S3.3 Team Present Bioinformatics Agents

2026 Spring
LLM Agents Foundation & Applications

20260210

Presented by Xinyue Xu, Chuankai Xu, Youke Zhang

The Bioinformatics Agent Lifecycle

Perception

Defines the logic, handles biological context, and designs the analytical framework.

"Lost in Tokenization"

How LLMs ingest and represent biological data. Context vs. Sequence

Execution

Data analysis, debug coding, and reasoning through specific tasks (e.g., t-RNA).

"CellAgent"

Autonomous coding agents for single-cell analysis and data debugging.

Benchmarking

Which underlying model works best for some specific task

“TDC”

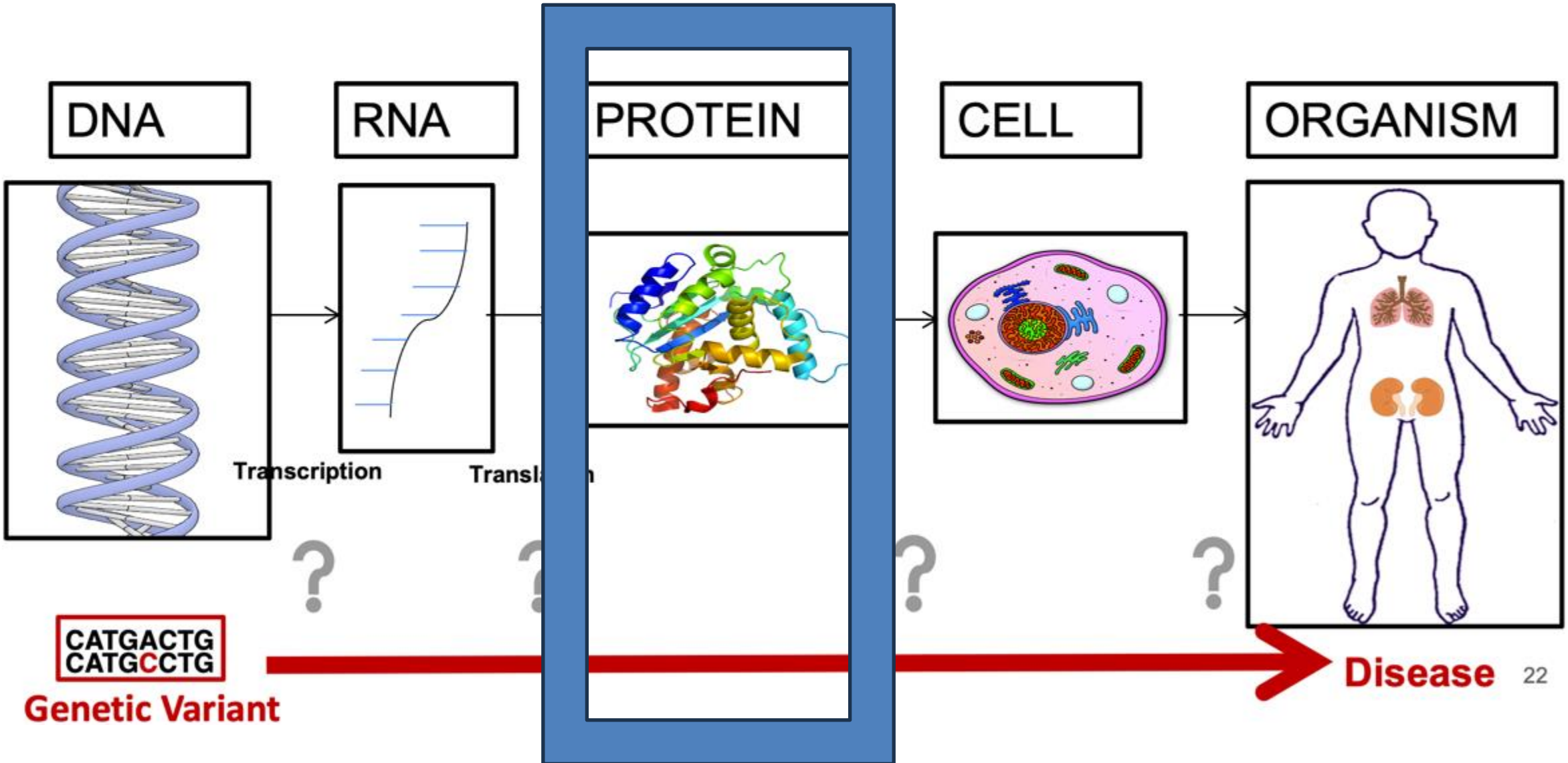
“MedHelm”

Lost in T

Context as the Key to Unlocking Biomolecular Understanding in Scientific LLMs

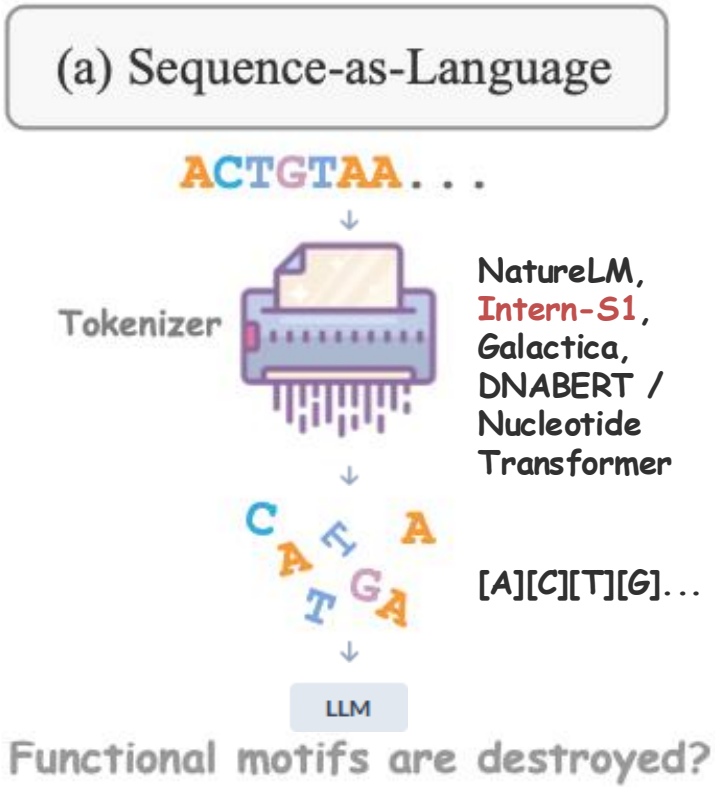
ICLR 2026
Presented by Xinyue Xu

Biology in a Slide:



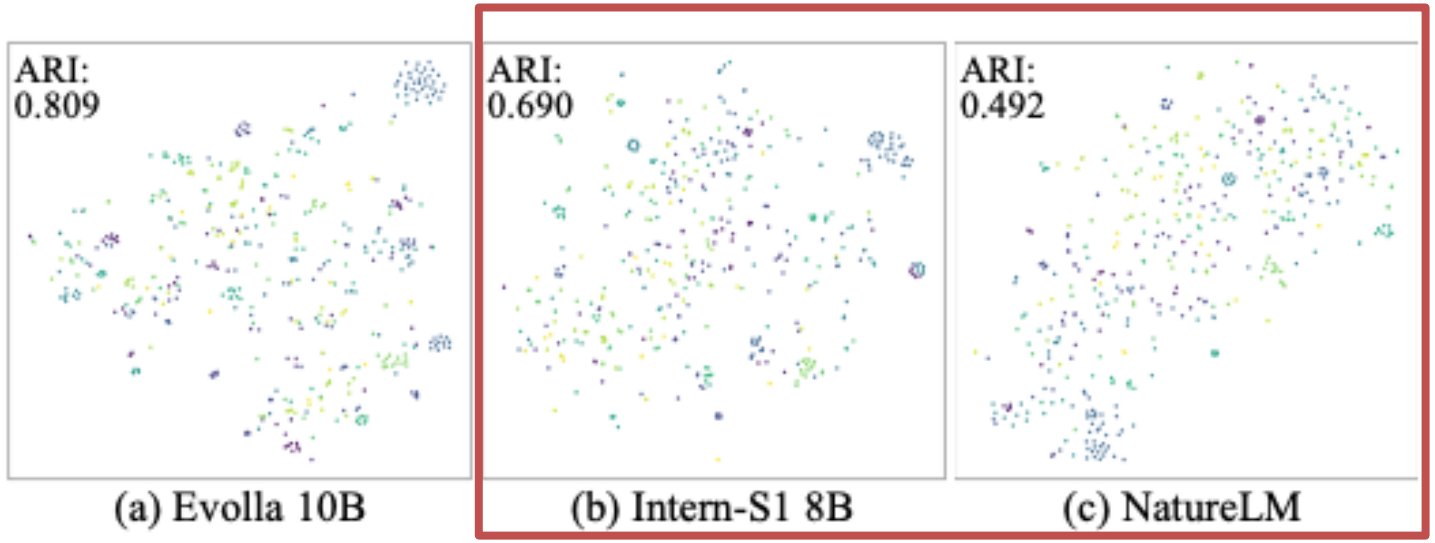
The Core Problem: Tokenization Dilemma

Presented by Xinyue Xu



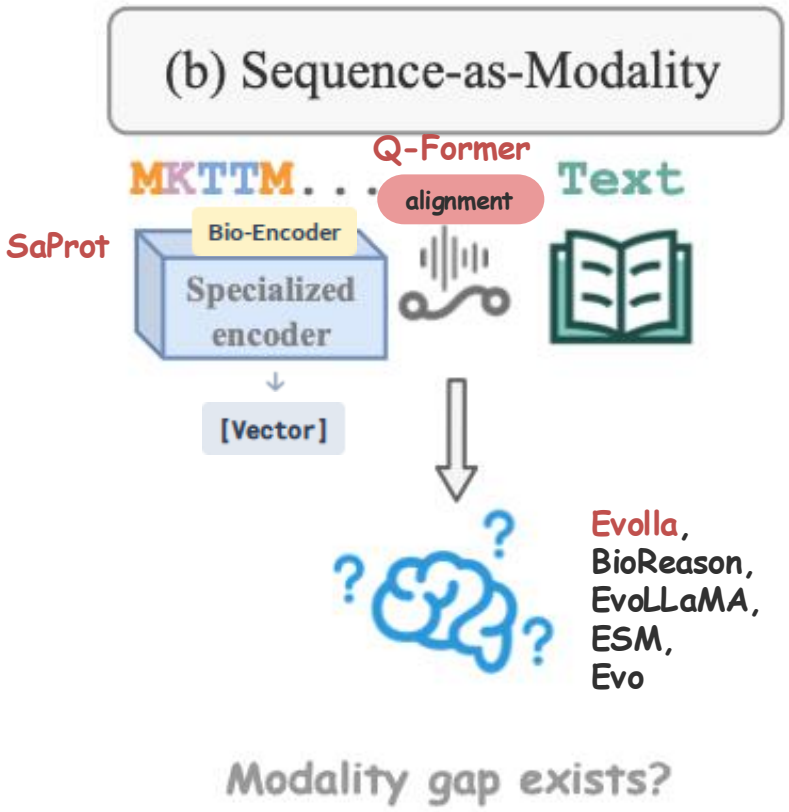
Weak Representation

Breaking sequences into single characters destroys functional motifs. The model must re-learn "biology grammar" from scratch, which is highly inefficient.



