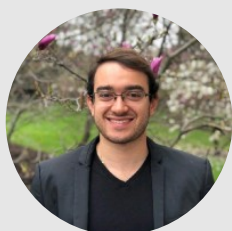# Andres F. **Cano Botero**
Researcher | Data Scientist

About me

I am a Data Scientist with a research background in quantum chemistry and plasma physics. My passion lies in exploring the synergy between data-driven methodologies and the natural sciences. Having witnessed the impact of AI/ML in computational chemistry, I am convinced that novel, physics-informed algorithms hold immense potential for simulating and extracting insights from our world. I am eager to contribute to this endeavor.

Contact

- 🧍 From Manizales, Colombia
- ✉️ andrescano30@gmail.com
- 📞 +1 404-956-0320
- 📍 Atlanta, GA
- 🔗 andresfelipecano
- 🔗 andresfcano

Languages

- 🇪🇸 Spanish - Native Language
- 🇬🇧 English - Professional Knowledge
- 🇫🇷 French - Conversational Knowledge

🎓 EDUCATION

08/2020-08/2022

M.S. Mechanical Engineering 📍 Boulder, CO
University of Colorado, Boulder
Thesis: *Alkyl End Group Effects on the Thermal Decomposition of Fuel Additives.*
Advisor: Dr. Nicole J. Labbe
My work focused on optimizing and leveraging quantum chemistry methods and kinetic modeling to identify the decomposition products of two potential diesel additives. Studying the unimolecular decomposition of these molecules provides valuable insight into the formation mechanism of soot particles in the exhaust of diesel engines.
GPA: 3.95/4.0

08/2015-05/2019

B.S. Astrophysics 📍 Rochester, NY
University of Rochester
Sigma Pi Sigma (National Physics Honor Society)
Lorenzo de Zavala Scholar, National Hispanic Institute
American Physics Society Member

💼 WORK EXPERIENCE

2022-present

Data Scientist 📍 Atlanta, GA (Remote)
Dataiku
- Scoped, built and deployed sophisticated production-level AI projects for customers spanning multiple domains including NLP, optimization, computer vision, LLM applications, and time series forecasting.
- Collaborated with R&D and Product teams to develop custom Python-based plugins that enhance Dataiku's functionality.
- Provided technical leadership to non-profit organizations, such as the ALMA observatory, that want to develop data-driven solutions.
- Offered strategic input, coaching, and training to customers across multiple industries (i.e. technology, pharma, banking) that want to implement AI solutions in their work.

2020-2022

Graduate Researcher 📍 Boulder, CO
Labbe Lab, University of Colorado Boulder
- Leveraged CU supercomputing resources, and legacy software (written in C++ and Python) to build reaction mechanisms that uncover the thermal decomposition of soot-reducing diesel additives.
- Optimized research methods by writing a backward-compatible high-performance computing pipeline that used Python and Unix shell scripts to generate, submit, and run quantum chemistry calculations to the RMACC Summit supercomputing cluster. and saves days to weeks of computational work.

2019-2020

Data Scientist 📍 Bogota, Colombia
Opera Urban Investment Firm
- Developed a Streamlit App that showed market trends, sales forecasts, and client activity from the company's website. This established direct communication between internal teams and improved client outreach.

2018-2019

Undergraduate Researcher 📍 Rochester, NY
Extreme State Physics Laboratory (XSPL), University of Rochester
- Built a high voltage transformer to supply ±200KV to the High Amperage Driver for Extreme States (HADES), a 250GW pulsed-power driver used to model matter at high temperatures and pressures.

📖 PUBLICATIONS

Journal Article
in progress

End Group Effects on the Thermal Decomposition of Oxymethylene Ethers, AFC Botero, P. Shah, O Aguirre, J Sampathkumar, NJ Labbe, *Aimed for: Fuel*

Journal Article
in progress

Temperature- and Pressure-Dependent Decomposition of Diethoxymethane and Dipropoxymethane Fuel Radicals., AFC Botero, P.Shah, O Aguirre, NJ Labbe, *Aimed for: PCCP*

——— **About me** ———

This page highlights the various projects I've worked on, both in academic research and industry settings. It showcases my experience in developing automation tools for computational chemistry, building machine learning models for data science operations, and creating applications that streamline processes like personal literature reviews and intricate MLOps pipelines. Each project demonstrates my ability to solve complex problems by leveraging computational tools and together they paint a picture of my professional journey so far.

