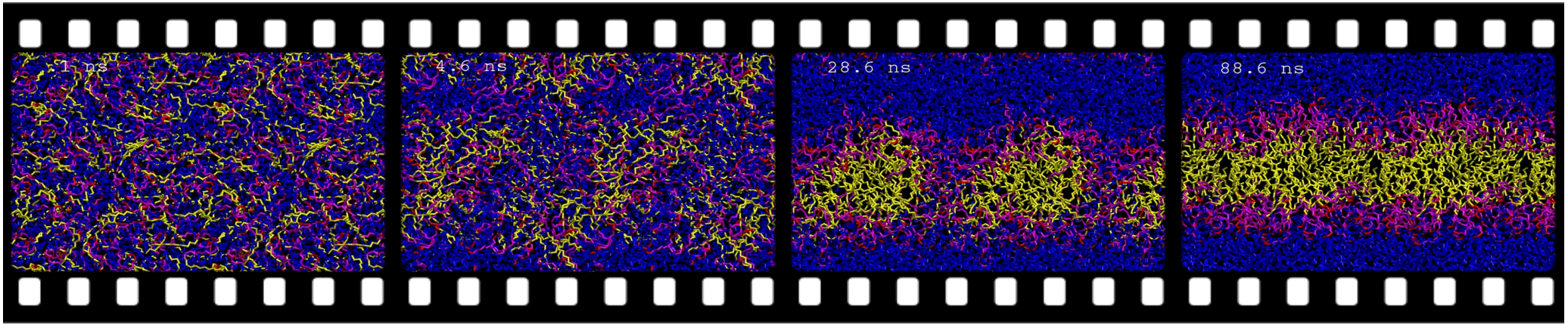
Science ouverte, données et IA : automatiser l'annotation scientifique

Pierre Poulain

Laboratoire de Biochimie Théorique
UMR 8266 CNRS & Université Paris Cité

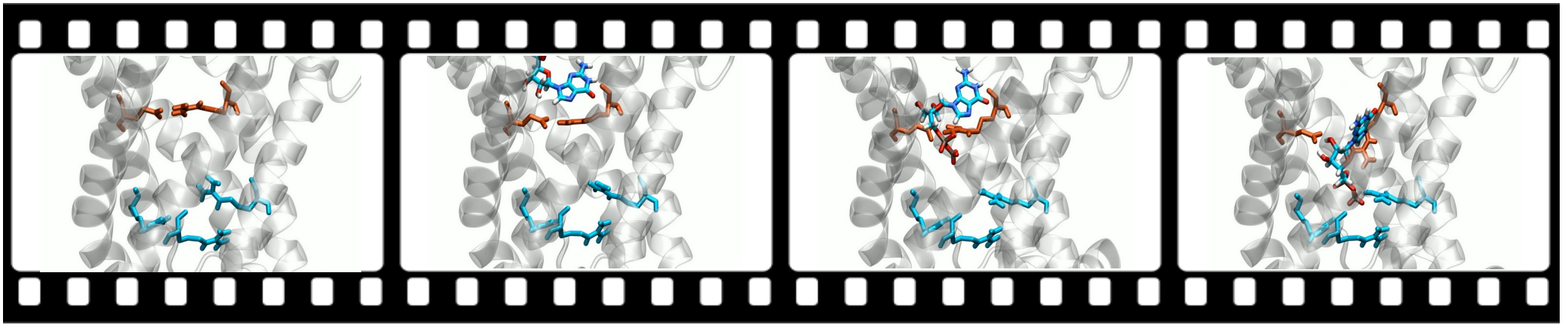
12/06/2026

What is molecular dynamics (MD)?



water + detergent

Senac et al, Langmuir, 2017. Movie by Patrick Fuchs.



Gagelin et al, Nature communications, 2023.

Supercomputers → high cost

Source : Photothèque CNRS/Cyril Frésillon (droits réservés)

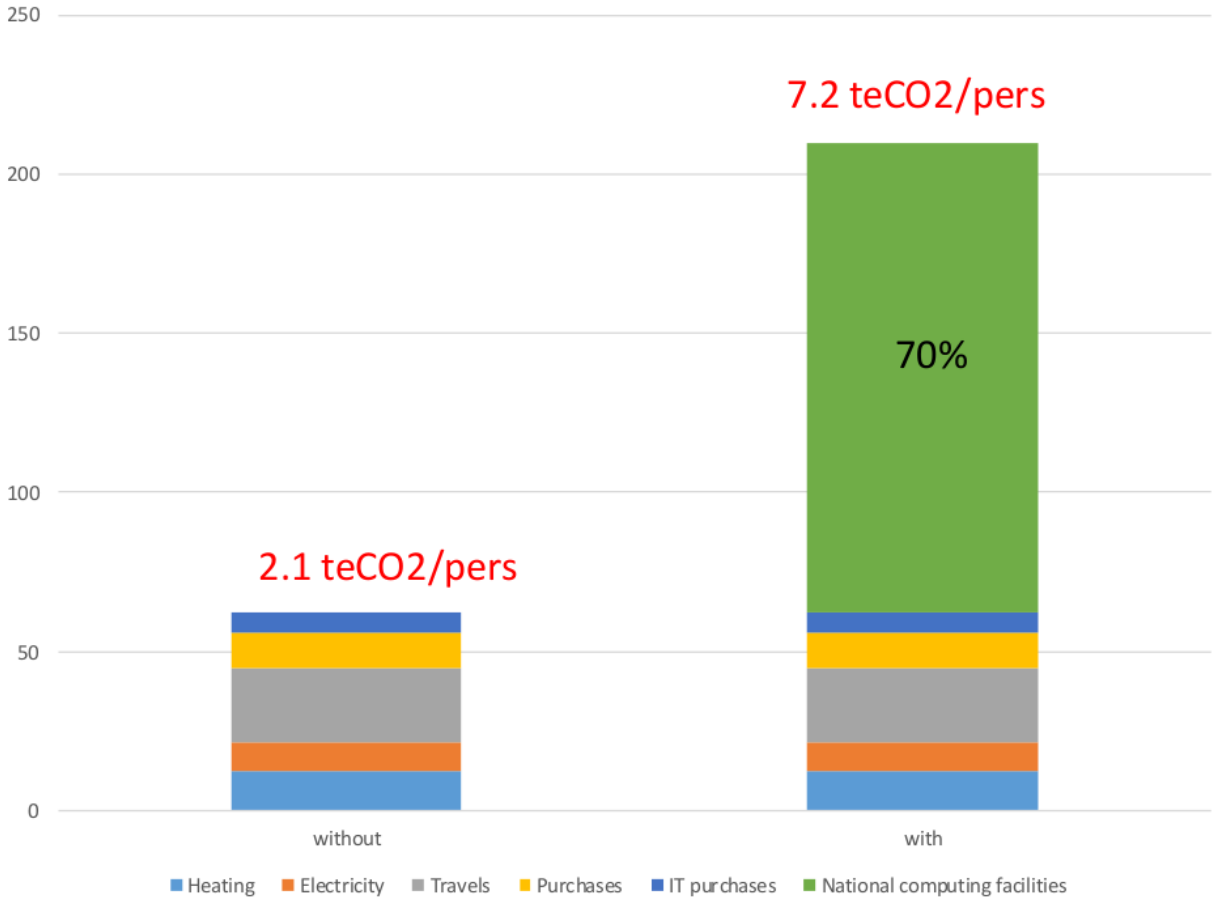
GENCI
HPC at the service of knowledge

In 2025:

- CT 7 (simulation in biology)
- 132 Mh CPU
- 8 Mh GPU
- **Total cost ~8 M€**

Patrick Fuchs, 2026.

