Geometry, Topology and Mechanistic Interpretability of Large Scale AI Models

Alessio Ansuini[1], Lorenzo Basile[1], Federica Bazzocchi[1], Matteo Biagetti[1], Alberto Cazzaniga[1], Stefano Cozzini[1], Francesca Cuturello[1], Diego Doimo[1], Yuri Gardinazzi[1,2], Ruggero Lot[1], Francesco Ortu[1,2], Emanuele Panizon[1], Valerio Piomponi[1], Tommaso Rodani[1,2], Alessandro Serra[1,3], Niccolò Tosato[1,2] and Lucrezia Valeriani[1,2]

[1]*Area Science Park, Trieste, Italy*
[2]*University of Trieste, Trieste, Italy*
[3]*Scuola Internazionale Superiore di Studi Avanzati, Trieste, Italy*

Abstract

We present the research framework and recent results of the Laboratory of Data Engineering (LADE) on advancing the interpretability and reliability of large-scale artificial intelligence models. Our work applies geometric, topological, and mechanistic analysis to modern neural networks. Using tools from data geometry and topological data analysis, we characterize the structure and evolution of internal representations in language and vision-language models. This approach reveals phase transitions, semantic organization, and invariant features that persist across layers and different learning regimes. We further develop mechanistic probes to disentangle behaviour of neural architectures at a fine-grained scale, also investigating how these models interact with factual and counterfactual information. Our methods also map and track information flow in multimodal systems. By integrating these mathematical and computational advances, LADE aims to contribute to safer and more trustworthy AI, strengthen model evaluation, and apply these insights to scientific domains, with a focus on life and material sciences.

Keywords

Interpretability, Transformers, Vision Language Models, Geometry, Topology, Mechanistic Interpretability

1. Memberships and Partnerships

LADE involves more than 10 members focusing on developing methods for interpreting modern AI systems. Current members rank as Staff researchers, Post-Docs, and PhD students, and are affiliated with the Institute of Research and Technological Innovation (RIT) at Area Science Park. The laboratory promotes visiting programs for senior researchers and internships for young researchers and has a vibrant schedule of seminars and scientific events. The laboratory collaborates closely with two experimental facilities within Area Science Park, the Laboratory of Genomics and Epigenomics (LAGE) and the Laboratory of Electron Microscopy (LAME). Moreover, it has tight relations with several local institutions, such as the University of Trieste, the International School of Advanced Studies (SISSA), and the International Center for Theoretical Physics (ICTP). It has a wide network of collaborations across Europe, including Spain, Netherlands, UK, Austria, Germany and several other countries.

2. Scientific Objectives

The core scientific objective of LADE is to advance the interpretability and reliability of Artificial Intelligence from a foundational perspective, developing rigorous mathematical, physical, and computational tools to understand what and how neural networks learn and to devise strategies that make AI systems transparent and safe by design. This theoretical focus is always pursued in synergy with

Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23-24, 2025, Trieste, Italy

alessio.ansuini@areasciencepark.it (A. Ansuini); alberto.cazzaniga@areasciencepark.it (A. Cazzaniga)
0000-0002-3117-3532 (A. Ansuini); 0000-0001-6271-3303 (A. Cazzaniga)

application-driven research: a parallel technological team at LADE is building expertise in constructing AI-augmented infrastructures that support cutting-edge scientific applications, spanning life and material sciences. Our applied and infrastructural vision is realized through active participation in major national initiatives, including two large Italian PNRR projects: one in the domain of life science, the Pathogen Readiness Platform (PRP) project, and the other in the domain of material science, the Nano Foundries and Fine Analysis Digital Infrastructure (NFFA-DI) project. In both projects, LADE plays a key role in the design and implementation of robust, FAIR-compliant digital ecosystems. These engagements enable us to bridge advances in AI interpretability with real-world scientific workflows, connecting foundational studies directly to structural biology, health research, and materials characterization. Furthermore, our involvement in the Malattie Rare project, funded by Regione Friuli Venezia Giulia, provides a direct pathway to apply interpretability methods to medical domains, fostering the translation of theoretical insights into practical, trustworthy AI tools in clinical settings. In this interdisciplinary and multi-project environment, LADE positions itself as a hub where fundamental advances in AI interpretation are integrated with AI-augmented infrastructure development and domain-specific applications.

