AI and Data Science Research in Life Science, System Biology and Medicine

Michele Alessi[1,2], Alessio Ansuini[1], Federico Barone[1,2], Alberto Cazzaniga[1], Stefano Cozzini[1], Francesca Cuturello[1], Fiorella Fabris[1], Valerio Piomponi[1], Marco Prenassi[1], Ruggero Lot[1], Lucrezia Valeriani[1,2] and Edith N. Villegas Garcia[1,2]

[1]*Area Science Park, Trieste, Italy*
[2]*University of Trieste, Trieste, Italy*

Abstract

The Laboratory of Data Engineering (LADE) conducts research at the intersection of machine learning, artificial intelligence, and the life sciences, with applications spanning systems biology, structural biology, genomics, and medicine. A key focus of our work is the development and application of advanced AI models to uncover meaningful insights from data, particularly through the careful analysis of hidden representations within deep neural networks.

Among our recent efforts, we include the use of protein language models to predict the effects of mutations, model protein–protein interactions, and enable high-throughput functional annotation of predicted protein structures—extending into the metagenomic domain. Through these innovative methodologies, LADE aims to deepen the understanding of biological systems, extract actionable knowledge from large-scale bioinformatics resources, enhance sequencing-based analyses, and support the development of novel medical applications.

Keywords: Deep Learning, Clustering, Protein Language Models, Structural Biology, Genomics, Metagenomics, Protein Mutations, Sparse Autoencoders, Clinical Stratification

1. Introduction

The Laboratory of Data Engineering (LADE) comprises around 10 members—including staff researchers, postdoctoral fellows, and PhD students—focused on applying artificial intelligence to life science. All members are affiliated with the Institute of Research and Technological Innovation at Area Science Park. LADE also hosts senior visiting researchers and offers internships for early-career scientists. The lab collaborates closely with two experimental facilities at Area Science Park: the Laboratory of Genomics and Epigenomics (LAGE) and the Laboratory of Electron Microscopy (LAME). It also maintains strong ties with regional institutions such as the University of Trieste, SISSA, and ICTP. In particular, LADE works with the ICGEB, the Burlo Garofolo Pediatric Institute, and the Oncology Reference Center (CRO). A core research area at LADE is advancing AI applications in life sciences, systems biology, and medicine. This work is closely integrated with foundational research on understanding how neural networks learn, what they represent, and how to ensure their outputs are transparent, reliable, and trustworthy, including in biomedical contexts. This focus on reliability is supported by LADE's strong tradition in designing FAIR, robust digital infrastructures for scientific data.

Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23-24, 2025, Trieste, Italy
*Corresponding authors: Alessio Ansuini, Alberto Cazzaniga.
michele.alessi@areasciencepark.it (M. Alessi); alessio.ansuini@areasciencepark.it (A. Ansuini); federico.barone@areasciencepark.it (F. Barone); alberto.cazzaniga@areasciencepark.it (A. Cazzaniga); stefano.cozzini@areasciencepark.it (S. Cozzini); francesca.cuturello@areasciencepark.it (F. Cuturello); fiorella.fabris@areasciencepark.it (F. Fabris); valerio.piomponi@areasciencepark.it (V. Piomponi); marco.prenassi@areasciencepark.it (M. Prenassi); ruggero.lot@areasciencepark.it (R. Lot); lucrezia.valeriani@areasciencepark.it (L. Valeriani); edith.villegas@areasciencepark.it (E. N. V. Garcia)
0009-0005-2365-6955 (M. Alessi); 0000-0002-3117-3532 (A. Ansuini); 0000-0001-5696-670X (F. Barone); 0000-0001-6271-3303 (A. Cazzaniga); 0000-0001-6049-5242 (S. Cozzini); 0000-0001-7242-7298 (F. Cuturello); 0000-0001-6922-4916 (F. Fabris); 0000-0003-0433-8319 (V. Piomponi); 0000-0003-0240-0860 (M. Prenassi); 0000-0001-5950-1270 (R. Lot); 0009-0005-4378-6044 (L. Valeriani); 0000-0002-7338-2068 (E. N. V. Garcia)

