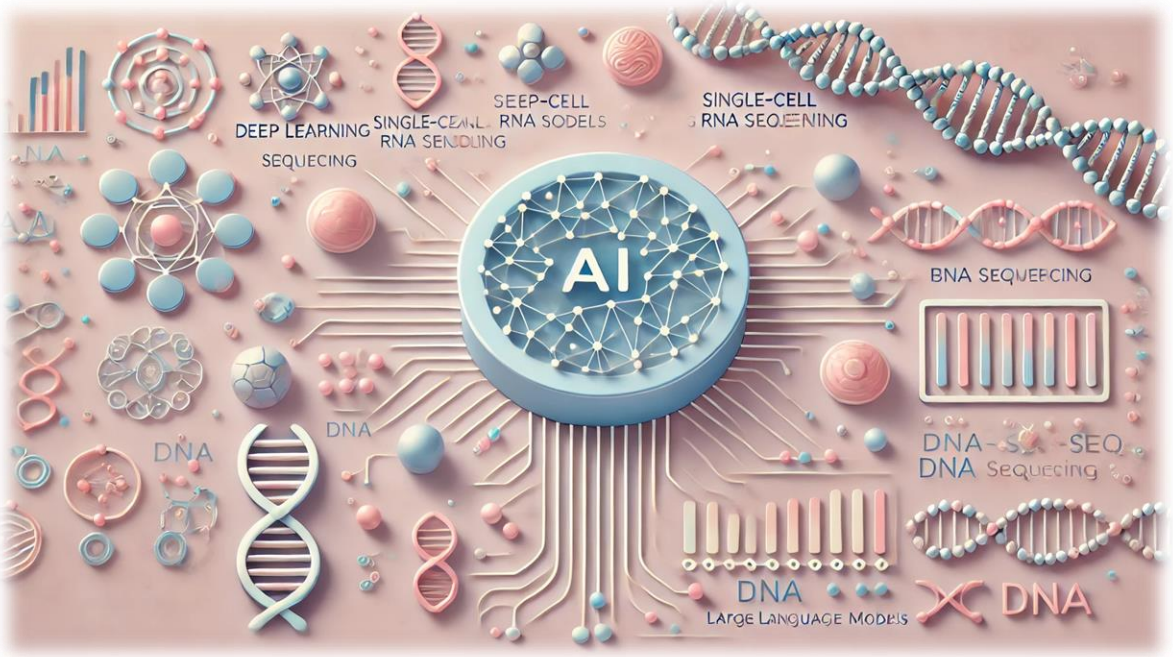
Applications of AI in Bioinformatics: Tools and Practical Implementation

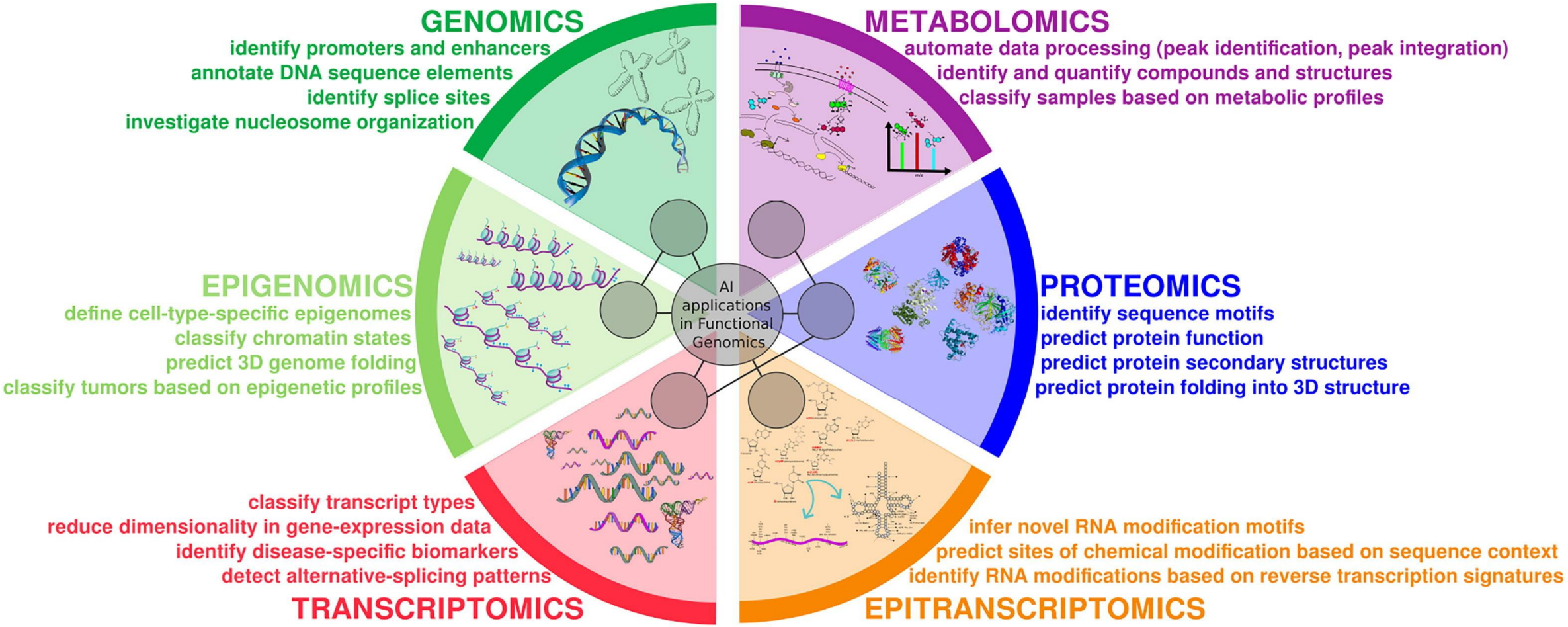


Chia-Jung Chang (Charlene)
2025/02/12

Keloid & Genodermatoses Lab(**KGD Lab**),
Department of Dermatology, College of Medicine, National Cheng Kung University

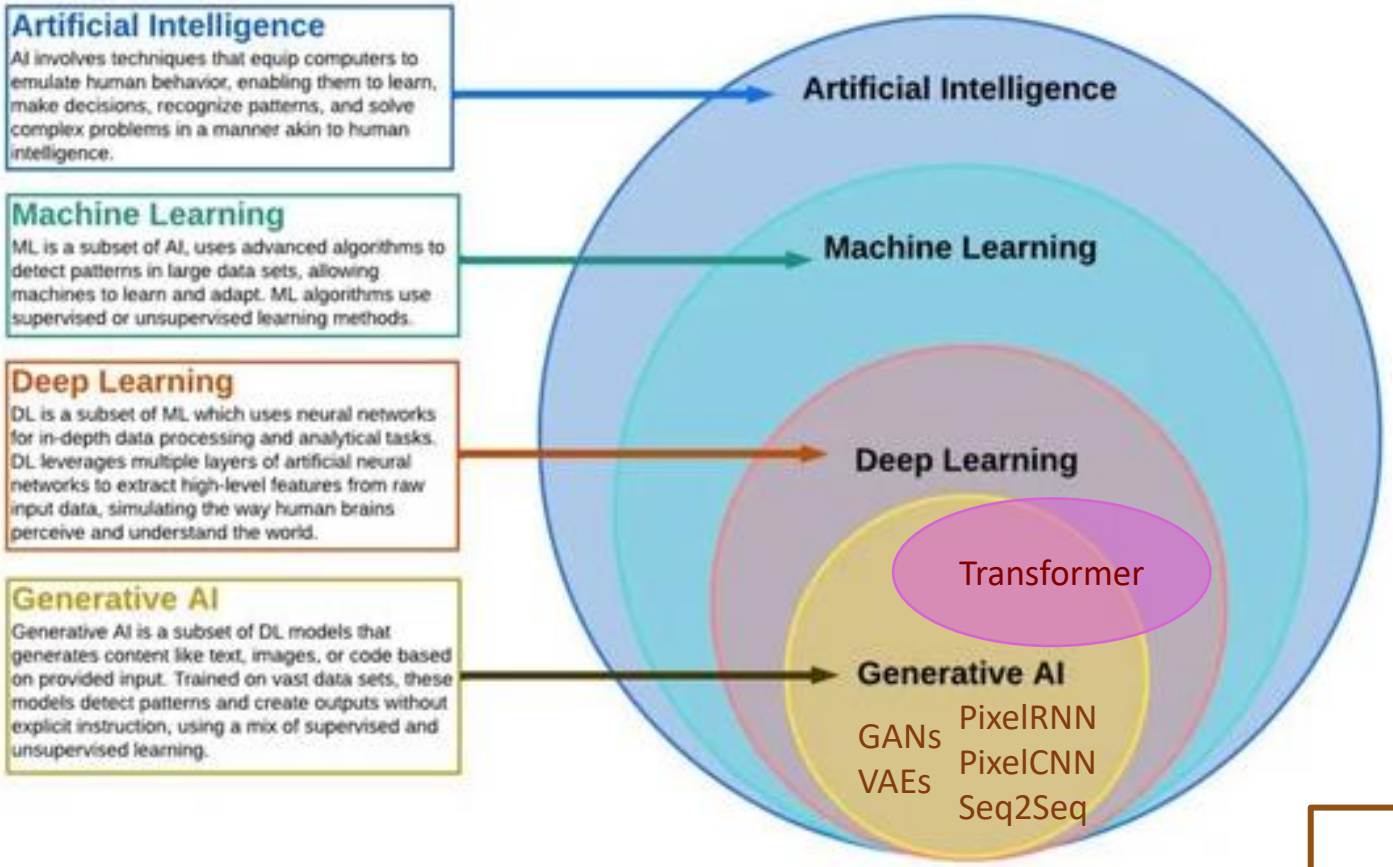
Artificial intelligence (AI) applications in functional genomics

AI, particularly deep learning and LLMs, is transforming bioinformatics by boosting analytical speed, accuracy, and scientific breakthroughs in research and healthcare.

