María José Molina Contreras, PhD

Sr. Data Scientist | NLP, SLM/LLMs, Agents & Applied Machine Learning | OSS contributor | PhD in Biotech

Berlin, Germany | mjmolinacontreras@gmail.com | linkedin.com/in/mjmolinacontreras
github.com/mjmolina

Summary

I am an enthusiastic professional with almost eight years applying machine learning in industry. Classical NLP, deep learning and time series, and LLM agents and retrieval systems in R&D setup. My main responsibilities to serve as project lead, oversee hiring, handle data processing, develop models, and deploy to production.

Research and Applied Experience

Independent Researcher, Independent Research – Berlin, Germany Nov 2025 – present

- Independent research and prototyping across several projects, including research on LLM agents as reliability, evaluation and failure analysis. Alongside a project applying machine learning to molecular detection in food chain.

Senior Research Data Scientist Researcher, Arbolitics – Berlin, Germany Oct 2024 – Nov 2025

Predictive modelling from geospatial, sensor and text data in R&D

- Designed and built forecasting models combining Convolutional Neural Networks with Adaptive Neuro-Fuzzy Inference Systems (ANFIS) to predict cherry tree yields, with explainability built into the model architecture rather than added post hoc.
- Achieved over 90% accuracy across varieties under field conditions by integrating satellite imagery, climate data and biomass sampling validated against ground-truth measurements.
- Built anomaly detection systems for agricultural fields, including evaluation of detection reliability under distribution shift.
- Led data strategy and dataset construction, including labelling protocols and quality control.

Senior Data Scientist Researcher, One Five – Berlin, Germany Oct 2023 – Oct 2024

R&D in applied NLP, LLMs and knowledge representation for circular packaging R&D.

- Built an agent-based system over open-source LLMs that generated queries against a graph database, giving non-technical researchers a natural-language interface to experimental data.
- Designed and built a custom Ontology and Knowledge graph representing experimental results, enabling structured exploration of an otherwise unstructured corpus.
- Built and evaluated RAG over a legal corpus, including analysis of silent retrieval failures, including ingestion, indexing and retrieval evaluation.
- Fine-tuned open-source models for domain-specific knowledge and evaluated relevance and factual accuracy of responses.
- Developed a multimodal approach combining image and text representations.

Senior Data Scientist, INFARM – Berlin, Germany Mar 2021 – Oct 2023

Machine learning for controlled-environment in agriculture, IoT sensor data at scale.

- Developed supervised and unsupervised deep learning pipelines for hardware anomaly detection in irrigation systems, deployed in production.
- Built an end-to-end pipeline from problem formulation to deployment for optimising light and irrigation, reducing resource consumption while maintaining plant health.
- Applied time-series forecasting (ARIMA/ARIMAX, gradient boosting, LSTM) and computer vision methods to sensor and imaging data.

Data Scientist and Analytics, ZAGENO – Berlin, Germany Dec 2018 – Mar 2021

NLP and information extraction for a scientific e-commerce platform.

- Applied NLP methods to product names, brands and supplier data to improve catalogue categorisation.
- Improved the recommendation system using relational models and text classification.
- Enriched a scientific product database by extracting attributes and linking to the primary literature.

Teaching

Invited Lecturer, Leuphana Universität Lüneburg – Lüneburg, Germany Dec 2020 – Jan 2021

- Designed and delivered lectures on networking, web development, Flask, IoT and microcontrollers, concluding with a project-based workshop.

Education

Universidad de Barcelona, PhD in Biotechnology (Molecular Biology) – Barcelona, Spain Dec 2011 – Sept 2017

- Thesis: "Plant proximity responses in *Arabidopsis* and related species"
- Integrated transcriptomic, proteomic, genomic and physiological data to characterise shade-avoidance signalling in Arabidopsis and Cardamine.

Universitat Autònoma de Barcelona, MSc in Plant Biology and Biotechnology – Barcelona, Spain Sept 2010 – Sept 2011

Universitat Autònoma de Barcelona, BSc in Biology – Barcelona, Spain Sept 2005 – Sept 2010

Publications

- ncbi.nlm.nih.gov/pubmed/?term=molina-contreras+MJ

Technical Skills

Languages: Python, SQL, Cypher

Machine Learning: scikit-learn, SciPy, statsmodels, supervised and unsupervised learning, gradient boosting, statistical and mechanistic modelling

Deep learning: PyTorch, TensorFlow, Keras, CNNs, RNNs, LSTMs, ANFIS, Transformers

NLP: spaCy, NLTK, Gensim, text classification, information extraction, data mining, scraping

LLMs and agents: Fine-tuning, RAG, agent frameworks, LangChain, LangGraph

Evaluation: Benchmark and test design, failures analysis, retrieval and generation evaluation, Giskard, LangSmith, SHAP

Data and infrastructure: PostgreSQL, Graph databases(Neo4J), vector databases(Qdrant), Google Cloud Platform (BigQuery,Vertex AI), Docker, Git, CI/CD, MLOps

Visualisation and apps: Plotly, Bokeh, Streamlit, Dash, Flask, Looker, Tableau

Honors

- Community Service Award from Python Software Foundation (PSF)
- Outstanding PyLady Award (PSF)
- Bertelsmann Data Analysis Challenge Scholarship (Udacity)
- "Climate Change and Biodiversity" (top 5 of 140 group projects)
- "Breast Cancer in Women" (honorary prize, top 10 of 100) individual projects

Languages

Spanish: Native

Catalan: Native

English: Fluent (technical and colloquial)

German: Elementary

French: Elementary