LADE plays a key role in two major Italian PNRR projects—the Pathogen Readiness Platform (PRP) for life sciences and the Nano Foundries and Fine Analysis Digital Infrastructure (NFFA-DI) for material sciences—where it contributes to the architecture and implementation of advanced data ecosystems. These initiatives create a bridge between interpretability research and practical impact in scientific research, and clinical data workflows. Finally, LADE's involvement in the project "Support to the diagnoses of rare diseases", funded by Regione Friuli Venezia Giulia, provides a direct pathway for applying interpretability-driven AI methods in clinical and translational settings. Through its integration of foundational, infrastructural, and applied research, LADE aims to become a key hub for AI transparency and digital innovation in biomedicine.

2. AI for Scientific Discovery and Tool Development in Life Sciences

AI is rapidly transforming scientific research, with AlphaFold standing as a landmark example of its impact on life sciences and medicine [1].

Large language models (LLMs) trained on biological sequences—such as proteins and DNA—have progressed from simple sequence predictors to advanced systems capable of extracting deep structural and functional insights. Models like the ESM family for proteins [2], along with newer architectures incorporating DNA and RNA data [3], enable high-resolution predictions of mutation effects, structural modeling, and large-scale functional annotation. Recent innovations also integrate structural data into training, improving the alignment between sequence and three-dimensional conformation [4]. These advances are reshaping bioinformatics, allowing researchers to derive meaningful biological knowledge directly from raw data with unprecedented accuracy and efficiency.

Complementing this, the integration of molecular dynamics (MD) simulations and AI offers powerful strategies for elucidating the structure and dynamics of biomolecules at high resolution. By incorporating experimental data—such as Small-Angle X-ray Scattering (SAXS), Nuclear Magnetic Resonance (NMR), and Cryogenic Electron Microscopy (cryo-EM)—researchers can derive dynamic structural models that are fully consistent with empirical observations. AI enhances this process by improving sampling efficiency, identifying metastable states, and extracting interpretable features from high-dimensional MD datasets [5]. These developments are expanding our capacity to investigate complex biological systems and foster innovation in biomedical applications.

Researchers at LADE bring expertise in both AI modeling and MD simulations to advance discovery in life sciences, systems biology, and medicine. In the following sections, we detail the key research directions pursued at LADE in this rapidly evolving domain.

2.1. Localizing and extracting information from LLMs to predict properties of biomolecules and guide protein design

Large language models (LLMs) have proven highly effective at extracting both local and global signals from biological sequences. These include chemical properties of amino acids, remote homology relationships between organisms, and protein structural features [2, 6, 7]. Building on LADE's expertise in identifying and interpreting relevant information within protein language models (PLMs) [8, 9], we have investigated the latent signals encoded in LLM representations for a variety of applications.

Predicting changes of protein stability under mutations.
Localized mutations in proteins—such as single amino acid substitutions—can significantly affect global properties like thermodynamic stability, a key factor under evolutionary selective pressure. Predicting these changes remains challenging due to the complex interplay of local and global structural effects, solvent interactions, and entropic contributions. Nevertheless, the integration of AI with biophysical modeling is steadily enhancing both the accuracy and generalizability of such predictions. Understanding the impact of mutations on thermodynamic stability is crucial for studying protein evolution, guiding

protein design, and interpreting disease-associated variants. In this context, we developed an AI-based method to predict the stability effects of localized mutations, achieving state-of-the-art performance [10].

Protein-protein interactions
Protein-protein interactions underlie the regulation of numerous cellular processes, yet accurately identifying the residues involved in contacts between chains remains an open challenge. Recent advances in PLMs, especially in their multimodal flavour [6], offer new opportunities for the structural characterization of these functional regions. We are currently investigating local properties promoting protein contact, leveraging representations in multimodal (sequence + structure) PLMs (Section 3.1).

Steering protein generation
One of the most recent advances in the mechanistic interpretability of large language models is the use of large, sparse autoencoders (SAEs) to disentangle their internal representations. This technique reveals how information is structured within hidden layers and offers actionable insights for model alignment, debiasing, and controlled generation (see [11] for a recent overview). Building on this framework, we developed a novel method for synthesizing protein sequences by steering protein language models (PLMs) along directions in representation space that correspond to specific SAE activation patterns. Preliminary results [12] demonstrate the potential of this approach, which is currently undergoing more extensive validation.