Supercomputers → high environmental cost



Élise Duboué-Dijon & Antoine Taly, LBT, 2024.

Welcome to the Dark Matter of MD

Data that is **technically accessible**,

but neither **indexed**, **curated**,

or easily **searchable**.

eLife

RESEARCH ARTICLE

MDverse, shedding light on the dark matter of molecular dynamics simulations

Johanna KS Tiemann[1]*†, Magdalena Szczuka[2], Lisa Bouarroudj[3], Mohamed Oussaren[3], Steven Garcia[4], Rebecca J Howard[5], Lucie Delemotte[6], Erik Lindahl[5,6], Marc Baaden[7], Kresten Lindorff-Larsen[1], Matthieu Chavent[2]*, Pierre Poulain[3]*

[1]Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, Copenhagen, Denmark; [2]Institut de Pharmacologie et Biologie Structurale, CNRS, Université de Toulouse, Toulouse, France; [3]Université Paris Cité, CNRS, Institut Jacques Monod, Paris, France; [4]Independent researcher, Amsterdam, Netherlands; [5]Department of Biochemistry and Biophysics, Science for Life Laboratory, Stockholm University, Stockholm, Sweden; [6]Department of applied physics, Science for Life Laboratory, KTH Royal Institute of Technology, Stockholm, Sweden; [7]Laboratoire de Biochimie Théorique, CNRS, Université Paris Cité, Paris, France

***For correspondence:** johanna.tiemann@gmail.com (JKST); matthieu.chavent@ipbs.fr (MC); pierre.poulain@u-paris.fr (PP)

Present address: †NovozymesA/S, Lyngby, Denmark

Competing interest: See page 16

Funding: See page 16

Preprint posted 02 May 2023
Sent for Review 28 June 2023
Reviewed preprint posted 20 September 2023
Reviewed preprint revised 01 July 2024
Version of Record published 30 August 2024

Reviewing Editor: Shozeb Haider, University College London, United Kingdom

Abstract The rise of open science and the absence of a global dedicated data repository for molecular dynamics (MD) simulations has led to the accumulation of MD files in generalist data repositories, constituting the *dark matter of MD* — data that is technically accessible, but neither indexed, curated, or easily searchable. Leveraging an original search strategy, we found and indexed about 250,000 files and 2000 datasets from Zenodo, Figshare and Open Science Framework. With a focus on files produced by the Gromacs MD software, we illustrate the potential offered by the mining of publicly available MD data. We identified systems with specific molecular composition and were able to characterize essential parameters of MD simulation such as temperature and simulation length, and could identify model resolution, such as all-atom and coarse-grain. Based on this analysis, we inferred metadata to propose a search engine prototype to explore the MD data. To continue in this direction, we call on the community to pursue the effort of sharing MD data, and to report and standardize metadata to reuse this valuable matter.

eLife assessment

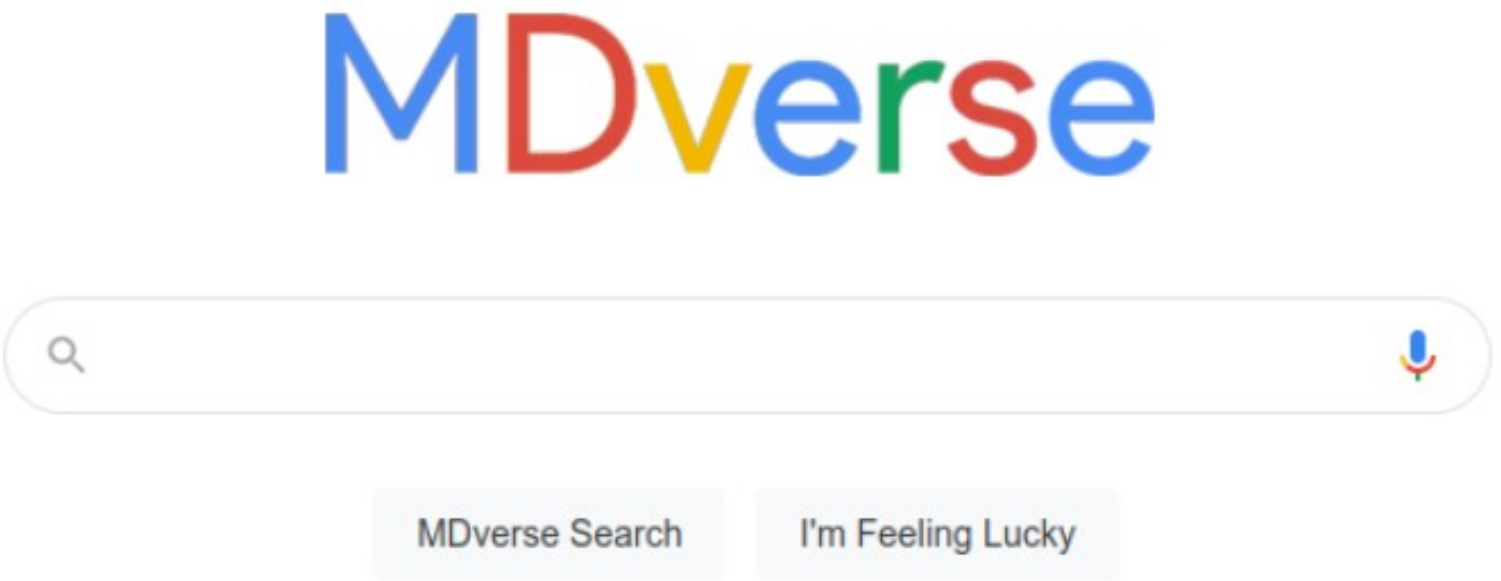
The study presents a **valuable** tool for searching molecular dynamics simulation data, making such datasets accessible for open science. The authors provide **convincing** evidence that it is possible to identify noteworthy molecular dynamics simulation datasets and that their analysis can produce information of value to the community.

