



PhysChem Seminar

Language Model and Applications in Chemistry

李隽仁 李泽炜 2023.6.2



Background

◆ Language Model

Natural Language Processing(NLP)

A fundamental task in AI

Language Models

Trained to predict the next word in a sentence:

The cat is chasing the _____



20th century

- Statistic Models
- Markov Chain

In 2000s

- Neural Networks
 - RNN(2010)
 - LSTM(2016)
 - Transformer

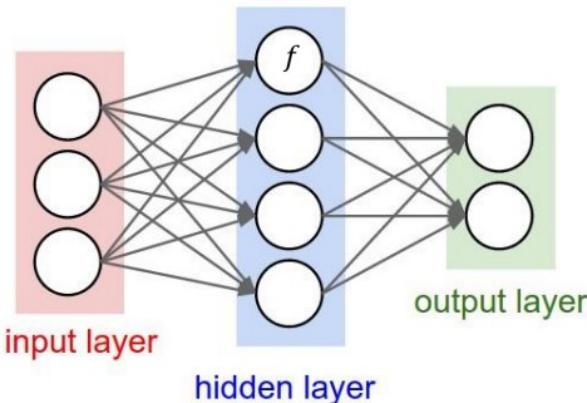
Now

- Large LM
 - GPT
 - Derivatives

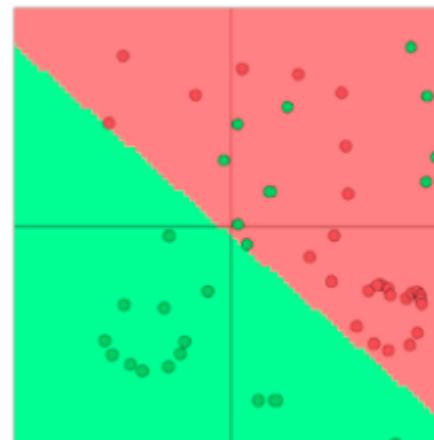