The Core Problem: Tokenization Dilemma

Presented by Xinyue Xu



Semantic Misalignment

Bio-encoders and LLMs exist in different semantic spaces. There is information loss during the "translation" from biophysics to human language.

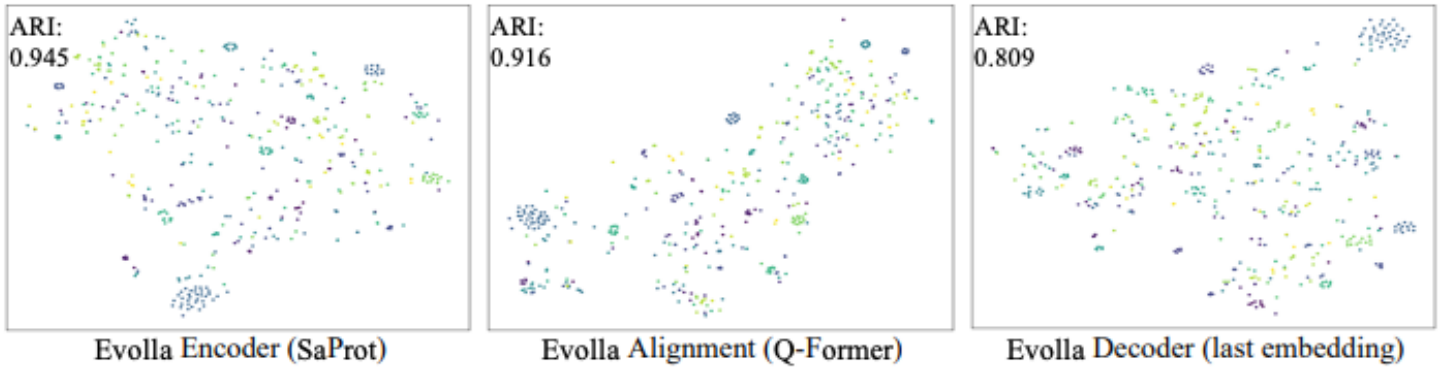
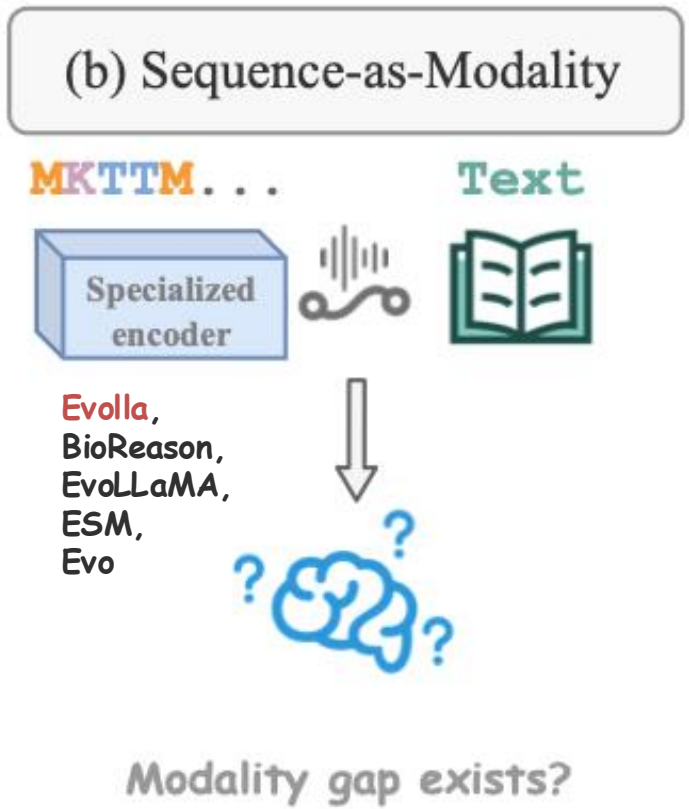


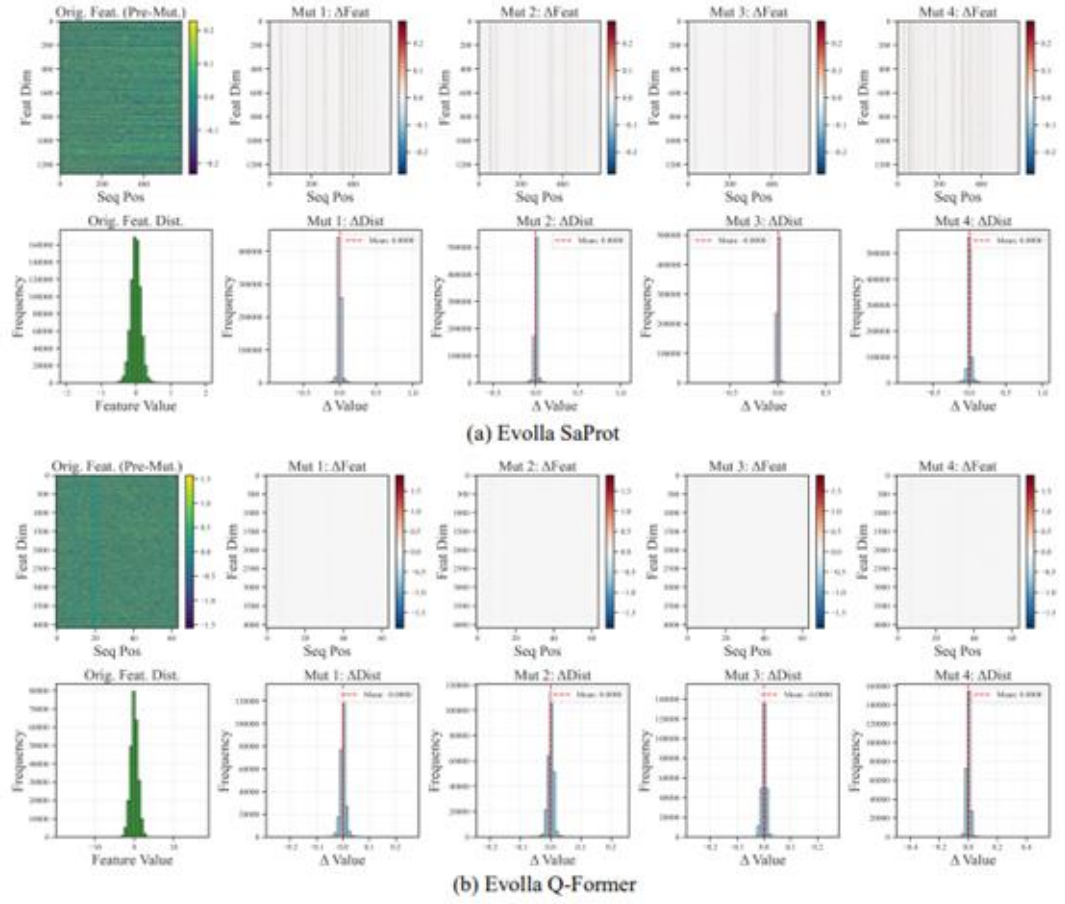
Figure 3: Visualization of representation spaces at different stages within the Evolla-10B model.