——— **Contact** ———

🧍 From Manizales, Colombia

✉ andrescano30@gmail.com

📞 +1 404-956-0320

📍 Atlanta, GA

🔗 andresfelipecano

🔗 andresfcano

🗀 RELEVANT RESEARCH & PROJECTS

Academic Projects - Labbe Lab

Optimizing Computational Chemistry Methods   📍 CU Boulder

Motivation: Modeling biofuels at a molecular level requires exploring molecular rotors during geometry optimization to find an accurate calculation of the ground state energies of the lowest energy conformer. Traditionally, this involves slow, manual work by computational scientists.

Solution: I developed a Python script to automate rotor identification and file preparation for supercomputing jobs. The script was later expanded to generate input files for ab initio and Density Functional Theory (DFT) calculations, streamlining the workflow and reducing manual effort. This automation saves weeks of computational work over time, improving efficiency in molecular modeling.

Unimolecular Decomposition of Diesel Additives   📍 CU Boulder

Motivation: The high compression ratios in diesel engines makes them prone to incomplete combustion, leading to the formation of harmful pollutants such as NOx. Efforts to reduce soot emissions have led to the exploration of fuel additives that blend well with diesel and minimize the pollutant production from diesel engines.

Solution: I conducted a study on the unimolecular decomposition of two potential diesel additives to gain insights into their behavior during combustion. This study was followed by a deeper analysis of the decomposition of the primary radicals generated in the initial stages of combustion. The results revealed the key reactions that occur during the onset of combustion, providing valuable knowledge for developing engine simulations of these fuels. More details in my thesis.

Industry Projects - Dataiku

Automating RFPs with Self-Hosted LLMs   📍 Dataiku

Motivation: Field experts at a major insurance company manually handled hundreds of Requests for Proposals (RFPs), many with recurring themes, leading to inefficiencies, missed deadlines, and ultimately loss in market competitiveness.

Solution: I developed a Retrieval-Augmented Generation (RAG) pipeline using open-source embedding and large language models (LLMs) running on local GPUs. A Dash web application enabled the experts to query the LLM in real-time with dozens of RFPs, significantly improving RFP response time.

MLOps and CI/CD of Data Science Projects   📍 Dataiku

Motivation: A U.S. airline's data science team was developing multiple projects using Dataiku but lacked a structured MLOps process. As a result, there was no separation between production and development environments, which led to security risks.

Solution: We collaborated with the airline to develop a clear MLOps strategy and created an Azure DevOps pipeline that automated project deployment. The pipeline also performed security and data quality checks before deployment, ensuring that only validated models were put into production. This greatly improved operational efficiency and project transparency across the organization.

AI for Science - ML for Observation Quality   📍 Dataiku

Motivation: After each science observation, Astronomers on Duty (AoD) had to manually assess the quality of the data, categorizing it as Pass, Semi-Pass, or Fail. This process was time-consuming, especially for uncertain cases.

Solution: We developed a machine learning model that automated the assessment of observational data quality. The model could recommend Pass, Semi-Pass, or Fail, allowing the AoD to focus only on uncertain cases, thereby improving the overall efficiency of the data review process and reducing workload.

Science RAG: Streamlining Literature Review with AI-Driven Analysis   📍 Personal

Motivation: Conducting literature reviews is often a time-consuming process, especially when searching for specific methods, insights, or discussions across multiple papers with abundant details.

Solution: I developed a local Python Retrieval-Augmented Generation (RAG) application, using Ollama as the backend and a Streamlit web application as the front end. The tool allows users to drag and drop PDFs of research papers and directly ask questions to a large language model (LLM). The application generates detailed responses with source citations, significantly speeding up the literature search process and helping users extract key information quickly.

− **Soft Skills and Strengths** −

Curiosity	Self Confidence
Self Driven	Problem Solving
Team Work	Constantly Learning
Good Communication

—— **Professional Skills** ——

| Data Science | Machine Learning |
| Computational Physics/Chemistry |
| Astrophysics & Mathematics |
| Engineering Design |

—— **Other Interests** ——

- Guitar
- Chess
- Gym
- Travel
- Triathlon
- Cooking
- Calisthenics
- Books

🖥 **INFORMATION TECHNOLOGY SKILLS**

Data Science and App Development	**Python**: *Higly Specialized* **HTML/CSS/JavaScript**: *Intermediate*
Version Control	**Git**: *Intermediate*
AI/ML	**Sci-kit Learn** : *Advanced* **Hugging Face & Langchain** : *Intermediate* **PyTorch & Keras** : *Intermediate*
Big Data	**Spark** : *Intermediate* **SQL** : *Intermediate* **Kubernetes & Cloud Deployments** : *Beginner*
Typesetting	**LaTeX**: *Intermediate*
Chemical Kinetic Modeling and Simulation	**Gaussian16** : *Advanced*