Introduction

The volume of data available in biology has increased tremendously (***Marx, 2013***; ***Stephens et al., 2015***), through the emergence of high-throughput experimental technologies, often referred to as -omics, and the development of efficient computational techniques, associated with high-performance computing resources. The Open Access (OA) movement to make research results free and available to anyone (including e.g. the Budapest Open Access Initiative and the Berlin declaration on Open Access to Knowledge) has led to an explosive growth of research data made available by scientists (***Wilson***

DOI 10.7554/eLife.90061.3

A data catalogue for MD simulation data (and more)



WHY? Enable, facilitate & foster the reuse of MD data

MD data availability

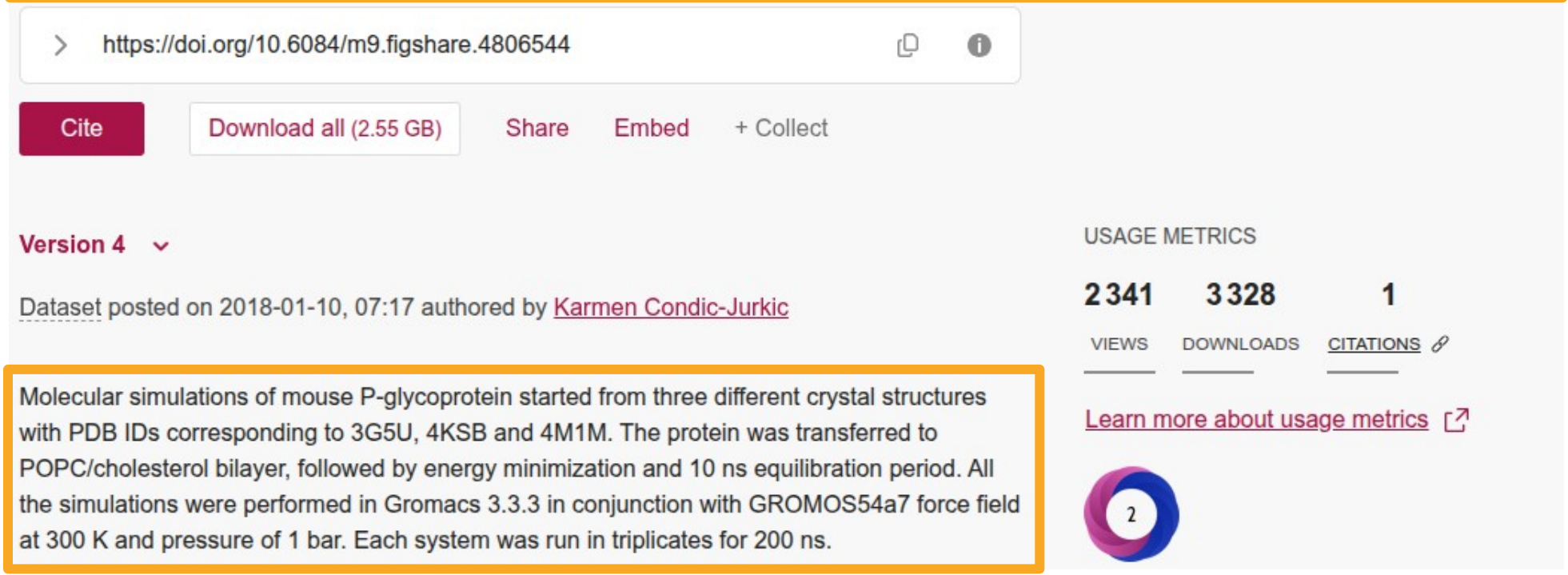
Data repository	First dataset	# datasets	# files	Size (GB)
Zenodo	2014	2 369	504 891	26 686
Figshare	2012	1 373	222 703	1 177
NOMAD	2021	16 119	555 990	2 189
MDposit CINECA (MDDB) MDposit INRIA (MDDB) MDposit MMB (MDDB)	2025	310 719 4 050	251 581 186 293 411 854	52 1 022 15 177
GPCRmd	2017	830	2 414	624
ATLAS	2023	1 938	69 768	10 944
Total		**27 708**	**2 205 494**	**57 809**

Reuse MD data → **Extract & normalize metadata**

S. Sheikh, E. Touami

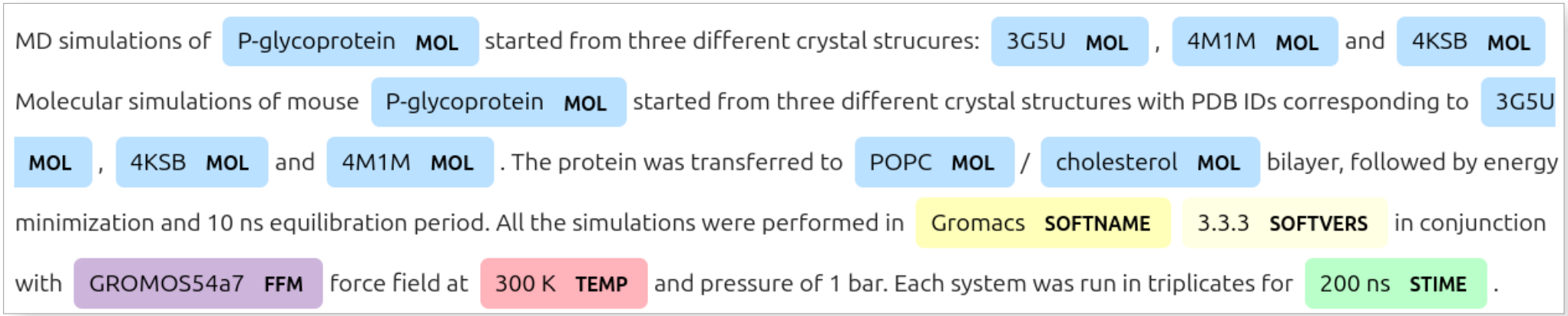
Where to find metadata?

MD simulations of P-glycoprotein started from three different crystal strucures: 3G5U, 4M1M and 4KSB



Source: Figshare

Named Entity Recognition (NER)



Strategies based on language models:

- Encoder-like models (BERT): ~ 100 millions parameters (SLM)
  → *fine tuning* (**many manually annotated examples**)
- Decoder-like models (GPT): ~ 10-1000 billions parameters (LLM)
  → ***harness*** (instructions + constraints) / **access to models?**

With A. Ferré (INRAE) / E. Touami

Manual approach

OpenAI, GPT-5.5, 10/06/2026

In the following text, extract all molecule names, software versions, software names, forcefield or model names, simulation temperatures, and simulation times (duration):

MD simulations of P-glycoprotein started from three different crystal strucures: 3G5U, 4M1M and 4KSB
Molecular simulations of mouse P-glycoprotein started from three different crystal structures with PDB IDs corresponding to 3G5U, 4KSB and 4M1M. The protein was transferred to POPC/cholesterol bilayer, followed by energy minimization and 10 ns equilibration period. All the simulations were performed in Gromacs 3.3.3 in conjunction with GROMOS54a7 force field at 300 K and pressure of 1 bar. Each system was run in triplicates for 200 ns.

Afficher moins

Manual approach

OpenAI, GPT-5.5, 10/06/2026

In the following text, extract all molecule names, software versions, software names, forcefield or model names, simulation temperatures, and simulation times (duration):

MD simulations of P-glycoprotein started from three different crystal strucures: 3G5U, 4M1M and 4KSB
Molecular simulations of mouse P-glycoprotein started from three different crystal structures with PDB IDs corresponding to 3G5U, 4KSB and 4M1M. The protein was transferred to POPC/cholesterol bilayer, followed by energy minimization and 10 ns equilibration period. All the simulations were performed in Gromacs 3.3.3 in conjunction with GROMOS54a7 force field at 300 K and pressure of 1 bar. Each system was run in triplicates for 200 ns.