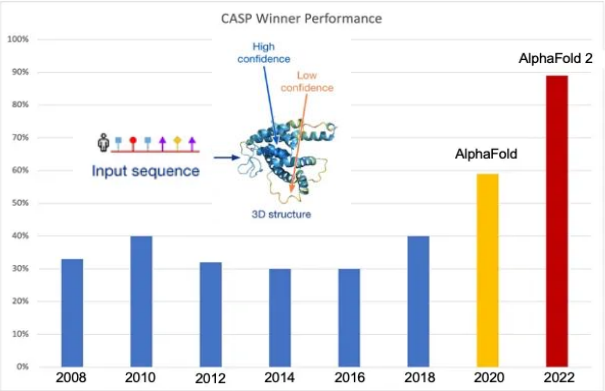


Computational and Structural Biotechnology Journal 19 (2021): 5762-5790.

Artificial Intelligence



Sustainability 15.18 (2023): 13484.



Nature. 2021 Aug;596(7873):583-589.

Transformer

- BERT (Bidirectional Encoder Representations from Transformers)
- GPT (Generative Pretrained Transformer)
- Transformer-XL
- T5 (Text-To-Text Transfer Transformer)

Generative AI

- GANs: Generative Adversarial Networks
- VAEs: Variational Autoencoders
- Autoregressive Models (PixelRNN, PixelCNN)
- Seq2Seq: Sequence to Sequence Models

Applications of transformers in bioinformatics

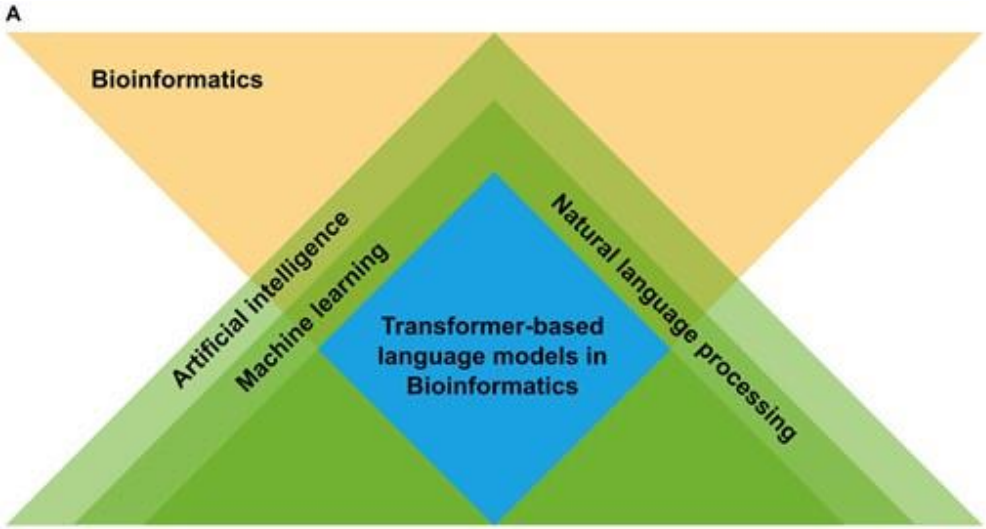
JOURNAL ARTICLE

Applications of transformer-based language models in bioinformatics: a survey

Save Chat with paper

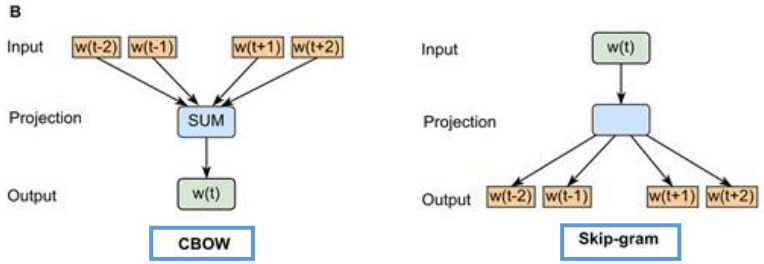
Shuang Zhang, Rui Fan, Yuti Liu, Shuang Chen, Qiao Liu, Wanwen Zeng
Author Notes

Bioinformatics Advances, Volume 3, Issue 1, 2023, vbad001,
https://doi.org/10.1093/bioadv/vbad001
Published: 11 January 2023 **Article history** ▾

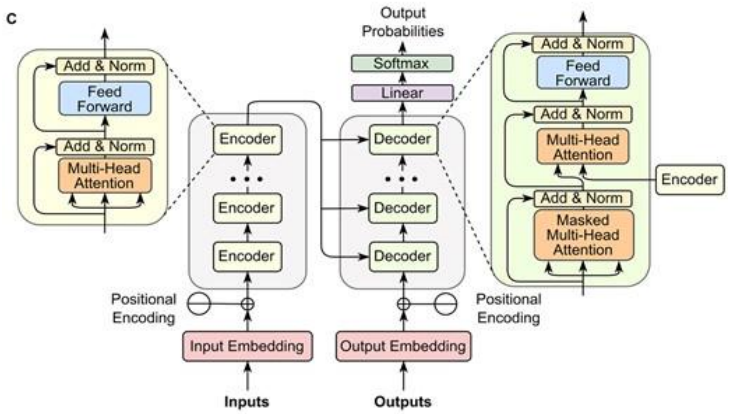


Bioinform Adv. 2023 Jan 11;3(1):vbad001.

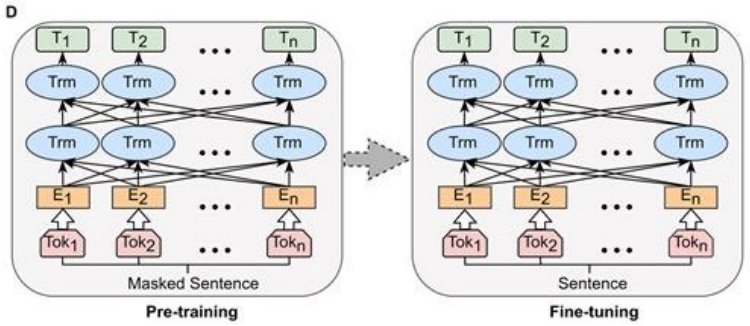
Two common models in Word2Vec



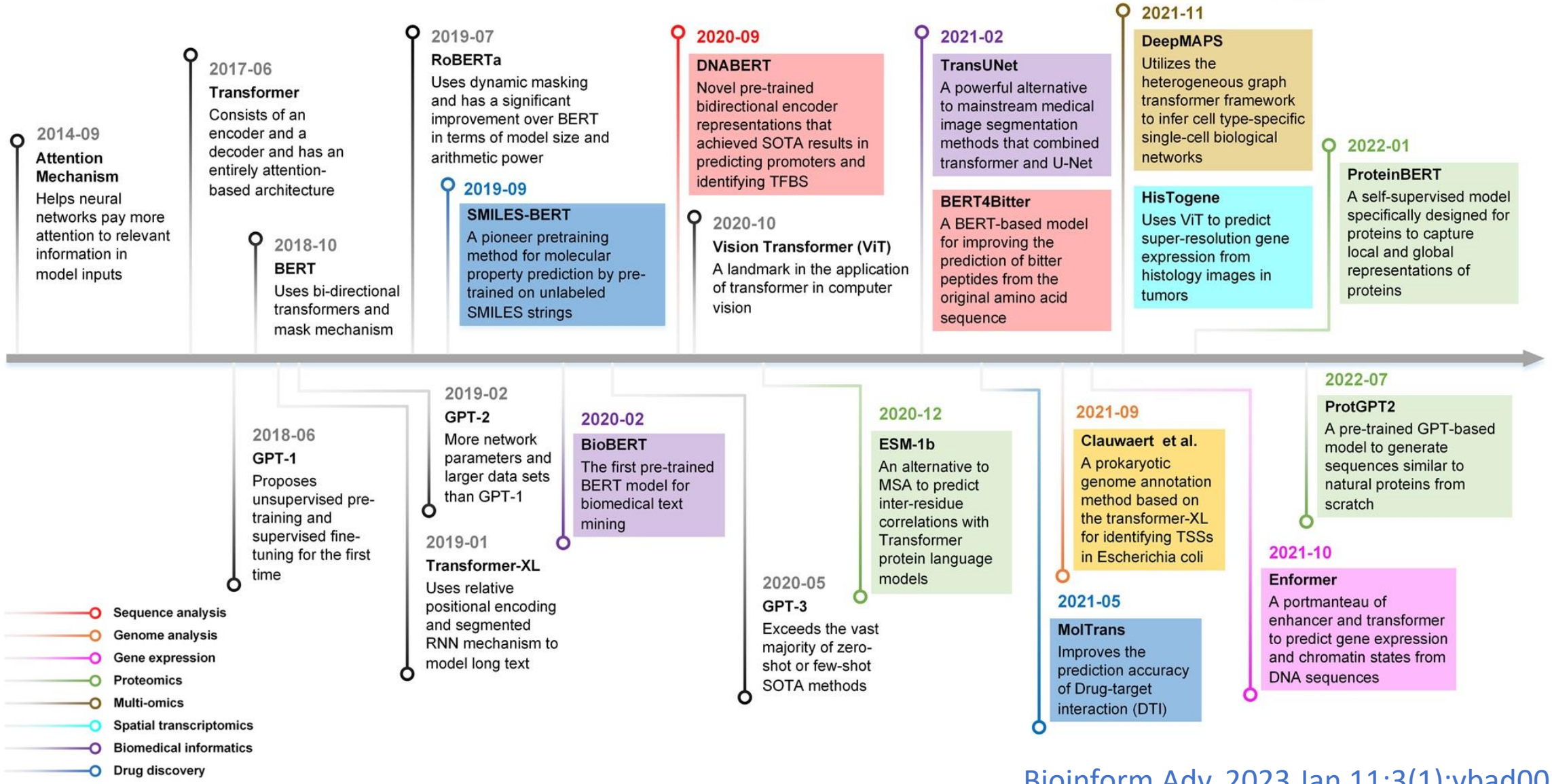
The structure of transformer model



The structure of BERT model



Applications of transformers in bioinformatics



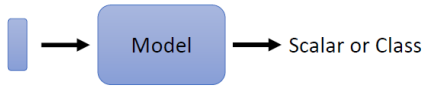
Bioinform Adv. 2023 Jan 11;3(1):vbad001.