Finding multiple conformational states
AlphaFold and related algorithms like RoseTTAFold and ESMFold have transformed structural biology by predicting protein structures directly from sequence. However, these models provide static snapshots, failing to capture the dynamic nature of proteins, particularly in fold-switching cases. We developed a framework to predict alternative conformations using AlphaFold2, guided by evolutionary signals from clustered MSAs and protein language model embeddings. By combining hierarchical clustering with Direct Coupling Analysis (DCA), we identify co-evolved residue pairs specific to alternative states and design stabilizing mutations, validated via molecular dynamics. This method reliably detects metastable conformations with high precision and extends to complex systems like GPCRs and kinases, illustrating the power of integrating evolutionary constraints into structure prediction [13].

2.2. Characterizing the protein landscape through clustering

The widespread availability of protein sequence data, combined with the release of the AlphaFold database of predicted structures, provides a unique opportunity to explore the protein universe from two complementary angles: sequence and structure.

Sequences
Although protein sequences are widely available, only a small fraction have known functional annotations. Grouping homologous regions into protein families can facilitate the generation of testable hypotheses by leveraging shared evolutionary features. To address the vast unclassified portion of sequence space, we developed an automated protocol to cluster over 20 million protein sequences from the UniRef50 database [14]. This method identified more than 14,000 candidate clusters not found in the Pfam database, potentially representing novel protein families. These clusters are publicly available and can support ongoing manual curation efforts in UniProtKB. The protocol is particularly valuable for exploring homology in poorly annotated datasets, such as those derived from metagenomic studies. In fact, we seized this opportunity and applied this methodology to the Unified Human Gastrointestinal Proteome [15].

Structures
The unprecedented scale of AlphaFoldDB—with over 214 million predicted protein structures—calls for new methods for large-scale, unbiased classification and annotation [7, 16]. In response, we developed

an unsupervised clustering approach aimed at domain-level classification of protein structures. Beyond automatic annotation, our method enables the discovery of remote homologies, identification of convergent evolution patterns, and the structural mapping of microbial proteins from gut metagenomic datasets, with promising applications in health and disease research [15].

3. Ongoing activities and future works

3.1. Protein-protein interaction from multimodal PLMs

ESM-3 is a protein language model that integrates sequence and atomic structure into a unified representation. Its structural tokenization captures each residue's local 3D context, enabling detection of geometric signals relevant to inter-chain interactions. Using a curated dataset of protein complexes, we annotate interface residues and use ESM-3 embeddings to train a classifier that distinguishes interacting from non-interacting residues. This approach uncovers recurring structural patterns at interfaces and highlights local features critical for protein contact and function.

3.2. Post-transcriptional regulatory networks altered in cancer: a Graph Attention Network approach for identifying functional interactions

In cancer, mRNA–microRNA interactions form dynamic, condition-specific regulatory networks that are not fully understood. Bulk RNA-seq data from TCGA, combined with curated databases of validated interactions, offer a rich framework for analysis. However, traditional static-network approaches often miss the functional reorganization occurring during disease progression.

We propose a Graph Attention Network (GAT) model where nodes represent mRNA and miRNA expression levels, and edges encode known regulatory links. Attention weights learned by the GAT act as proxies for condition-specific affinities, revealing the relative importance of interactions in healthy versus tumor contexts. This enables the identification of deregulated nodes, potential regulatory hubs, and network rewiring events associated with cancer.

3.3. Copy Number Alteration Calling

Copy number alterations (CNAs)—amplifications or deletions of genomic regions—are key structural variants that influence gene dosage and phenotype, playing a prominent role in diseases like cancer. Accurately inferring allele-specific CNAs (ASCNA), which resolve copy number changes at the level of individual parental alleles, remains a challenge for conventional short-read and bulk sequencing approaches. To address this, we leverage long-read sequencing and develop novel machine learning and probabilistic models that phase information from long reads. Our Bayesian and machine-learning-based models capture the complexity of subclonal architectures by deconvolving allele-specific copy number profiles and clustering cells into subpopulations with shared ASCNA patterns. Furthermore, since modern technologies provide additional layers of information, such as DNA methylation, we are working towards integrating genomic assays with epigenomic information to achieve a more comprehensive understanding of regulatory heterogeneity. Applied to malignancies such as colorectal cancer, these methods reveal fine-grained insights into genomic instability and tumor evolution. Extending this approach to multi-sample and longitudinal datasets will enable the study of ASCNA dynamics over time and space, ultimately advancing cancer subtype classification and personalized treatment strategies.