3. Scientific Context: Interpretability for Artificial Intelligence

As artificial intelligence systems grow in scale and usage, questions of transparency, interpretability, and reliability have become central for both scientific and industrial applications. Large neural models—including language models (LLMs) and vision-language models (VLMs)—now achieve remarkable accuracy in many domains, yet their decision-making processes and internal representations often remain obscure. Globally, the research community is addressing these challenges with interdisciplinary approaches, including mathematics, computer science, and domain applications.

The work at LADE aligns with these efforts, aiming to understand and explain how modern AI systems learn, generalize, and interact with data. Our research is broadly organized around the following themes:

- **Representation geometry and topology:** Uncovering the structure, evolution, and invariants of learned representations in large models;
- **Learning dynamics:** Exploring how different training regimes shape the emergence and transformation of internal features;
- **Mechanistic interpretability:** Developing methods to reveal and intervene on the underlying mechanisms driving predictions;
- **Multimodal AI:** Investigating information flow and interpretability in systems that integrate multiple data types and modalities.

In the following sections, we discuss recent advances and ongoing initiatives in each of these areas.

4. Research Topics

4.1. Geometric and Topological Structure of Neural Representations

The high-dimensional representations formed in hidden layers are the principal building blocks of AI models. LADE research investigates global and local geometric properties of these representations, most notably through the lens of intrinsic dimension, clustering structure, and topological transformations across layers. The first study investigating geometric properties of transformer representations [1] outlined an explicit strategy based on intrinsic dimension. Intrinsic dimension measures the minimum number of variables needed to describe the structure of data confined to a lower-dimensional manifold within a high-dimensional space. Applying intrinsic dimension analysis to neural network representations helps identify where networks compress and organize meaningful features, revealing how and where abstraction and semantic structure emerge across layers. This provides a principled way to assess information complexity and guide the selection of layers for transfer learning or model interpretation. In

[1], computing intrinsic dimension across layers helped to identify intermediate representations that are more suitable for downstream learning tasks. Along these directions [2] shows that deep architectures organize the processing of data point clouds in their internal layers into phases, with transitions marked by sudden changes in intrinsic dimensionality and the emergence of semantic clustering. Furthermore, [3, 4] leverage the intrinsic dimension to identify a distinct phase corresponding to language abstraction, emerging both from the geometric organization of whole sentences and token representations within a paragraph. Using a different perspective, but with a similar goal, [5] builds a framework based on topological data analysis to track the birth and death of topological cycles across model layers, revealing robust universal features that persist across different datasets, architectures, and scales. These techniques contribute to systematizing an effective mathematical language to compare layers, identify redundancy, and support principled model pruning without significantly degrading performance. All these results reveal a remarkable universality in the strategies employed by transformers, which can be used to improve downstream applications in several domains, including proteomics and electron microscopy.

Applications. In the context of proteomics, it was already shown in Valeriani et al. [1, Figure S2] that fundamental geometric considerations lead to the improvement of bioinformatics pipelines for protein remote homology detection without further training. The results on transformer representations have been crucial to improving the description of important properties of biological systems. Insight into transformer behavior led to the development of a state-of-the-art pipeline for the prediction of changes in protein stability induced by single mutations using MSA-based language models in [6]. More recently, based on the clustering of protein language model representations, we developed a method to enhance the prediction of alternative folds for several protein families, and to reveal key coevolutionary signals that inform the design of mutations that stabilize specific conformations [7]. In the context of electron microscopy, the geometry of representations of deep learning models has been at the core of a pipeline leveraging human annotation, machine learning techniques, and instrument metadata filtering to improve annotation that is at the core of the "STM explorer" within the Trieste Advanced Data services (TriDAS) website [8]. Furthermore, vision transformer representation has been used to improve multi-tip artifacts in Scanning Tunneling Microscopy (STM) images [9].