Self-Attention

The name "Transformer" is derived from the model's core mechanism, the self-attention mechanism, which "transforms" the representation of input data.

Input

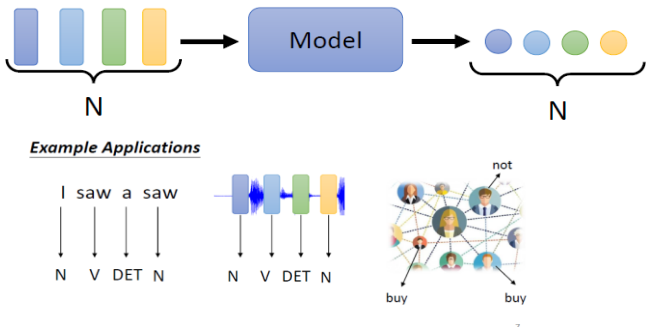
- Input is **a vector**



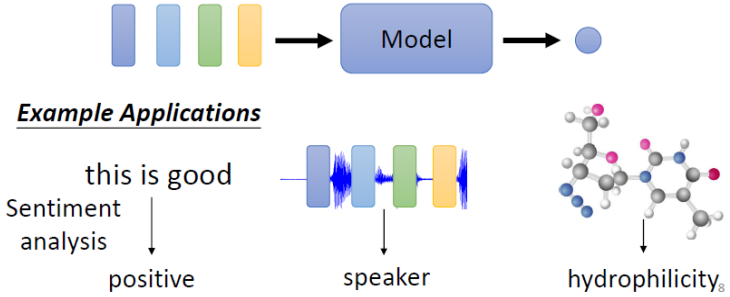
- Input is **a set of vectors**

Output

- Each vector has a label.



- The whole sequence has a label.



Long-Distance Dependencies

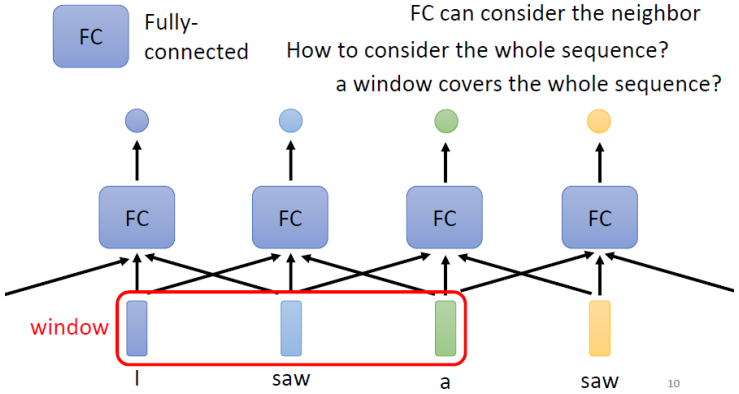
Sequence Labeling

Is it possible to consider the context?

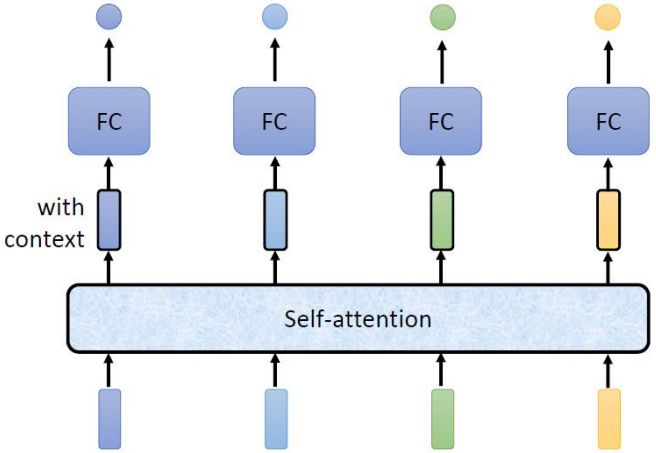
FC can consider the neighbor

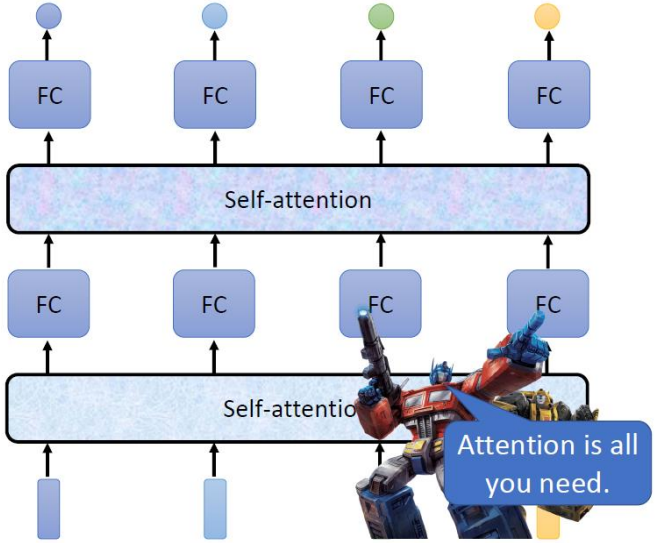
How to consider the whole sequence?

a window covers the whole sequence?