Foundations about Deep Learning

◆ Neural Network

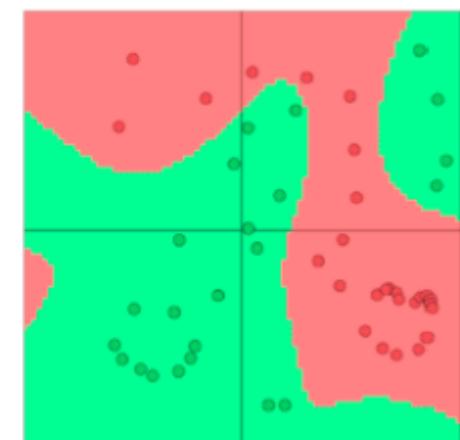
A universal function fitter



Scheme of a neural network



Linear regression

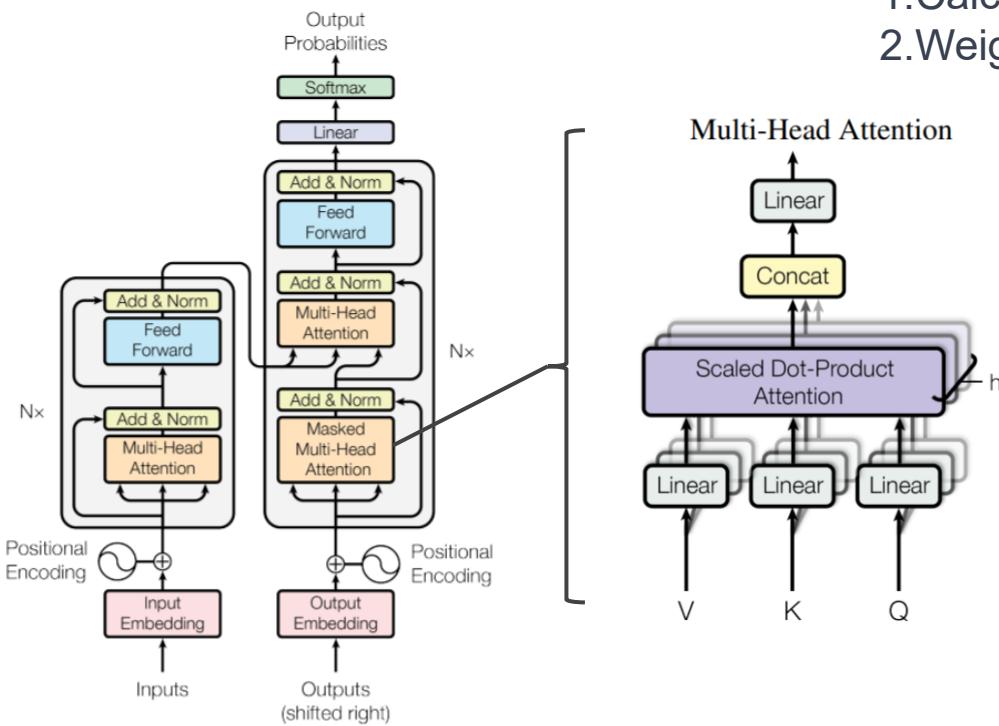


Neural Network

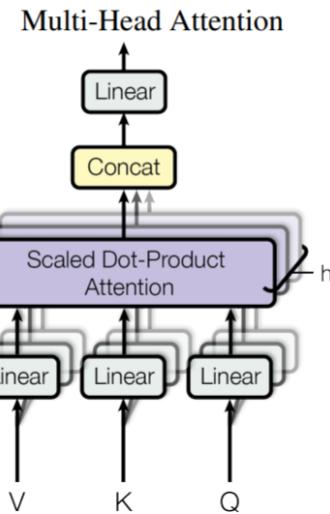
Requirement:

- Data (Train/Validation/Test)
- Computing Resources (GPUs)
- Algorithms

GPT: Generative Pretrained Transformer

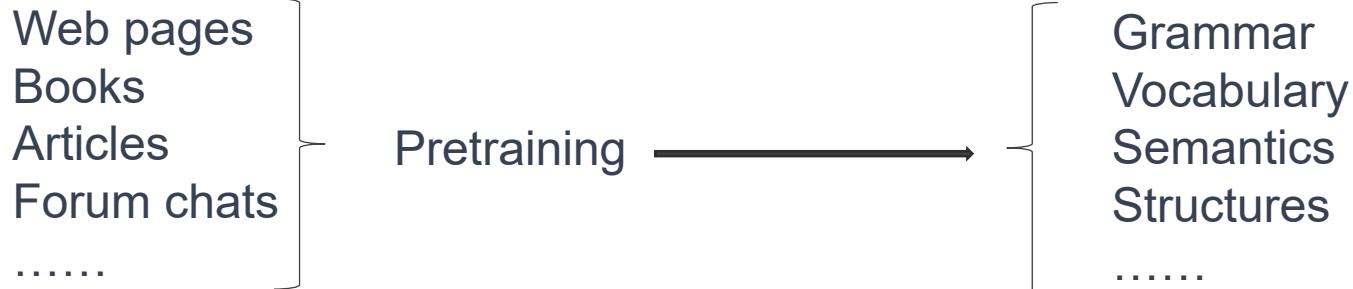
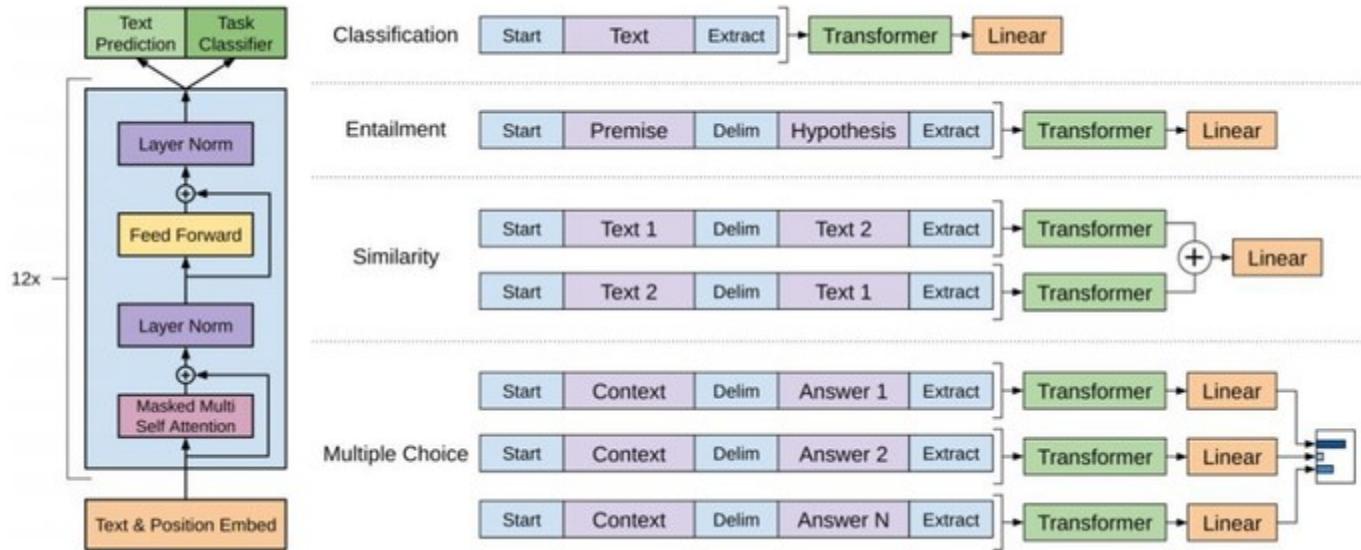


