Yves Gaetan Nana Teukam

Phone: +41 76 728 31 21
Email: yves.g.nana@gmail.com
Website: https://yvesnana.github.io
Passport: Italian
Swiss Permit: B

PROFESSIONAL SUMMARY

Detail-oriented Data Scientist with 4 years of experience developing AI/ML-based tools for complex biological problems. Currently completing a Ph.D. in Biomedical Engineering (expected January 2025) at IBM Research Zürich and Eindhoven University of Technology, focusing on Language Modelling for Protein Design. Expert in machine learning (e.g. generative modelling and language models), Omics, and bioinformatics, with a proven track record of optimizing biomolecules and enhancing model performance for biocatalysis, drug discovery, and green chemistry. Proficient in advanced ML frameworks (e.g. DeepSpeed, Transformers, and Peft) for analysing complex biological/chemical datasets. Contributed to open-source projects like GT4SD and published in high-impact journals like Nature Communications. Fluent in English, French, Italian, and Spanish. Passionate about driving innovative projects, fostering collaboration, and achieving team goals.

WORK HISTORY

PRE-DOCTORAL RESEARCHER | 01/2022 to Current
IBM Research - Zürich, Switzerland - Zürich, Switzerland

- Lead developer of Enzeptional, a framework integrating genetic algorithms (GAs) and large language models (LLMs) for enhanced enzyme design, improving their feasibility and stability.
- Designed and implemented a molecular dynamics simulation framework (using GROMACS) to validate the structural integrity and stability of AI-generated enzyme designs. This assessment of protein stability and dynamics provided valuable insights to guide further optimization and enhance the likelihood of real-world functionality and applicability of the optimized enzymes.
- Lead developer of the RXNAAMapper, a state-of-the-art tool utilizing transformer-based language models and Byte Pair Encoding to accurately predict enzymatic binding sites in protein sequences. This approach outperformed existing transformer-based methods, achieving a 38% improvement in accuracy and a 30% reduction in false positives with respect to the baseline.
- Contributed to GT4SD, an open-source library for training and fine-tuning generative models (e.g., VAE and GAN) to accelerate scientific discovery.

RESEARCH INTERN | 02/2021 to 07/2021
IBM Research - Zürich, Switzerland - Zürich, Switzerland

- Developed a synthesis planning approach integrating biocatalysis with data-driven learning, enhancing efficiency and sustainability in chemical synthesis. Applied transfer learning techniques using OpenNMT-py for effective model training and leveraged IBM's RXN4Chemistry API for accurate reaction prediction and data processing.
- Achieved a 49.6% top-1 accuracy in biocatalyzed forward predictions using a Transformer model, marking a major advancement in computational chemistry and AI-driven drug discovery.
- Conducted in-depth attention mechanism analysis in the encoder/decoder layers, providing insights into model interpretability and decision-making processes.
- Leveraged Python tools for data analysis and modelling, including Pandas, TensorFlow, Keras, RDKIT, and Biopython, to process and analyse large-scale biological datasets.

DATA SCIENCE AND BIOINFORMATICS PROJECT LEAD | 05/2020 to 09/2020
StemAway - California, USA - California, USA (Remote)

- Mentored 30 students from various countries and academic backgrounds through all stages of gene expression analysis, covering data collection, processing, and analysis.
- Utilized Python and R, focusing on bioinformatics tools such as DESeq2, Bioconductor, edgeR, and limma for comprehensive gene expression analysis.

RESEARCH INTERN | 04/2019 to 07/2019
Sequentia Biotech - Barcelona, Spain - Barcelona, Spain

- Collected microbe data from NCBI and integrated it with data from various experiments and research studies.
- Conducted human gut microbiome analysis using bioinformatics tools, including Samtools, BLAST, and Bowtie for sequence alignment and variant calling.

SKILLS

- **Programming Languages & Version Control**: Python, R, Bash, Linux, Git, GitHub
- **Data Analysis and Tools**: Statistical Analysis, Data Mining, Data Visualization, Pandas, NumPy, SciPy
- **Machine Learning & AI**:
  - **Core Skills**: Machine Learning, Deep Learning, Generative Modelling, Natural Language Processing (NLP)
  - **Specific Techniques**: Transformers, LSTMs, Attention Mechanisms, Sequence-to-Sequence Models, Few-Shot learning, Zero-Shot learning
  - **Frameworks & Libraries**: TensorFlow, Keras, PyTorch, Scikit-learn, XGBoost, SFTrainer, Hugging Face Transformers, OpenNMT, SpaCy, NLTK, Fairseq
  - **Models & Architectures**: BERT, ProTrans, ESM2, T5, XLNet
  - **Optimization & Training**: Hyperparameter Tuning, Model Compression (Quantization, Pruning)
  - **Advanced Techniques**: Transfer Learning, Reinforcement Learning, Self-Supervised Learning
- **Bioinformatics & Computational Biology Tools**: Protein Optimization, Molecular Dynamics (Gromacs), Evolutionary Algorithms, Samtools, BLAST, Bowtie, DESeq2, edgeR, Rosetta, AlphaFold
- **Development Tools & Experiment Tracking**: Jupyter Notebooks, VS Code, MLflow, Weights & Biases
- **Soft Skills**: Team Collaboration, Leadership, Mentoring, SCRUM, Agile Methodologies
- **Communication Skills**: Conferences Presentations and Scientific Divulgation

EDUCATION

Ph.D. in Biomedical Engineering
IBM Research Zürich & Eindhoven University of Technology – Zürich, Switzerland & Eindhoven, Netherlands
01/2022 to 01/2025 (expected graduation)

Master of Science in Data Science
University of Rome La Sapienza – Roma, Italy
09/2019 to 10/2021

Exchange Program Erasmus
ESCI-Universitat Pompeu Fabra – Barcelona, Spain
09/2018 to 02/2019

Bachelor of Science in Bioinformatics
University of Rome La Sapienza – Roma, Italy
09/2016 to 06/2019

AWARDS

- 1st IEEE Open Software Service Awards as part of the GT4SD team. 2022.
- Sandmeyer Award of the Swiss Chemical Society as part of the RXN for Chemistry project team. 2022.

PUBLICATIONS

- **Teukam, Yves Gaetan Nana**, et al. "Language models can identify enzymatic binding sites in protein sequences." *Computational and Structural Biotechnology Journal* 23 (2024): 1929-1937.
- **Teukam, Yves Gaetan Nana**, et al. "Integrating Genetic Algorithms and Language Models for Enhanced Enzyme Design." (2024), (Preprint).
- Manica, Matteo, Jannis Born, Joris Cadow, Dimitrios Christofidellis, Ashish Dave, Dean Clarke, **Yves Gaetan Nana Teukam** et al. "Accelerating material design with the generative toolkit for scientific discovery." *npj Computational Materials* 9, no. 1 (2023): 69.
- Probst, Daniel, Matteo Manica, **Yves Gaetan Nana Teukam**, Alessandro Castrogiovanni, Federico Paratore, and Teodoro Laino. "Biocatalysed synthesis planning using data-driven learning." *Nature communications* 13, no. 1 (2022): 964.