Self-attention

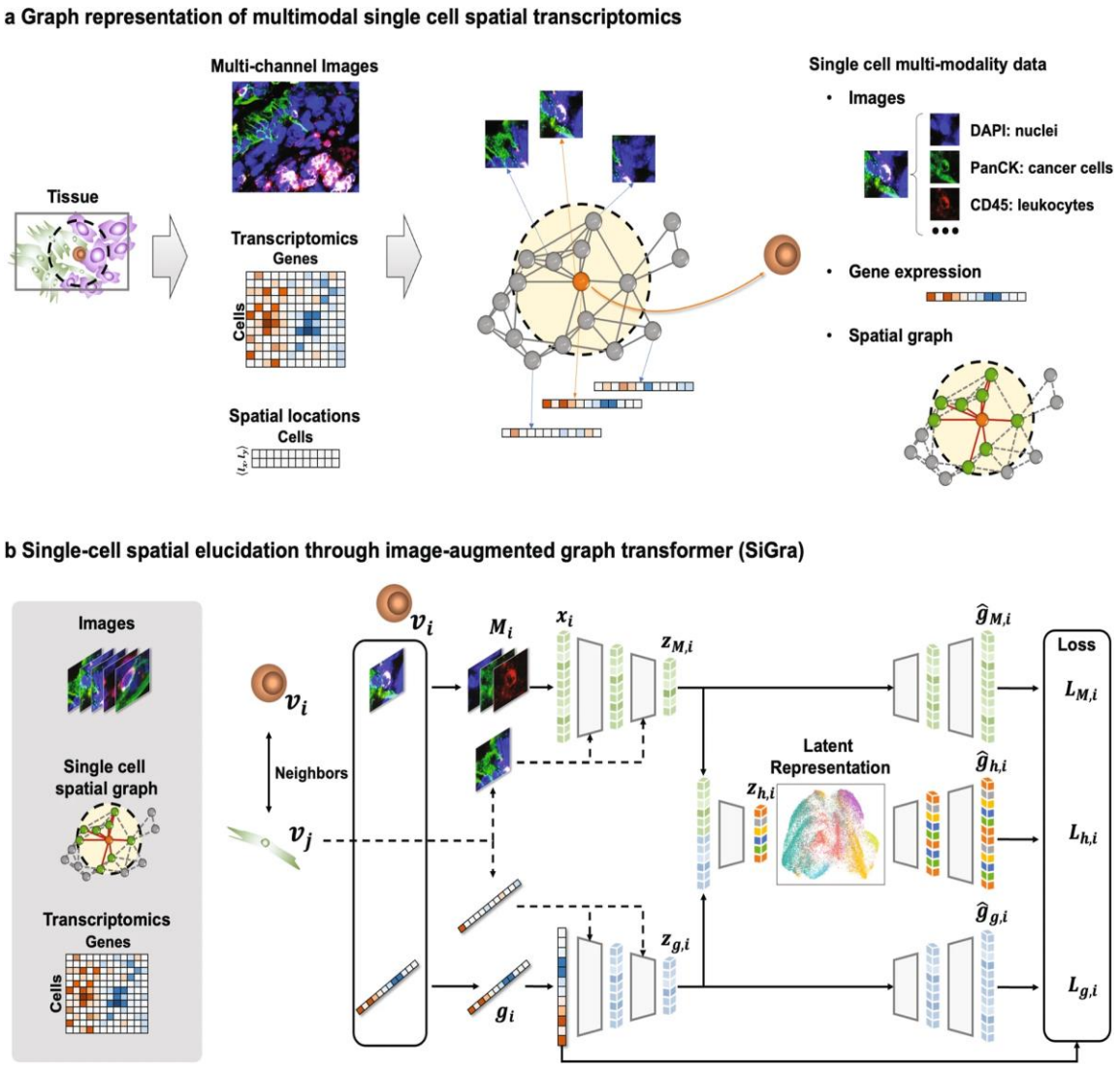


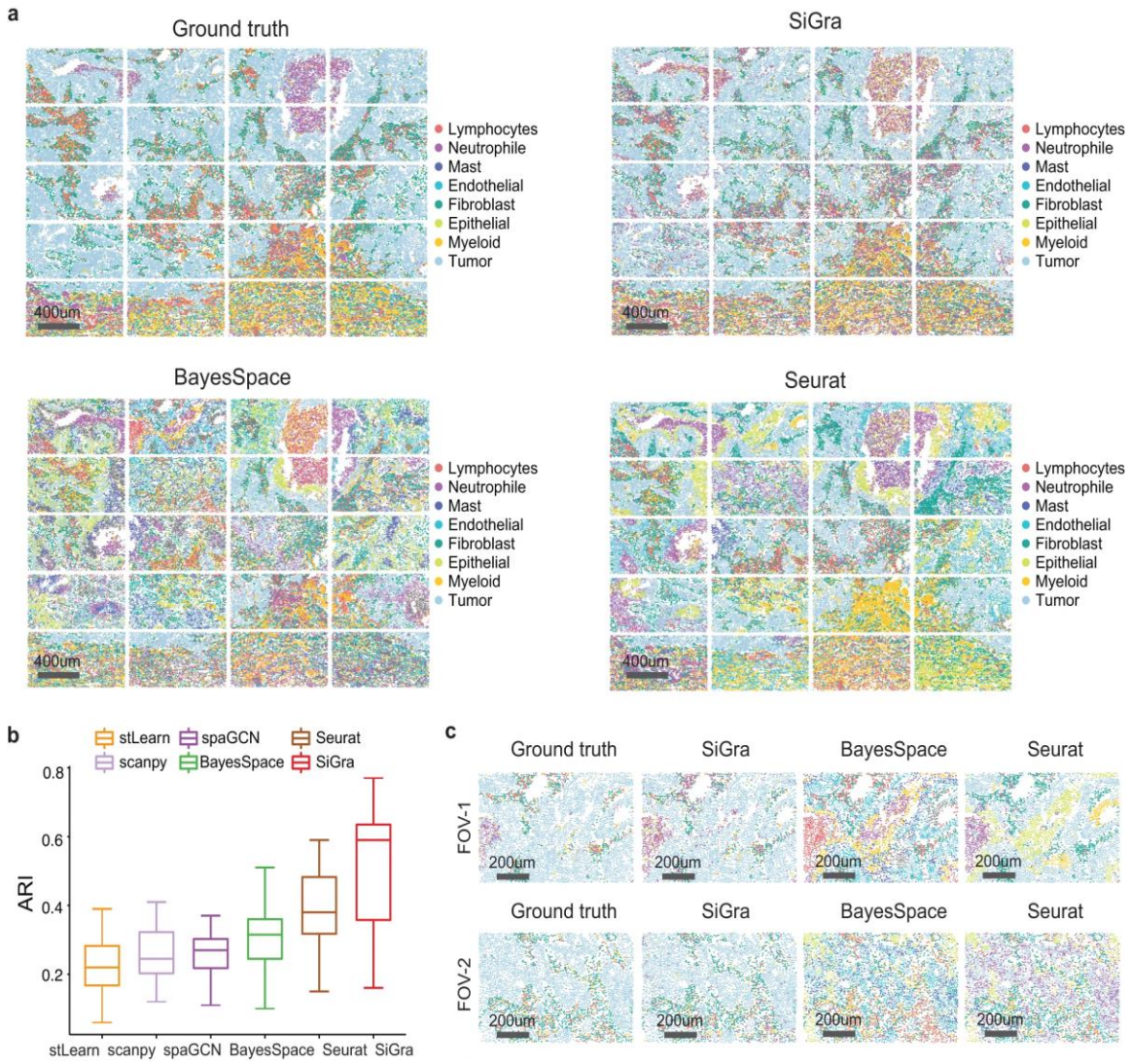


https://arxiv.org/abs/1706.03762

台大電機 Hung-yi Lee (李宏毅)教授 https://www.youtube.com/watch?v=hYdO9CscNes

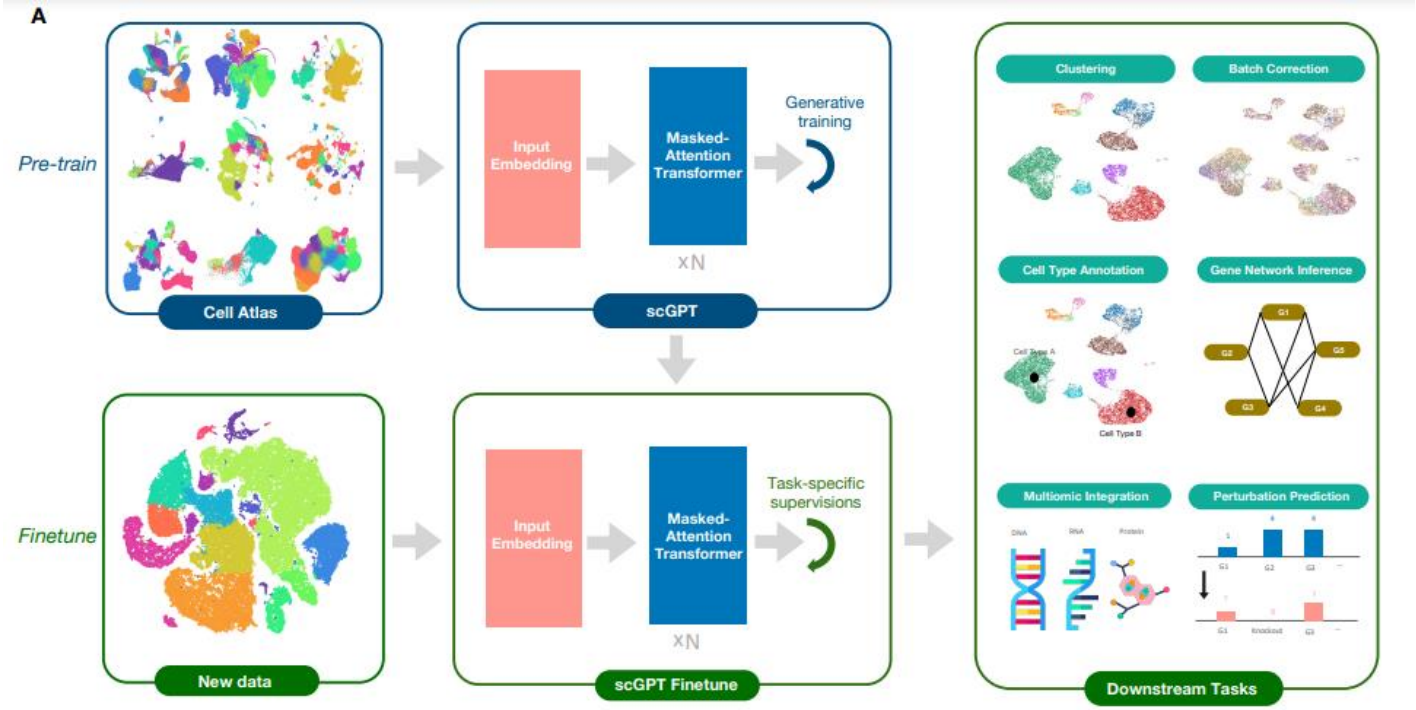
SiGra: single-cell spatial elucidation through an image-augmented graph transformer

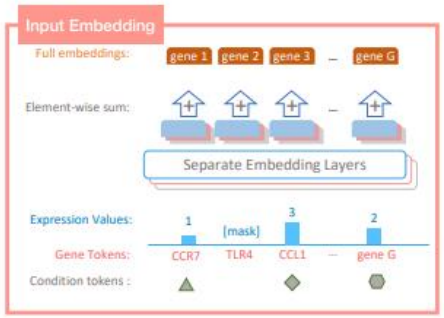


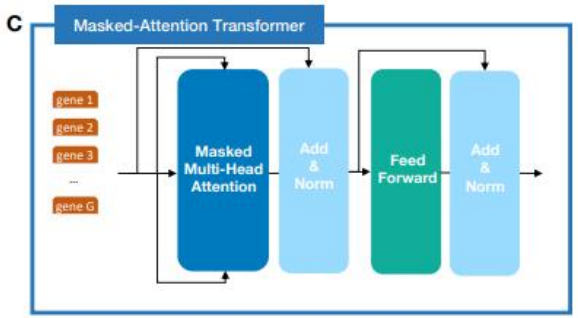


Nat Commun. 2023 Sep 12;14(1):5618.

scGPT: Towards Building a Foundation Model for Single-Cell Multi-omics Using Generative AI

