# Machine Learning Data Scientist

**Location: EMBL-EBI, Hinxton near Cambridge, UK**
**Staff Category: Staff Member**
**Contract Duration: 3 years**
**Grading: Grade 6 (monthly salary starting from £3,456.96 after tax) plus other paid benefits**
**Closing Date:**
**Reference Number: EBI02201**

Open Targets (OT) is a unique public-private partnership working to deliver experimental data and informatics platforms that enable researchers to make more informed decisions about target selection for drug discovery. OT is a shared initiative between the European Bioinformatics Institute (EMBL-EBI), a global leader in the management, integration and analysis of public domain life science data; world-leading pharmaceutical companies GSK, Sanofi, Bristol Myers Squibb and Pfizer; and the Wellcome Sanger Institute.

Scientific inquiry depends on knowledge, but there are many ways in which this knowledge is represented, one of which is the unstructured knowledge source of the scientific literature. Driven by the recent progress in Large Language Models (LLMs), knowledge extraction via natural language processing (NLP) will be utilised and this project will automate knowledge extraction, representation and usages via NLP-Knowledge Graph-AI to deliver more human-interpretable outputs to OT users.

We are looking for an enthusiastic and talented machine learning data scientist to join the AI knowledge management project, which is for a period of 3 years. This position will be situated in the Literature Services team, which manages Europe PMC. We are interested in applications of machine learning to mine the research literature for additional types of entities relevant to drug discovery not already available to OT (such as variants, biomarkers, tissues/cell types, adverse events, and assay conditions). This position provides a real opportunity to make a significant impact on a critical problem in drug discovery for the many users of the OT Platform and an opportunity to contribute to the open source models and code associated with biological entities.

You will be embedded into a knowledge management project team that includes a biological expert from the CHEMBL team and two data scientists/engineers from the OT team. You will need to be able to demonstrate the ability to work well with colleagues and to collaborate with external partners. You must have excellent communication and interpersonal skills and enjoy working in a stimulating, international environment.

## Your role

Key responsibilities include:

- To find and use suitable pre-trained NLP models from the public domain (e.g.HuggingFace)
- To retrain them on the open scientific literature available in Europe PMC to ensure they are optimised for the project's need.
- To modernise and extend the current entity recognition workflows to cover an array of additional types of entities relevant to drug discovery
- Apply quantization methods to optimise model performance at production and efficiency, especially when dealing with large datasets.
- Utilise and refine vector embedding techniques to improve the representation of scientific textual data, enhancing the accuracy of entity extraction and relationship mapping.
- Development of new machine learning, deep learning or NLP protocols to enhance curation workflows
- Develop interfaces for deep searching of Europe PMC, customised for curator groups
- Collaboration with EMBL-EBI curator groups to collect specifications, test prototypes and deliver tools that meet their needs
- Work with the EBI ChEMBL team to develop and utilise statistically robust methods for data analysis and benchmarking
- Collaborate with the OT partners to assess, prioritise, validate and refine the developed methods
- Work closely with the OT core team for the seamless integration of data and workflows into the OT Platform
- Actively disseminate the outcomes of the project to the scientific community and stakeholders through well-crafted presentations and publications

## You have

- Advanced degree (MSc, PhD) in data science or related discipline
- Proficiency in at least one programming/scripting language (e.g. Python)
- Proficiency in using deep learning frameworks like PyTorch, crucial for advanced NLP and machine learning tasks

- Experience with advanced big data preprocessing, cleaning, and transformation techniques specific to textual data including ontologies
- Good understanding of statistical methods and their application to data analysis and use of data visualisation tools and libraries (such as Matplotlib, Seaborn) to effectively communicate data insights.
- Knowledge of version control systems (e.g., GitHub)
- Excellent attention to detail
- Strong communication skills, both presentations and verbal
- Experience working in a team-oriented environment
- Able to work independently, to manage your time and work to deadlines

### You might also have

- Experience of biological data curation and knowledge of bioinformatics databases
- Experience working in a drug discovery and development environment
- Knowledge and practical experience with bioinformatics methods including systems biology and genetics analysis

### Why join us

**Do something meaningful**
At EMBL-EBI you can apply your talent and passion to accelerate science and tackle some of humankind's greatest challenges. EMBL-EBI, part of the European Molecular Biology Laboratory, is a worldwide leader in the storage, analysis and dissemination of large biological datasets. We provide the global research community with access to publicly available databases and tools which are crucial for the advancement of healthcare, food security, and biodiversity.

**Join a culture of innovation**
We are located on the Wellcome Genome Campus, alongside other prominent research and biotech organisations, and surrounded by beautiful Cambridgeshire countryside. This is a highly collaborative and inclusive community where our employees enjoy a relaxed atmosphere. We are committed to ensuring our employees feel valued, supported and empowered to reach their professional potential.

**Enjoy lots of benefits:**

- **Financial incentives:** Monthly family, child and non-resident allowances, annual salary review, pension scheme, death benefit, long-term care, accident-at-work and unemployment insurance
- **Flexible working arrangements**
- **Private medical insurance** for you and your immediate family (including all prescriptions and generous dental and optical cover)
- **Generous time off:** 30 days annual leave per year, in addition to eight bank holidays
- **Relocation package,** including installation grant (if required)
- **Campus life:** Free shuttle bus to and from work, on-site library, subsidised on-site gym and cafeteria, casual dress code, extensive sports and social club activities (on campus and remotely)
- **Family benefits**: On-site nursery, ten days of child sick leave, generous parental leave, holiday clubs on campus and monthly family and child allowances
- **Benefits for non-UK residents:** Visa exemption, education grant for private schooling, financial support to travel back to your home country every second year and a monthly non-resident allowance.

For more details, please see our employee benefits page.

### What else you need to know

- **Contract duration:** This position is a 3-year fixed-term contract (grant-limited).
- **International applicants:** We recruit internationally, and successful candidates are offered visa exemptions. Read more on our page for international applicants.
- **Diversity and inclusion**: At EMBL-EBI, we firmly believe that inclusive and diverse teams benefit from higher innovation and creative thought levels. We encourage applications from women, LGBTQ+ and individuals from all nationalities.
- **EMBL is a signatory of DORA.** Find out how we implement best practices in research assessment in our recruitment processes here.
- **Job location:** This role is based in Hinxton, near Cambridge, UK. If you are based overseas, you will be required to relocate and receive a generous relocation package to support you.
- **How to apply:** Please submit a cover letter and a CV through our online system.