split for ES during k-fold CV. A portion of the training data is reserved for monitoring the ES criterion, while the remaining data are used for model fitting. This separation helps prevent information leakage and biased performance estimates in high-dimensional, low-sample-size settings.*

2.2 Hybrid Neural Networks as Multi-Objective Systems: Balancing Data Fidelity and Mechanistic Constraints

Purely data-driven NN models offer substantial flexibility for learning complex relationships from omics data, but they may produce solutions that are difficult to interpret or that violate known biological or physical constraints. In response to these limitations, hybrid modeling approaches have been proposed that integrate mechanistic knowledge directly into the learning process [16, 30]. In the context of omics data analysis, such approaches aim to combine the predictive power of NNs with established domain models, such as constraint-based representations of biological systems.

This work investigates hybrid NN architectures that incorporate mechanistic constraints as additional learning objectives. Rather than treating mechanistic consistency as a post hoc validation step, these constraints are embedded directly into the training process, resulting in models that must simultaneously satisfy data-driven and mechanistic requirements. From a methodological perspective, this setup gives rise to a multi-objective optimization problem, where competing objectives may influence training dynamics, convergence behavior, and generalization [26].

A key contribution of this subsection is the analysis of hybrid models through the lens of multitask learning. By framing data fidelity and mechanistic consistency as distinct but related objectives, the study examines how loss formulation, weighting strategies, and scheduling choices affect model stability and performance. The results highlight trade-offs between predictive accuracy and adherence to mechanistic constraints, and show that inappropriate balancing of objectives can lead to unstable training or degraded performance on one or both tasks.

The findings demonstrate that hybrid models can improve robustness and interpretability when methodological choices are carefully designed. In particular, adaptive loss weighting and constraint-aware training strategies are shown to help reconcile competing objectives, enabling models that remain predictive while producing biologically meaningful outputs. These results provide practical guidance for the development and evaluation of hybrid ML models in omics research.

3 Applied Machine Learning, Explainability, and Workflow Engineering in Foodomics

This thesis group focuses on the application of ML methods to real-world omics problems, with particular emphasis on robustness, interpretability, and experimental organization. While methodological soundness is a prerequisite for reliable modeling, applied omics studies introduce additional challenges related to data heterogeneity, domain-specific constraints, and the practical management of complex experimental pipelines [21, 25]. Addressing these issues is essential for translating methodological advances into scientifically useful and reusable results.

A central theme of this thesis group is the development and evaluation of robust DL models for applied omics tasks. In foodomics and agronomy, predictive models must operate under challenging conditions, including limited labeled data, strong feature correlations, and varying experimental setups across seasons or studies. The work investigates training strategies and evaluation practices that improve model stability and generalization, with a focus on ensuring that reported performance reflects genuine predictive capability rather than artifacts of experimental design.

In addition to predictive performance, this thesis group addresses the issue of model explainability. For many omics applications, understanding which molecular features drive model predictions is critical for biological interpretation and for building trust in model outputs [22, 27]. The work therefore integrates explainable DL techniques into applied modeling pipelines, enabling the systematic analysis of feature importance and the assessment of model behavior across different data partitions and experimental conditions.

Finally, this thesis group examines workflow engineering and experiment management in

applied ML research. As omics studies grow in scale and complexity, ad hoc experimentation becomes increasingly difficult to reproduce and compare. The thesis explores the use of research-stage Machine Learning Operations (MLOps) frameworks to structure data processing, model training, Hyper-Parameter Optimization (HPO), and result tracking [15, 28]. By organizing experiments in a transparent and reproducible manner, these workflows support more reliable comparison across models and contribute to improved reproducibility in applied omics studies.

The contributions presented in this thesis group are developed in Chapter 3 of the dissertation. Through applied case studies in foodomics and agronomy, the chapter demonstrates how robust modeling practices, explainable learning approaches, and structured experimental workflows can be combined to support reliable and interpretable ML applications in omics research.

