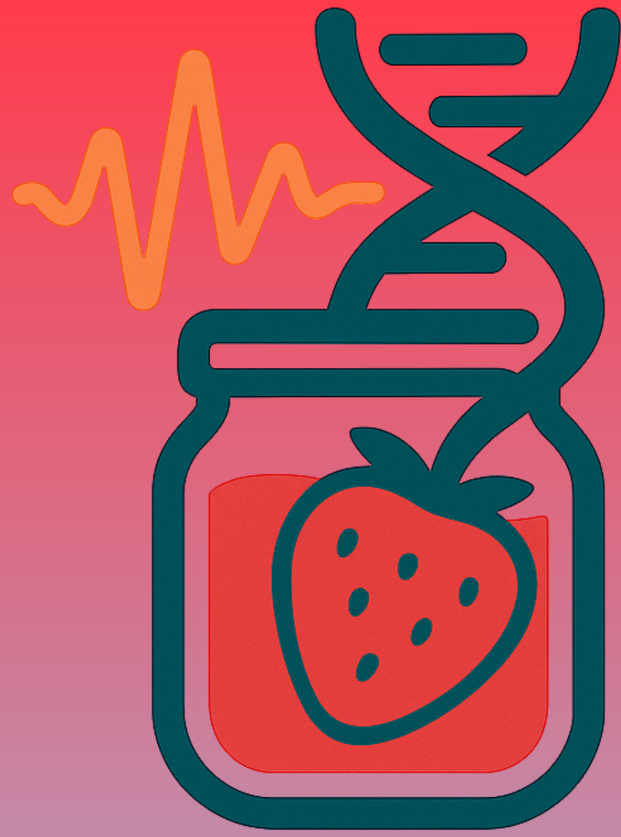
JAMSFetch

Joint Automated Multi-source Sequence Fetcher



Joanna Dąbrowska
Anna Szymik
Michał Stanowski
Stanisław Gołębiewski

Architecture of large projects in bioinformatics 2025

INTRODUCTION

GOALS

01 Automate the retrieval of sequence and structure data from major bioinformatics databases, reducing manual effort

02 Create a unified Python interface that intelligently handles multiple data sources based on ID type

03 Simplify bioinformatics workflows by automatically organizing downloaded data into standard formats and user-defined directories for easy access and management.

WHY?

01

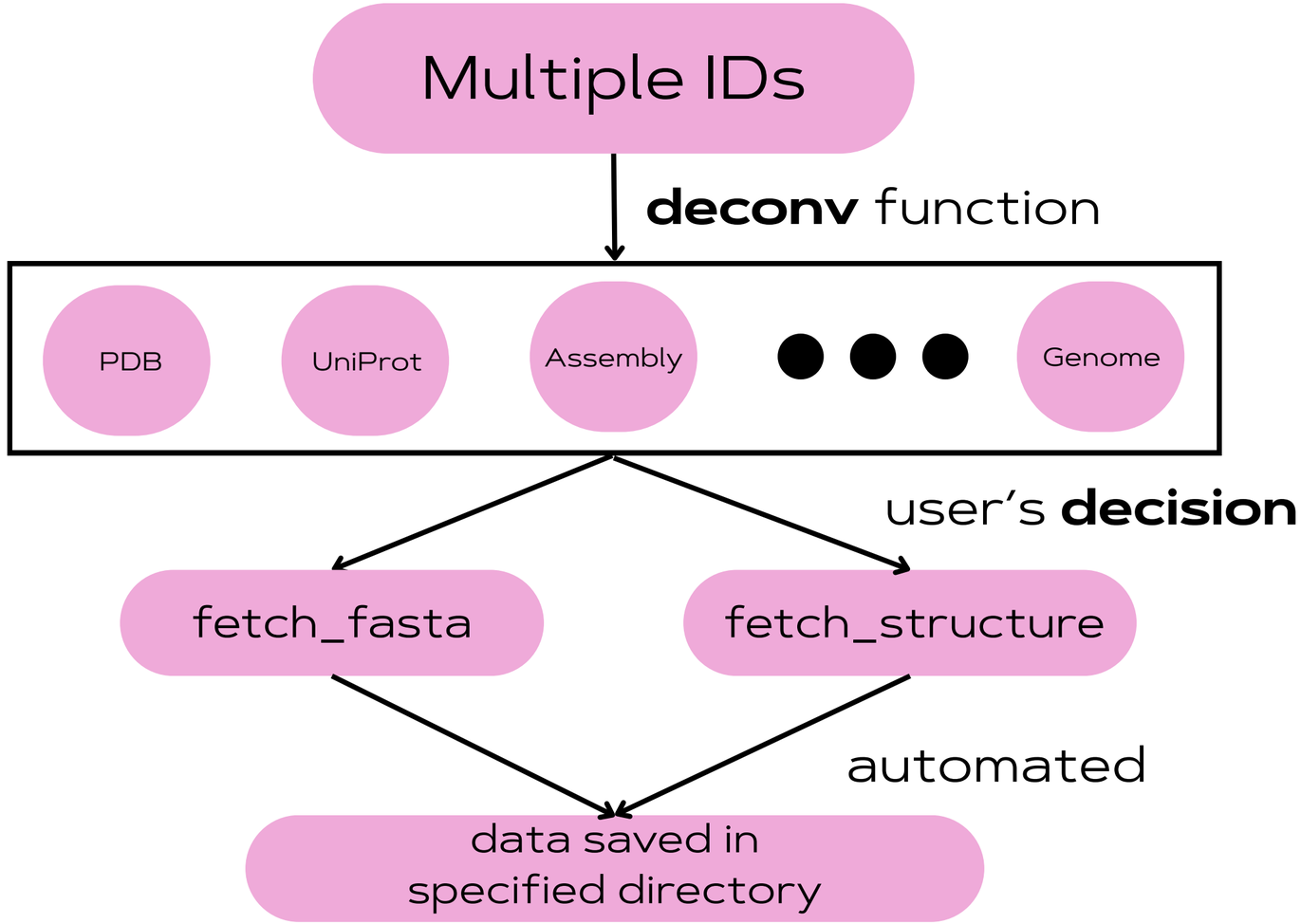
Managing and downloading biological sequence and structure data from multiple databases manually is time-consuming and error-prone, so an automated tool greatly improves efficiency and accuracy.