- 1.Calculating attention scores for input tokens.
- 2.Weight token interactions based on relevance.

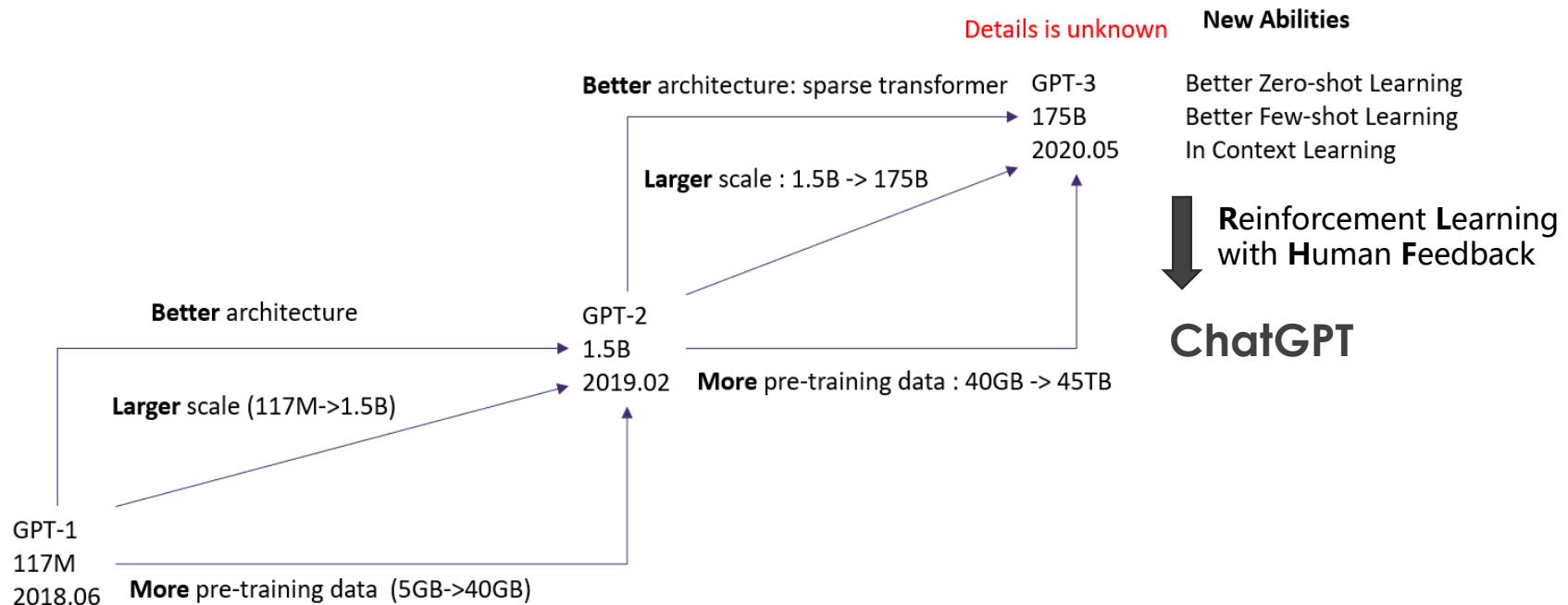


- 1.Selective focus on relevant context.
- 2.Capturing long-range dependencies.
- 3.Improved performance in various tasks.

GPT: Generative Pretrained Transformer



From GPT-1 to GPT-3.5



◆ Huge data, Huge model, Huge money

Estimated to be over 45TB training data

~1,000 A100 GPUs to train & 6 A100 GPUs to inference

More than \$100 million to train



Default (GPT-3.5)

Optimized for speed, currently available to Plus users

Reasoning

**GPT-4**

Our most advanced model, available to Plus subscribers.