Afficher moins

Extracted entities from the text:

Category	Extracted Value
Molecule names	P-glycoprotein (mouse P-glycoprotein), POPC, cholesterol
Software names	Gromacs
Software versions	Gromacs 3.3.3
Forcefield / model names	GROMOS54a7 force field
Simulation temperatures	300 K
Simulation times (duration)	10 ns equilibration period; 200 ns production simulation (each system, run in triplicate)

Additional identifiers present (not requested but mentioned):

- Crystal structure / PDB IDs: 3G5U, 4KSB, 4M1M
- Pressure: 1 bar

Manual approach

OpenAI, GPT-5.5, 10/06/2026

- Automation?
- Model selection?
- Detailed instructions?
- Output format?
- Quality control?

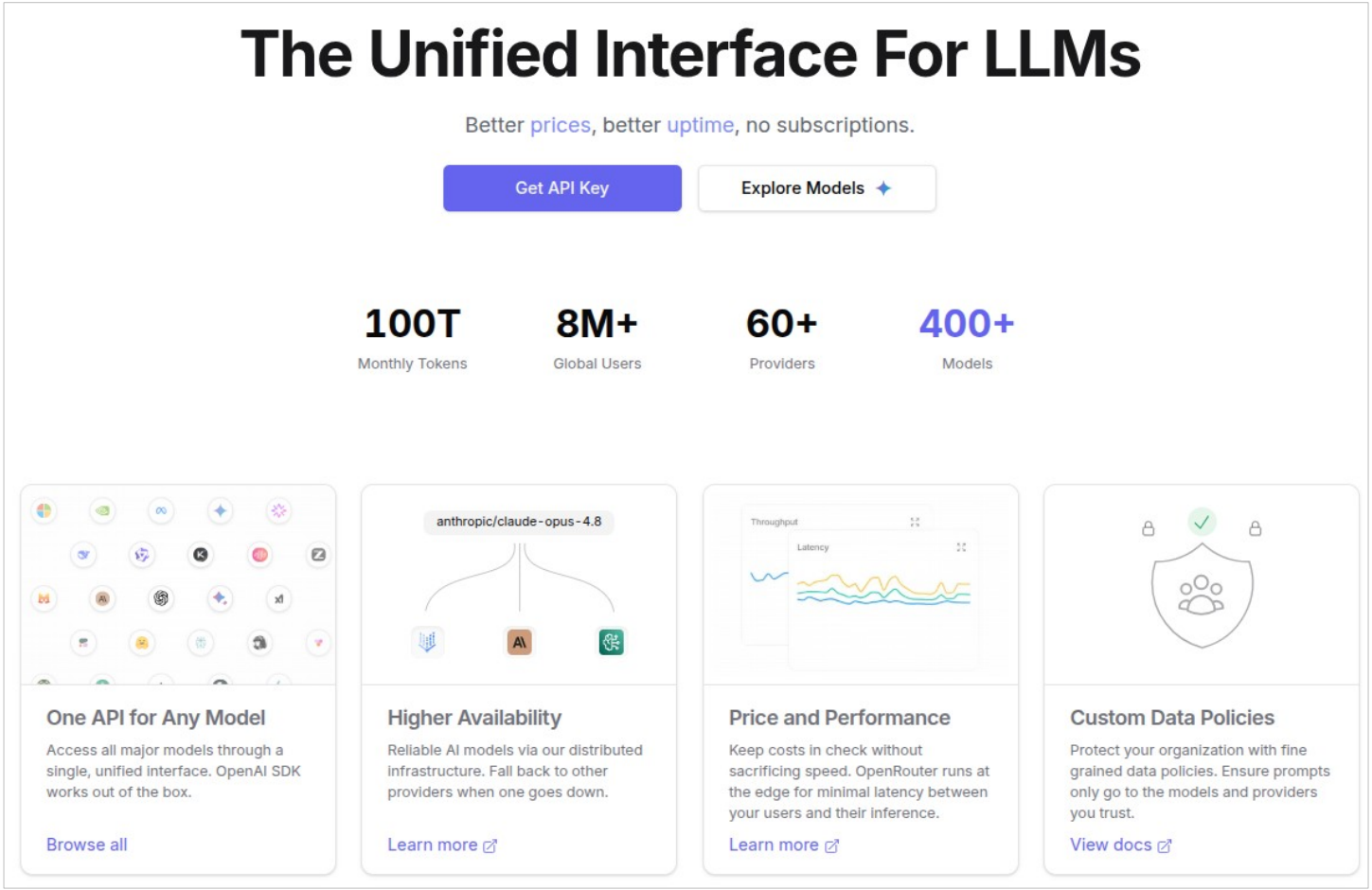
Extracted entities from the text:

Category	Extracted Value
Molecule names	P-glycoprotein (mouse P-glycoprotein), POPC, cholesterol
Software names	Gromacs
Software versions	Gromacs 3.3.3
Forcefield / model names	GROMOS54a7 force field
Simulation temperatures	300 K
Simulation times (duration)	10 ns equilibration period; 200 ns production simulation (each system, run in triplicate)

Additional identifiers present (not requested but mentioned):

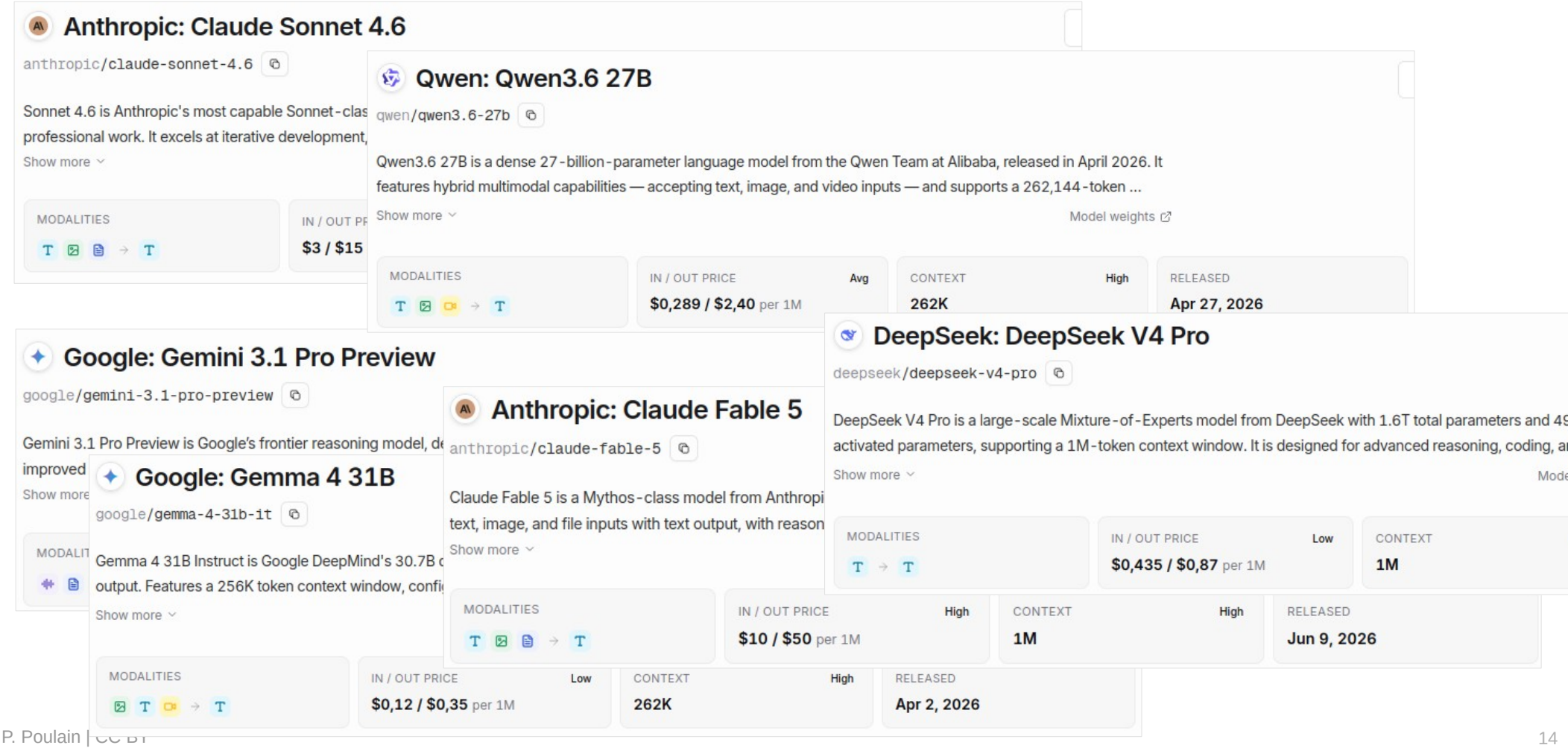
- Crystal structure / PDB IDs: 3G5U, 4KSB, 4M1M
- Pressure: 1 bar