02

Researchers need a streamlined, unified solution to easily access diverse data sources without dealing with different formats and interfaces, enabling faster and more reproducible analyses.

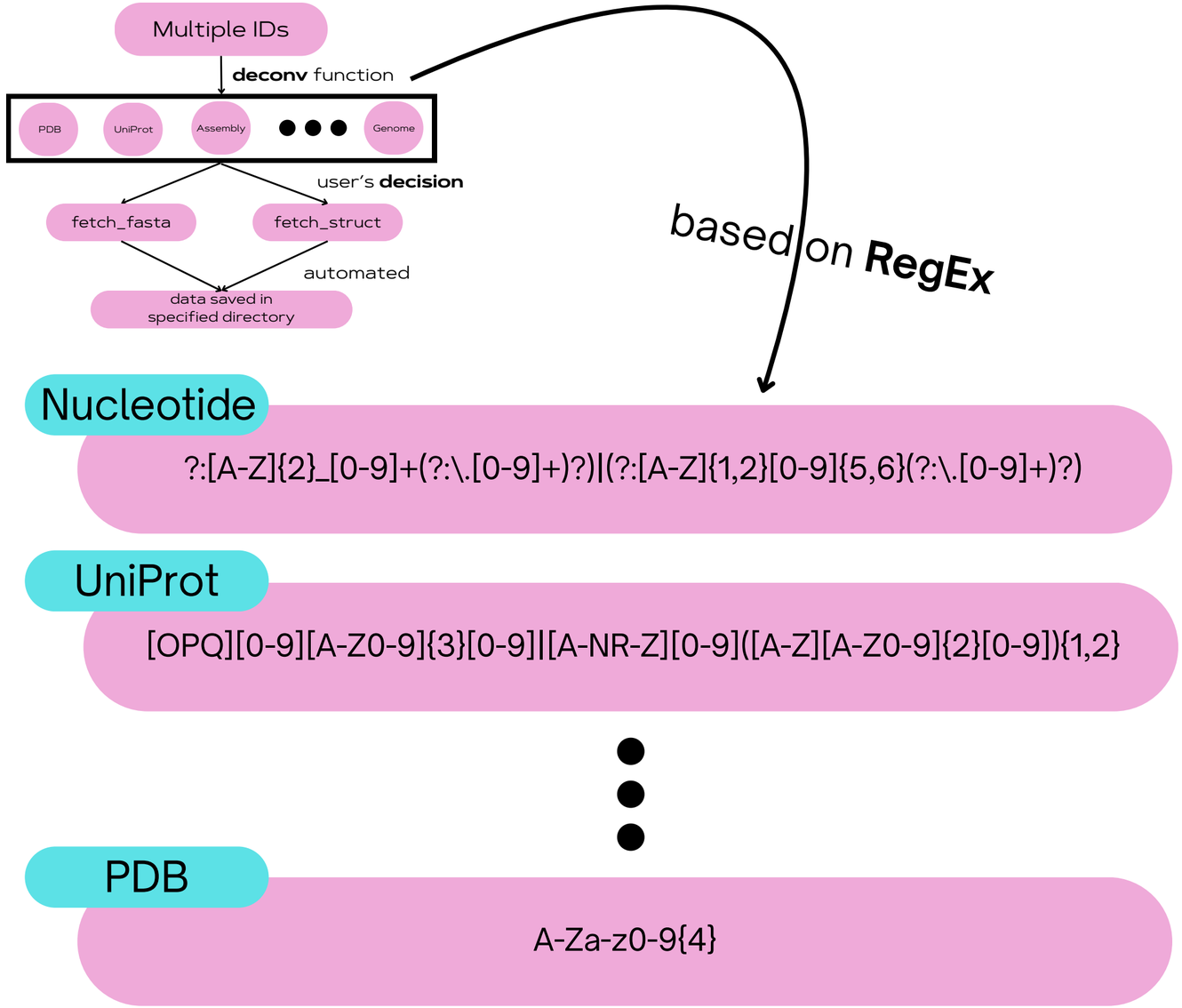
JAMSFetch

Design of the tool

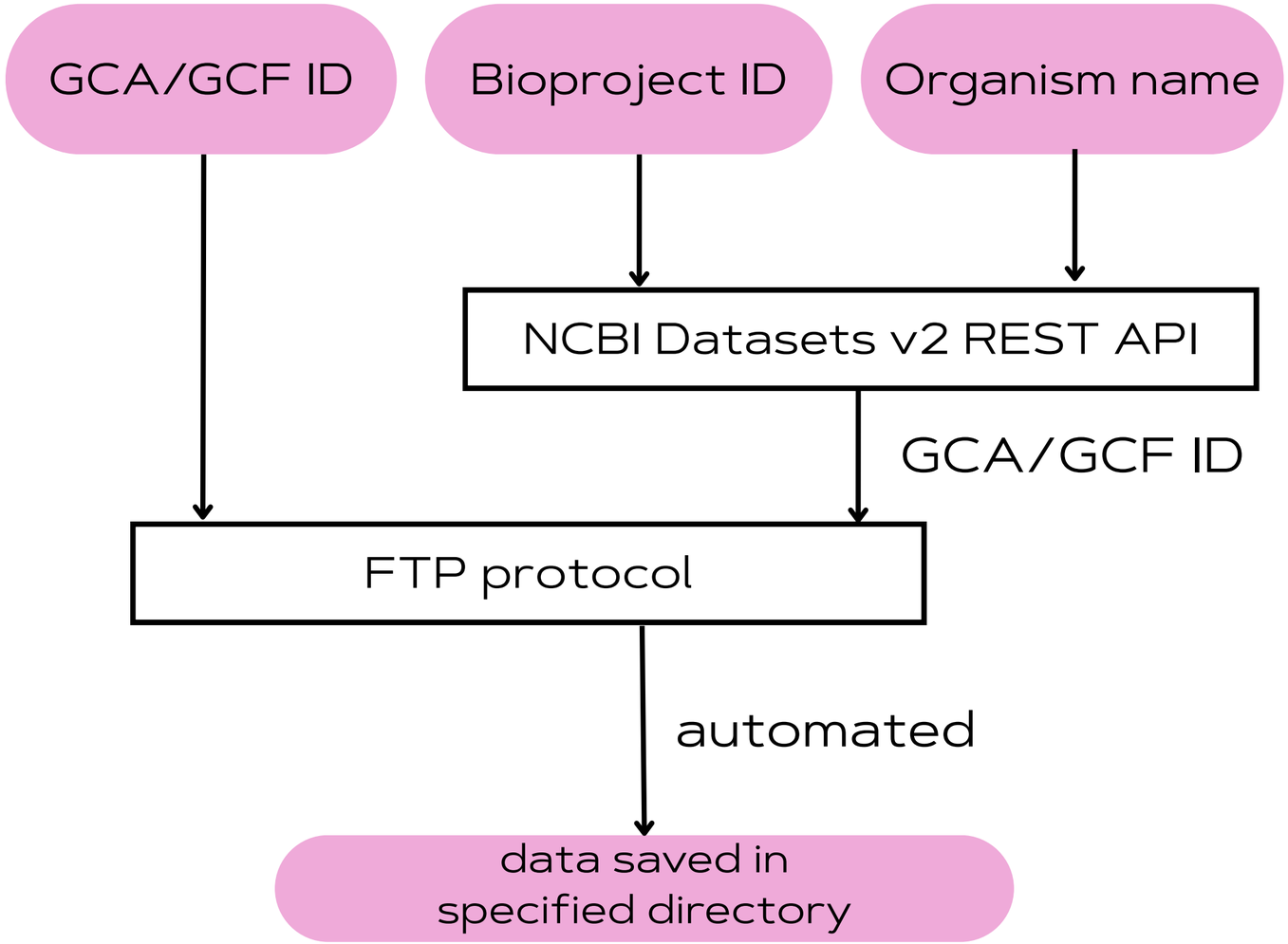


JAMSFetch

Deconvolute function



JAMSFETCH.
UTILS.
GET_ASSEMBLY



JAMSFetch

Usage example – fetching sequences/structures from **unspecified** databases

```
>>> from jamsfetch import fetch_fasta
>>> ids = ["P12345", "NM_001200.2", "GCF_000006945"]
... fetch_fasta(
...     id_list=ids,
...     output_dir="downloads/",
...     assembly_data_type="genomic"  # or "protein"
... )
...

📋 Sorted identifiers by category:

Uniprot (1):
  - P12345

Pdb (0):

Nucleotide (1):
  - NM_001200.2

Assembly (1):
  - GCF_000006945
➡️Downloading amino acid seuqences from Uniprot for: P12345
Saved: downloads/P12345.fasta
➡️Downloading nucleotide sequences from NCBI:nucleotide for: NM_001200.2
➡️Downloading genome from Genome Assembly for: GCF_000006945
--2025-06-09 18:22:41--  https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/006/945/GCF_000006945.2_ASM694v2/GCF_000006945.2_ASM694v2_genomic
.fna.gz
Loaded CA certificate '/etc/ssl/certs/ca-certificates.crt'
Resolving ftp.ncbi.nlm.nih.gov (ftp.ncbi.nlm.nih.gov)... 130.14.250.31, 130.14.250.10, 130.14.250.7, ...
Connecting to ftp.ncbi.nlm.nih.gov (ftp.ncbi.nlm.nih.gov)|130.14.250.31|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 1471540 (1,4M) [application/x-gzip]
Saving to: ‘downloads/GCF_000006945.2_ASM694v2_genomic.fna.gz’

GCF_000006945.2_ASM694v2_genomic.f 100%[=================================================================>]   1,40M   312KB/s    in 4,8s

2025-06-09 18:22:46 (302 KB/s) - ‘downloads/GCF_000006945.2_ASM694v2_genomic.fna.gz’ saved [1471540/1471540]

✅ Finished fetching all FASTA sequences.
```