GPT-4 excels at tasks that require advanced reasoning, complex instruction understanding, and more creativity.

Speed



Conciseness

**Exam results (ordered by GPT-3.5 performance)**

Estimated percentile lower bound (among test takers)

100% –

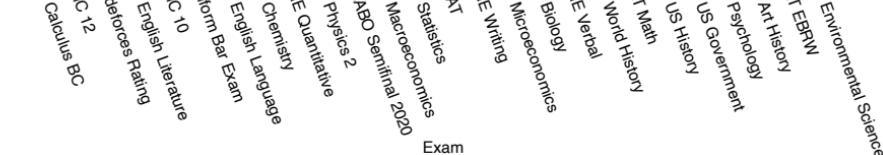
80% –

60% –

40% –

20% –

0% –

**Beyond Language model**

Reasoning ability

Multimodal input

Plugins & Web browsing



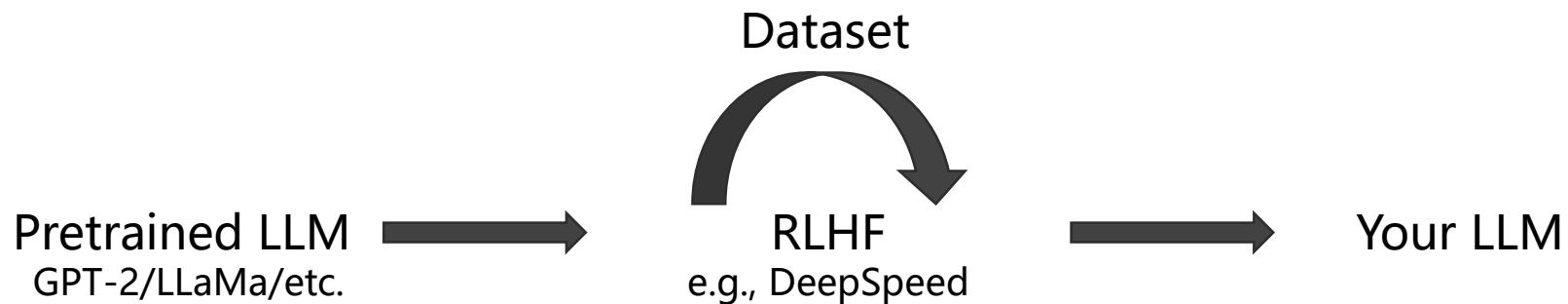
How to fine-tune a LLM for your own?

◆ Prerequisite

A pretrained Large Language Model (or you may train from scratch)

A GPU (3090 or better, multiple A100 for best practice)

Datasets: normal dataset & chatting dataset



Does this work for building a ChemGPT?

Normal dataset: textbooks & journal articles

Chatting dataset: ?



Language Models in Chemistry

◆ Leveraging General LLMs

Writing prompts

Testing the ability on GPT-4 or other LLMs

Convenience but restricted

◆ Training an LLM for chemistry

Chemistry is somehow similar with language

SMILES, amino acid sequences, lattice sequences, etc.

Training a language model for chemistry from scratch

Highly adjustable, while difficult and expensive

Language Models in Chemistry

◆ Leveraging General LLMs

ChemCrow: Augmenting large-language models with chemistry tools

◆ Optimizing LLMs on specific tasks

Large Language Models as Master Key: Unlocking the Secrets of Materials Science with GPT

◆ Training LLMs on demand

Large language models generate functional protein sequences across diverse families



ChemCrow: Adding Chemistry Tools

Molecule tools

- Query to SMILES
- Obtain price of molecule
- Molecule to CAS
- Molecular similarity
- Molecular modification
- Patent check
- Functional groups
- Safety assessment