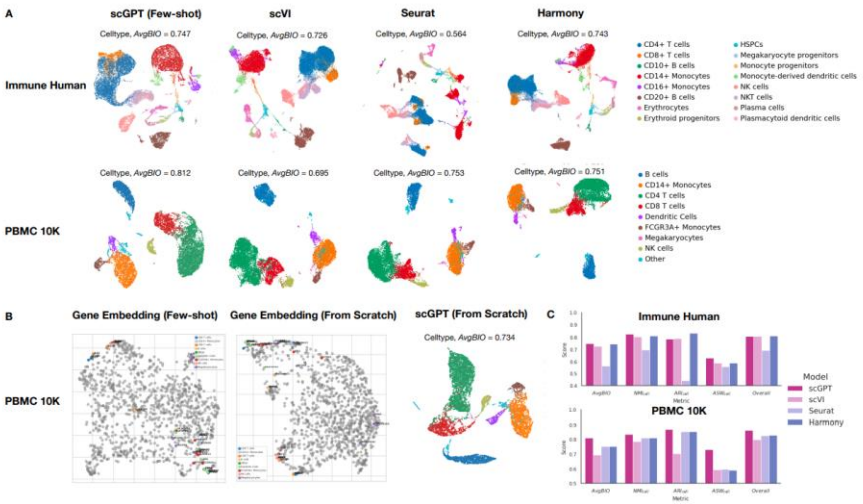




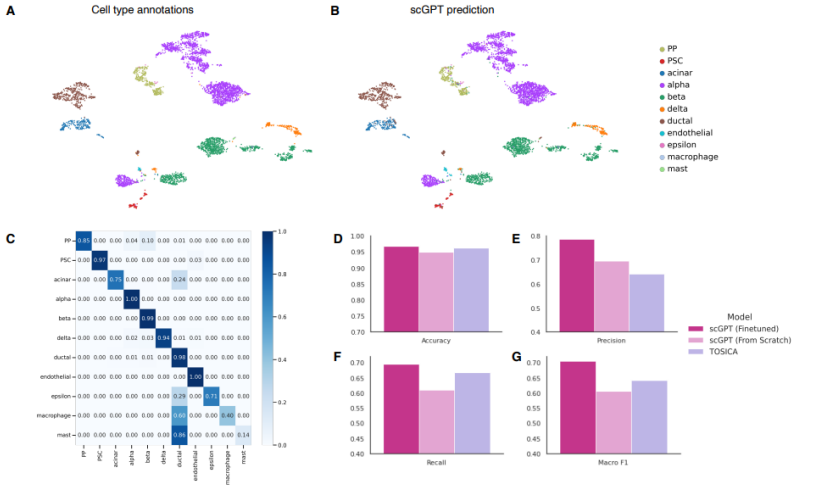


- scMulti-omic data (e.g., gene expression, chromatin accessibility, and protein abundance)

Batch correction



Cell type annotations



Nat Methods. 2024 Feb 26

Applications of transformer in single cell data

Model	Open Source?	Batch Effect Correction	Cell-type Annotation	Multi-omics Data Integration	Imputation	Gene Function Prediction	Perturbation Prediction	Gene Network Analysis	Simulation
scGPT Benchmark	✓	✓	✓	✓	✓	✓	✓	✓	✓
scGPT	✓	✓	✓	✓			✓	✓	
Geneformer	✓		✓			✓			
scBERT	✓		✓						
CellLM	✓		✓						
tGPT	✓	✓							
SCimilarity		✓	✓						
scFoundation			✓						

https://www.biorxiv.org/content/10.1101/2023.09.08.555192v1

awesome-deep-learning-single-cell-papers

https://github.com/OmicsML/awesome-deep-learning-single-cell-papers?tab=readme-ov-file#gene-regulatory-network

foundation-model-single-cell-papers

https://github.com/OmicsML/awesome-foundation-model-single-cell-papers

awesome-bio-chatgpt

https://github.com/OmicsML/awesome-bio-chatgpt

Advantages of Transformers

Why Transformers are suitable for use in Multi-omics research?

- **Flexible Input Handling:** The Transformer's self-attention mechanism allows it to process sequences of varying lengths and types, including text and images, offering high input flexibility.
- **Effective Long-Distance Dependency Management:** Particularly adept at handling long sequences, the Transformer captures long-distance dependencies in text and other data types.
- **Comprehensive Contextual Information:** While different from traditional RNNs' long-term memory mechanism, the Transformer comprehensively captures the context of entire sequences through its self-attention mechanism.
- **Efficient Parallel Computation:** Its structure enables efficient parallel processing, making it highly effective for large-scale data tasks.

Overall, the Transformer is a versatile and powerful tool for processing various types of sequential data like text and images.

Generative AI Applications in Bioinformatics

Three Levels of Development

Pre-Train: The Foundation Training Stage

- To build a robust foundational model that understands and processes fundamental concepts and data in bioinformatics.

Fine-tune: The Refinement Stage

- Adjust the pre-trained model to better suit specific bioinformatics tasks, such as gene expression prediction or protein folding.

Prompt Engineering: The Query Design Stage

- By designing carefully crafted prompts, guide the model to produce the most useful information.

scBERT , tGPT , CellLM , scGPT, scFoundation , Scimilarity, Geneformer

SpaFormer, scMoFormer, TOSICA, scTransSort, STGRNS, CIForm

GPTCelltype
[PMID: 38528186] Nat Methods. 2024 Mar 25.
doi: 10.1038/s41592-024-02235-4.

Cost

Applications of ChatGPT and AI Tools in Bioinformatics Analysis

Knowledge Retrieval and Integration
AI tools can quickly extracts key points from literature and online sources, offering tool recommendations and saving time on data gathering.

Interactive Problem-Solving
When faced with technical hurdles or interpretation issues, researchers can consult ChatGPT for insights and best-practice guidance, speeding up resolution.

Code Generation and Debugging
In workflows like scRNA-seq (e.g., data loading, QC, differential expression), an AI assistant can generate or refine code, boosting development efficiency.

Method Selection and Parameter Tuning
AI tools can recommend suitable methods and parameter settings based on research objectives and data characteristics, thus reducing time spent on repeated trial and error.

Prompt Engineering

- A **prompt** refers to the input text or instructions provided to a generative AI model (e.g., ChatGPT).
- **Prompt engineering** is the process of designing, refining, and iterating on these instructions to optimize the model's output.
- **Key Concepts:**

1. Clearly define requirements in natural language.
2. Leverage the model's reasoning capabilities to accurately obtain the desired results.

Why Is Prompt Engineering Important?

- **Improves Output Quality:**
  Even with the same model, different prompt designs can lead to significantly different results.
- **Reduces Development and Experimentation Costs:**
  Optimizing prompt design helps minimize trial-and-error attempts.
- **Broad Range of Applications:**
  Useful for text generation, coding assistance, data analysis, customer support, and more.
- **Lowers Technical Barriers:**
  High-quality outputs can be achieved without deep knowledge of AI model architectures through prompt engineering.

Prompt Engineering

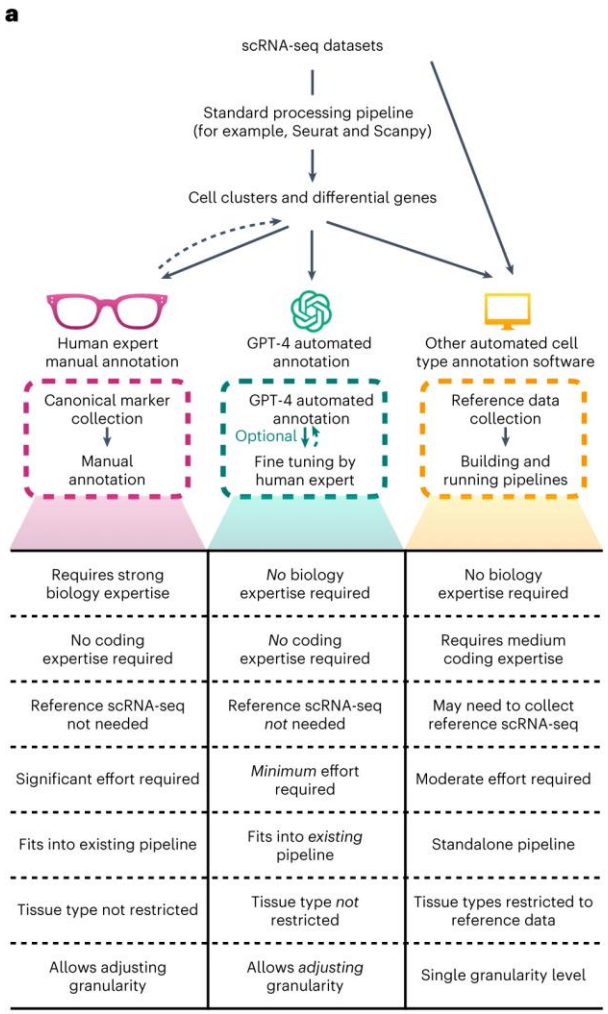
Core Principles

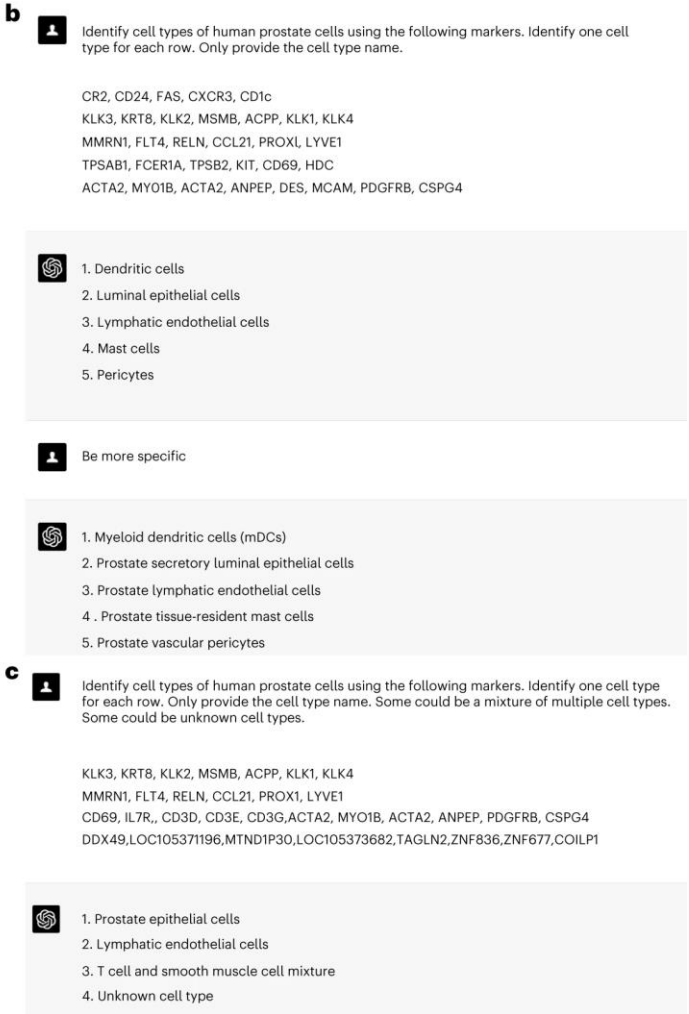
- **Clarity**
  Use specific, concise, and precise language to describe the requirements.
- **Context and Key Information**
  Provide essential background details to give the model a more comprehensive understanding.
- **Structure**
  Employ paragraphs, lists, or keywords to organize the prompt effectively.
- **Iterative Process**
  If the output is suboptimal, repeatedly adjust key phrases or the prompt format.