Back-end

GitHub

New realese

GitHub worklow

python Package Index

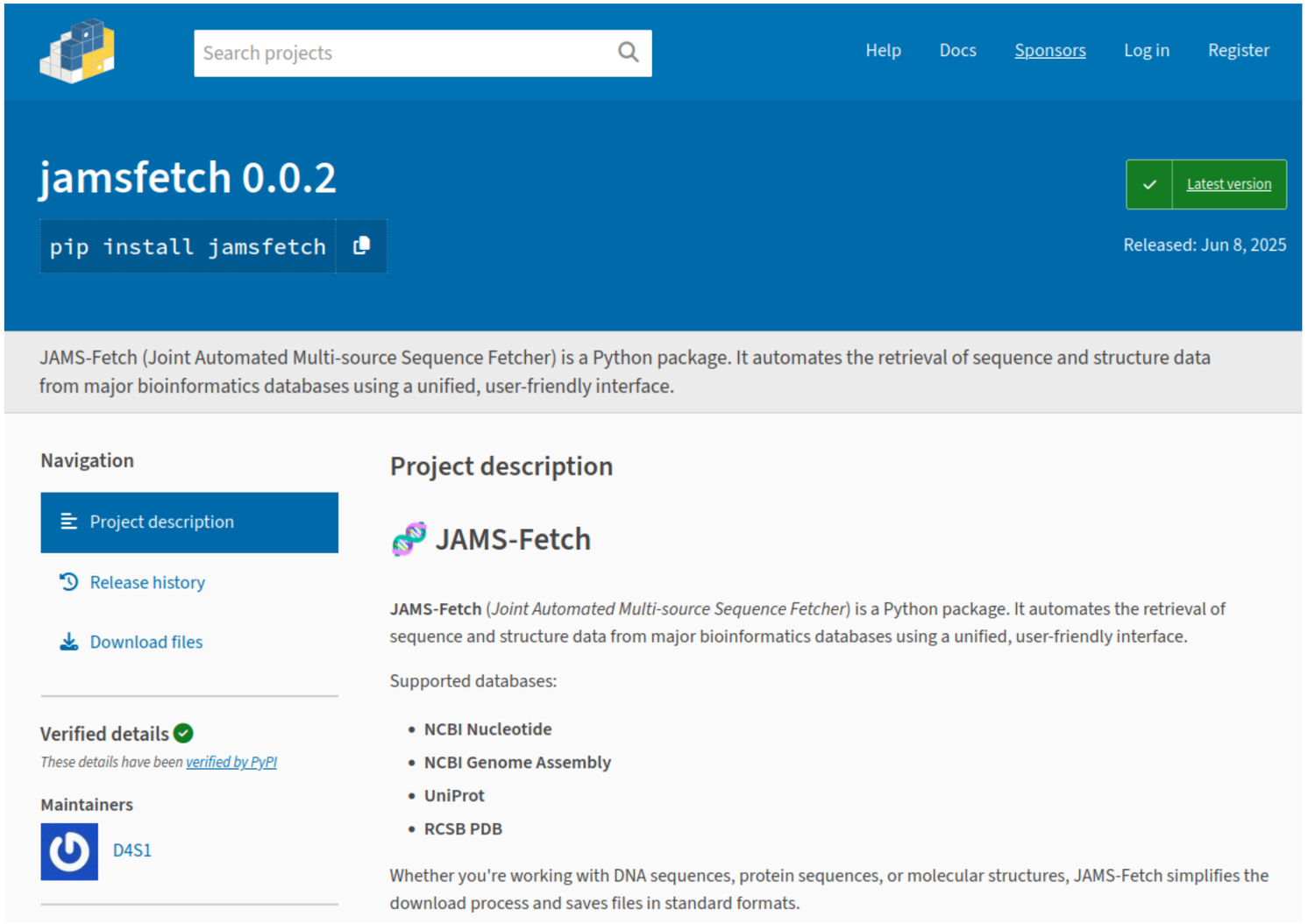
pytest

```
platform linux -- Python 3.12.3, pytest-8.4.0, pluggy-1.6.0
rootdir: /home/lambi/Desktop/studia/ArchProj/JAMS
configfile: pyproject.toml
collected 15 items

tests/test_alphafold.py ..
tests/test_assembly.py ..
tests/test_auto.py ..
tests/test_deconvolute.py .
tests/test_esm.py ..
tests/test_nucleotide.py ..
tests/test_pdb.py ..
tests/test_uniprot.py ..
```

JAMSFetch

How to install?



JAMSFetch

Which databases are supported?

- UniProt
- NCBI Nucleotide
- NCBI Genome Assembly
- PDB
- AlphaFold DB
- ESM Metagenomic Atlas

PDB

Download desired protein structures:

```
from jamsfetch.utils import get_pdb

get_pdb(
    pdb_ids=['1TUP', '9E2J'],       # specify a single ID or