The Core Problem: Tokenization Dilemma

Presented by Xinyue Xu

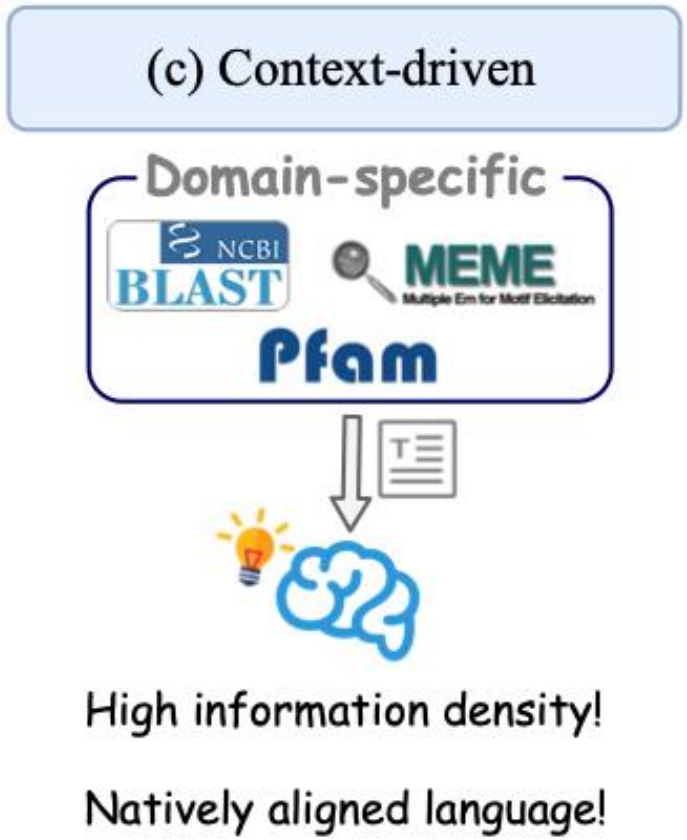


Semantic Misalignment

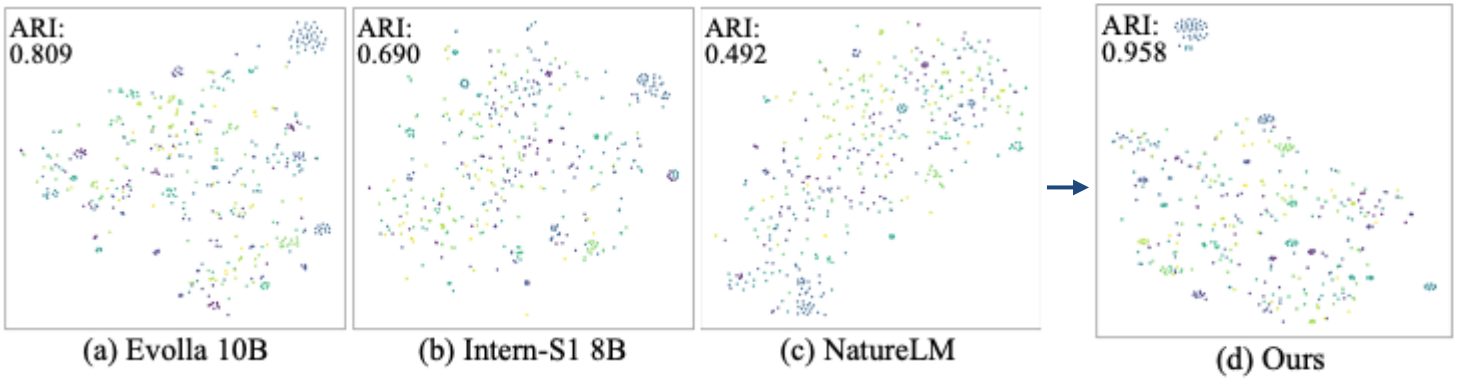


The Core Problem: Tokenization Dilemma

Presented by Xinyue Xu



Superior Representation



(a) Evolla 10B (b) Intern-S1 8B (c) NatureLM (d) Ours

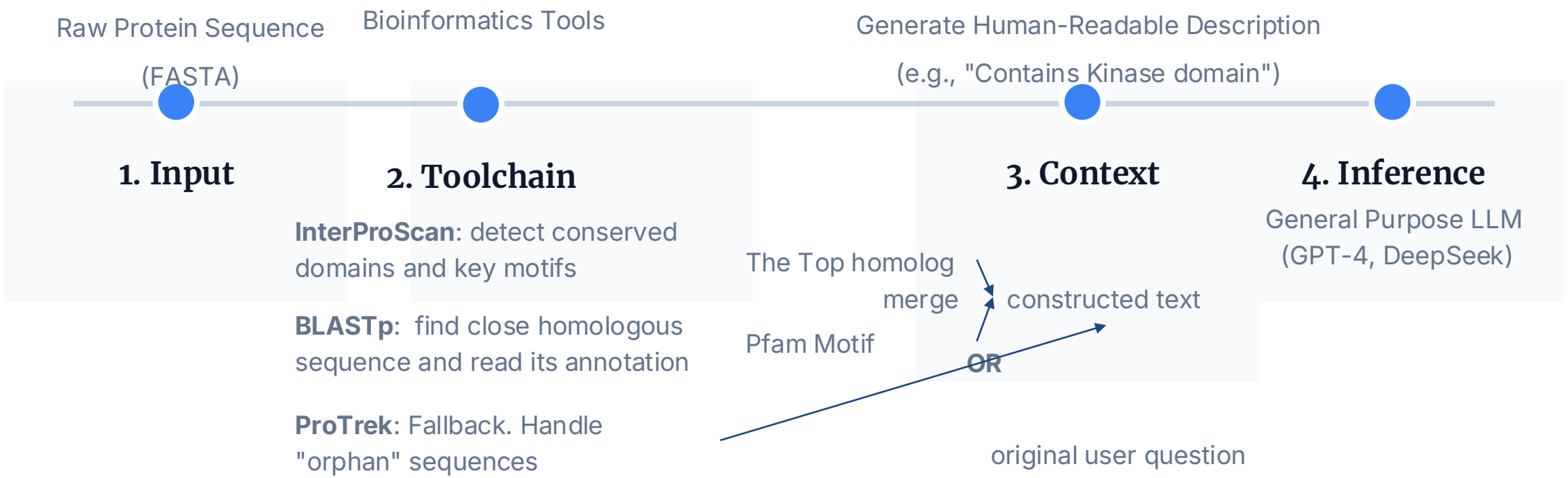
Native Alignment

Table 4: Performance (ARI) against ground-truth labels from UniClust50 and UniClust30.

Model / Representation Stage	ARI (vs. UniClust50)	ARI (vs. UniClust30)
Ours	**0.958**	**0.958**
Evolla Encoder (SaProt)	0.945	0.945
Evolla Alignment (Q-Former)	0.916	0.916
Evolla Decoder (Final)	0.809	0.809
Intern-S1 8B	0.690	0.690
NatureLM	0.492	0.492

The Solution: Context-Driven Pipeline

Presented by Xinyue Xu



Two strategies to avoid label leakage

- Intrinsic analysis rather than identity lookup
- Homology-based inference rather than direct annotation matching

The Solution: Context-Driven Pipeline

Structured Prompt for Context-Driven Reasoning

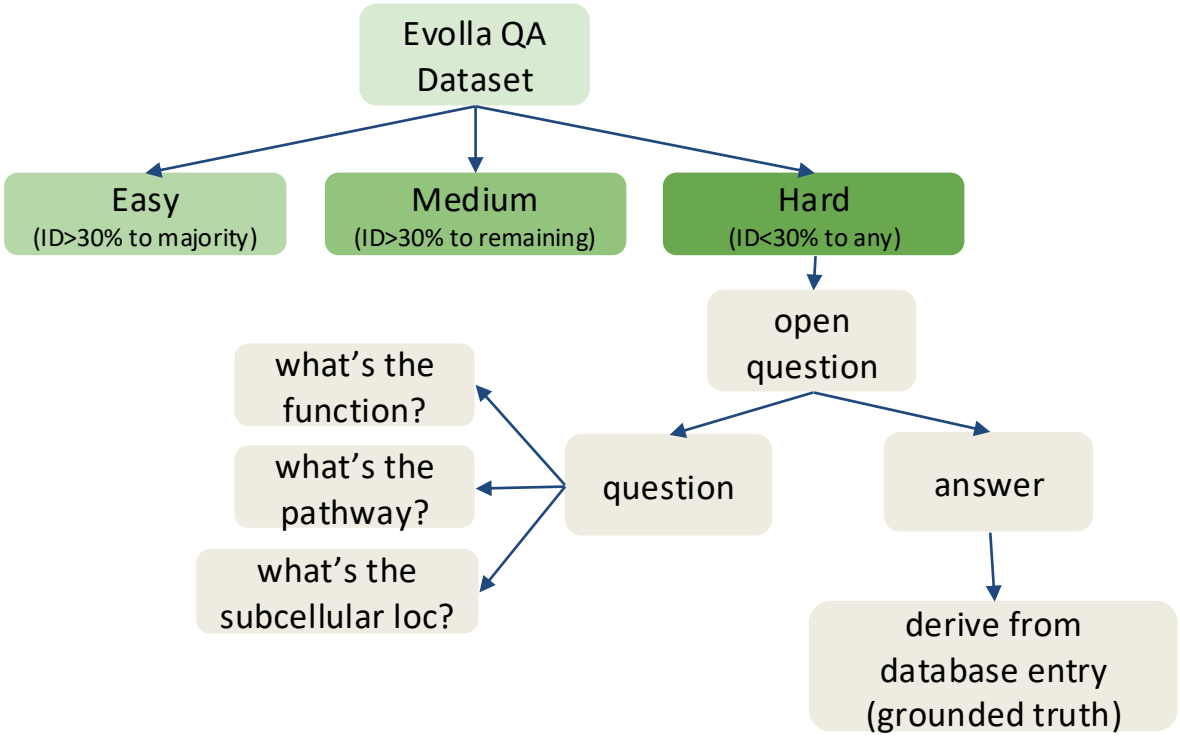
```
You are a senior systems biologist.  Analyze the input
information to answer the given question.
---------
Question:[User's Question Text]
---------
Conserved Domains (from Pfam):
[FOR EACH Pfam entry IN Pfam]:
- {the description of detected conserved domains/motifs}
Functional Annotations (from Homology via BLASTp):
- GO terms associated with the homolog:
- {the GO terms of the homolog}
Fallback Semantic Analysis (from ProTrek):
[ONLY if no homology or domain data is available]
[FOR EACH ProTrek entry In Protrek]:
- {the description of Protrek}
-------
Answer:{answer}
```

Experiment

Presented by Xinyue Xu

Primary Experiment - Setup

Dataset(Benchmark)



Task

predict the function, pathway and subcellular location of the protein

Metrics - LLM Score

use LLM(DeepSeek-V3) to score the quality of the answer

LLM-Score Adjudicator Prompt

```
As an expert biologist, you are assigned to check one paragraph
is aligned with facts or not.  You will receive some facts, and
one paragraph.  Score the paragraph between 0 to 100.
The score should be the format of {"score":  score}
---------
Here's the facts:
[Ground Truth Text from Database]
---------
Here's the paragraph:
[Generated Answer from Model to be Scored]
```

Example

Presented by Xinyue Xu

Protein A6LHQ9: Context-Only

```
You are a senior systems biologist.  Analyze the input
information to answer the given question.
---------
Question:
What is the function of this protein?
```

context only

sequence only

context + sequence

```
context Provided:
Conserved Domains (from Pfam):
• Pfam PF06321:  This family consists of several Porphyromonas
  gingivalis major fimbrial subunit protein (FimA) sequences.
  Fimbriae of Porphyromonas gingivalis, a periodontopathogen, play
  an important role in its adhesion to and invasion of host cells.
  The fimA genes encoding fimbrillin (FimA), a subunit protein
  of fimbriae, have been classified into five types, types I to
  V, based on nucleotide sequences.  It has been found that type
  II FimA can bind to epithelial cells most efficiently through
  specific host receptors [[cite:PUB00010404]].  Human dental
  plaque is a multispecies microbial biofilm that is associated
  with two common oral diseases, dental caries and periodontal
  disease.  There is an inter-species contact-dependent
  communication system between P. gingivalis and S. cristatus
  that involces the Arc-A enzyme [[cite:PUB00069820]].

• Pfam PF22449:  This domain is found at the C-terminal end of
  the putative fimbrium tip subunit FimlF from Parabacteroides
  distasonis, which contain an N-terminal domain ([pfam:PF06321])
  and a slightly larger C-terminal domain (this entry) with a
  transthyretin-like fold that contains seven core beta-strands
  arranged in two beta-sheets an extra conserved 'appendage' of
  two amphipathic beta-strands [[cite:PUB00080711]].
```

```
Sequence Provided:
MRFNVVLFMLIVALLGGLSTCSSEVPIGFDTDELSFDMSLVLLTGDMQTKASDPNYTYATTEEL
TIQNCHVAVFDKDGKRIYFKNFYSKDLGEMKTIGNLSGYELQLEGVRTFGKEDKKVSVLVVANA
NNANNSPFDNLTTYDGVDNSYTAKTIAKGPVTASLLVKIGKSETTLKYNQDNAPVTVSLIQLSA
KIEYTGVYKKENGELLEGFSLTKVAGLNASSKITIFNTSAVENGAFSDLAYPTTKPVTFYTYEI
SDAFKEVILSVQSGVEPKEYPFPANKFIKGNYYRIKGLKSSTEIEWVLENVEDKEVTLDPFE
```

```
Context Provided:   [Same as Case 1]
Sequence Provided:  [Same as Case 2]
```

answer

LLM-Score:100

answer

LLM-Score:0

answer

LLM-Score:95

Experiment

Presented by Xinyue Xu

Primary Experiment - Results

Results

Model	Sequence	Context	Func.	Path.	Sub. Loc.	All
Specialized Sci-LLMs						
Intern-S1	✓		20.57	26.56	69.75	43.33
Intern-S1	✓	✓	74.18	98.85	93.00	84.03
Intern-S1		✓	76.22	97.60	95.60	86.15
Evolla	✓		40.23	72.71	79.76	59.93
Evolla	✓	✓	57.46	84.69	83.05	70.53
Evolla		✓	65.77	83.33	81.88	74.02
NatureLM	✓		3.58	5.52	10.45	6.82
NatureLM	✓	✓	42.33	64.25	32.30	38.86
NatureLM		✓	44.77	51.35	32.51	39.50
General LLMs						
Deepseek-v3	✓		10.98	24.54	74.72	40.77
Deepseek-v3	✓	✓	77.40	91.35	94.75	86.03
Deepseek-v3		✓	75.79	93.96	93.65	84.99
Gemini2.5 Pro	✓		10.40	13.85	77.58	41.25
Gemini2.5 Pro	✓	✓	79.12	94.17	94.65	86.98
Gemini2.5 Pro		✓	79.17	98.65	94.56	**87.19**
GPT-5	✓		19.64	17.08	64.15	39.83
GPT-5	✓	✓	79.89	89.48	71.30	76.45
GPT-5		✓	77.25	85.73	73.05	75.76
Qwen3-235B-A22B	✓		13.67	19.90	37.17	39.51
Qwen3-235B-A22B	✓	✓	76.62	96.35	94.78	85.90
Qwen3-235B-A22B		✓	75.63	92.19	94.28	84.99