4.2. Evolution of Representations Across Learning Paradigms

The organization of the geometry of representations is not static; it evolves dynamically as a function of training methodology and inference strategies. Our work in [2] demonstrates that supervised fine-tuning and in-context learning (few-shot) regimes induce markedly different geometric organizations in LLMs when solving multiple choice Question Answering tasks. In-context learning leads to the emergence of interpretable semantic structures recovering the subject of the questions in the early layer of the network, while fine-tuning sharpens and separates in clusters the representations of the answer in deeper layers of the model. Furthermore, we show in [4] that the intrinsic dimension of token representations directly correlates with cross-entropy loss and prompt difficulty—the model struggles in high-dimensional regions. This offers a geometric explanation for generalization bottlenecks and may support the development of training-time diagnostics for reliability. Finally, beyond the classical supervised and self-supervised frameworks, we study emergent representation phenomena in models trained by non-standard, biologically-inspired approaches. The Forward-Forward algorithm [10], which substitutes backpropagation with layerwise local objectives, yields qualitatively different representational geometry, with clear implications for both interpretability and computational efficiency. This work shows that biologically inspired, layer-local objectives can produce both generic and distinctive structures, highlighting alternative routes to compositionality and abstraction beyond standard back-propagation. The study of this learning paradigm may point the way forward to more transparent and energy-efficient AI.

4.3. Mechanistic interpretability of LLMs and VLMs

Interpretability requires not only descriptive metrics but also the understanding of model behavior. Mechanistic interpretability aims to do this at a fine-grained level by describing the role of each model component and their interaction. In the context of transformers, it is particularly important to interpret which among the many attention heads operating in the model is responsible for the solution of a specific task. In [11], we develop probing techniques that intervene on the hidden states to separate the competing (or cooperating) mechanisms by which autoregressive LLMs manage factual and counterfactual knowledge. This work shows that facts and counterfactuals are stored, disentangled, and retrieved along specific pathways distributed across layers and heads, and reveals how they interact or "compete" to drive final predictions. These mechanistic insights are essential for trust in tasks such as question-answering or decision support. In [12], the focus is shifted to vision-language transformers, where it is shown that attention heads specialize on distinct visual attributes, encoding task-relevant information in low-dimensional subspaces. By identifying and reweighting these specialized components, we boost model performance without full fine-tuning. This leads to ResiDual, a simple and interpretable method that selectively amplifies relevant residual directions, unlocking the model's latent capabilities and achieving strong zero-shot results with minimal parameter updates. Moving from language-only models to vision-language models (VLMs), the interpretability challenge is amplified by the need to disentangle and audit cross-modal communication. Our work [13] uncovers a dichotomy in multimodal architectures: in some models, all information from the image stream is funneled through a small set of "narrow gate" tokens, which can be precisely localized, intervened upon, and even manipulated; in contrast, other designs distribute the cross-modal interaction more gradually, offering robustness at the expense of fine-grained control. These findings open avenues for the intentional design of interpretable, controllable, and robust multimodal interfaces.

5. Ongoing Activities and Projects

Probing Attention Specialization in Multimodal Transformers. Current work is exploring how specific components within large language and vision-language models contribute to the generation of particular types of content, such as colors, emotions, or toxic information. In particular, applying methods from signal processing and extending the approach of [12] to generative models, we are developing algorithms to identify and rank attention heads in VLMs based on their alignment with interpretable concepts, enabling targeted analysis and intervention. By uncovering this internal specialization, our studies opens the door to practical applications such as improving content safety (e.g., reducing toxicity), enhancing model transparency, or enabling fine-grained control over model outputs.

Disentangling Knowledge Conflicts in Vision-Language Models. We are investigating how vision-language models (VLMs) handle conflicts between their internal factual knowledge and contradictory visual inputs. In particular, we are constructing a counterfactual multimodal dataset to analyze how specific attention heads in the model mediate this tension. We show that a small, localized subset of components is causally responsible for aligning the model's predictions with either visual context or internal knowledge, enabling interpretable and controllable interventions in multimodal reasoning. Our studies will enhance the interpretability and controllability of vision-language models by identifying how they resolve conflicts between visual input and internal knowledge. It will enable more reliable AI systems, supports targeted model editing, and offer tools for stress-testing and benchmarking multimodal behavior in complex scenarios.

