# Oualid Bencheikh Lehocine

AI & Data Science Research Candidate

Algeria, Constantine | bencheikhlehocineoualidmouatez@gmail.com | 0671263226

linkedin.com/in/oualid-bencheikh-lehocine-86a914280 | github.com/wbench12

## RESEARCH PROFILE

Final-year Computer Science engineering student at the Higher National School of Computer Science (ESI-SBA), seeking an international research internship or final-year research project in Artificial Intelligence and Data Science. Hands-on experience turning research ideas into reproducible experiments: dataset curation, feature engineering, strong baselines, model comparison, error analysis, and explainability. Particularly interested in data-efficient learning, representation and graph learning, and trustworthy model evaluation. Backed by professional software-engineering experience that makes research prototypes clean, reproducible, and deployable.

## RESEARCH INTERESTS

Machine Learning • Data-Efficient Learning (Semi-Supervised, Active Learning) • Representation Learning • Graph Neural Networks • Multimodal Learning • Explainable and Trustworthy AI • Model Evaluation, Generalization and Robustness • AI for Science

## RESEARCH EXPERIENCE

**Research Intern (AI/ML Focus)** September 2025

**Centre de Recherche en Biotechnologie (CRBT) – Constantine, Algeria**

- Built and evaluated QSAR pipelines for molecular property prediction using molecular descriptors and fingerprints, focused on reproducibility and fast experimentation.
- Reproduced published (2020–2021) SMILES-based NLP QSAR workflows and extended them by replacing Word2Vec + ConvLSTM representations with pretrained ChemBERTa embeddings.
- Benchmarked downstream models (Multiple Linear Regression, Random Forest, XGBoost, SVM) on LD50 and IGC50 toxicity datasets.

## RESEARCH PROJECTS

**Data-Efficient and Explainable Machine Learning for Molecular Prediction** 2025 – 2026

*Collaborative research project; molecular datasets used as experimental benchmarks*

- **Research question:** How learning from limited labels and combining complementary data representations can improve predictive performance, while keeping predictions interpretable and reliable.
- **Semi-supervised learning:** Compared self-training (pseudo-labeling), co-training, label propagation, and label spreading on FDA-approval, ClinTox, and Tox21 under varying labeled/unlabeled ratios; self-training performed best on approval and clinical-toxicity tasks, co-training on Tox21.
- **Feature selection & reliability:** Designed a genetic-algorithm feature-selection pipeline for AMES mutagenicity, reducing descriptors from 154 to 60 while improving ROC-AUC from 0.8267 to **0.8694**. Defined the applicability domain with PCA, K-means, and distance thresholds, covering **95.2%** of test compounds.
- **Representation learning:** Built a multi-target benchmark of 8,409 curated molecules across five protein targets, comparing Morgan fingerprints and ChemBERTa baselines against graph models (GIN, GAT) and a hybrid ChemBERTa–GIN model with LoRA adaptation and late fusion.
- **Rigorous evaluation:** Used molecule-level train/validation/test splits and a masked multi-target loss for missing labels. Morgan + MLP reached a macro ROC-AUC of **0.9475** vs. 0.8654 for the hybrid GIN, showing the value of strong, well-tuned baselines.
- **Explainability:** Implemented gradient-based atom attribution, GNNExplainer-style node/bond masks, ChemBERTa token attribution with token-to-atom alignment, and fused atom-level explanations, visualized as activity heatmaps and similarity graphs.
- **Tools:** Python, PyTorch, scikit-learn, RDKit, NumPy, pandas, SciPy, Matplotlib, Jupyter, Git.

## EDUCATION

**Engineering Degree & Master's in Computer Science** Expected 2027

*Higher National School of Computer Science (ESI-SBA), Sidi Bel Abbès*

- Solid mathematical foundations in Probability, Statistics, Linear Algebra, Multivariate Calculus, and Algorithms.
- Focus on Software Engineering and System Design; active participation in multidisciplinary projects.

**Baccalaureate** 2019

*Ibn Taymiya High School, Constantine, Algeria*

- Graduated among the top of my class with excellence.

## TECHNICAL SKILLS

- **Machine Learning:** Supervised & semi-supervised learning, classification/regression, Random Forest, XGBoost, SVM, MLP, feature selection, cross-validation, ROC-AUC and precision-recall analysis
- **Deep Learning:** PyTorch, transformers (ChemBERTa), graph neural networks (GIN, GAT), LoRA fine-tuning, multimodal/late fusion, gradient-based attribution
- **Data Science:** Python, NumPy, pandas, SciPy, scikit-learn, Matplotlib, RDKit, Jupyter, EDA, preprocessing, feature engineering
- **Research Computing:** Git, GitHub, Linux, Bash, Docker, Google Colab, experiment scripting, REST-based model serving
- **Software Engineering:** FastAPI, Django REST Framework, React/Next.js, TypeScript, PostgreSQL, Redis, CI/CD, AWS

## SELECTED ENGINEERING EXPERIENCE

**Fullstack Developer – Zalisa (zalisa.io/en)** 2026 – Present

- Contribute to a production platform built on FastAPI microservices, React/Next.js, TypeScript, PostgreSQL, and AWS, designing authentication, authorization, and access-control subsystems.

**Backend / Fullstack Developer – Re:Healthify & State Library of Tissemsilt** January 2025 – July 2025

- Built secure APIs (2FA, RBAC, social auth) for a UK health startup and a high-traffic public library platform; cut response times by $\sim$40% with Redis caching.
- Developed a university final-year-project management platform used by 200+ students, professors, and administrators.

## LANGUAGES

Arabic (Native) · English (Fluent) · French (Conversational)