Internship Opportunity for Master's in Research or Engineering School Students – Université Côte d'Azur

Internship Topic: **Scientific Data Extraction in Chemical Ecology Using Autonomous AI Agents**

Host Laboratory: Institut de Chimie de Nice (ICN) - UMR7272 CNRS – UniCA

Contacts: Dr. Sébastien Fiorucci et Dr. André Lanrezac (UniCA)
sebastien.fiorucci@univ-cotedazur.fr ; andre.lanrezac@univ-cotedazur.fr

Internship Funding: Stipend of approximately 630 euros per month



Duration: 6 months starting from early 2027

Scientific Description: Understanding the chemical ecology of an insect requires in-depth knowledge of the odors it perceives and their effects. This internship is part of a broader project whose main objective is to predict the response of an insect olfactory receptor using computational approaches. By deciphering the molecular basis of insect olfaction, we can identify new active semiochemicals that can manipulate insect behavior and develop more effective biocontrol methods. Currently, the M2iOR database[1] contains more than 35,000 insect receptor-odorant interactions manually extracted from the literature. To accelerate its enrichment while ensuring data quality, the objective of the internship is to develop autonomous AI agents for extraction, based on advanced language models (LLMs run locally or hosted remotely, for example via the Albert API). The intern will adapt existing approaches to the M2iOR data schema to extract structured information from text, tables, figures, and supplementary materials [2-5]. The work will include the use of reference databases (PubChem, UniProt, NCBI), evaluation of agents' performance, depending on the LLM used, against the curated dataset, and integration of the agents into a litterature → M2iOR pipeline with confidence scores and human validation. Depending on the progress of the internship, the impact of the new data on predictive models of olfactory activity may be evaluated.

References:

[1] M2iOR: A Quantitative, Context-Resolved Atlas of Insect Odorant Receptor-Odorant Interactions. D. Geslin, T. Pitakbut, A. Lanrezac, A. De Fouchier, Z. Bouchouireb, A.T. Thorhallsson, D. Gilardot, C. Meslin, N. Montagné, E. Jacquin-Joly, S. Fiorucci, A. Groot, J. Topin, Nucleic. Acids. Res. 2026. *accepted*. https://m2ior.chemsensim.fr/

[2] From Papers to Property Tables: A Priority-Based LLM Workflow for Materials Data Extraction. arXiv:2604.07584. https://doi.org/10.48550/arXiv.2604.07584

[3] BioMiner: A Multi-modal System for Automated Mining of Protein-Ligand Bioactivity Data from Literature, arxiv:2604.21508. https://doi.org/10.48550/arXiv.2604.21508

[4] Detailed curation of biological samples and experimental designs for genomics using LLM-supported agentic workflows. bioRxiv:2026.07.30.741874 https://doi.org/10.64898/2026.07.30.741874

[5] Benchmarking LLM Agents on Real-World Biological Database Curation for Data-Driven Scientific Discovery. https://doi.org/10.1145/3770855.3817556

Keywords: AI agent, LLM, information extraction, database, semio-chemical, olfactory receptor.

Methodology: Python, machine learning.

Skills sought: The candidate should have a degree in physical chemistry, physics, biochemistry or computer science and have a strong interest in biomolecular simulations. Knowledge of structural biology is a plus for understanding the systems under study. Experience with Linux environments and Python programming is required.