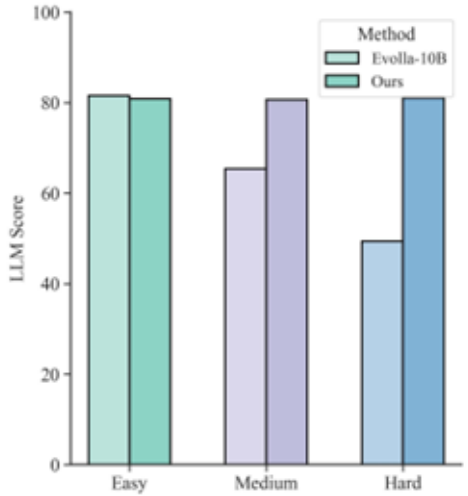


Figure 4: Comparison of Evolla-10B and our approach across the easy, medium, and hard subsets.

Take away: Sequence is Noise. Raw biomolecular sequences, when provided alone, offer limited utility and,when combined with context, consistently act as informational noise.

Experiment

Presented by Xinyue Xu

Performance on a specific bioinformatics task

Dataset

Evolla EC Number Dataset

Task

Predict the functional class of enzymes.

Metric

a hierarchical F1-score metric that evaluates performance at each of the four levels of the EC number

ref: 1.2.3.4

ans1: 1.2.3.8

ans2: 1.3.2.1

$$\text{Precision}_{\text{micro}} = \frac{\sum \text{TP}_i}{\sum \text{TP}_i + \sum \text{FP}_i}$$

$$\text{Recall}_{\text{micro}} = \frac{\sum \text{TP}_i}{\sum \text{TP}_i + \sum \text{FN}_i}$$

$$\text{F1}_{\text{micro}} = 2 \cdot \frac{\text{Precision}_{\text{micro}} \cdot \text{Recall}_{\text{micro}}}{\text{Precision}_{\text{micro}} + \text{Recall}_{\text{micro}}}$$

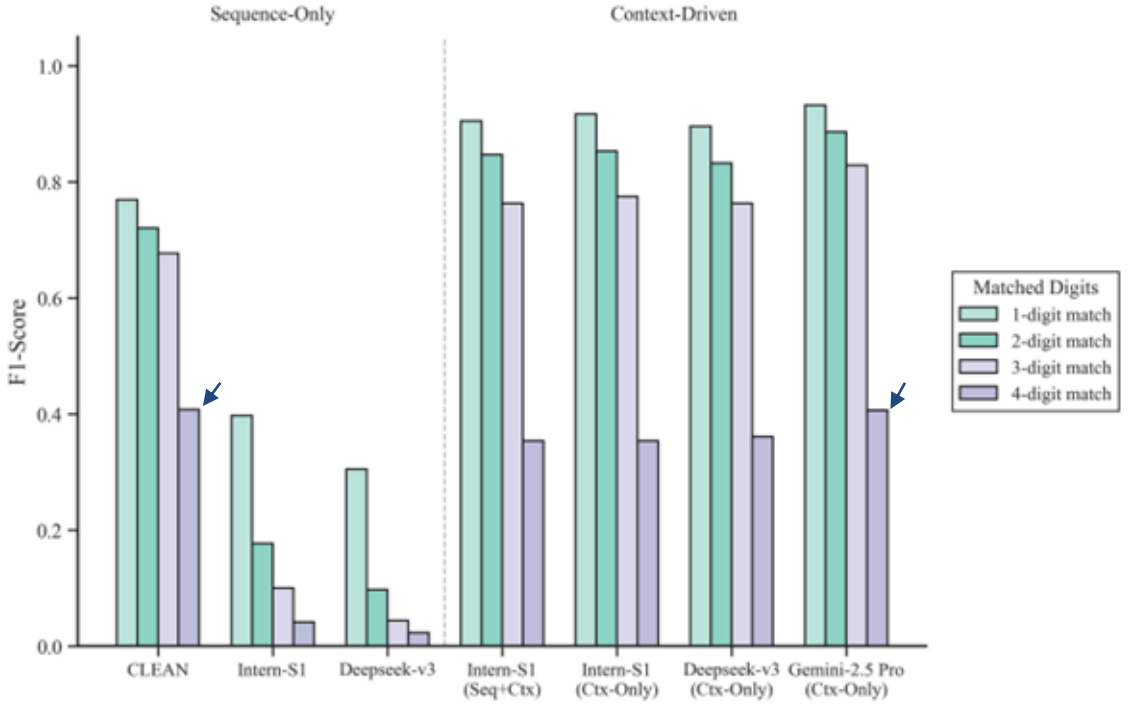


Figure 9: **Performance on EC Number Prediction (F1-Score).** The plot is divided into Sequence-Only models (left) and Context-Driven models (right). A clear and dramatic performance gap is visible between the two groups. Context-driven approaches significantly outperform even specialized sequence-based models like CLEAN, especially at higher levels (3 and 4 digits).

Experiment

Presented by Xinyue Xu

Performance over time and impact of sequence novelty

Dataset

random 100 proteins/year
from 1995-2024

Task

predict the function, pathway and subcellular location of the protein

Metric

LLM Scores

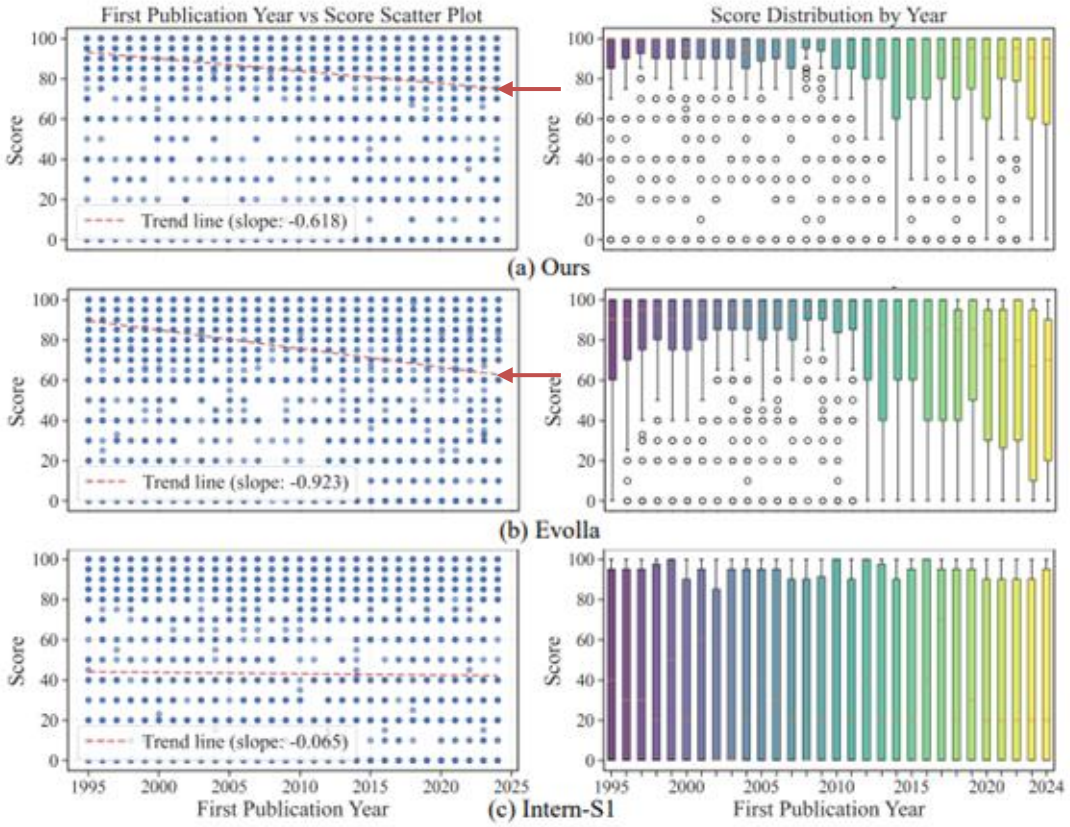


Figure 5: Analysis of model performance versus protein's first publication year.

Experiment

Presented by Xinyue Xu

Wet-Lab Validation on Novel Sequence

Dataset

Novel Sequence
(Unpublished and unseen data)

Task

Classification.Predict if it is belong to Rhodopsin and PETase family.

Metric

Accuracy

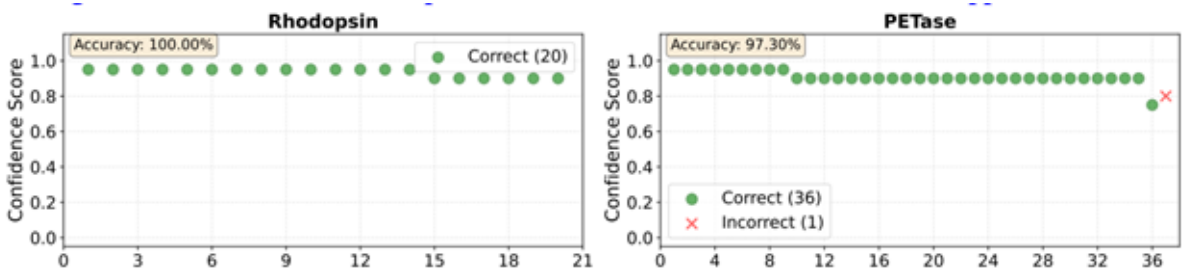


Figure 6: Sample-level performance of our context-driven method + Deepseek-v3.

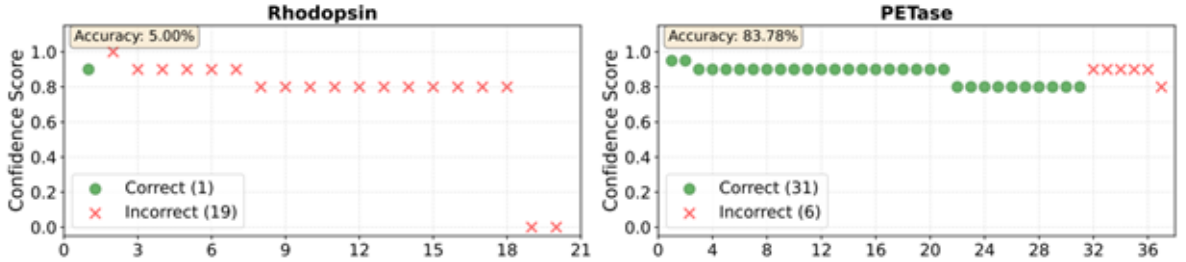


Figure 7: Sample-level performance of Evolla.