3.1 Explainable Deep Learning for SNP-Based Prediction in Agronomy

The application of DL models to Single-Nucleotide Polymorphism (SNP) based prediction tasks in agronomy presents a challenging setting characterized by extremely high-dimensional feature spaces, limited sample sizes, and strong correlations among input variables. While NNs offer the capacity to model complex, non-linear relationships in such data, their flexibility also increases the risk of overfitting and limits the interpretability of predictions. These issues are particularly relevant in agronomic applications, where model outputs are expected to support biological insight and practical decision-making [21, 24].

This work investigates the use of explainable DL models for SNP-based phenotype prediction, with a focus on robustness, evaluation reliability, and interpretability. The proposed modeling framework combines carefully designed training and evaluation strategies with post hoc explainability techniques to analyze feature contributions across different data partitions and experimental conditions [22]. Particular attention is given to avoiding optimistic performance estimates by enforcing strict separation between training, validation, and evaluation data.

The study demonstrates that the proposed training design leads to consistent and measurable performance improvements over existing benchmarks in SNP-based prediction tasks. Rather than relying solely on architectural complexity, the gains arise from the combined effect of adaptive optimization, noise-aware data augmentation, and carefully controlled evaluation protocols. These methodological choices improve generalization in high-dimensional, low-sample-size settings, reducing overfitting and stabilizing performance across cross-validation folds. As a result, the model achieves statistically significant improvements not only in predictive accuracy but also in performance robustness, indicating that the observed gains reflect genuine generalization rather than favorable experimental conditions.

Beyond predictive performance, explainability analyses provide insight into the distribution and stability of feature importance, revealing how a small subset of features contributes disproportionately to model predictions. These analyses support the biological plausibility of the learned models and help identify features that are consistently influential across CV folds and experimental settings. To improve interpretability, we used SHapley Additive exPlanations (SHAP) feature attribution and aggregated explanations across the k-fold test folds. Figure 2 presents a SHAP summary (beeswarm) plot of the top-ranked SNP markers:

each dot represents a test sample, the horizontal position indicates the signed contribution of the feature to the prediction, and the color encodes the SNP value (blue for low, red for high). A wider horizontal spread reflects stronger influence on the model output, while a clear separation of colors suggests stable directional effects across samples.

Overall, this results demonstrate that explainable DL approaches can be effectively applied to high-dimensional SNP data when methodological rigor is combined with interpretability-focused analysis. The results highlight the importance of integrating robust evaluation practices with explainability techniques to produce models that are not only accurate, but also transparent and scientifically informative.

3.2 Research-Stage MLOps for Reproducible Omics Experiments

Applied ML studies in omics research often involve complex experimental pipelines, including data preprocessing, feature engineering, model training, HPO, and evaluation across multiple data splits. As the number of experimental variants grows, managing these components using ad hoc scripts becomes increasingly difficult, limiting reproducibility, transparency, and systematic comparison across models.

This work investigates the use of research-stage MLOps practices to structure and manage applied omics experiments. Rather than targeting large-scale industrial deployment, the focus is on supporting exploratory and iterative scientific workflows, where rapid experimentation must be balanced with traceability and reproducibility [20, 28]. The proposed approach organizes experiments into modular, versioned components that explicitly capture data provenance, model configurations, training procedures, and evaluation results.

Through a case study in agronomic ML, the study demonstrates how an MLOps framework can be integrated into an applied omics workflow, organizing data processing, model training, HPO, and evaluation into a structured and traceable pipeline (Figure 3) [11]. The framework enables consistent tracking of experiments, automated execution of HPO procedures, and systematic aggregation of results across multiple experimental replicates. This structure facilitates fair comparison between models and helps identify sources of performance variability that may otherwise remain hidden.

The results show that structured workflow engineering improves not only reproducibility, but also the reliability of experimental conclusions. By making experimental decisions explicit and traceable, research-stage MLOps practices support more transparent reporting and reduce the risk of unintentional bias or irreproducible findings. These insights highlight the value of workflow-level contributions as an integral component of applied ML research in omics.

4 Contributions of the thesis

In the **first thesis group**, my contributions are related to methodological aspects of ML for omics data, including evaluation strategies for DL models and hybrid approaches integrating mechanistic constraints. Detailed discussion can be found in Chapter 2.