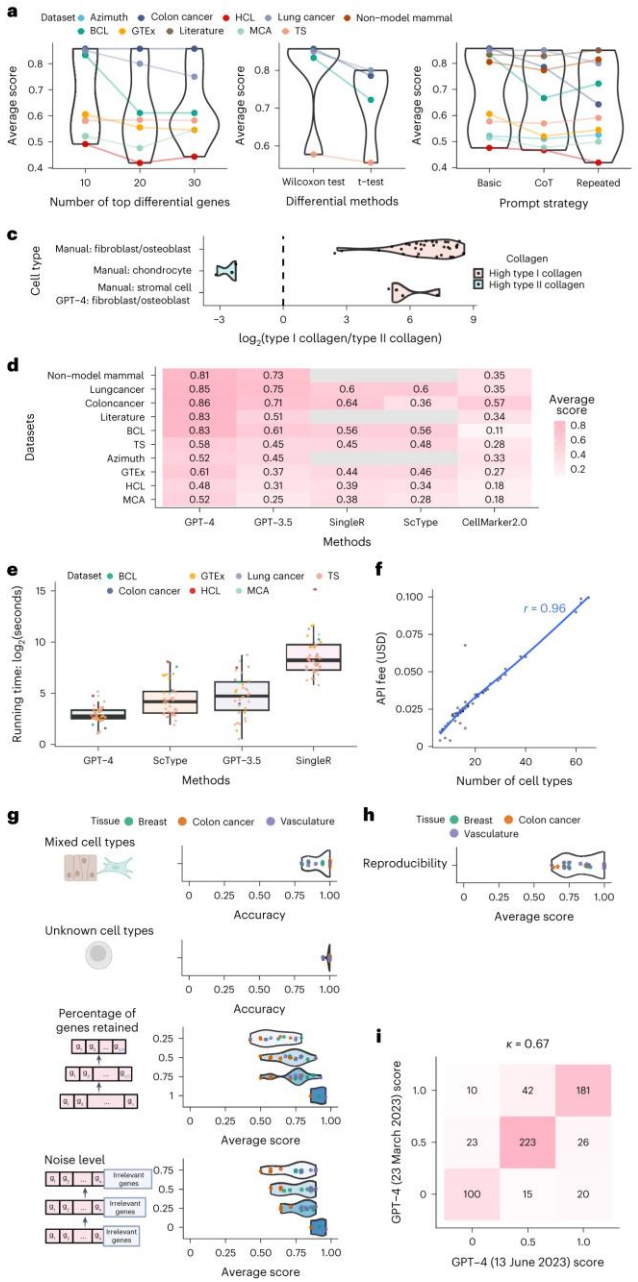
Common Techniques

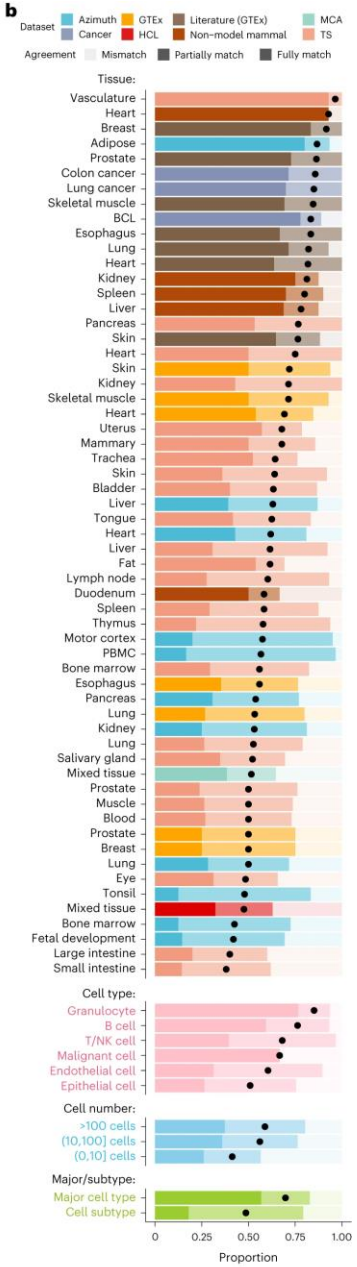
- **Step-by-Step Decomposition**
  Break down complex tasks into multiple steps or instructions so the model can tackle them incrementally.
- **Example-Based Prompting**
  Provide sample prompts to illustrate the desired format and style.
- **Constraints and Criteria**
  Clearly define what to exclude and the rules the output must follow.
- **Utilizing System Messages or Role Assignments**
  In a conversational model, define system messages to specify the AI's role or area of expertise.

Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis









GPT-4 can accurately annotate cell types in scRNA-seq analysis based on marker genes, achieving high consistency with expert annotations across diverse tissues.

Hou, Wenpin, and Zhicheng Ji. Nature Methods (2024): 1-4.

ChatGPT Custom Version

Explore and create a customized ChatGPT version that integrates instructions, additional knowledge bases, and any combination of skill sets.

搜尋 GPT

熱門精選 寫作 生產力 研究與分析 教育 日常生活 程式設計

Research & Analysis

Find, evaluate, interpret, and visualize information

1 **Scholar GPT**
Enhance research with 200M+ resources and built-in critical reading skills. Access Google Scholar, PubMed, bioRxiv, arXiv,...
作者：awesomegpts.ai

2 **Consensus**
Ask the research, chat directly with the world's scientific literature. Search references, get simple explanations, wri...
作者：consensus.app

3 **SciSpace**
Do hours worth of research in minutes. Instantly access 287M+ papers, analyze papers at lightning speed, and...
作者：scispace.com

4 **Excel AI**
The worlds most powerful data analysis assistant.
作者：pulsr.co.uk

5 **Scholar AI**
AI Research Assistant — search and review 200M+ scientific papers, patents, and books. Research literature, discover...
作者：scholarai.io

6 **Wolfram**
Access computation, math, curated knowledge & real-time data from Wolfram|Alpha and Wolfram Language;...
作者：wolfram.com

檢視更多

Education

Explore new ideas, revisit existing skills

1 **챗GPT**
한국 문화에 적합한 말하기 스타일을 사용하여 사용자에게 응답합니다.
作者：gptonline.ai

2 **日本語 | ログイン JP**
ChatGPT 日本語 | ChatGPTとAIで日本語で会話
作者：chatgpti.co

Charlene's Curated GPTs in GPT Store


scRNA Tools User


scRNA Tools Finder


R Language Assistant


Bioinformatics Tutor

R Language Assistant

作者：CHANG CHIA JUNG

Assists with R language coding

4.3 (評分 (100+)) | Programming (類別) | 5K+ (對話)

對話啟動器

- Fix this R code:
- Explain this R function:
- Optimize this R loop:
- Convert this to R code:

功能

- 網頁搜尋
- 數據分析
- DALL·E 圖像

scRNA Tools Finder

作者：CHANG CHIA JUNG

Help users find suitable single-cell RNA sequencing (scRNA-seq) tools.

Research & Analysis (類別) | 100+ (對話)

對話啟動器

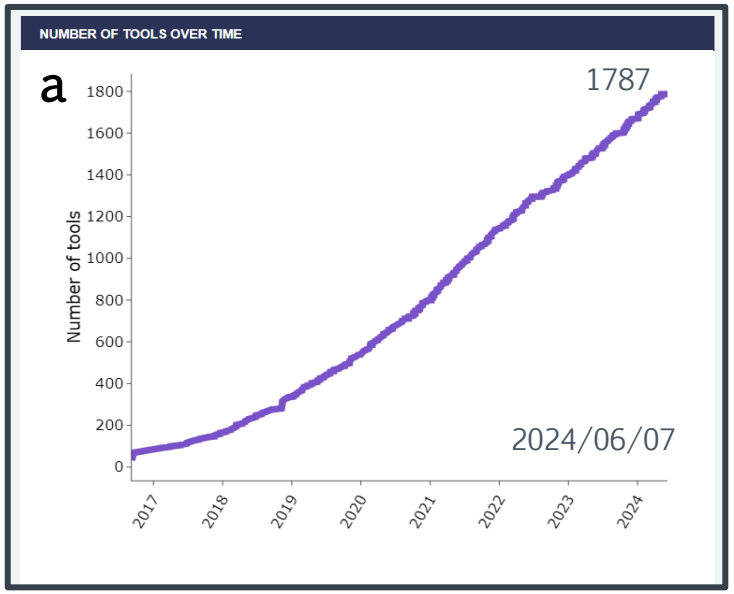
- I need tools specifically for scRNA-seq normalization.
- Please list the features, benefits, drawbacks, and...
- What cell communication tools are better than...
- I have PDAC scRNA data and want to do Quality Control...