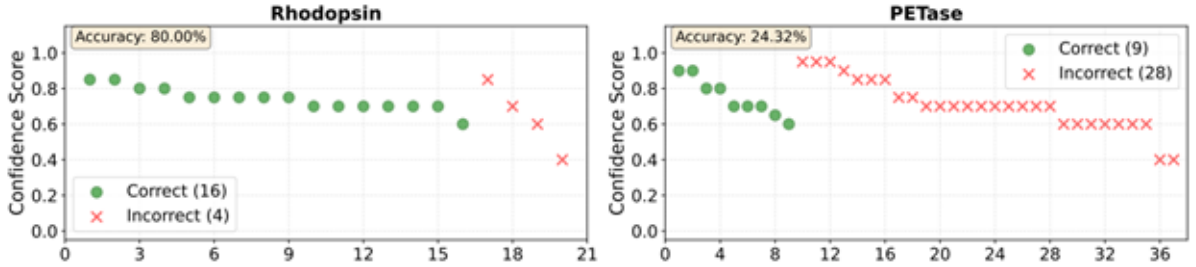


Figure 13: Sample-level performance on classification accuracy of Intern-S1.

Experiment

Presented by Xinyue Xu

Performance on Other Standard Benchmark Dataset

Dataset

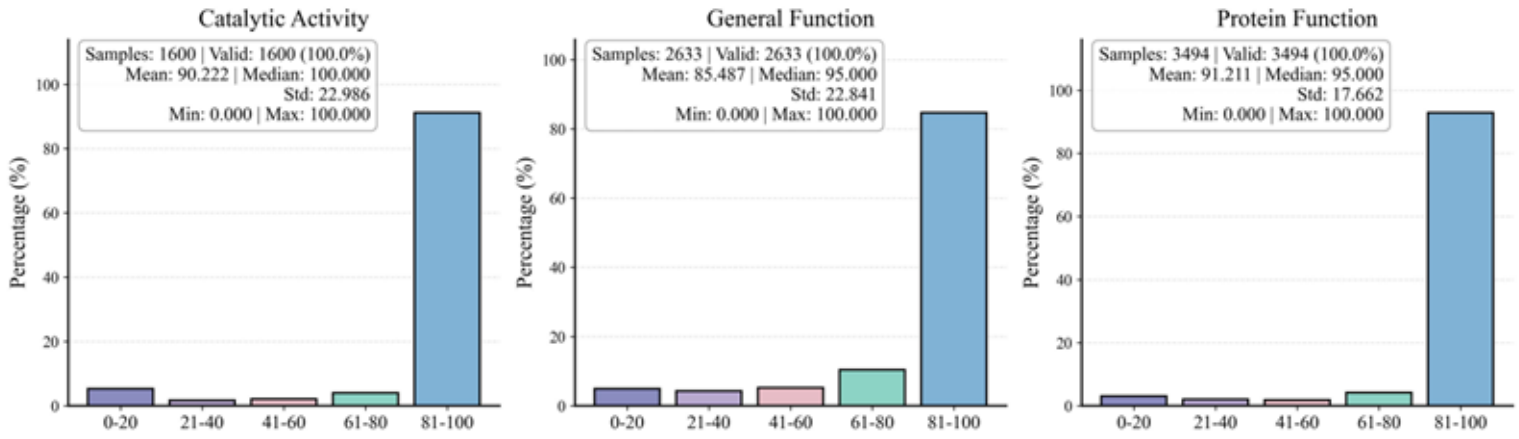
Mol-Instruction benchmark

Task

Predict molecular catalytic activity, general function and function of the protein.

Metric

LLM-Score



Experiment

Presented by Xinyue Xu

Other Biomolecular Types - DNA

Dataset

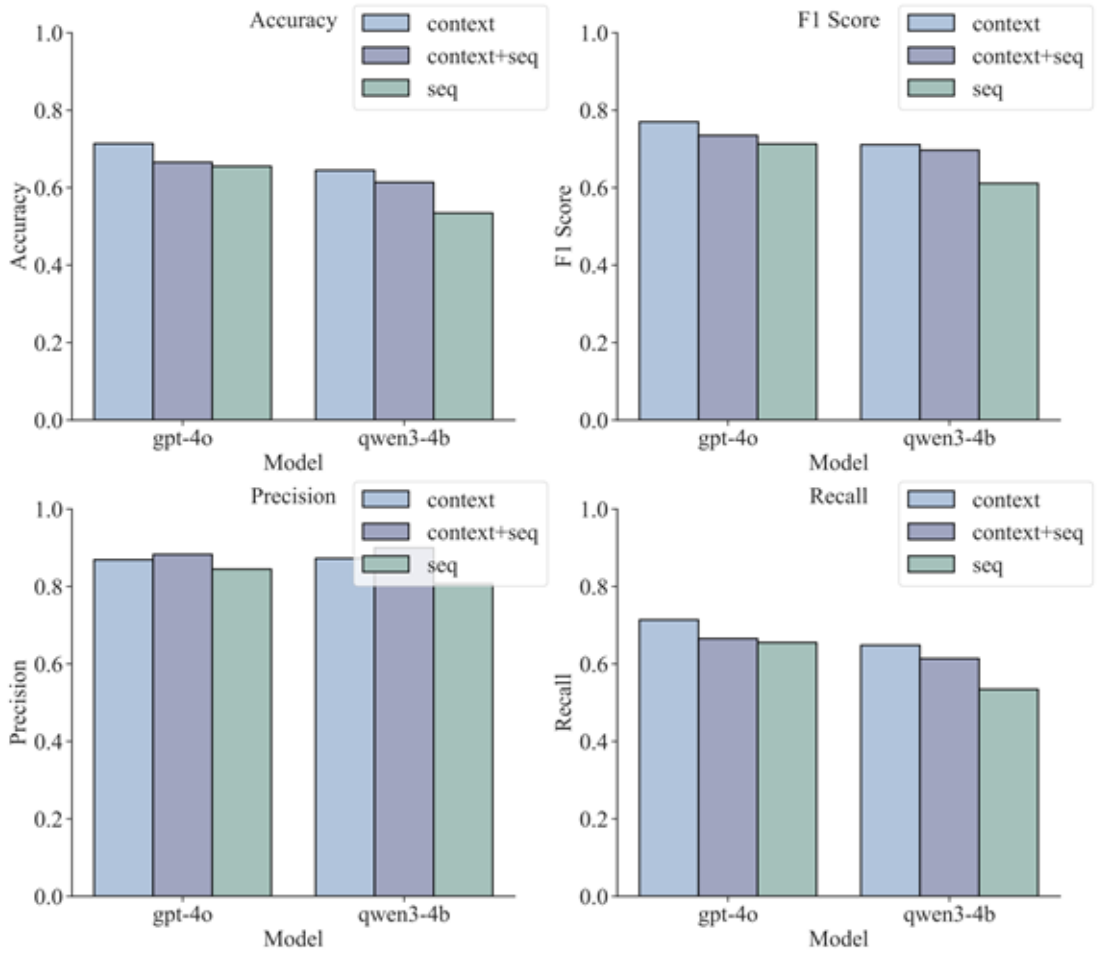
KEGG Disease Dataset
(Reference DNA,Variant DNA,KEGG Pathway, label:Desease)

Task

Reason from the mutation and its functional context to predict the correct disease

Metric

Accuracy, F1 Score, Precision, Recall



Evaluation

Presented by Xinyue Xu

Efficiency

Table 2: Comparative analysis of inference efficiency.

Method	Mode	Input to LLM	Avg. Time	Avg. Cost
Deepseek-v3	Single	Raw sequence	~30s	$0.0005
Evolla GPU	Single	Raw sequence	~90s	$0.0690
Our Method CPU	Single	Context	~70s	$0.0030
Evolla	Batch	Raw sequence	~20s	$0.0152
Our Method	Batch	Context	~0.13s	$0.0005

score : 40.77

score : 59.93

score : 84.99

Evaluation

Presented by Xinyue Xu

Ablation Study

Table 3: Ablation study of context components on our benchmark dataset. Scores reflect the model's performance when provided with different combinations of contextual information. Our final, conditional approach yields the best result.

Context Components Provided	LLM Score
Single Components	
Pfam only	74.90
GO only	84.02
ProTrek only	66.44
Pairwise Combinations	
Pfam + GO	84.60
Pfam + ProTrek	77.00
GO + ProTrek	77.78
Full Combinations	
Pfam + GO + ProTrek (Unconditional)	81.56
Pfam + GO + ProTrek (Conditional) fallback	**84.99**

Limitation

Presented by Xinyue Xu

- InterProScan and BLAST are not sensitive to subtle mutations at a single or few amino acid positions.
- For truly novel orphan proteins from unexplored regions of the protein universe, this method's performance may be constrained.

Conserved Domains (from Pfam):
- **PF00732**
- **PF05199**

Functional Annotations (from Homology via BLASTp):
- **GO 0005737**
- **GO 0005576**
- **GO 0046562**
- **GO 0050660**
- **GO 0044550**

Sequence:
GIEASLLTDPKEVAGRTVDYIIAGGGLTGLTTAARLTENPDITVLVIESGSYES
DRGPIIEDLNAYGDIFGSSVDHAYETVELATNNQTALIRSGNGLGGSTLVNGGT
WTRPHKAQVDSWETVFGNEGWNWDSVAAYSLQAERARAPNAKQIAAGHYFNASC
HGINGTVHAGPRDTGDDYSPIVKALMSAVEDRGVPTKKDLGCGDPHGVSMFPNT
LHEDQVRSDAAREWLLPNYQRPNLQVLTGQYVGKVLLSQNATTPRAVGVEFGTH
KGNTHNVYAKHEVLLAAGSAVSPTILEYSGIGMKSILEPLGIDTVVDLPVGLNL
QDQTTSTVRSRITSAGAGQGQAAWFATFNETFGDYTEKAHELLNTKLEQWAEEA
VARGGFHNTTALLIQYENYRDWIVKDNVAYSELFLDTAGVASFDVWDLLPFTRG
YVHILDKDPYLRHFAYDPQYFLNELDLLGQAAATQLARNISNSGAMQTYFAGET
IPGDNLAYDADLRAWTEYIPYNFRPNYHGVGTCSMMPKEMGGVVDNAARVYGVQ
GLRVIDGSIPPTQMSSHVMTVFYAMALKIADAVLADYASMQ

