

Ruggero Cortini, PhD

Data Scientist | Computational Biology & Bioinformatics | NLP for Life Sciences

+34 652 02 65 67 | ruggero.cortini@gmail.com | [linkedin.com/in/ruggerocortini](https://www.linkedin.com/in/ruggerocortini) | github.com/rcortini | Barcelona, Spain

PROFILE

Data Scientist and Consultant with 5+ years of experience applying **NLP, machine learning, and advanced analytics to life sciences and biomedical research challenges**. Built and delivered scalable pipelines for mining scientific literature, patents, and clinical trial data at scale. Deep domain expertise in **computational biology and biophysics** (PhD from Imperial College London, postdoctoral research at CRG Barcelona), with hands-on experience analyzing **omics datasets (single-cell RNA-seq, ChIP-seq), molecular simulations, and heterogeneous biomedical data sources**. Proven track record of **end-to-end consulting project delivery** for international clients, translating complex scientific data into actionable insights through dashboards, visualizations, and evidence-based reports. Experienced in client engagement, cross-functional team leadership, and mentoring.

PROFESSIONAL EXPERIENCE

Data Scientist & Consultant | SIRIS Academic

Oct 2019 – Apr 2025

Barcelona-based consulting firm delivering data-driven strategy for research institutions, public bodies, and industry partners

- **Led 10+ cross-functional consulting projects** for international clients (Google DeepMind, Sorbonne Université, Generalitat de Catalunya), developing NLP and AI-powered decision-support tools. Scoped projects by understanding client needs, developed analytical strategies, and delivered evidence-based recommendations.
- **Designed and deployed scalable Python pipelines for NLP analysis of biomedical and scientific corpora** (tens of millions of documents including publications, patents, and clinical trials), performing entity extraction, document classification, topic modeling, and semantic search to map research strengths and emerging therapeutic areas.
- **Integrated heterogeneous data sources** (publications, patents, clinical trials, project databases, experimental outputs) into coherent analytical frameworks using ETL pipelines, data normalization, and deduplication techniques, producing interactive dashboards and reports for strategic decision-making.
- Applied **ML and statistical methods** (classification, clustering, topic modeling, dimensionality reduction) in Python and R to validate results, assess model robustness, and generate predictive insights across biomedical, environmental, and technological domains.
- **Coached and mentored junior data scientists**, guiding best practices in NLP workflows, code quality, and reproducible analytics.

Marie Curie Post-doctoral Fellow | Centre for Genomic Regulation (CRG)

Jan 2016 – Sep 2019

World-class research centre in genomics and molecular biology, Barcelona

- **Developed computational models of DNA-protein interactions and chromatin dynamics**, combining analytical theory with large-scale molecular simulations on HPC clusters (CPUs and GPUs) to generate insights into transcription factor binding, gene regulation, and 3D genome organization. Results published in *Nature Communications*.
- **Analyzed multi-omics experimental data**: processed and interpreted single-cell RNA-seq datasets (gene expression quantification, clustering, differential expression), ChIP-seq data (peak calling, genome alignment), and DNA sequence data to establish correlations between chromatin structure and gene regulation.
- Built **reproducible Python analysis pipelines** to process hundreds of thousands of simulations and experimental datasets, with strong attention to scalability, version control, and documentation.

Post-doctoral Fellow | Sorbonne Université / ORNL

Jul 2013 – Dec 2015

Paris, France; visiting researcher at Oak Ridge National Laboratory, USA

- Mathematical modelling of DNA mechanical properties using Molecular Dynamics, Monte Carlo simulations, and analytical theory. Performed **large-scale all-atom molecular dynamics simulations** on HPC infrastructure at Oak Ridge National Laboratory.

TECHNICAL SKILLS

Programming: Python (primary), R, SQL, C/C++, Bash. Contributions to open-source projects (MDAnalysis, OpenAlex).

AI, NLP & ML: Pre-trained LLMs (OpenAI, HuggingFace Transformers), topic modeling (BERTopic, TF-IDF), entity recognition and text vectorization (spaCy, NLTK), semantic search, document classification and clustering, logistic regression, random forests, SVMs, k-means, model evaluation.

Life Sciences & Bioinformatics: Single-cell RNA-seq analysis, ChIP-seq processing, genome alignment, gene expression quantification, molecular simulations, omics data integration, biomedical text mining, scientific literature curation.

Data Engineering & DevOps: ETL pipelines, PostgreSQL, data normalization, Docker, CI/CD (GitHub Actions, Bitbucket), Git, HPC (SLURM, PBS).

Visualization & Reporting: Matplotlib, Plotly, R/ggplot2, Apache Superset, Google Looker Studio, scientific dashboards.

EDUCATION

PhD in Physics (DNA Biophysics) | *Imperial College London* Nov 2009 – Jul 2013
Theoretical and computational study of electrostatic interactions in DNA systems

MSc in Condensed Matter Physics | *Università di Napoli "Federico II"* 2006 – 2009
110/110 cum laude

BSc in Physics | *Università di Napoli "Federico II"* 2003 – 2006
110/110 cum laude

SELECTED PROJECTS (GITHUB)

- **sc_hiv** (R, Python): End-to-end analysis of single-cell RNA-seq data from cell lines expressing HIV proteins. Performed gene expression quantification, clustering, differential expression analysis, and established correlations between HIV integration locus and transcriptional patterns. Demonstrates omics data analysis and translational research application.
 - **VocTagger** (Python): High-performance NLP library for entity extraction and thematic mapping in large scientific corpora (spaCy, multiprocessing). Used to classify 30,000+ Horizon 2020 projects against UN SDGs via controlled vocabulary matching. Demonstrates biomedical text mining at scale.
 - **sbs_tracers** (Python): Scalable simulation and analysis pipeline for coarse-grained models of DNA-protein interactions on HPC clusters. Published in *Nature Communications*. Demonstrates computational biology workflows and large-scale data processing.
-

SELECTED PUBLICATIONS

- Theoretical principles of transcription factor traffic on folded chromatin, *Nature Communications* (2018)
 - The Physics of Epigenetics, *Reviews of Modern Physics* (2016)
 - Theory and Simulations of Toroidal and Rod-Like Structures in DNA Condensation, *J. Chem. Phys.* (2015)
-

ADDITIONAL

Languages: Italian (native); full professional proficiency in English, Spanish, Catalan, and French

Training: Professional trainer at The Paper Mill (2018–2019): workshops on scientific communication for PhD students at ICTA and IDIBAPS