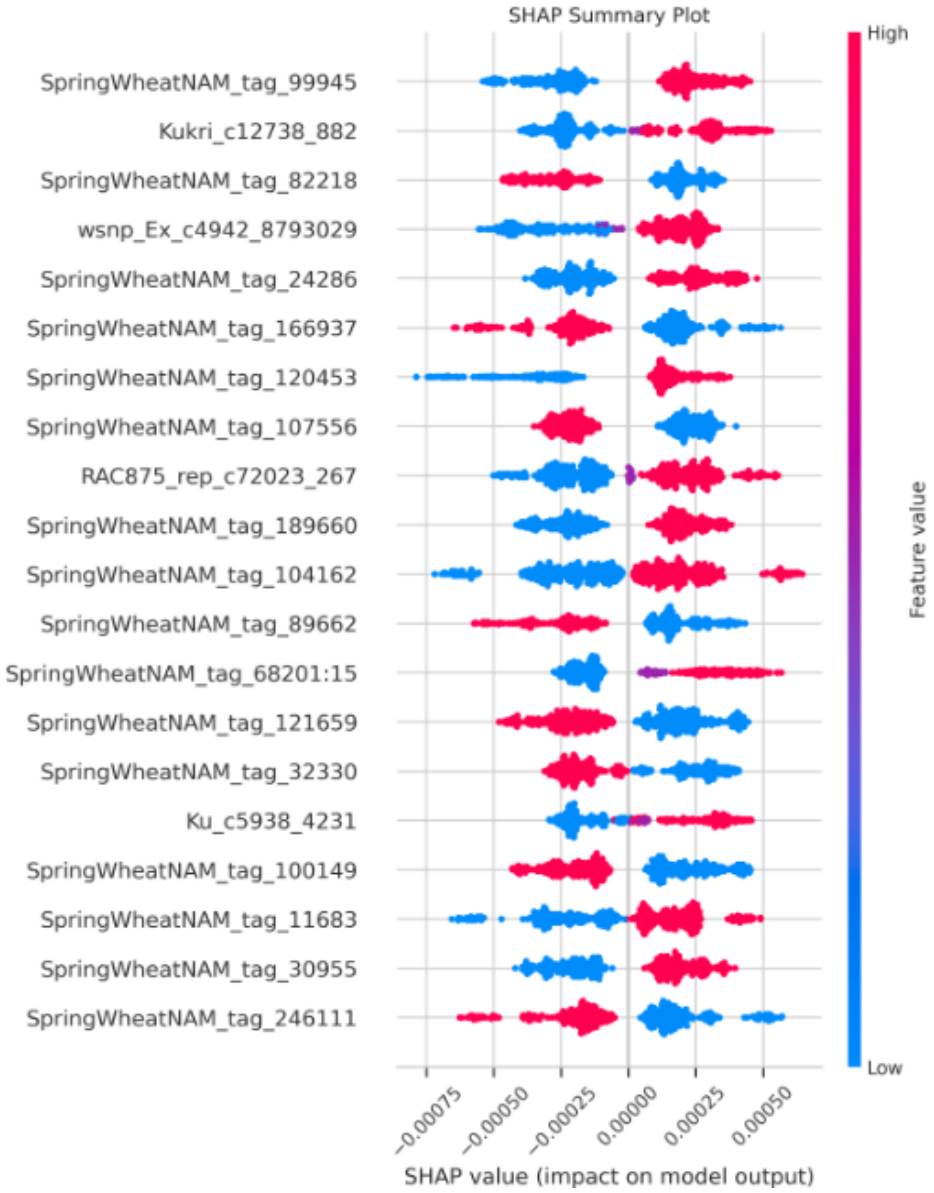

Figure 2: *SHAP summary (beeswarm) plot of the top 20 SNP features for the SNP-based DL model (example shown for the test weight trait in 2015). Each dot represents a test sample aggregated across k-fold splits. The x-axis shows the SHAP value (signed impact on the model output), and color indicates the SNP feature value (low to high). A clear separation of high and low feature values along the x-axis indicates a consistent directional effect of a marker on the prediction.*