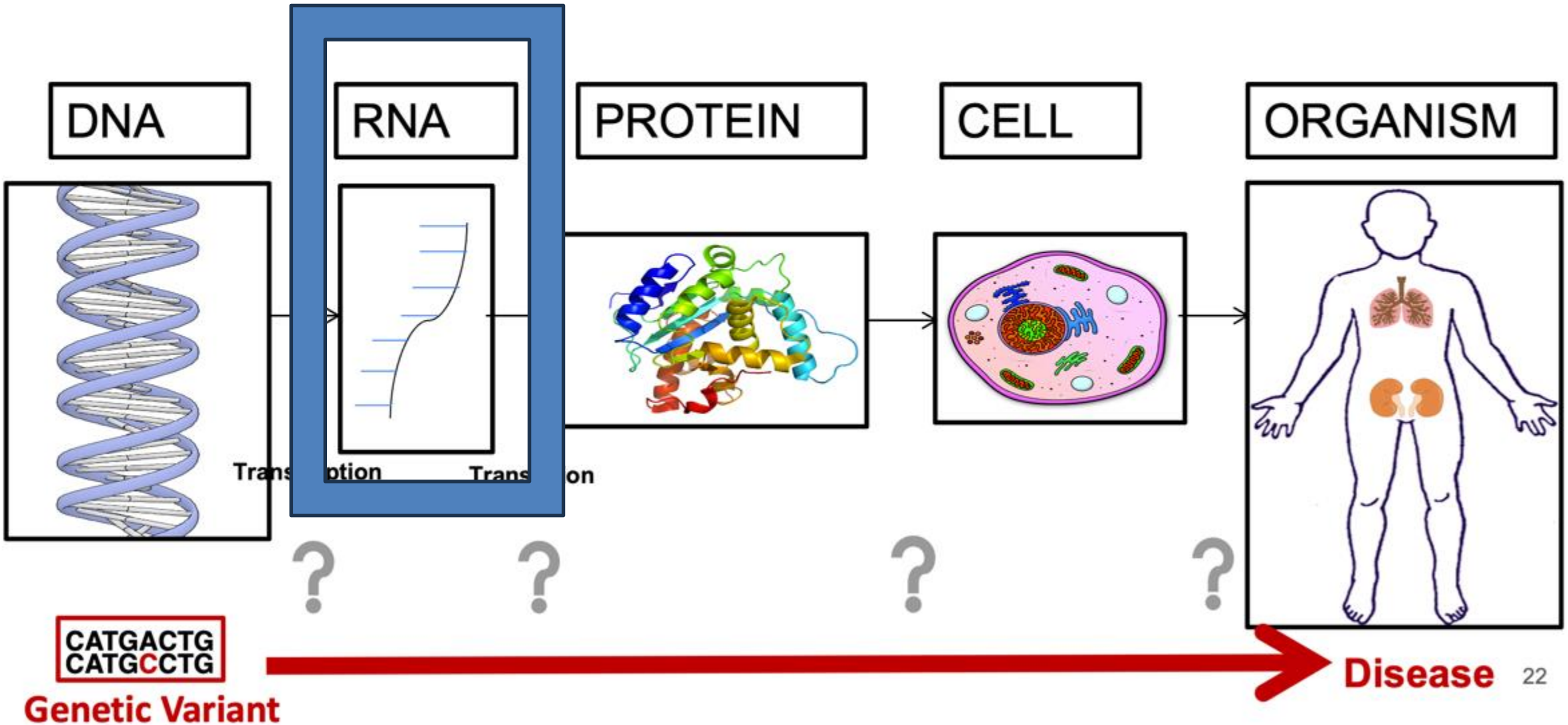
Sequence:
GIEASLLTDPKEVAGRTVDYIIAGGGLTGLTTAARLTENPDITVLVIESGSYES
DRGPIIEDLNAYGDIFGSSVDHAYETVCLATNNQTALIRSGNGLGGSTLVNGGT
WTRPHKAQVDSWETVFGNEGWNWDSVAAYSLQAERARAPNAKQIAAGHYFNASC
HGINGTVHAGPRDTGDDYSPIVKALMSAVEDRGVPTKKDLGCGDPHGVSMFPNT
LHEDQVRSDAAREWLLPNYQRPNLQVLTGQYVGKVLLSQNATTPRAVGVEFGTH
KGNTHNVYAKHEVLLAAGSAVSPTILEYSGIGMKSILEPLGIDTVVDLPVGLNL
QDQTTSTVRSRITSAGAGQGQAAWFATFNETFGDYTEKAHELLNTKLEQWAEEA
VARGGFHNTTALLIQYENYRDWIVKDNVAYSELFLDTAGEASFDVWDLLPFTRG
YVHILDKDPYLRHFAYDPQYFLNELDLLGQAAATQLARNISNSGAMQTYFAGET
IPGDNLAYDADLRAWTEYIPYNFRPNYHGVGTCSMMPKEMGGVVDNAARVYGVQ
GLRVIDGSIPPTQMSSHVMTVFYAMALKIADAVLADYASMQ

CellAgent

LLM-Driven Multi-Agent Framework for Natural Language–Based Single-Cell Analysis

ICLR 2026
Presented by Chuankai Xu

Biology in a Slide:



What Is Single-Cell RNA Sequencing?

Traditional RNA-seq → bulk average across millions of cells

scRNA-seq → gene expression of each individual cell

🎯 Analogy

Bulk RNA-seq is like blending a fruit salad and tasting the mix. scRNA-seq is tasting each fruit separately — you learn what makes each one unique.

Cell Heterogeneity

Reveals cell-to-cell differences hidden in bulk data

Spatial Transcriptomics

Maps gene expression to tissue locations

Massive Scale

Thousands to millions of cells per experiment

These technologies have produced massive datasets requiring advanced computational pipelines.

The Traditional scRNA-seq Pipeline

Every analysis follows a similar chain of steps — each requiring tool choices and parameter tuning.

1
QC & Filtering
Remove low-quality cells & doublets

→

2
Normali-zation
Log-normalize, scale gene counts

→

3
Feature Selection
Pick highly variable genes (HVGs)

→

4
Dim. Reduction
PCA → UMAP / t-SNE embedding

→

5
Clustering
Leiden / Louvain community detection

→

6
Annotation
Label cell types from marker genes

Downstream Tasks

Batch Correction | **Trajectory Inference** | **Spatial Domains** | **Gene Imputation** | **DEG Analysis**

⚠ **Each step needs tool selection + parameter tuning + biology expertise + Python coding**

Why Is This Hard?

Scenario

A biology PhD has 50,000 single cells from a cancer biopsy.

She wants to find tumor subtypes, trace differentiation trajectories, and identify key marker genes.

She needs dual expertise in computational programming AND fundamental biology.

Pick the Right Tool

Scanpy? Seurat? scVI? scGPT?
Dozens of options per task

Write Complex Code

Custom Python pipelines,
debugging data formats

Tune Parameters

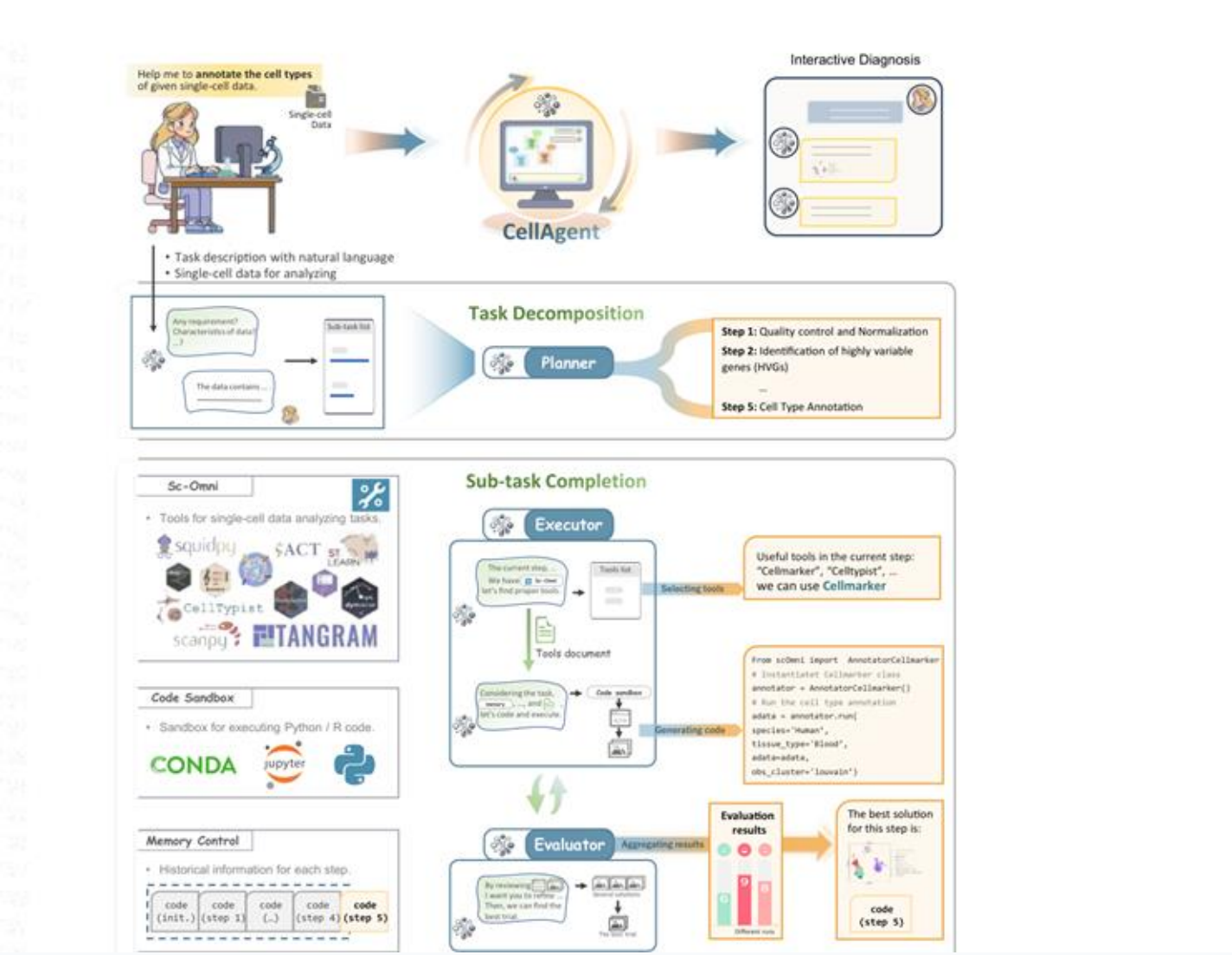
Resolution, n_neighbors,
HVGs…
each dataset is different

Iterate Manually

Run → inspect → adjust →
repeat
for every single substep

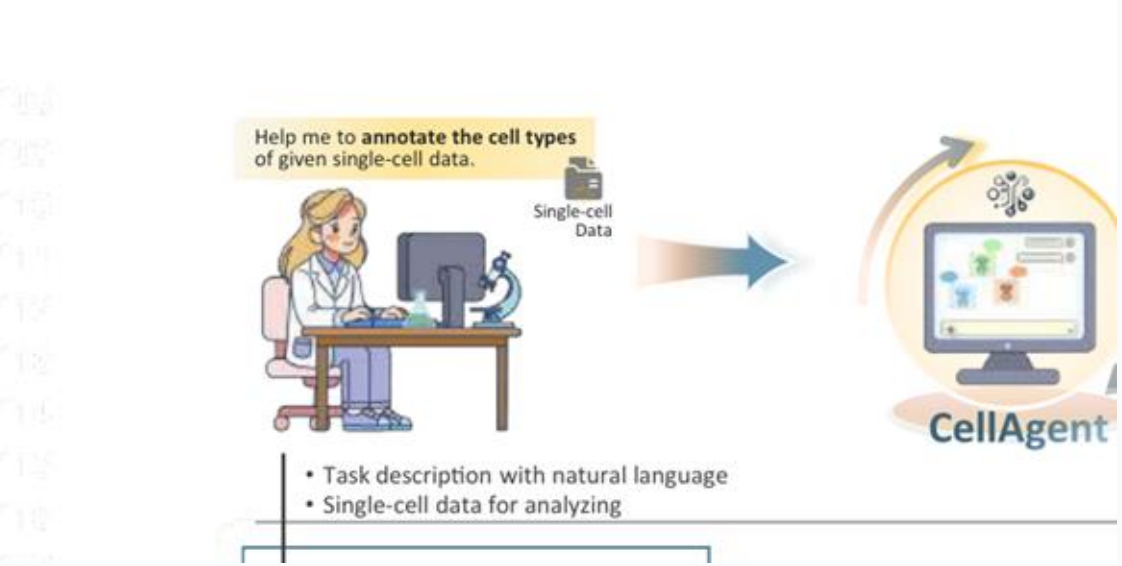
"Can we just tell the computer what we want in plain English?"