3.4. Isoform Discovery and Quantification

RNA molecules undergo alternative splicing, generating distinct transcript isoforms from the same gene locus and contributing to cellular and tissue complexity. Traditional short-read RNA-seq methods often fall short in reconstructing full-length isoforms due to read length limitations. Long-read sequencing—especially in single-cell contexts [17]—now enables direct observation of complete isoforms at single-cell resolution.

In our ongoing PRIN project SCOLORINA (Single Cell Long Reads inference for Nanopore), we leverage the extended read lengths and phasing capabilities of SCLRS technologies to develop ML/AI tools for isoform discovery and quantification. Our computational framework assembles and classifies full-length isoforms from noisy long-read data, estimates their abundance in individual cells, and investigates allele-specific expression. By combining statistical modeling with deep learning-based noise correction, we aim to achieve a more accurate reconstruction of the transcriptomic landscape and a deeper understanding of alternative splicing in health and disease.

3.5. Unsupervised learning for CLL patients stratification

We apply unsupervised machine learning to stratify chronic lymphocytic leukemia (CLL) patients using multi-modal clinical and laboratory data, as described in [18]. A k-means-based approach enables the objective identification of 6 risk clusters validated on multiple cohorts and independent of classical clinical staging, providing potential guidance for early intervention and clinical trials.

3.6. Integrating AI models and MD simulations for prediction of the dynamics of biomolecules

AI-based structural prediction algorithms are limited to static models, which constrains their ability to capture flexible or intrinsically disordered regions—often crucial to biological function. Environmental factors like pH, temperature, and ligand presence are typically ignored, though they can significantly affect protein conformation. While AlphaFold-Multimer extends predictions to complexes, its accuracy declines for transient interactions or systems involving non-protein components.

Molecular dynamics (MD) simulations naturally complement AI approaches by modeling proteins in motion, capturing conformational flexibility and environmental effects. MD is especially useful for studying mutations, ligand binding, and allosteric regulation through atomistic, time-resolved simulations. Enhanced sampling methods further enable the exploration of rare events and long-timescale transitions, making MD and AI together a powerful toolkit for understanding protein behavior. We aim to explore and advance the integration of these approaches in the coming years.

Acknowledgements

The authors acknowledge the AREA Science Park supercomputing platform ORFEO made available for conducting the research discussed in this paper and the technical support of the LADE staff. We acknowledge financial support from the following souces of funding: “Supporto alla diagnosi di malattie rare tramite l’intelligenza artificiale"- CUP: F53C22001770002; NextGenerationEU within the project PNRR "PRP@CERIC" IR0000028 - Mission 4 Component 2 Investment 3.1 Action 3.1.1; PON “BIO Open Lab (BOL) - Raforzamento del capitale umano”- CUP: J72F20000940007; National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.1, Call for tender No. 1409 published on 14.9.2022 by the Italian Ministry of University and Research (MUR), funded by the European Union – NextGenerationEU– Project Title "Machine Learning algorithms for single-cell genomics from long-reads sequencing technologies" – CUP J53D23015070001.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-4.1 for: Grammar and spelling check, Paraphrase and reword. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication’s content.

References

[1] Z. Yang, X. Zeng, Y. Zhao, R. Chen, Alphafold2 and its applications in the fields of biology and medicine, Signal Transduction and Targeted Therapy 8 (2023) 115.

[2] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, et al., Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences, Proceedings of the National Academy of Sciences 118 (2021) e2016239118.

[3] E. Nguyen, M. Poli, M. G. Durrant, B. Kang, D. Katrekar, D. B. Li, L. J. Bartie, A. W. Thomas, S. H. King, G. Brixi, et al., Sequence modeling and design from molecular to genome scale with evo, Science 386 (2024) eado9336.