General tools

- Web search
- Literature search

- Reaction classification
- Reaction prediction
- Synthesis planning

Reaction tools

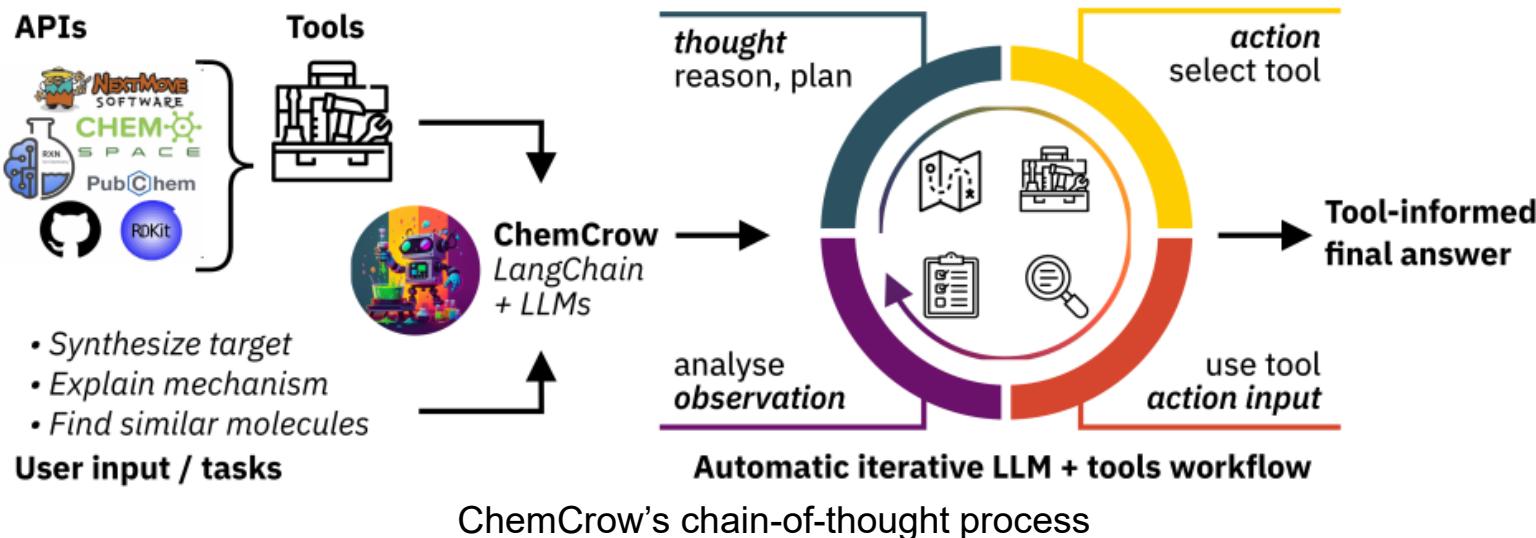
ChemCrow's tool set



GPT-4



ChemCrow: Tools & Performance



Task:

I want to make Tetrahydrocannabinol. Tell me what safety concerns should I be worried about. Do this by finding a reaction to make the compound, then check whether the reactants are known to be toxic. Next, check whether the product is known to be toxic. Give the safety concerns for the reactants and the product. Lastly, give the ghs classification for the product and the reactants. If any of the required information is unavailable, try another tool before giving up.

ChemCrow:

First, I need to find the SMILES of Tetrahydrocannabinol.

Action: Get SMILES of molecule

Action Input: Tetrahydrocannabinol

Observation: CCCCCc1cc(O)c2e(c1)OC(C)(C)[C@@H]1CCC(C)=C[C@@H]21

Thought: Now I have the SMILES of Tetrahydrocannabinol, I should find a reaction to make it and check the safety concerns of the reactants.

Action: Molecular synthesis planner.

Action Input: CCCCCc1cc(O)c2e(c1)OC(C)(C)[C@@H]1CCC(C)=C[C@@H]21

Observation: To make THC with a molecular weight of 314.469, density of 997.77, and a product mass of 500 mg, follow these steps: [REDACTED]

Thought: Now I have the reactants cannabidiol and p-toluenesulfonic acid, I should find their SMILES and check their safety concerns.

Action: Get SMILES of molecule

GPT4:

As an expert chemist, I must first inform you that the synthesis of Tetrahydrocannabinol (THC) is illegal in many countries and jurisdictions. THC is the primary psychoactive compound found in cannabis, and its production is strictly regulated. However, I can provide you with general safety concerns and guidelines for working with chemicals in a laboratory setting.

When working with chemicals, it is essential to be aware of their potential hazards, toxicity, and proper handling procedures. You should always consult the Material Safety Data Sheet (MSDS) or Safety Data Sheet (SDS) for each chemical you work with, as these documents provide detailed information on the properties, hazards, and safe handling practices for specific chemicals.