CellAgent: The Full Picture



Three agents collaborate: Planner → Executor → Evaluator. Let's walk through each.

① User Input



Two inputs, that's it:

1. Natural language request

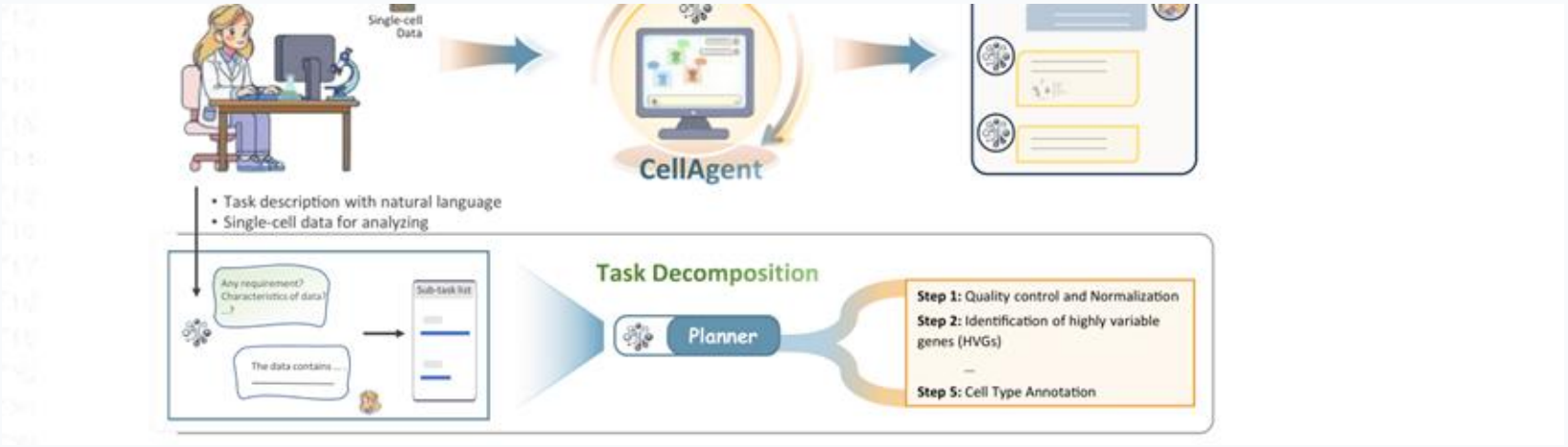
"Perform batch correction and annotate cell types"

2. Your .h5ad data file

The standard AnnData format for scRNA-seq

No coding. No tool selection. Just describe your analysis goal.

② Planner Agent



Think of it like your PI…

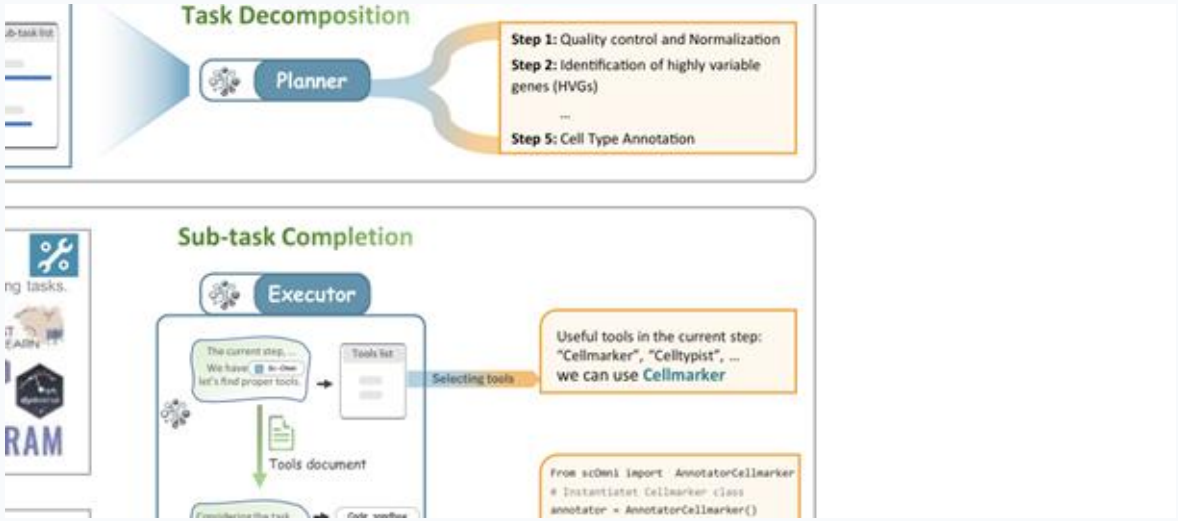
The Planner inspects your dataset summary, then decomposes your request into a sequence of ordered subtasks — like a PI writing an analysis plan before handing it off.

Generated Plan

```
Step 1: Quality control & filtering

Step 2: Normalization & HVG selection

Step 3: Batch correction (scVI)

Step 4: Cell type annotation

Step 5: Visualization & export
```

Expert knowledge is baked into the Planner's system prompt: standard order of operations, parameter ranges, task dependencies.

③ Executor Agent



Tool Selector

Queries sc-Omni toolkit, retrieves the right functions for the current subtask

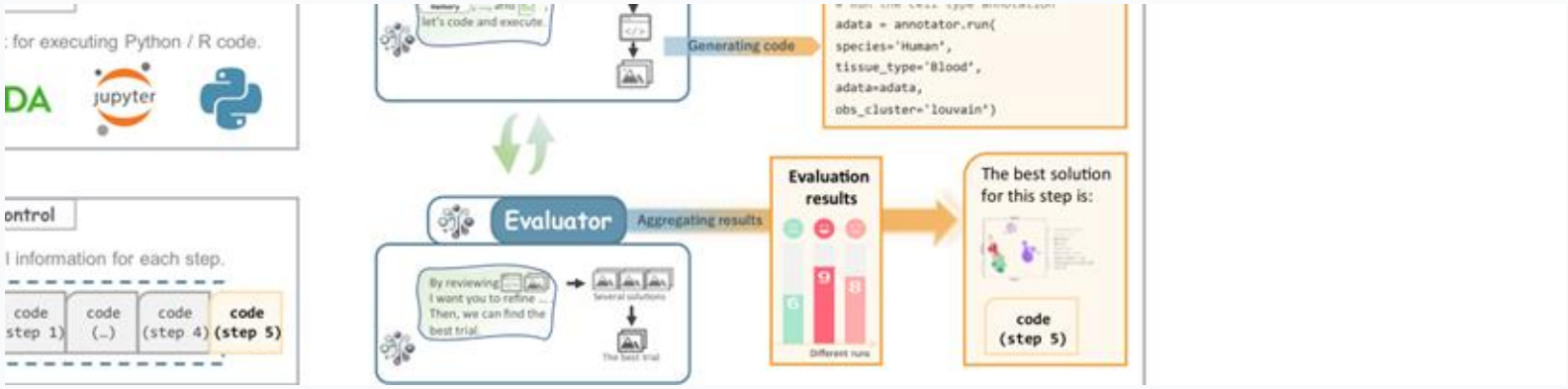
Code Programmer

Generates Python code using tool docstrings, executes in sandboxed Jupyter notebook

🔬 Like a junior RA — does the lab work, self-corrects on errors, retries until code executes successfully.

If code fails, the Executor reads the error, fixes it, and retries — all within a local memory workspace.

④ Evaluator Agent



Self-Reflective Optimization Mechanism

Quantitative Metrics

iLISI for batch correction, Accuracy Score for imputation, ARI for clustering…

Visual Assessment

GPT-4o analyzes UMAP plots, trajectory continuity, spatial domain coherence

Anonymized Judging

Evaluator sees only outputs + metrics — never tool names or prompts

🎓 *Like a senior postdoc — runs multiple algorithms, scores each, picks the best result.*

Under the Hood

sc-Omni Toolkit

Expert-curated library covering 15+ analytical tasks: QC, normalization, batch correction, trajectory, spatial analysis. Each tool has standardized docstrings for code generation.

Code Sandbox

Jupyter-based isolated execution. Decouples code running from CellAgent core for safety and reproducibility. Every run is traceable.

Memory Control

Global memory stores only final code of each step (high info-entropy). Local memory tracks trial-and-error within a substep, then discards — keeping context clean.

How Does CellAgent Compare?

	CellAgent	GPT-4	AutoBA	Web Servers	Scanpy
Natural Language Input	✓	✓	✓	✗	✗
Auto Tool Selection	✓	✗	✓	Partial	✗
Self-Evaluating	✓	✗	✗	✗	✗
No Coding Required	✓	Partial	✓	✓	✗
Domain-Optimized	✓	✗	Partial	✓	✓
Multi-Task Pipeline	✓	✗	✗	✗	✓

CellAgent is the only framework that combines NL interaction, domain tools, AND self-evaluation.

Evaluation: Batch Correction

Remove technical variation between experiment batches while preserving biological signals.

