STUDENT INTERNSHIP

Topic: **AI-Driven Drug Discovery in Oncology: Harnessing Large Language Models for Molecular Insights**

Duration: 3 to 6 months start on February
Location: Oncodesign HQ – Dijon
Benefits: Monthly indemnity 1000€ + meal Ticket

Our Company

OPM is a biopharmaceutical company specialized in precision medicine. OPM's mission is to bring innovative therapeutic and diagnostic solutions to treat therapeutic resistance and metastasis evolution. The patient is at the center of our reflection, of our unique innovative model, and our investments. For OPM "our collective success is paramount", there can be no value creation without exchange, without dialogue. The value creation resulting for us from reciprocity, i.e. balanced and fair exchanges at all levels, whether between internal collaborators, or with our partners, therapists, patients, experts and investors.

Context

The language of molecular biology, once deciphered only through rigorous experimentation, is on the brink of a revolution. As we enter the realm of large language models (LLMs) and deep learning, the promise of highly accurate in silico models for comprehending the intricate world of molecular biology, from DNA to gene expression to proteins, is within reach.

The fusion of cutting-edge artificial intelligence and molecular biology holds the potential to reshape medicine and pharmaceutical discovery. For our biopharmaceutical company, Oncodesign Precision Medicine (OPM), dedicated to identifying new therapeutic targets and developing drugs for the treatment of advanced, resistant, and metastatic cancers, harnessing the power of LLMs is not just a possibility; it's a necessity. LLMs emerge as a crucial tool in our pursuit of breakthroughs in oncology, enabling us to explore the complex molecular landscape of cancer, understand its underlying mechanisms, and discover novel avenues for treatment. Some examples of recent LLMs that our team is investigating:

- scGPT (1), a foundation model designed for single-cell transcriptomics, chromatin accessibility, and protein abundance.
- ESM1b (2) and Alpha Missense (3), LLMs that predict the pathogenic effect of missense variants in the human genome.
- Geneformer (4), a foundation model pretrained on ~30 million single cell transcriptomes from a broad range of human tissues to enable context-aware predictions in settings with limited data in network biology.

Missions & activities of the internship

Under co-supervision of two Senior Data Scientists holding PhD titles and interdisciplinary background in artificial intelligence, immunology, mathematics, genetics, genomics and bioinformatics, your duties is to deliver:

- **Evaluation of State-of-the-Art LLMs in Molecular Biology:** Conduct an extensive review of the latest developments in LLMs, assessing their performance and identifying the most promising LLMs that can be leveraged for new therapeutic targets discovery.
- **Model Implementation:** Develop the technical proficiency to apply LLMs. This involves understanding the practical aspects of model implementation and finetuning.
- **Data preprocessing and analysis:** extract meaningful insights from public and/or proprietary datasets.
- Git code repositories with well-documented scripts in Python and notebooks with any conducted analysis.
- A report summarizing your findings and contributions.

Student expected background/Knowledge.

M2 student or last year in Engineer School with educational background in a relevant field (Computational biology, bioinformatics, artificial intelligence or related).
Essential skills include programming, machine learning, understanding of key concepts of molecular biology.
Familiarity with NLP and LLMs is a significant advantage.
Fluent in French & English languages

How apply?

Contact: Thierry Billoué – Chief Human Resources Officer – Oncodesign Precision Medicine
Send your application (resume & cover letter) under ref "LLM4Molins" to tbilloue@oncodesign.com
Candidates will be ranked based on their CV and cover letter.
Best candidates will be invited to an interview on site or through Teams.

References

1. Haotian Cui, Chloe Wang, Hassaan Maan, Bo Wang, scGPT: Towards Building a Foundation Model for Single-Cell Multi-omics Using Generative AI, bioRxiv 2023.04.30.538439; doi: https://doi.org/10.1101/2023.04.30.538439
2. Brandes, N., Goldman, G., Wang, C.H. *et al.* Genome-wide prediction of disease variant effects with a deep protein language model. *Nat Genet* **55**, 1512–1522 (2023). https://doi.org/10.1038/s41588-023-01465-0
3. Jun Cheng et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science 381, eadg7492 (2023). DOI:10.1126/science.adg7492
4. Theodoris, C.V., Xiao, L., Chopra, A. *et al.* Transfer learning enables predictions in network biology. *Nature* **618**, 616–624 (2023). https://doi.org/10.1038/s41586-023-06139-9

STUDENT INTERNSHIP

Topic: Computational identification of off-target proteins for drug candidates

Duration: 3 to 6 months start on February
Location: Oncodesign HQ – Dijon
Benefits: Monthly indemnity 1000€ + meal Ticket

Our Company

OPM is a biopharmaceutical company specialized in precision medicine. OPM's mission is to bring innovative therapeutic and diagnostic solutions to treat therapeutic resistance and metastasis evolution. The patient is at the center of our reflection, of our unique innovative model, and our investments. For OPM "our collective success is paramount", there can be no value creation without exchange, without dialogue. The value creation resulting for us from reciprocity, i.e. balanced and fair exchanges at all levels, whether between internal collaborators, or with our partners, therapists, patients, experts and investors.