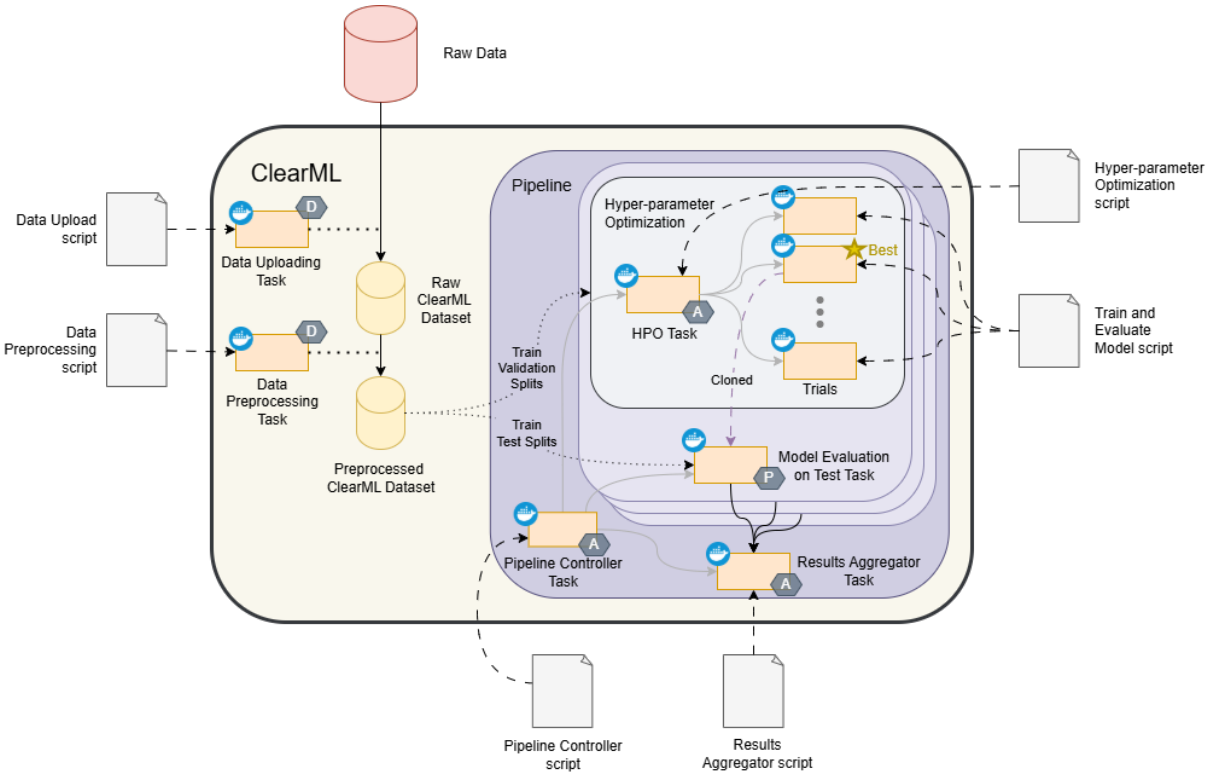

Figure 3: *High-level overview of the research-stage MLOps workflow used in the applied omics studies. Data preparation, model training, HPO, and evaluation steps are organized into modular tasks, enabling systematic experiment tracking, reproducibility, and transparent comparison across experimental variants.*

I / 1. I conducted a systematic review and methodological analysis of different strategies for integrating CV and ES in NNs applied to high-dimensional omics data, identifying common pitfalls and showing how certain integration choices can introduce bias and information leakage.

I / 2. I proposed an alternative integration strategy for combining CV and ES that prevents evaluation inconsistencies, improves methodological soundness, and enables more reliable estimation of model generalization performance in data-scarce settings.

I / 3. I investigated knowledge-informed hybrid NNs from a multi-objective optimization perspective, analyzing the trade-off between predictive accuracy and mechanistic consistency, and identifying training strategies that achieve a stable balance between the two objectives.