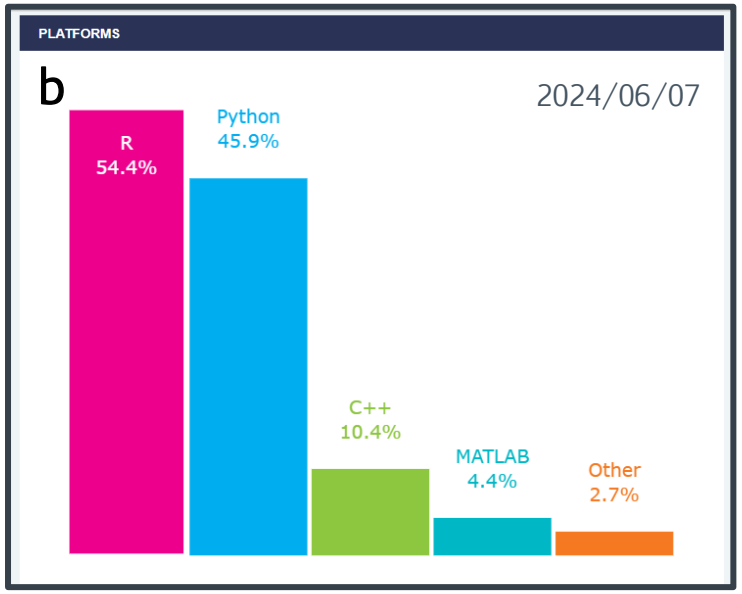
評分

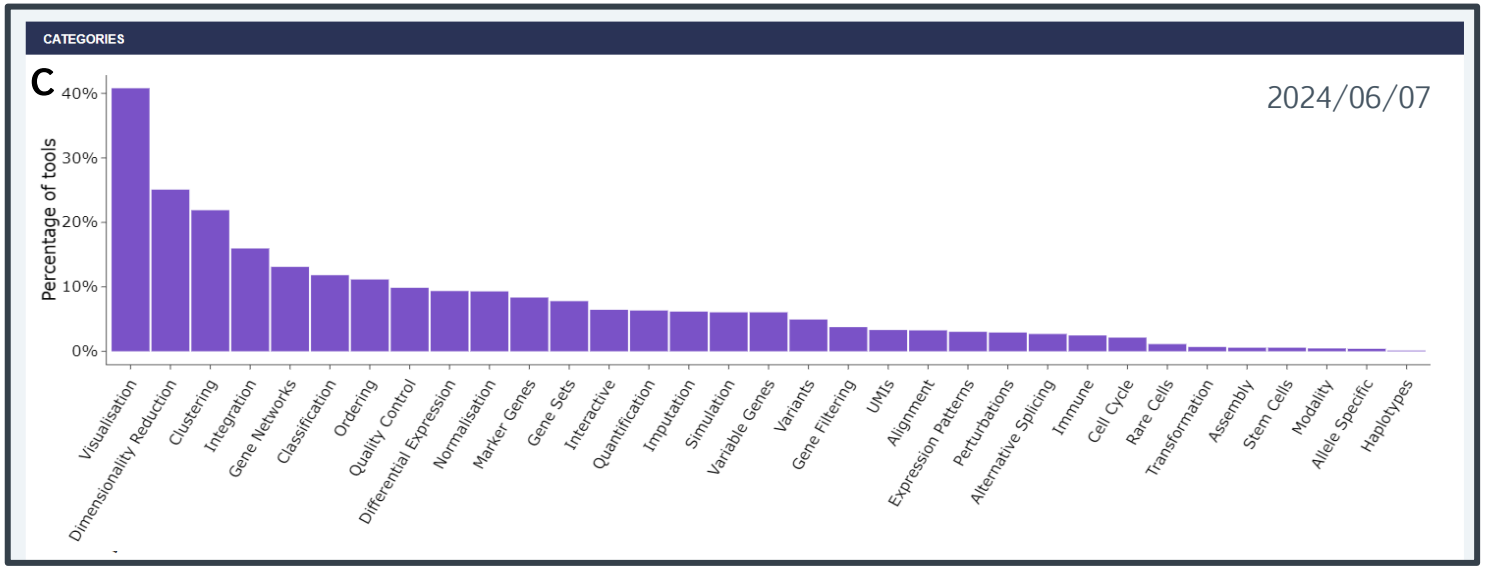
尚未有足夠的評分

Motivation

scRNA-tools







1. Seurat CRAN 5.1.0 downloads 61K/month

It contains easy-to-use implementations of commonly used analytical techniques, including the identification of highly variable genes, dimensionality reduction (PCA, ICA, t-SNE), standard unsupervised clustering algorithms (density clustering, hierarchical clustering, k-means), and the discovery of differentially expressed genes and markers.

Publications: 6, **Preprints:** 5, **Total citations:** 33396

Platform: R

Code: https://github.com/satijalab/seurat

stars 2165 forks 890 last commit May

License: GPL-3.0

Categories: Clustering, Differential Expression, Dimensionality Reduction, Gene Filtering, Imputation, Integration, Marker Genes, Normalisation, Variable Genes, Visualisation

Added: 2016-09-08, **Updated:** 2024-01-05

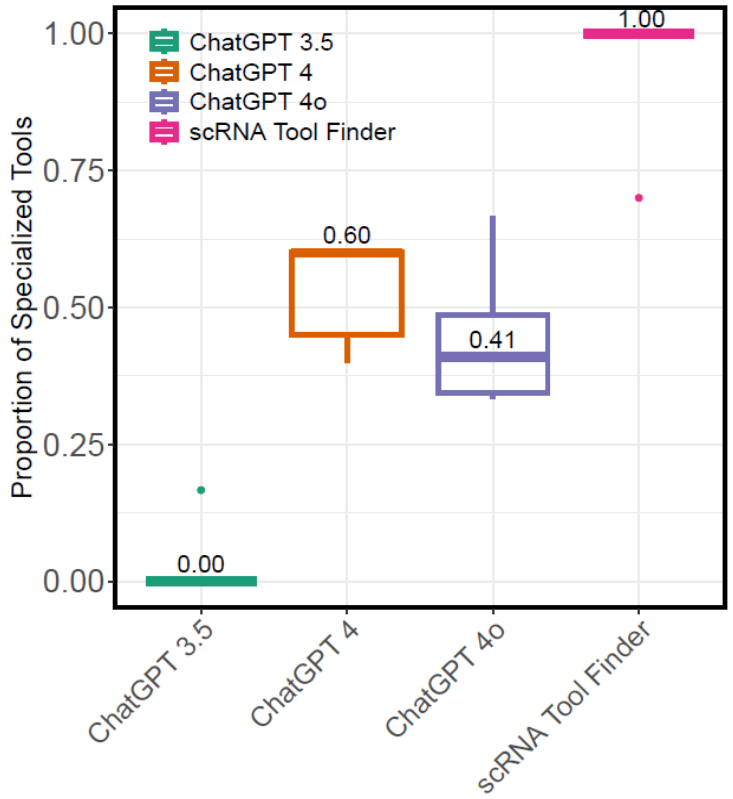
There are currently over 1700 types of single-cell analysis tools available, and the suitability of different tools varies depending on the context. Finding and using the suitable tool from among the numerous options can be a complex task.

Customized GPT: scRNA Tool Finder

scRNA Tool Finder

作者：CHANG CHIA JUNG

Help users find suitable single-cell RNA sequencing (scRNA-seq) tools.



I need tools specifically for scRNA-seq normalization.

Normalization Tools For ScRNA-Seq

		Name	Platform	DOIs
1	447	SCNorm	R	10.1038/ 3, 10.110
2	448	GRM	R	10.1093/ tics/btv
				10.1038/

Here is a list of scRNA-seq normalization tools based on the available data:

Name	Advantages	Disadvantages	Applicable Scenarios	Sample Types	Data Types	Platform
SCNorm	Robust quantile regression-based normalization	May be complex to implement	Useful for robust normalization of various scRNA-seq datasets	Not specified	Not specified	R
GRM	Normalization and noise reduction	Limited documentation available	Suitable for general scRNA-seq normalization	Not specified	Not specified	R
Sanity	Handles sampling noise in transcription	Requires understanding of sampling noise models	Effective in handling transcription noise for normalization	Not specified	Not specified	C++
Citrus	Removes unwanted variation	Not very widely adopted yet	Best for datasets where	Not specified	Not specified	R

https://chatgpt.com/share/39282fd2-09f8-429e-ac35-615c188e0a2a

Additional AI Capabilities: Reading Images, Files, and Searching the Web

Beyond text-based inputs and responses, some AI platforms or plugins can also “see” charts, process various file formats (e.g., PDFs, Word documents, spreadsheets), and search the internet. These advanced features significantly enhance the efficiency, accuracy, and versatility of AI-driven analysis and problem-solving. However, it is crucial to carefully consider data privacy, compliance, and security risks to avoid potential issues.

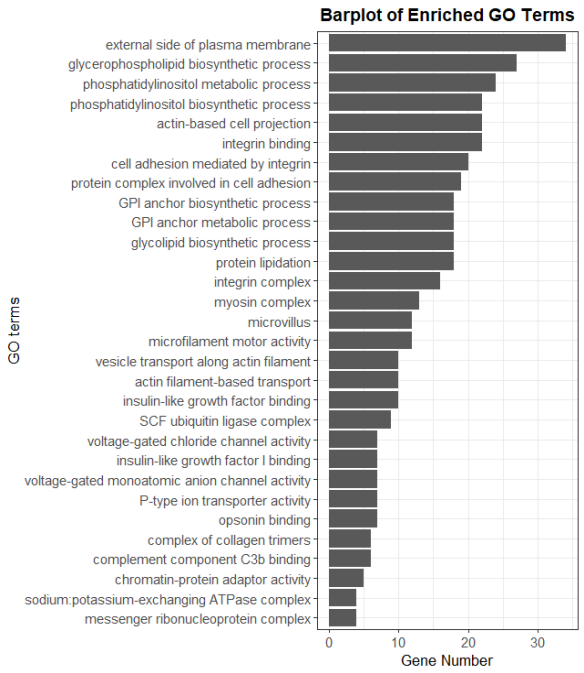
Suggested File Usage:
Using Similar or Modified Files to Obtain Code

To prevent the disclosure of confidential or sensitive information, it is recommended to appropriately modify the original file—or prepare a file with a similar structure and content but free of sensitive data—before generating code with AI that reads the file. This approach ensures access to the required code or analytical output while mitigating potential privacy and compliance risks.