6. Conclusion

In this work, we have outlined LADE's ongoing efforts toward geometric, topological, and mechanistic methods to interpret large-scale AI models. Our results suggest that mathematical and mechanistic analyses provide valuable tools for clarifying how neural networks organize and transform information, across a range of architectures, training regimes, and application domains. By increasing our understanding of the inner workings of these models we aim not only to increase their performance in downstream applications but also to make them more interpretable and trustworthy.

Acknowledgments

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.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-4.1 for: Grammar and spelling check, paragraph rephrasing. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] 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.

[2] D. Doimo, A. P. Serra, A. ansuini, A. Cazzaniga, The representation landscape of few-shot learning and fine-tuning in large language models, in: The Thirty-eighth Annual Conference on Neural Information Processing Systems, 2024. URL: https://openreview.net/forum?id=nmUkwoOHFO.

[3] E. Cheng, D. Doimo, C. Kervadec, I. Macocco, L. Yu, A. Laio, M. Baroni, Emergence of a high-dimensional abstraction phase in language transformers, in: The Thirteenth International Conference on Learning Representations, 2025. URL: https://openreview.net/forum?id=0fD3iIBhlV.

[4] K. Viswanathan, Y. Gardinazzi, G. Panerai, A. Cazzaniga, M. Biagetti, The geometry of tokens in internal representations of large language models, 2025. URL: https://arxiv.org/abs/2501.10573. arXiv:2501.10573.

[5] Y. Gardinazzi, G. Panerai, K. Viswanathan, A. Ansuini, A. Cazzaniga, M. Biagetti, Persistent topological features in large language models, 2024. URL: https://arxiv.org/abs/2410.11042. arXiv:2410.11042.

[6] 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.

[7] V. Piomponi, A. Cazzaniga, F. Cuturello, Evolutionary constraints guide alphafold2 prediction of alternative conformations and inform rational mutation design, bioRxiv (2025) 2025–04.

[8] T. Rodani, E. Osmenaj, A. Cazzaniga, M. Panighel, A. Cristina, S. Cozzini, Towards the fairification of scanning tunneling microscopy images, Data Intelligence 5 (2023) 27–42.

[9] T. Rodani, A. Cazzaniga, Enhancing multi-tip artifact detection in STM images using fourier transform and vision transformers, in: ICML'24 Workshop ML for Life and Material Science: From Theory to Industry Applications, 2024. URL: https://openreview.net/forum?id=D84n98dXJU.

[10] N. Tosato, L. Basile, E. Ballarin, G. D. Alteriis, A. Cazzaniga, A. Ansuini, Emergent representations in networks trained with the forward-forward algorithm, Transactions on Machine Learning Research (2025). URL: https://openreview.net/forum?id=JhYbGiFn3Y.

[11] F. Ortu, Z. Jin, D. Doimo, M. Sachan, A. Cazzaniga, B. Schölkopf, Competition of mechanisms: Tracing how language models handle facts and counterfactuals, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguis-

tics, Bangkok, Thailand, 2024, pp. 8420–8436. URL: https://aclanthology.org/2024.acl-long.458/. doi:10.18653/v1/2024.acl-long.458.

[12] L. Basile, V. Maiorca, L. Bortolussi, E. Rodolà, F. Locatello, Residual transformer alignment with spectral decomposition, Transactions on Machine Learning Research (2025). URL: https://openreview.net/forum?id=z37LCgSIzI.

[13] A. Serra, F. Ortu, E. Panizon, L. Valeriani, L. Basile, A. Ansuini, D. Doimo, A. Cazzaniga, The narrow gate: Localized image-text communication in vision-language models, 2025. URL: https://arxiv.org/abs/2412.06646. arXiv:2412.06646.