In the **second thesis group**, my contributions are related to applied ML problems in foodomics and agronomy, with emphasis on robustness, explainability, and reproducible experimental workflows. Detailed discussion can be found in Chapter 3.

II / 1. I developed and evaluated robust DL training strategies for SNP-based prediction tasks in agronomy, integrating adaptive optimization, regularization, and noise-aware data augmentation. I showed that these methodological choices improve generalization and lead to statistically significant performance gains over existing benchmarks and baselines.

II / 2. I introduced an explainability-driven analysis framework for DL models applied to omics data, using SHAP-based feature attribution to investigate the stability and biological plausibility of model predictions. I demonstrated how aggregating explanations across cross-validation folds provides deeper insight into feature importance and supports transparent interpretation of model behavior.

II / 3. I designed and implemented a research-stage MLOps workflow for applied omics ML experiments, developing a structured pipeline for data processing, model training, HPO, and evaluation, and showed that structured experiment management improves reproducibility, transparency, and comparability across experiments.

Table 1 summarizes the relation between the thesis points and the corresponding publications.

Table 1: *Correspondence between the thesis points and my publications.*

Publication	Thesis point					
	I/1	I/2	I/3	II/1	II/2	II/3
[1]				•	•	
[2]						•
[3]	•	•				
[4]						
[5]			•			
[6]						•

The author's publications on the subjects of the thesis

Journal publications

[1] **D. Ruggeri** and L. Vidács. Advancing Wheat Single-Nucleotide Polymorphism Data Analysis with Explainable Deep Learning Models. *Applied Artificial Intelligence*, 39(1):2565169, 2025.

[2] **D. Ruggeri**, G. Tazza, and L. Vidács. Introducing MLOps to Facilitate the Development of Machine Learning Models in Agronomy: A Case Study. *IEEE Access*, 13:122059–122070, 2025.

[3] **D. Ruggeri** and L. Vidács. K-Fold Cross-Validation and Early Stopping in Foodomics Neural Networks: Practices, Pitfalls, and Recommendations. *IEEE Access*, 13:190820–190832, 2025.

[4] G. Tazza, **D. Ruggeri**, and L. Vidács. Improving Microbiome-Based Disease Prediction With SuperTML and Data Augmentation. *IEEE Access*, 13:144505–144515, 2025.

[5] G. Tazza, F. Moro, **D. Ruggeri**, B. Teusink, and L. Vidács. MINN: A Metabolic-Informed Neural Network for Integrating Omics Data into Genome-Scale Metabolic Modeling. *Computational and Structural Biotechnology Journal*, 27:3609–3617, 2025.

Full papers in conference proceedings

[6] **D. Ruggeri** and L. Vidács. Introducing MLOps to Facilitate Reproducible Model Development on Omics Data. In *Proceedings of the 13th International Conference on Simulation and Modelling in the Food and Bio-Industry (FOODSIM)*, pp. 198–203, 2024.

Other References

[11] Lisana Berberi, Valentin Kozlov, Giang Nguyen, Judith Sáinz-Pardo Díaz, Amanda Calatrava, Germán Moltó, Viet Tran, and Álvaro López Garc\ á. Machine Learning Operations Landscape: Platforms and Tools. *Artificial Intelligence Review*, 58(6):167, March 2025.

[12] Francesco Capozzi and Alessandra Bordoni. Foodomics: A new Comprehensive Approach to Food and Nutrition. *Genes & Nutrition*, 8(1):1–4, January 2013.

[13] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient Normalization for Adaptive Loss Balancing in Deep Multitask Networks. In *Proceedings of the 35th International Conference on Machine Learning*, pages 794–803. PMLR, July 2018.

[14] Chandra Mohan Dasari and Raju Bhukya. Explainable Deep Neural Networks for Novel Viral Genome Prediction. 52(3):3002–3017, 2022.

[15] Beyza Eken, Samodha Pallewatta, Nguyen Khoi Tran, Ayse Tosun, and Muhammad Ali Babar. A multivocal Review of Mlops Practices, Challenges and Open Issues, June 2024.