    output_dir="structures/",       # specify output director
    file_format="pdb",              # or "cif"
    unzip=True                      # whether to unzip or lea
)
```

The package also supports automatic fetching of identifiers from specific databases.

SUMMARY

1. One tool, all major sources

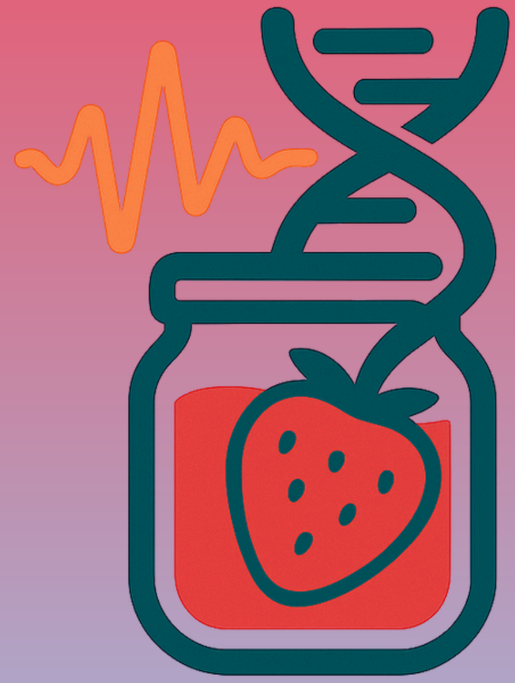
Automatically detects and fetches data from UniProt, PDB, NCBI, AlphaFold, and ESM - no manual API juggling needed.

2. Streamlined batch workflows

Download sequences and structures in bulk with minimal code — perfect for pipelines, scripts, or notebooks.

3. Standardized formats, organized output

Saves data in FASTA, PDB, or CIF formats and neatly organizes it into user-defined directories.



JAMS-Fetch