Annotating Molecular Dynamics simulation data using Artificial Intelligence

Touami, E.[1,2], Di Gennaro G.[2], Chavent M.[3], Poulain, P. [2]
[1]touami@ipbc.fr
[2] *Université Paris Cité, CNRS, Laboratoire de Biochimie Théorique, 13 rue Pierre et Marie Curie, F-75005, Paris, France*
[3] *Laboratoire de Microbiologie et Génétique Moléculaires (LMGM), Centre de Biologie Intégrative (CBI), Université de Toulouse, CNRS, France*

In the context of Open Science, large volumes of research data are now publicly available across scientific domains. However, their reuse remains difficult, due to missing or inconsistent metadata. Within the European project LUMEN (https://lumenproject.eu/), we propose an Artificial Intelligence (AI)-based methodology to automatically annotate molecular dynamics simulation descriptions using large language models (LLMs). Our approach relies on Named Entity Recognition (NER)[1] to extract key metadata entities from natural-language dataset descriptions, such as those found in dataset repositories, or the Materials and Methods sections of scientific publications. The extracted entities include molecule names (e.g., “alanine dipeptide”), simulation times (e.g., “87 microsecond”), force fields (e.g., “AMBER99SB-ILDN”), temperatures (e.g., “300 K”), and software information (e.g., “GROMACS 2021.4”) (Figure 1). The originality of our pipeline lies in enforcing a structured output that is automatically validated to reduce hallucinations. We evaluated several proprietary and open-weight LLMs (such as GPT-5.4, Claude 4.6, Gemma-4-31b, Glm-5.1, Qwen 3.5...) based on performance, robustness, and cost. We benchmarked our approach against GLiNER2[2], a state-of-the-art model optimized for entity recognition.

These annotations will be integrated into MDverse[3], a FAIR data catalogue for molecular simulations datasets, awarded in 2024 by the French Ministry of Higher Education and Research for Open Science in research data, and which currently indexes 27,000 datasets and more than 2 millions files. While demonstrated on molecular simulations data, the methodology is domain-agnostic and can be easily transferred to other scientific disciplines.

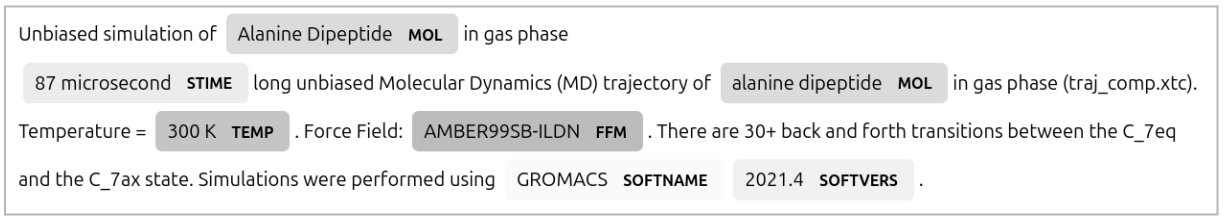

Figure 1. Example of a molecular dynamics simulation dataset description from Zenodo (https://zenodo.org/records/7323535), with extracted entities highlighted and labeled according to their types: MOL (molecule name), STIME (simulation time), TEMP (temperature), FFM (force field), SOFTNAME (software name), SOFTVER (software version).

Bibliography:

[1] Pakhale K. arXiv preprint, 2023, *arXiv:2309.14084*.
[2] Zaratiana, Urchade, et al. arXiv preprint, 2025, *arXiv:2507.18546*.
[3] Tiemann J.; Szczuka M. et al. eLife, 2024, https://doi.org/10.7554/eLife.90061.3