[16] 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.

[17] Ziwei Ji, Justin Li, and Matus Telgarsky. Early-Stopped Neural Networks Are Consistent. In *Advances in Neural Information Processing Systems*, volume 34, pages 1805–1817. Curran Associates, Inc., 2021.

[18] Tuija Leinonen, David Wong, Antti Vasankari, Ali Wahab, Ramesh Nadarajah, Matti Kaisti, and Antti Airola. Empirical Investigation of Multi-Source Cross-Validation in Clinical Ecg Classification. 183:109271, 2024.

[19] Qi Liu, Shi-min Zuo, Shasha Peng, Hao Zhang, Ye Peng, Wei Li, Yehui Xiong, Runmao Lin, Zhiming Feng, Huihui Li, Jun Yang, Guo-Liang Wang, and Houxiang Kang. Development of Machine Learning Methods for Accurate Prediction of Plant Disease Resistance. 40:100–110, 2024.

[20] Beatriz M. A. Matsui and Denise H. Goya. MLops: A guide to Its Adoption in the Context of Responsible Ai. In *2022 IEEE/ACM 1st International Workshop on Software Engineering for Responsible Artificial Intelligence (SE4RAI)*, pages 45–49, May 2022.

[21] Osval A. Montesinos-López, Abelardo Montesinos-López, Roberto Tuberosa, Marco Maccaferri, Giuseppe Sciara, Karim Ammar, and José Crossa. Multi-Trait, Multi-Environment Genomic Prediction of Durum Wheat With Genomic Best Linear Unbiased Predictor and Deep Learning Methods. *Frontiers in Plant Science*, 10, 2019.

[22] Pierfrancesco Novielli, Donato Romano, Stefano Pavan, Pasquale Losciale, Anna Maria Stellacci, Domenico Diacono, Roberto Bellotti, and Sabina Tangaro. Explainable

Artificial Intelligence for Genotype-to-Phenotype Prediction in Plant Breeding: A case Study with a Dataset from an Almond Germplasm Collection. 15.

[23] Jeffrey D. Orth, Ines Thiele, and Bernhard Ø Palsson. What is flux balance analysis? *Nature Biotechnology*, 28(3):245–248, March 2010. Number: 3 Publisher: Nature Publishing Group.

[24] Karansher Sandhu, Shruti Sunil Patil, Michael Pumphrey, and Arron Carter. Multitrait Machine- and Deep-Learning Models for Genomic Selection Using Spectral Information in a Wheat Breeding Program. *The Plant Genome*, 14(3):e20119, 2021.

[25] Karansher S. Sandhu, Dennis N. Lozada, Zhiwu Zhang, Michael O. Pumphrey, and Arron H. Carter. Deep Learning for Predicting Complex Traits in Spring Wheat Breeding Program. *Frontiers in Plant Science*, 11, 2021.

[26] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 525–536, Red Hook, NY, USA, December 2018. Curran Associates Inc.

[27] Peipei Wang, Melissa D. Lehti-Shiu, Serena Lotreck, Kenia Segura Abá, Patrick J. Krysan, and Shin-Han Shiu. Prediction of Plant Complex Traits via Integration of Multi-Omics Data. 15(1):6856.

[28] Samar Wazir, Gautam Siddharth Kashyap, and Parag Saxena. MLops: A review, August 2023.

[29] Suqin Yuan, Runqi Lin, Lei Feng, Bo Han, and Tongliang Liu. Instance-Dependent Early Stopping, February 2025.

[30] Guido Zampieri, Supreeta Vijayakumar, Elisabeth Yaneske, and Claudio Angione. Machine and deep learning meet genome-scale metabolic modeling. *PLoS Computational Biology*, 15(7), July 2019.

[31] Yongli Zhang and Yuhong Yang. Cross-Validation for Selecting a Model Selection Procedure. *Journal of Econometrics*, 187(1):95–112, July 2015.