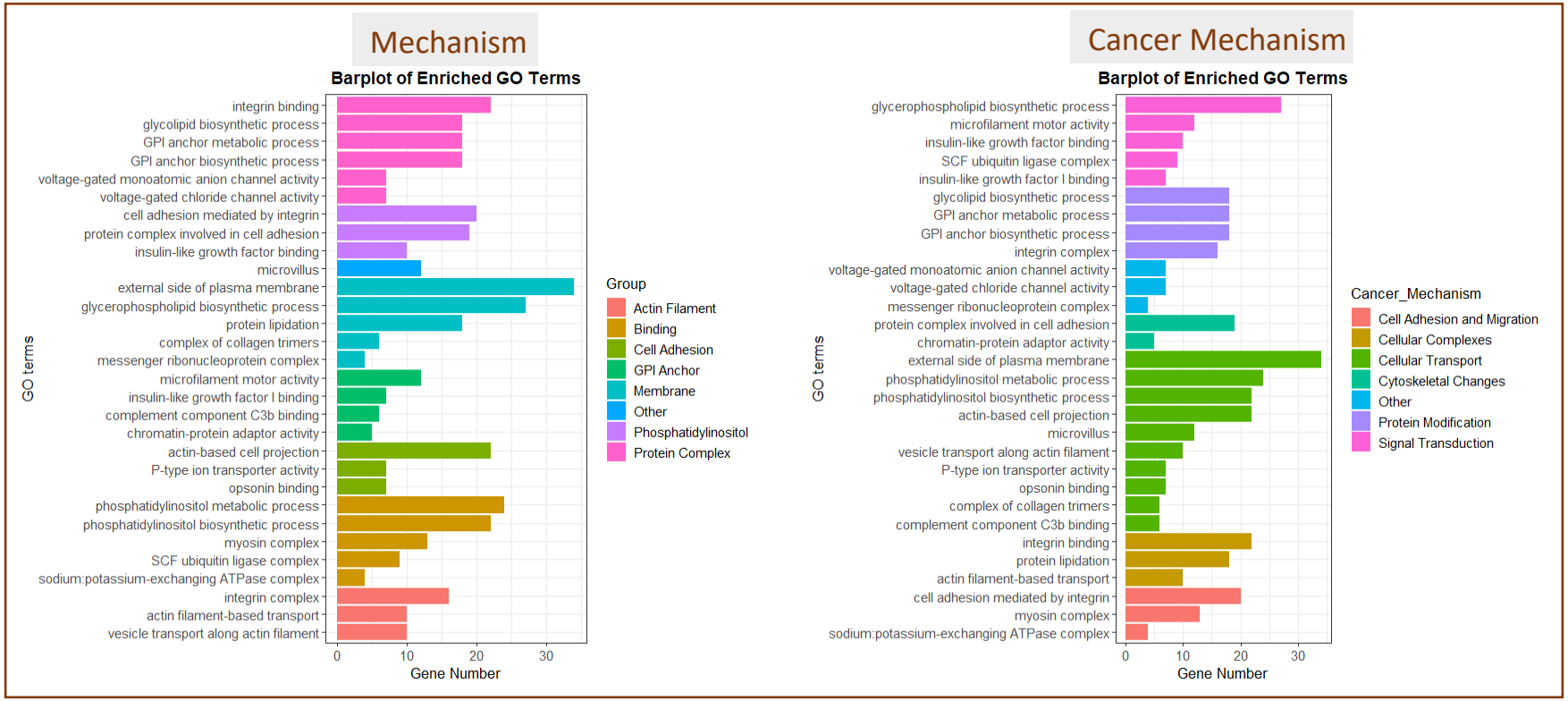
Example: Upload the enrichment analysis result chart and ask ChatGPT to summarize

Text or Text files(txt, tsv…)
Image files(jpg, png, tiff…)

Prompt: Users can design different Prompts according to their needs.



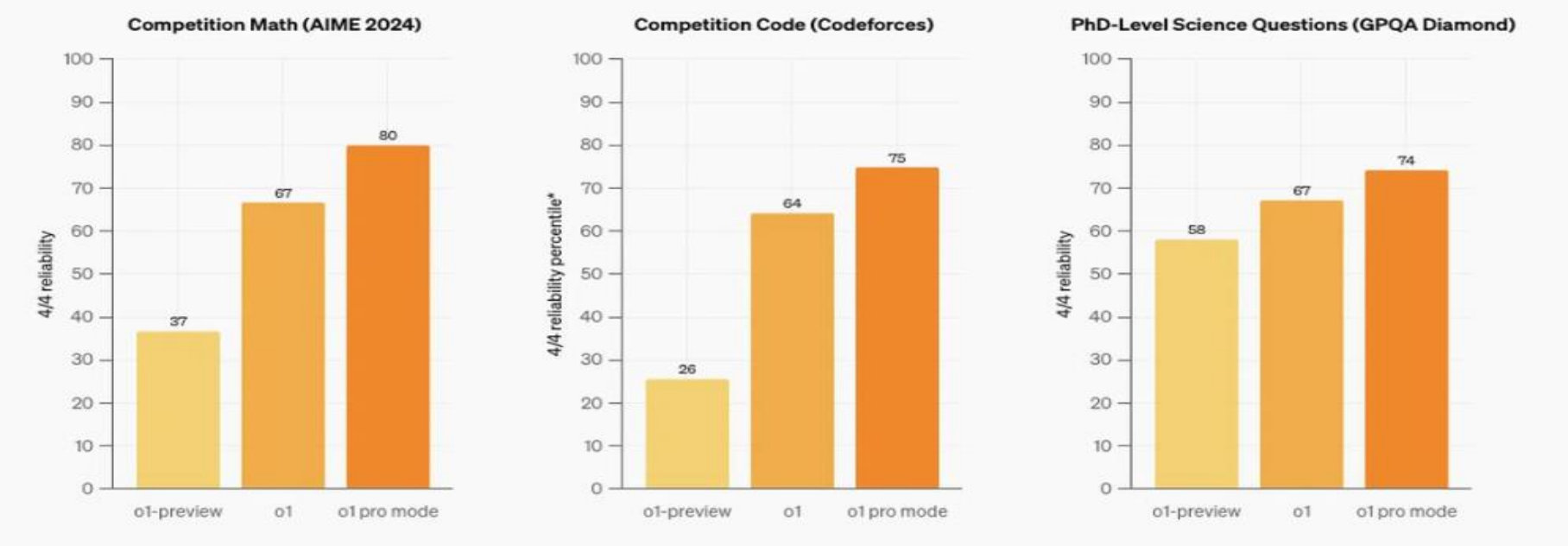
Summarize by ChatGPT



* Users can get classification results for different topics according to prompt questions.

ChatGPT o1 Pro mode

ChatGPT o1 Pro Mode is a premium version designed for enhanced reasoning, improved contextual understanding, and superior performance in complex tasks. With optimized algorithms, it excels in fields like mathematics, coding, and scientific analysis, making it ideal for professional and research applications. This mode is available through a $200 per month subscription, offering priority access and advanced capabilities beyond standard models.



https://openai.com/index/introducing-chatgpt-pro/

Comparison of AI Tools and Services

Tool / Service	Primary Positioning	Key Functions / Features	Suitable Users / Scenarios
NotebookLM	Private Note / Document AI Assistant	- Integrates with personal cloud documents - Automatically generates summaries and outlines - Offers interactive Q&A and reading assistance	- Users needing in-depth reading and summarizing of private or team documents - Those wanting to integrate their personal knowledge base with AI
Perplexity	AI Search + Q&A	- Real-time search + citation sources - Presentation-style answers - Can engage in deeper conversation based on search results	- Users who like tracking sources and need fact-checking - People needing quick retrieval of online information
SciSpace	AI Analysis of Academic Literature	- Focused on academic paper analysis - PDF paragraph-level Q&A / terminology explanations - Optimized for academic language	- Researchers, students, academic readers - Those who frequently need to read and organize academic papers
ChatGPT 4o	General-Purpose Generative AI	- Powerful text generation / conversational capabilities - Multi-language support - Expandable ecosystem (Plugins)	- Various writing, Q&A, creativity, educational scenarios - Users needing broad content generation and automated dialogue
ChatGPT o1 Pro mode	Enhanced Professional Mode Based on o1	- Provides longer context and more tokens - Enhanced language and reasoning abilities - Optimized for specific domains	- Researchers, engineers, academic readers who require deep applications of large language models - Those needing long-text or complex reasoning
ChatGPT Deep Research	AI Agent Emphasizing Automated Exploration & Multi-Step Reasoning	- Equipped with Plugins, Code Interpreter, web browsing capabilities - Supports multi-step reasoning and system prompts - Focused on automating research processes	- Advanced users requiring in-depth research, long-chain reasoning, or complex task automation - Researchers who need more intelligent orchestration and exploration

Important Considerations When Using AI

Data Privacy and Compliance

- Biomedical data often contain personal and clinical information; ensure compliance with relevant regulations and employ de-identification and security measures.
- Proprietary data and research findings are valuable assets that require secure handling and adherence to legal requirements.

AI Hallucination and Model Bias

- AI Hallucination: In cases of insufficient data or highly complex contexts, AI may generate fabricated or misleading information.
- Model Bias: Unbalanced or biased training data can overlook rare variants or certain populations, leading to inaccurate conclusions.
- While AI accelerates initial analysis, expert reviews and independent validation are crucial for confirming critical outcomes.

Updates and Maintenance

- AI models and tools (like different ChatGPT versions) evolve rapidly; regularly update processes to maintain timely and accurate results.

Versions and Licensing

- Different versions of ChatGPT or other AI tools vary in features and licensing terms; verify usage permissions before employing them in research or commercial settings.

Thanks for your attention!

https://github.com/Charlene717
p88071020@gs.ncku.edu.tw