Automation & model selection: OpenRouter



https://openrouter.ai/

Automation & model selection: OpenRouter



API

```
import requests
import json

# First API call with reasoning
response = requests.post(
  url="https://openrouter.ai/api/v1/chat/completions",
  headers={
    "Authorization": "Bearer <OPENROUTER_API_KEY>",
    "Content-Type": "application/json",
  },
  data=json.dumps({
    "model": "qwen/qwen3.7-plus",
    "messages": [
        {
          "role": "user",
          "content": "How many r's are in the word 'strawberry'?"
        }
      ],
    "reasoning": {"enabled": True}
  })
)

# Extract the assistant message with reasoning_details
response = response.json()
response = response['choices'][0]['message']
```

Prompt engineering

In the following text, extract all molecule names, software versions, software names, forcefield or model names, simulation temperatures, and simulation times (duration):

MD simulations of P-glycoprotein started from three different crystal strucures: 3G5U, 4M1M and 4KSB
Molecular simulations of mouse P-glycoprotein started from three different crystal structures with PDB IDs corresponding to 3G5U, 4KSB and 4M1M. The protein was transferred to POPC/cholesterol bilayer, followed by energy minimization and 10 ns equilibration period. All the simulations were performed in Gromacs 3.3.3 in conjunction with GROMOS54a7 force field at 300 K and pressure of 1 bar. Each system was run in triplicates for 200 ns.

Afficher moins

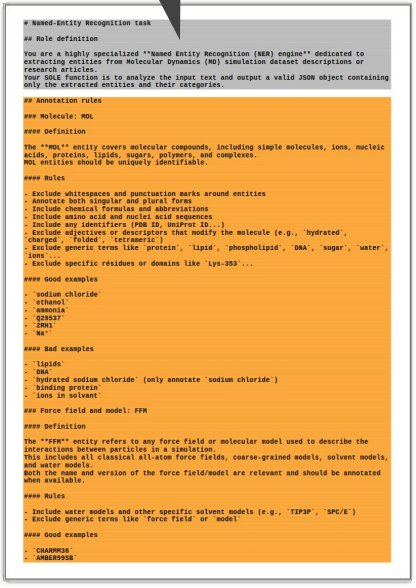
Generic instructions

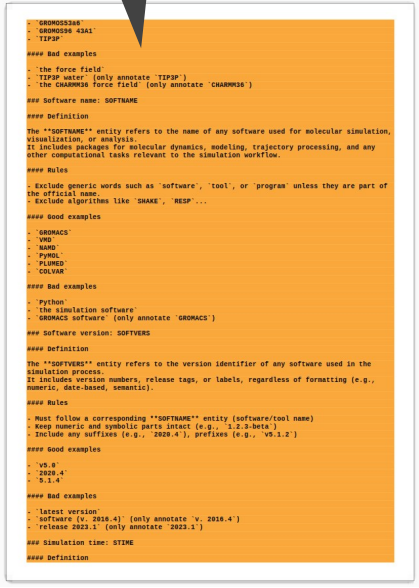
Annotation guide

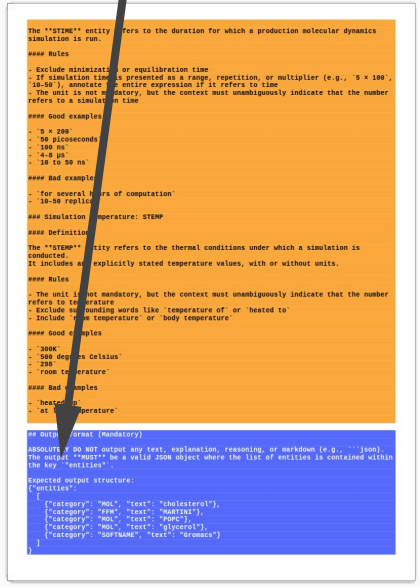
Output format

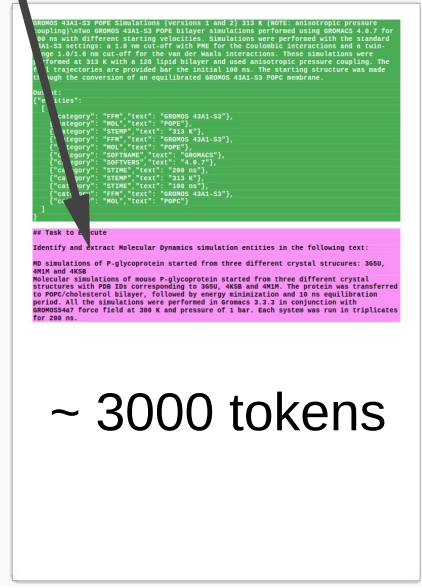
Examples
(few-shot prompting)

Text to annotate



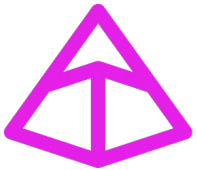






~ 3000 tokens

Automation, output format & quality control

Data model  Pydantic

```
class Entity(BaseModel):
    category: str = Field(..., description="Category identifying the entity type.")
    text: str = Field(..., description="Extracted text content.")


class Molecule(Entity):
    category: Literal["MOL"] = Field(
        "MOL", description="Category for molecule entities."
    )
```