5 Összefoglalás

Az értekezés a gépi tanulási (ML) és mélytanulási (DL) módszerek megbízható alkalmazását vizsgálja omikai adatok elemzésében, különös tekintettel a nagy dimenziószámú, kis elemszámú adathalmazokra, amelyek az agronómiai és élelmiszer-omikai kutatásban egyaránt jellemzőek. A bemutatott megközelítések közös vonása a módszertani megbízhatóság, a magyarázhatóság és a reprodukálhatóság következetes érvényesítése.

A munka két fő témakört dolgoz fel. Az első rész módszertani kérdésekkel foglalkozik, különös tekintettel az modell értékelési stratégiákra és a hibrid modellezésre, míg a második rész esettanulmányokon keresztül mutatja be a magyarázható mélytanulási modellek és a kísérletmenedzsment szerepét agronómiai kutatási környezetben.

Az *Evaluation Strategies and Hybrid Models for Omics Machine Learning* című fejezetben a dolgozat a keresztvalidáció és a korai leállítás (early stopping) együttes alkalmazásának módszertani kérdéseit tárgyalja, rámutatva az adatszivárgás és a torzított teljesítménybecslés lehetséges problémáira. Emellett olyan hibrid neurális hálózati megközelítéseket ismertet, amelyek mechanisztikus tudást integrálnak az adatvezérelt tanulásba, és elemzi a többcélú optimalizáció hatását a modell stabilitására és magyarázhatóságára.

Az *Explainable Deep Learning for SNP-Based Prediction in Agronomy* című fejezet egy magyarázható neurális hálózati keretrendszert mutat be búza SNP-adatok elemzésére. Korszerű mélytanulási eljárások (regularizáció, adataugmentáció és hiperparaméter-optimalizálás) alkalmazásával a modell statisztikailag szignifikáns teljesítményjavulást ér el a korábbi módszerekhez képest, miközben SHAP-alapú magyarázhatósági elemzések segítségével azonosítja a predikciókat meghatározó genetikai markereket.

A *Research-Stage MLOps for Reproducible Omics Experiments*című fejezet a kutatási fázisban alkalmazott MLOps-gyakorlatok szerepét vizsgálja az omikai gépi tanulási kísérletek szervezésében. Az eredmények azt mutatják, hogy a strukturált munkafolyamatok érdemben javítják a reprodukálhatóságot, az átláthatóságot és az egyes modellek közötti összehasonlíthatóságot.

Összességében a disszertáció hozzájárul a megbízhatóbb és magyarázhatóbb ML-alapú omikai elemzések módszertanának fejlesztéséhez, ötvözve a módszertani alaposságot az alkalmazott kutatási szemlélettel.

Declaration

In the PhD dissertation of Dario Ruggeri entitled *Reliable Machine Learning for Omics Data: Evaluation Protocols, Hybrid Models, and Applications in Foodomics*, with the list of publications:

[J2] Dario Ruggeri, Gabriele Tazza, and László Vidács. *Introducing MLOps to Facilitate the Development of Machine Learning Models in Agronomy: A Case Study*. IEEE Access, 13:122059–122070, 2025.

[J5] 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*. Computational and Structural Biotechnology Journal, 27:3609–3617, 2025.

Dario Ruggeri's contribution was decisive in the following results:

- In the thesis point related to *Research-Stage MLOps for Reproducible Omics Experiments*, the author was the principal contributor and led the development of the research-stage MLOps framework for agronomic machine learning studies. The author conceived and implemented the overall experimental design, developed the experiment tracking and evaluation workflow, conducted the experiments, and carried out the analysis related to reproducibility and experiment management. [J2]
- In the thesis point related to *Hybrid Neural Networks as Multi-Objective Systems*, the author was the principal contributor to the methodological development of the hybrid multi-objective framework. The author designed and implemented the multi-task learning formulation, developed the objective balancing strategy and evaluation considerations, and led the methodological investigation of hybrid neural network models integrating mechanistic constraints. [J5]

These results cannot be used to obtain an academic research degree, other than the submitted PhD thesis of Dario Ruggeri.

Date: _____ Signature of candidate: _____ Signature of supervisor: _____

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: _____ Signature of Head of Doctoral School: _____