[4] T. Hayes, R. Rao, H. Akin, N. J. Sofroniew, D. Oktay, Z. Lin, R. Verkuil, V. Q. Tran, J. Deaton, M. Wiggert, et al., Simulating 500 million years of evolution with a language model, Science (2025) eads0018.

[5] F. Noé, A. Tkatchenko, K.-R. Müller, C. Clementi, Machine learning for molecular simulation, Annual review of physical chemistry 71 (2020) 361–390.

[6] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, et al., Evolutionary-scale prediction of atomic-level protein structure with a language model, Science 379 (2023) 1123–1130.

[7] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al., Highly accurate protein structure prediction with alphafold, nature 596 (2021) 583–589.

[8] L. Valeriani, F. Cuturello, A. Ansuini, A. Cazzaniga, The geometry of hidden representations of protein language models, "Machine Learning in Structural Biology", Workshop at the 36th Conference on Neural Information Processing Systems (NeurIPS) (2022).

[9] L. Valeriani, D. Doimo, F. Cuturello, A. Laio, A. Ansuini, A. Cazzaniga, The geometry of hidden representations of large transformer models, in: Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23, Curran Associates Inc., Red Hook, NY, USA, 2023.

[10] F. Cuturello, M. Celoria, A. Ansuini, A. Cazzaniga, Enhancing predictions of protein stability changes induced by single mutations using msa-based language models, Bioinformatics 40 (2024) btae447.

[11] L. Gao, T. D. la Tour, H. Tillman, G. Goh, R. Troll, A. Radford, I. Sutskever, J. Leike, J. Wu, Scaling and evaluating sparse autoencoders, arXiv preprint arXiv:2406.04093 (2024).

[12] E. N. V. Garcia, A. Ansuini, Interpreting and steering protein language models through sparse autoencoders, in: ICLR 2025 Workshop Learning Meaningful Representations of Life, 2025. URL: https://openreview.net/forum?id=jYPlsr1VHF.

[13] V. Piomponi, A. Cazzaniga, F. Cuturello, Evolutionary constraints guide AlphaFold2 in predicting alternative conformations and inform rational mutation design 65 (2025) 9459–9468. URL: https://doi.org/10.1021/acs.jcim.5c01090. doi:10.1021/acs.jcim.5c01090, publisher: American Chemical Society.

[14] E. T. Russo, F. Barone, A. Bateman, S. Cozzini, M. Punta, A. Laio, Dpcfam: unsupervised protein family classification by density peak clustering of large sequence datasets, PLOS Computational Biology 18 (2022) e1010610.

[15] F. Barone, E. T. Russo, E. N. Villegas Garcia, M. Punta, S. Cozzini, A. Ansuini, A. Cazzaniga, Protein family annotation for the unified human gastrointestinal proteome by dpcfam clustering, Scientific Data 11 (2024) 568.

[16] M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon, A. Žídek, T. Green, K. Tunyasuvunakool, S. Petersen, J. Jumper, E. Clancy, R. Green, A. Vora, M. Lutfi, M. Figurnov, A. Cowie, N. Hobbs, P. Kohli, G. Kleywegt, E. Birney, D. Hassabis, S. Velankar, Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models, Nucleic Acids Research 50 (2021) D439–D444. URL: https://doi.org/10.1093/nar/gkab1061. doi:10.1093/nar/gkab1061.

[17] Z. Xu, H.-Q. Qu, J. Chan, S. Mu, C. Kao, H. Hakonarson, K. Wang, Single-cell omics for transcriptome characterization (scotch): isoform-level characterization of gene expression through long-read single-cell rna sequencing, bioRxiv (2025). URL: https://www.biorxiv.org/content/early/2025/02/06/2024.04.29.590597. doi:10.1101/2024.04.29.590597.

[18] F. Cuturello, F. Pozzo, E. N. Villegas Garcia, F. M. Rossi, M. Degan, P. Nanni, I. Cattarossi, E. Zaina, P. Varaschin, A. Braida, et al., An unsupervised machine learning method stratifies chronic lymphocytic leukemia patients in novel categories with different risk of early treatment, Blood 140 (2022) 4111–4112.