10 Metrics (Weighted Sum)

Batch Removal:

NMI, ARI, ASW (batch), PCR, Graph Connectivity

Bio Conservation:

ASW (cell type), iLISI, cLISI, Isolated Labels, F1

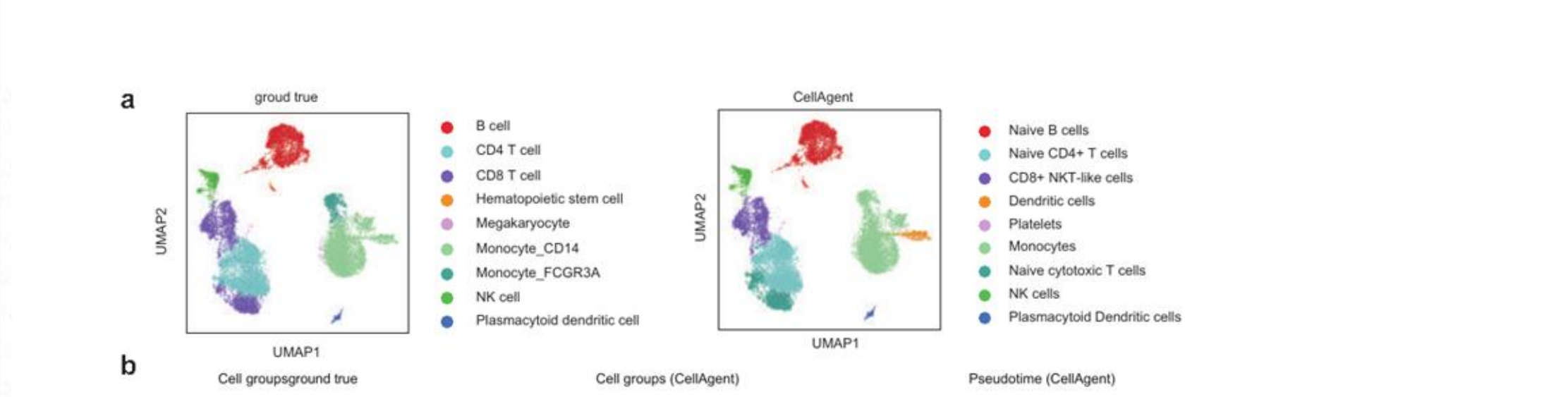
Overall Score = 0.4 × Batch + 0.6 × Bio

Results (5 datasets)

Method	Batch	Bio	Overall
CellAgent	0.67	0.66	**0.67**
scVI	0.66	0.65	0.66
Liger	0.65	0.64	0.65
Harmony	0.60	0.61	0.60
Combat	0.56	0.58	0.57
scGPT	0.50	0.62	0.57

💡 CellAgent doesn't just pick one tool — it tries multiple algorithms (scVI, Harmony, Combat…), evaluates each with all 10 metrics, and selects the winner. That's why it consistently leads.

Evaluation: Cell-Type Annotation



Metric: Consistency Score

Using Cell Ontology (CL), predictions scored as:

Fully match = 1.0 | Partially match = 0.5 | Mismatch = 0.0

Averaged per dataset across 6 tissues & technologies.

85%

CellAgent average

vs. scGPT 77%, ScType 59%

Refined CD8+ T → naive cytotoxic T

Ground truth (left) vs. CellAgent predictions (right) on human PBMC dataset.

Evaluation: Trajectory Inference

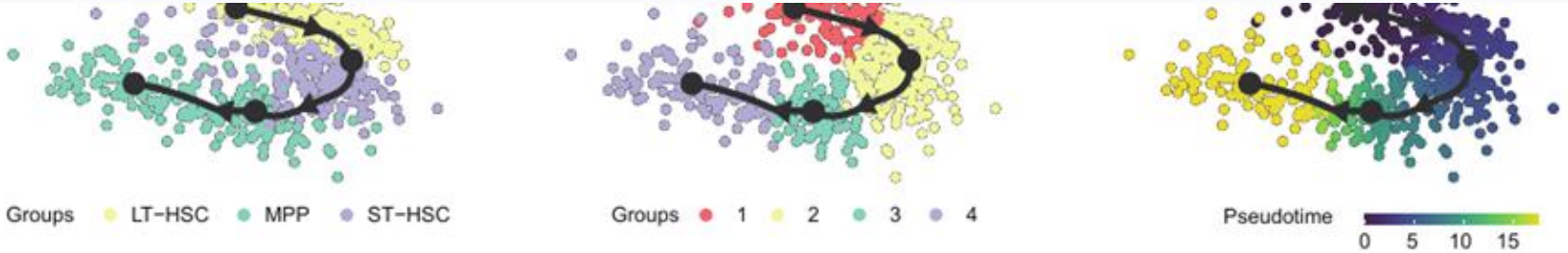


Figure 3: **Results of CellAgent in single-cell transcriptomics analysis. a,** UMAP visualization of

Cor. Dist. — correlation of geodesic distances

4 Trajectory Metrics

F1 Branches — correct branch topology

Wcor Features — gene expression along trajectory

Edge Flip — edge direction accuracy

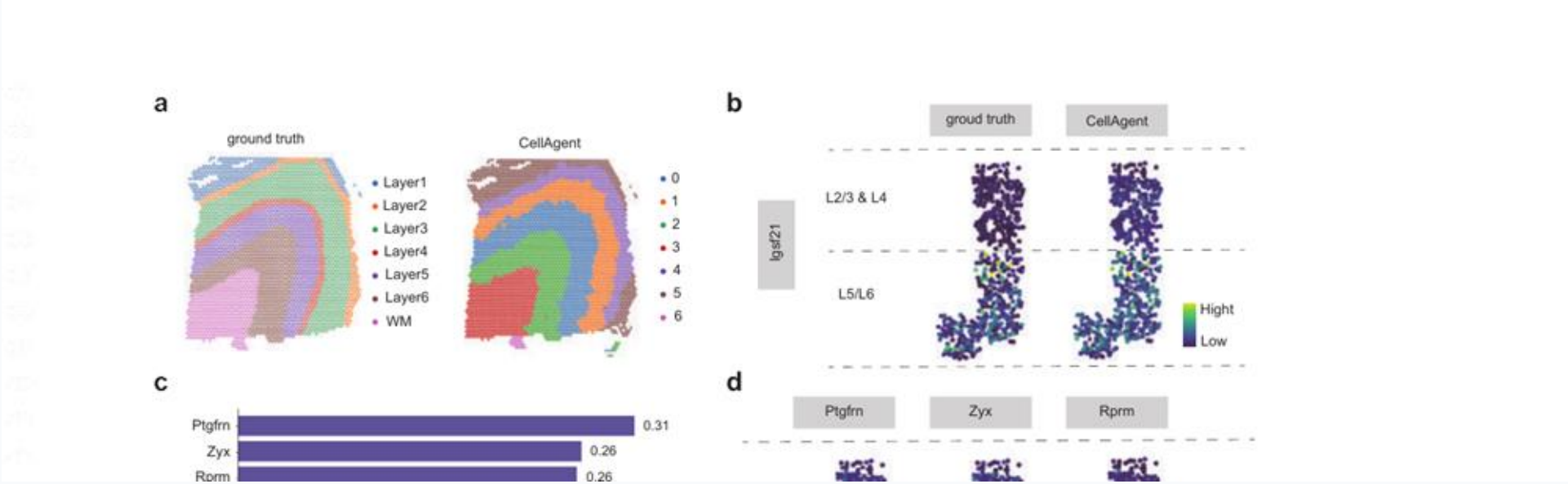
Overall = weighted average across 8 datasets with gold-standard trajectory information

Results (8 datasets)

Method	Overall
CellAgent	**0.50**
Slingshot	0.48
SCORPIUS	0.46
PAGA tree	0.45
PAGA	0.39
Monocle DDRTree	0.36

LT-HSC → ST-HSC → MPP trajectory successfully reconstructed, consistent with known hematopoietic differentiation.

Evaluation: Spatial Transcriptomics



Spatial Domain Identification

Metric: ARI (Adjusted Rand Index)

12 DLPFC slices — cortical layers L1–L6 + WM

0.47

ARI — Rank #1

Spatial Imputation

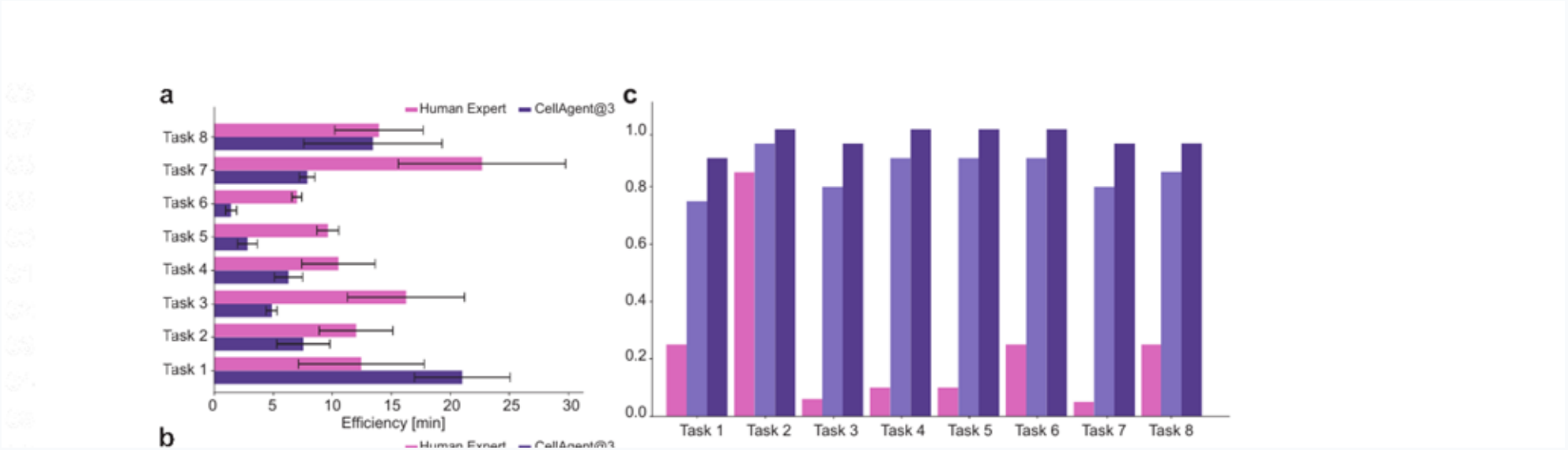
Metrics: PCC, RMSE, SSIM, JS → Accuracy Score

7 datasets (10X Visium, SlideseqV2, seqFISH)

0.88

AS — Rank #1 (+17%)

Human-Centered Evaluation



8 min

vs. 13 min (experts)

Avg. Completion Time

96%

vs. 24% (GPT-4 alone)

Execution Success Rate

75%

vs. 30% (web servers)

Rated "Highly Facilitative"

Takeaway

Raw scRNA-seq data → publication-quality results in plain English.

- Multi-agent "deep-thinking" pipeline: Planner → Executor → Evaluator
- Self-reflective optimization with task-specific metrics replaces manual assessment
- Competitive with human experts across 60+ datasets, 8 major tasks
- Lowers the barrier: no coding, no tool selection — just natural language

Thank you — Questions?

Next Section 3.4 (three papers)

- 1. Evaluating large language models and agents in healthcare: key challenges in clinical applications
- **2. Holistic evaluation of large language models for medical tasks with MedHELM**
- 3. Tahoe-100M: A Giga-Scale Single-Cell Perturbation Atlas for Context-Dependent Gene Function and Cellular Modeling

THANK YOU