UNIVERSITÀ
DEGLI STUDI
DI MILANO

Doctoral Dissertation
Doctoral Program in Artificial Intelligence (38^{th}cycle)

Theory and Tools for Structure-Aware Machine Learning on Biological Data

By

Davide D'Ascenzo

******

Supervisor(s):
Prof. Sebastiano Vigna, Supervisor
Prof. Nicolò Cesa-Bianchi, Co-Supervisor

Doctoral Examination Committee:
Prof. Maurizio Parton, Referee, University of Chieti-Pescara
Prof. Hernan Makse, Referee, City University of New York
Prof. Alessandro Rizzo, Politecnico di Torino

Politecnico di Torino ◇ Università degli Studi di Milano
2025

Declaration

I hereby declare that, the contents and organization of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

Davide D'Ascenzo
2025

* This dissertation is presented in partial fulfillment of the requirements for **Ph.D. degree** in the Graduate School of Politecnico di Torino (ScuDo).

Theory and Tools for Structure-Aware Machine Learning on Biological Data

Davide D'Ascenzo

The exponential growth of biological data has created unprecedented opportunities for machine learning to advance our understanding of complex biological systems. However, traditional machine learning approaches often treat biological data as collections of independent features, ignoring the rich structural relationships that characterize biological systems across all scales of organization. This dissertation develops theory and methods for structure-aware machine learning, demonstrating how incorporating structural knowledge into deep learning models can improve both theoretical understanding and practical performance.

We begin by studying centrality measures under network growth, proving that closeness, harmonic centrality, and betweenness satisfy rank semi-monotonicity when an edge is added to an undirected network. We then address the "curse of dimensionality", proving that the success of deep learning stems from its ability to exploit compositional sparsity, a hierarchical property inherent to all efficiently Turing-computable functions. Moving from theory to practice, we demonstrate that structural knowledge can be encoded directly into training objectives. We introduce a hierarchical cross-entropy loss that embeds the cell ontology into the learning process, improving out-of-distribution generalization by 12-15% for atlas-scale single-cell annotation. Evaluated on over 6 million cells across diverse architectures, this approach recovers roughly half of the performance drop observed when models are applied to newly released studies. Finally, we address the practical bottlenecks of training on large-scale datasets by developing a high-throughput data loading solution that uses quasi-random sampling to enable efficient training on disk-resident datasets comprising hundreds of millions of cells.