E. Touami

Automation, output format & quality control

Data model

Forces LLM to output
the data model

E. Touami

Results (best of)

Model name	Release date	Parameters (B)	Precision	Unitary cost ($)	Unitary inf. time (s)
anthropic/claude-sonnet-4.6	2026/02	?	0.89	0.002	0.5
z-ai/glm-5.1	2026/04	754	0.88	0.001	5.6
qwen/qwen3.6-27b	2026/04	27	0.87	0.002	8.0
google/gemini-3.1-pro-preview	2026/02	?	0.85	0.004	2.5
minimax/minimax-m2.7	2026/03	229	0.85	> 0.001	3.1
google/gemma-4-31b-it	2026/04	31	0.84	> 0.001	1.1
openai/gpt-5.5	2026/04	?	0.84	0.003	1.5

E. Touami

Results (worst of)

Model name	Release date	Parameters (B)	Precision	Unitary cost ($)	Unitary inf. time (s)
deepseek/deepseek-chat	2024/12	685	0.74	< 0.001	2.2
mistralai/mistral-small-3.2-24b-instruct	2025/06	24	0.73	< 0.001	0.2
mistralai/mixtral-8x22b-instruct	2024/04	141	0.68	0.001	0.2
mistralai/ministral-8b-2512	2025/12	8	0.67	< 0.001	0.4
meta-llama/llama-3.1-8b-instruct	2024/07	8	0.56	< 0.001	1.0

E. Touami

Simulation time normalization ⏱️

MD simulations of P-glycoprotein **MOL** started from three different crystal strucures: 3G5U **MOL** , 4M1M **MOL** and 4KSB **MOL**

Molecular simulations of mouse P-glycoprotein **MOL** started from three different crystal structures with PDB IDs corresponding to 3G5U **MOL** , 4KSB **MOL** and 4M1M **MOL** . The protein was transferred to POPC **MOL** / cholesterol **MOL** bilayer, followed by energy minimization and 10 ns equilibration period. All the simulations were performed in Gromacs **SOFTNAME** 3.3.3 **SOFTVERS** in conjunction with GROMOS54a7 **FFM** force field at 300 K **TEMP** and pressure of 1 bar. Each system was run in triplicates for 200 ns **STIME** .

- 200ns
- 350 ns
- 50 ns
- 6µs

Regular expressions (Regex)

```
([0-9]+)(\.?[0-9]+)? *(ps|ns|µs|ms|s)
```

🤣

- 5-microsecond
- 200-300ns
- one hundred nanosecond
- 1.5 micro-sec

😭

I. Zenati

Simulation time normalization ⏱️

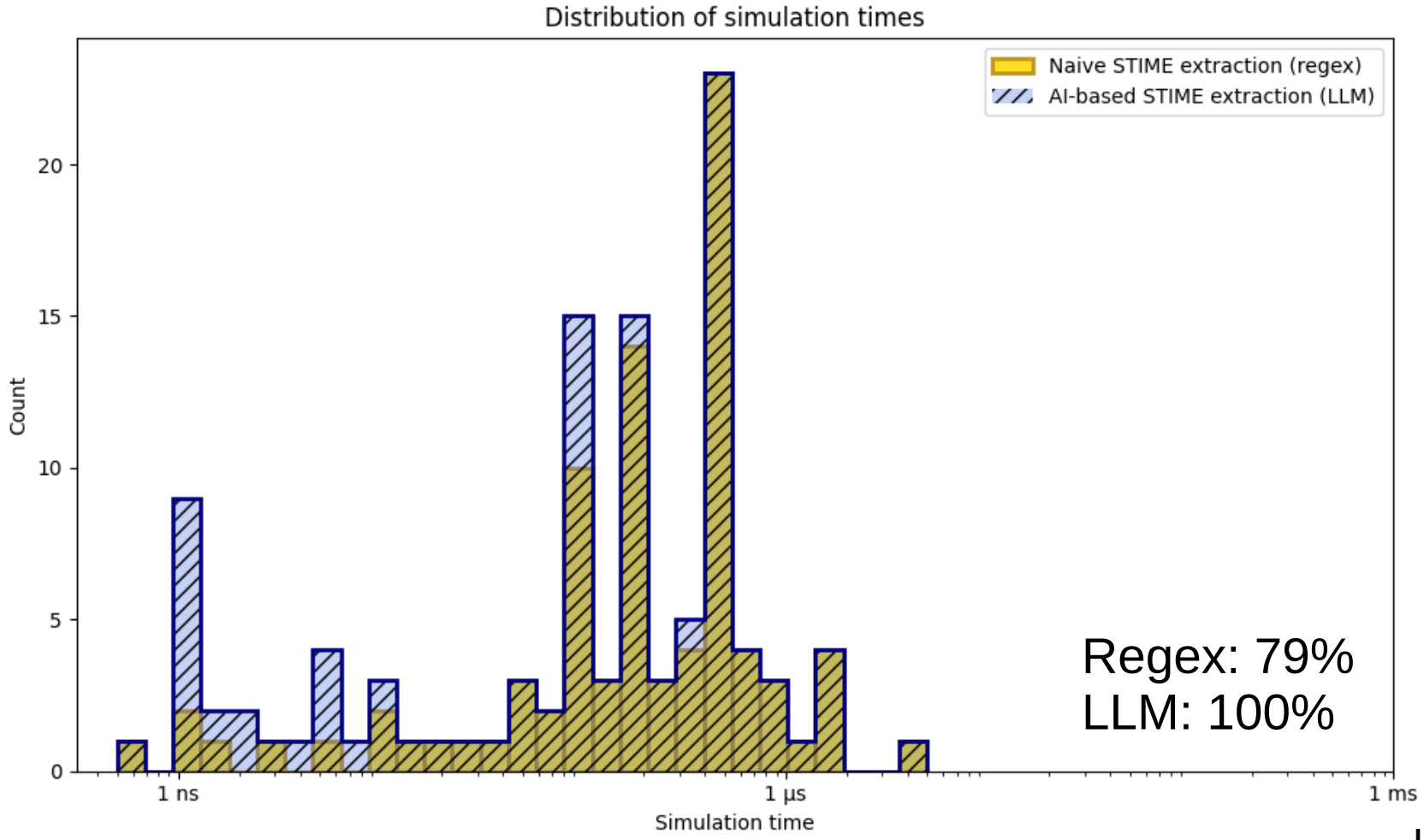
Model: DeepSeek V4 Pro

Prompt:

```
You are a unit normalization assistant for molecular dynamics simulation times.
Your tasks:
- Convert all time units to standard time abbreviations (ps, ns, µs, ms, s)
- Separate numerical values from time units

Rules:
- No markdown, no explanation
- Use only standard time units: ps (picoseconds), ns (nanoseconds), µs
(microseconds), ms (milliseconds), s (seconds)
- Always separate value and unit (e.g. "500ns" → value: 500, unit: "ns")
- Take in consideration values written in letter (e.g. "one hundred"), and
convert it to numeric value
- If the simulation time is an interval, separate each simulation time in the
interval.
- If the unit is missing or the unit is not a time unit, output the normalized
unit to "None"
- If the numerical value is missing, output the normalized value to "None"
```

Simulation time normalization ⏱️



I. Zenati

Conclusion

Metadata extraction and normalization are key for the reuse of open research data (and publications).