Context

Obtaining the structure of a protein is a challenge: experimental methods such as x-rays are expensive, laborious and it is not always possible to crystallize the protein. On the other hand, considering that proteins can be composed of hundreds of amino acids, generating an algorithm capable of predicting the structure of a protein is a rather a complex task. Proteins play a fundamental role in living beings and are the main target for therapeutic molecules. Thus, the folding problem (how to obtain the structure of a protein from its sequence) has occupied the minds of researchers for most of the 20th century.

AlphaFold2 won the main competition for protein structure prediction (CASP14) in 2021. AlphaFold2's predictions were considered to be almost at the level of those determined experimentally. DeepMind has recently made both the code and the model available on GitHub as open source, allowing the community to be able to use the model both for the prediction of structures from an amino acid sequence and to incorporate it into other models for other applications.

The possibility of accurately predicting the structure of a protein opens up different applications (from the possibility of designing new enzymes for the food industry, to nanotechnology for medicine). The pharmaceutical industry and especially the drug discovery field is a domain where the greatest effects are expected. Indeed, most approved drugs are small molecules and biologics that interact with a protein. Typically for small molecules, having identified a target of interest and the structure of a protein, molecular modeling could be used to design virtual compounds that can bind to the protein's active site and modulate its function.

To date, only 10 % of drug candidates make it through the clinical trial stages and reach the market. The main reason for the failure of clinical trials is safety. While in general a drug candidate is selected to have high affinity for its target, it could potentially bind to other targets (off-targets) resulting in secondary effects. Indeed, drugs often have off-targets that lead to unwanted effects that can be serious. This is why it is important to identify potential off targets before a molecule enters the clinical trial phase, which can cost up to a billion dollars.

One of the most important class of target is the kinases protein family. Kinase is an enzyme that catalyzes the transfer of a phosphate group to a specific substrate. This mechanism has different functions and it involved in fundamental process, which can be often dysregulated in cancer. There are known 500 kinases, while they can be expressed in different sites, they have a similar structure with an ATP binding domain. OPM has developed a class of molecules that are highly specific for kinases, flat molecules called macrocycles and has developed Nanocyclix® a specific & proprietary platform. In most cases, off-targets are proteins that display similarity in the region of the ATP-binding domain to the active site of the protein of interest. The aim of this training is to develop an algorithm that can identify potential off-target sites and develop a similarity scoring function.

The objective of the internship is to use an algorithm based on AlphaFold2 to be able to identify potential off-target proteins.

Missions & activities of the internship

Under co-supervision by a Senior Data Scientist and a Medicinal Chemist holding PhD titles and interdisciplinary background in artificial intelligence, medicinal chemistry, and bioinformatics, your duties will be the following one.

- Build an algorithm based on AlphaFold2 source code to generate embedding representation of protein kinases active sites.
- Testing potential other algorithms as RosettaFold
- Identity potential candidates for off-target in a case study
- Modeling of the active site and the interaction of small molecules

Keywords: Python, structural biology, deep learning

Student expected background/Knowledge.

M2 or last year of engineer school with specialty/knowledge in Computer Science / Bioinformatics / Structural Biology/Statistics Biology with knowledge in programming (R / Python).
Docking knowledge, working with computer clusters is a plus.
Fluent in French & English languages

How apply?

Contact: Thierry Billoué – Chief Human Resources Officer – Oncodesign Precision Medicine
Send your application (resume & cover letter) under ref “ComputID” to tbilloue@oncodesign.com
Candidates will be ranked based on their CV and cover letter.
Best candidates will be invited to an interview on site or through Teams.

References

1. Jumper et al., 2021 **Highly accurate protein structure prediction with AlphaFold**. Nature
2. Baek et al., 2021. **Accurate prediction of protein structures and interactions using a three-track neural network**. Science

STUDENT INTERNSHIP

Topic: **Large Language Models for Drug Discovery: Your AI Biologist Assistant**
Duration: 3 to 6 months start on February
Location: Oncodesign HQ – Dijon
Benefits: Monthly indemnity 1000€ + meal Ticket

Our Company

OPM is a biopharmaceutical company specialized in precision medicine. OPM's mission is to bring innovative therapeutic and diagnostic solutions to treat therapeutic resistance and metastasis evolution. The patient is at the center of our reflection, of our unique innovative model, and our investments. For OPM "our collective success is paramount", there can be no value creation without exchange, without dialogue. The value creation resulting for us from reciprocity, i.e. balanced and fair exchanges at all levels, whether between internal collaborators, or with our partners, therapists, patients, experts and investors.

Context

Over the past year, Large Language Models (LLMs) have taken the world by storm [1]. It was clear early on that the use of these large models could revolutionize different disciplines and fields of application (education, finance, law, and so on). The scientific community was immediately excited about the potential breakthrough that these models could bring to the medical field (disease understanding, target discovery, drug design, and so on). Applying these models to the medical field, though, presents additional challenges (specific language, particular complexity, specific knowledge) [2-3].

The advent of open-source models (LLaMA, Falcon) [4-5] has allowed the scientific comity to investigate how to adapt these models to the biomedical field. Notable examples are PMC-LLaMA and Chat-Doctor [6-7] where pre-trained models were further refined by training them with scientific abstracts or medical documents. To truly realize the potential of this scientific revolution, it is necessary to have quality datasets with which to train these models for specific medical applications.

Oncodesign Precision Medicine (OPM) is a company focused on the identification of therapeutical targets in oncology and the development of medical drugs against resistant cancers. Thus, At OPM we have collected a large corpus containing medical information (articles, patents, clinical trials, etc...) as well as possessing a vast amount of oncology patient data collected over the years. This year we trained a LLM model specifically on pancreatic cancer, showing how such a model can reach the state of the art on medical questions.

Building on this success, OPM has an interest in continuing to investigate the use of LLM for the identification of new tumor therapeutic targets. The final goal is to train a model capable of providing information about a specific target, potential therapeutic opportunities, and disease understanding. The model is meant to assist in future target selection, investigation, and analysis. This model will be tested in real-case scenarios as a biologist assistant and for completing a target dossier.

The objectives of the internship are to deliver:

- a state-of-the-art analysis for LLM training, quantization, and deployment technologies.
- an extensive review of methods that can be implemented.
- Model fine-tuning and deployment. The obtained model should be able to respond to drug discovery-related questions, competition landscape, disease-related information, and so on.
- Pipeline to integrate the model with different databases. Test tasks to show the LLM's capabilities in finding information from different sources.

Missions & activities of the internship

Under supervision of a Senior Data Scientist holding PhD title and an interdisciplinary background in artificial intelligence, immunology, mathematics, genetics, genomics, and bioinformatics, your duties will be the following one.

- Evaluate the state-of-the-art per language model and fine-tuning of LLM. Starting from the obtained baseline, we want to refine the LLM by implementing the latest technologies. We will explore strategies for tailoring a model focused on drug discovery, target, and disease knowledge.
- Deployment of the model. Establish a pipeline for monitoring and deployment of the model (running on a server or alternatives)
- Biology database integration. Integrate the model with a wide range of biological databases (mutations, pathways, patient data, and so on), so that it can talk to them and extract information from them, and therefore perform complex tasks.
- Git code repositories with well-documented scripts in Python and notebooks with any conducted analysis.
- A report summarizing your findings and contributions.

Student expected background/Knowledge.

M2 student with educational background in a relevant field (Computational biology, bioinformatics, artificial intelligence or related).
Essential skills include programming, machine learning, understanding of key concepts of molecular biology.
Fluent in French & English languages

How apply?

Contact: Thierry Billoué – Chief Human Resources Officer – Oncodesign Precision Medicine
Send your application (resume & cover letter) under ref "LLM4DD" to tbilloue@oncodesign.com
Candidates will be ranked based on their CV and cover letter.
Best candidates will be invited to an interview on site or through Teams.

References

1. Zhao et al., 2023, A Survey of Large Language Models, https://arxiv.org/abs/2303.18223
2. Singhal et al., 2022, Large Language Models Encode Clinical Knowledge, https://arxiv.org/abs/2212.13138
3. Singhal et al., 2023, Towards Expert-Level Medical Question Answering with Large Language Models, https://arxiv.org/abs/2305.09617
4. Touvron et al., 2023, LLaMA: Open and Efficient Foundation Language Models, https://arxiv.org/abs/2302.13971
5. Touvron et al., 2023, Llama 2: Open Foundation and Fine-Tuned Chat Models, https://arxiv.org/abs/2307.09288
6. Wu et al., 2023, PMC-LLaMA: Towards Building Open-source Language Models for Medicine. https://arxiv.org/abs/2304.14454
7. Li et al., 2023, ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge, https://arxiv.org/abs/2303.14070