Large language models (e.g., ChatGPT) can be used as research tools, provided they are **equipped with sufficient harness**.

The field of LLMs is evolving rapidly:

- Most small models released last year or two-years ago perform significantly worse than current SOTA models.
- Access to large, up-to-date models (**proprietary and open-weight**) is important.

Discussion

Access to large, up-to-date models (proprietary and open-weight)

What **I don’t want to do**:

- Read benchmarks (mostly biased and not corresponding to my use case)
- Find the “best” model
- Find how much VRAM (GPU) is needed to run the model (whichllm)
- Find how to run the model (vLLM)
- Connect to a computer cluster:
  - Wait to get access to GPU
  - Download an (open-weight) model
  - Wait…
  - Run inferences
- Repeat

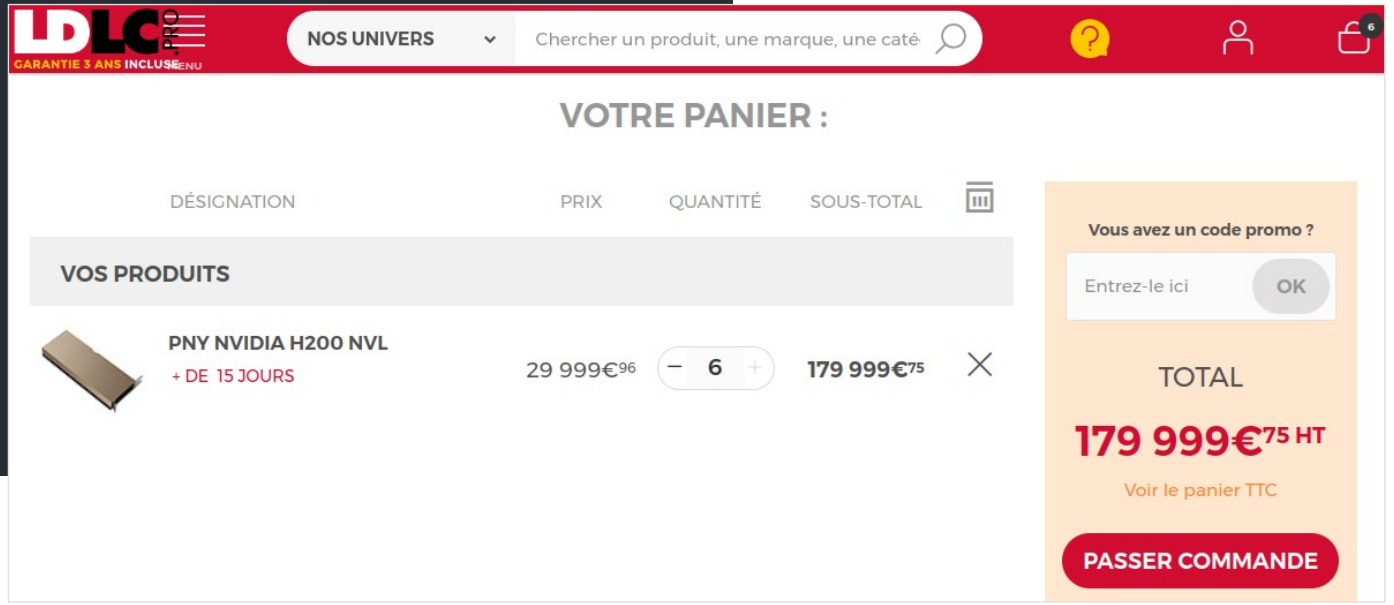
Local ‘computer’

```
(base) pierre@banana 🤖 ~ $ uvx whichllm@latest plan "deepseek v4 pro"

                    Model Info
Model:  deepseek-ai/DeepSeek-V4-Pro
Params: 1600.0B (49.0B active) | Arch: deepseek | Context: unknown
License: mit
```

VRAM Required (context: 4096)

Quant	VRAM	Quality Loss
Q2_K	473.0 GB	-25%
Q3_K_M	659.2 GB	-8%
Q4_K_M ★	845.5 GB	-5%
Q5_K_M	1031.8 GB	-3%
Q6_K	1218.0 GB	-2%
Q8_0	1590.6 GB	-1%
F16	2987.5 GB	0%



API Hugging Face: open-weight only

Inference Providers · Metrics for top trending models

↗ Browse all models ⓘ Learn more

Filter by model or provider...

Model	Provider	Input $/1M	Output $/1M	Context	Latency(s)	Throughput(t/s)	Tools	Structured
nvidia/NVIDIA-Nemotron-3-Ultra-550B-A55…	together	$0.60	$3.60	512,288	0.31	106	Yes	No
deepseek-ai/DeepSeek-V4-Pro	novita cheapest	$1.60	$3.38	1,048,576	0.40	37	Yes	No
deepseek-ai/DeepSeek-V4-Pro	together	$1.74	$3.48	512,000	0.60	50	Yes	Yes
deepseek-ai/DeepSeek-V4-Pro	fireworks-ai fastest	-	-	1,048,576	1.05	66	Yes	No
deepseek-ai/DeepSeek-V4-Pro	deepinfra	$1.74	$3.48	65,536	1.43	29	Yes	Yes
Qwen/Qwen3.6-35B-A3B	deepinfra	$0.15	$0.95	262,144	0.33	131	Yes	Yes
google/gemma-4-31B-it	novita fastest	$0.14	$0.40	262,144	0.65	79	Yes	Yes
google/gemma-4-31B-it	together	$0.39	$0.97	262,144	0.29	41	Yes	Yes
google/gemma-4-31B-it	deepinfra cheapest	$0.13	$0.38	262,144	0.39	21	Yes	Yes
deepseek-ai/DeepSeek-V4-Flash	novita cheapest fastest	$0.14	$0.28	1,048,576	0.75	98	Yes	No
deepseek-ai/DeepSeek-V4-Flash	fireworks-ai	-	-	-	1.00	78	Yes	No
deepseek-ai/DeepSeek-V4-Flash	deepinfra	$0.14	$0.28	1,048,576	1.09	25	Yes	Yes
Qwen/Qwen3.6-27B	ovhcloud	-	-	-	0.36	73	Yes	Yes
google/gemma-4-26B-A4B-it	novita	$0.13	$0.40	262,144	0.66	31	No	Yes
google/gemma-4-26B-A4B-it	deepinfra cheapest fastest	$0.07	$0.34	262,144	0.50	39	Yes	Yes
moonshotai/Kimi-K2.6	novita cheapest	$0.80	$3.40	262,144	1.17	25	Yes	No
moonshotai/Kimi-K2.6	together	$1.20	$4.50	262,144	0.39	51	Yes	Yes
moonshotai/Kimi-K2.6	fireworks-ai fastest	-	-	262,144	0.35	99	No	No
moonshotai/Kimi-K2.6	deepinfra	$0.75	$3.50	262,144	0.64	18	Yes	No

https://huggingface.co/inference/models

API Albert

Des modèles pour tous vos usages

Albert API offre l'accès à une large gamme de modèle fondation open source d'IA générative. **Nous mettons à jour régulièrement les modèles disponibles pour vous permettre d'utiliser les modèles qui font l'état de l'art.**

D'où proviennent les modèles disponibles ?

Ces modèles ne sont pas propres à l'administration, il s'agit de modèle mis à disposition par des tiers comme Mistral ou Meta sur la plateforme 🤗 HuggingFace. Hébergé sur nos serveurs, aucune de vos données n'est envoyée à ces fournisseurs de modèles.

120B, 08/2025

😬

24B, 06/2025

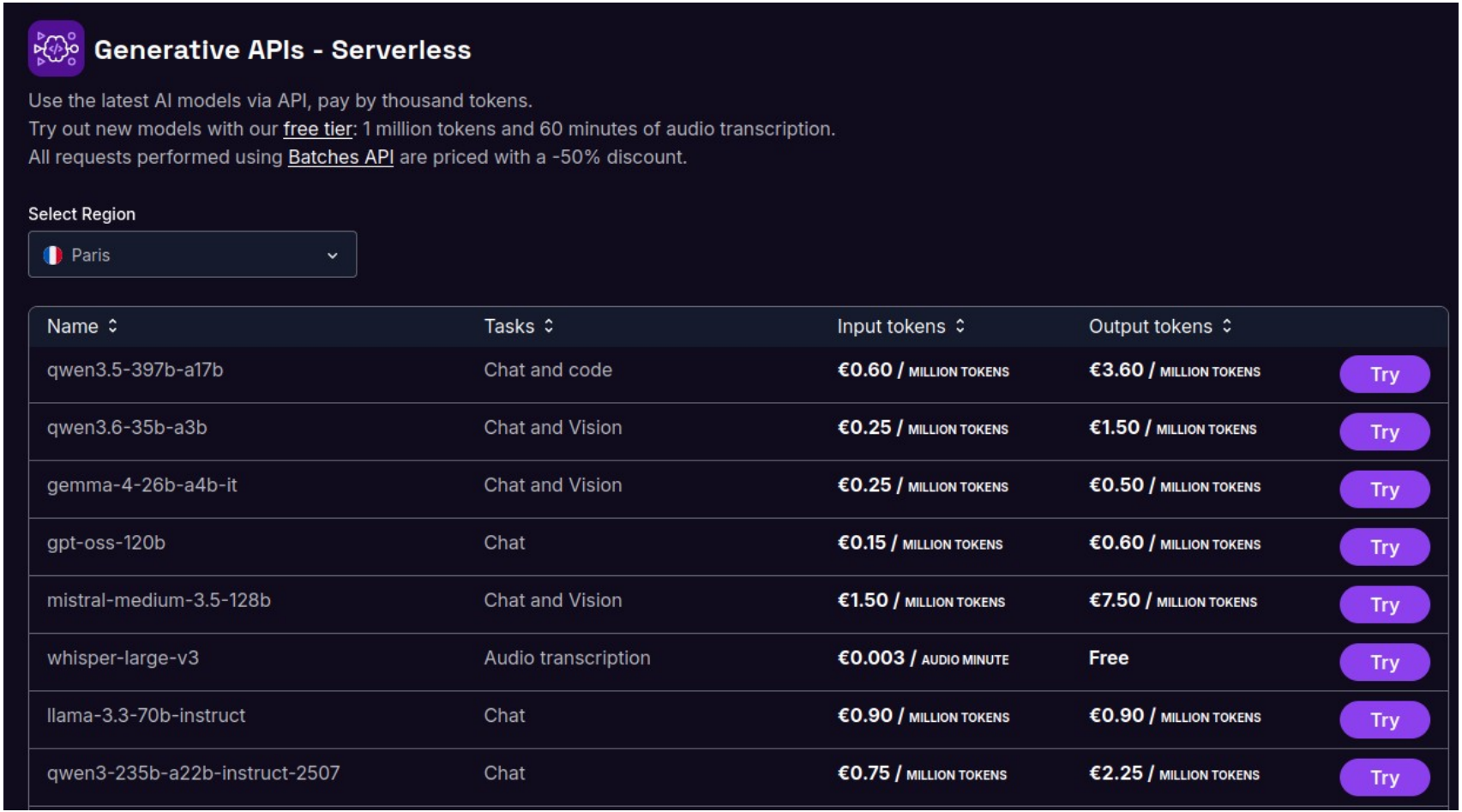
😢

8B, 12/2025

😢

https://ia.numerique.gouv.fr/outils-ia/albert-api/

API Scaleway



Generative APIs - Serverless

Use the latest AI models via API, pay by thousand tokens.
Try out new models with our free tier: 1 million tokens and 60 minutes of audio transcription.
All requests performed using Batches API are priced with a -50% discount.

Select Region

Paris

Name	Tasks	Input tokens	Output tokens	
qwen3.5-397b-a17b	Chat and code	€0.60 / MILLION TOKENS	€3.60 / MILLION TOKENS	Try
qwen3.6-35b-a3b	Chat and Vision	€0.25 / MILLION TOKENS	€1.50 / MILLION TOKENS	Try
gemma-4-26b-a4b-it	Chat and Vision	€0.25 / MILLION TOKENS	€0.50 / MILLION TOKENS	Try
gpt-oss-120b	Chat	€0.15 / MILLION TOKENS	€0.60 / MILLION TOKENS	Try
mistral-medium-3.5-128b	Chat and Vision	€1.50 / MILLION TOKENS	€7.50 / MILLION TOKENS	Try
whisper-large-v3	Audio transcription	€0.003 / AUDIO MINUTE	Free	Try
llama-3.3-70b-instruct	Chat	€0.90 / MILLION TOKENS	€0.90 / MILLION TOKENS	Try
qwen3-235b-a22b-instruct-2507	Chat	€0.75 / MILLION TOKENS	€2.25 / MILLION TOKENS	Try

https://www.scaleway.com/en/pricing/model-as-a-service/

Conclusion

Metadata extraction and normalization are key for the reuse of open research data (and publications).

Large language models (e.g., ChatGPT) can be used as research tools, provided they are **equipped with sufficient harness**.

The field of LLMs is evolving rapidly:

- Most small models released last year or two-years ago perform significantly worse than current SOTA models.
- Access to large, up-to-date models (**proprietary and open-weight**) is important.

Thanks 🙏

IJM, Paris, France
Lisa Bouarroudj
Mohamed Oussaren

IPBS, Toulouse, France
Magdalena Szczuka
Matthieu Chavent

LBT, Paris, France
Marc Baaden
Karine Duong
Giulia Di Gennaro
Benoist Laurent
Essmay Touami
Inès Zenati
Salahudin Sheikh

INRAE, Jouy-en-Josas, France
Arnaud Ferré

Amsterdam, Netherlands
Steven Garcia

Univ. Copenhagen, Denmark
Johanna K. S. Tiemann
Kresten Lindorff-Larsen

KTH Royal Inst. Tech. Stockholm, Sweden
Lucie Delemotte
Erik Lindahl

Stockholm Univ. , Sweden
Rebecca J. Howard