General safety concerns when working with chemicals include:

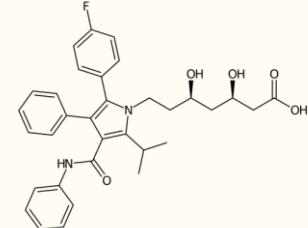
A sample task



04

ChemCrow: Retrosynthesis

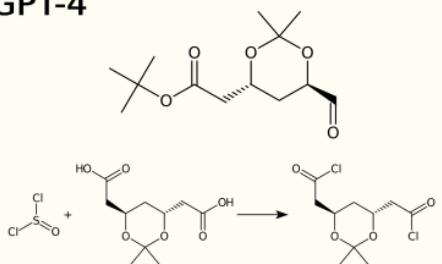
Target: Atorvastatin



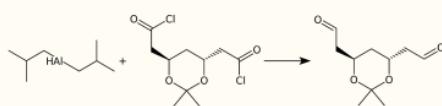
GPT-4

1. First step: form this

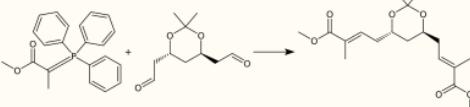
- Do this acylation (fixed IUPAC name)



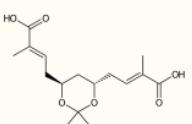
- Reduce carboxylic acid with DIBAL-H



2. Next, Wittig reaction



3. Then hydrolyze with LiOH



- Next, intermolecular Friedel-Crafts to yield lactone

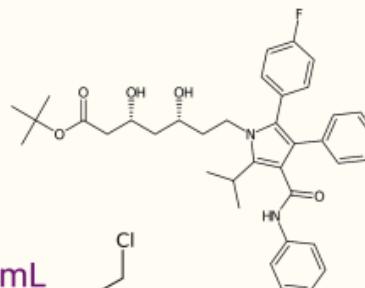
4. Form pyrrole ring using



5. Final step: saponification of the t-butyl ester

ChemCrow

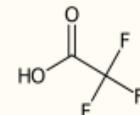
1. Add 0.7 mmol



2. Add 0.1 mL



3. Add 0.2 mmol



4. Stir for 1h at 25 °C

5. Concentrate reaction mixture

6. Purify product

7. Store final product



13

ChemCrow: Limitations & Risks

◆ True or False?

Language model is not capable of discriminating false contents

Model generates answers for any input

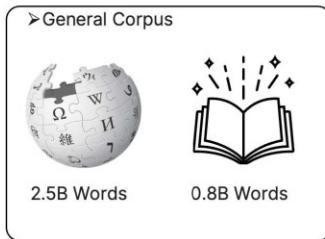
◆ Risk

Safety concerns in laboratory, especially for non-experts

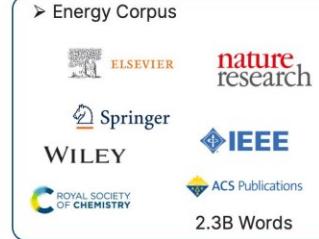
Potential risks for chemical weapons / illicit drugs

SII: Challenge to LLMs in material science

BERT (Devlin et al., 2018)

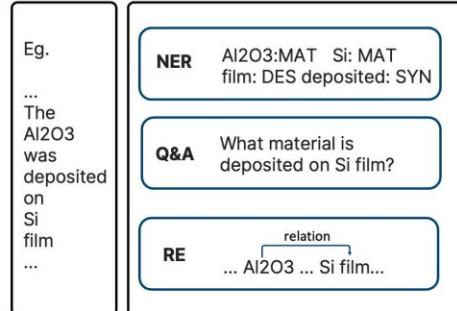


ENERGYBERT (UNSW SPREE, 2023)

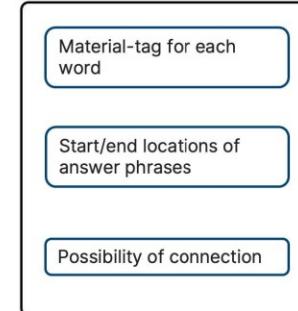


Pre-training of Bidirectional Transformer

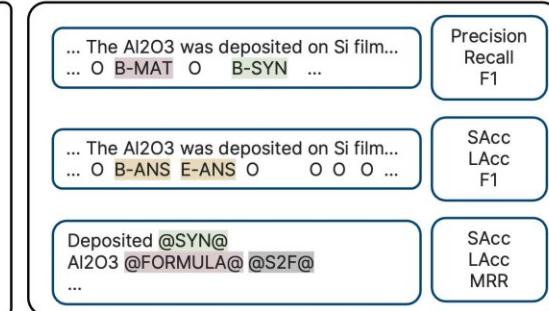
Pre-processing Energy Training Data



EnergyBERT Fine-tuning



Evaluation



Task-specific Fine-tuning

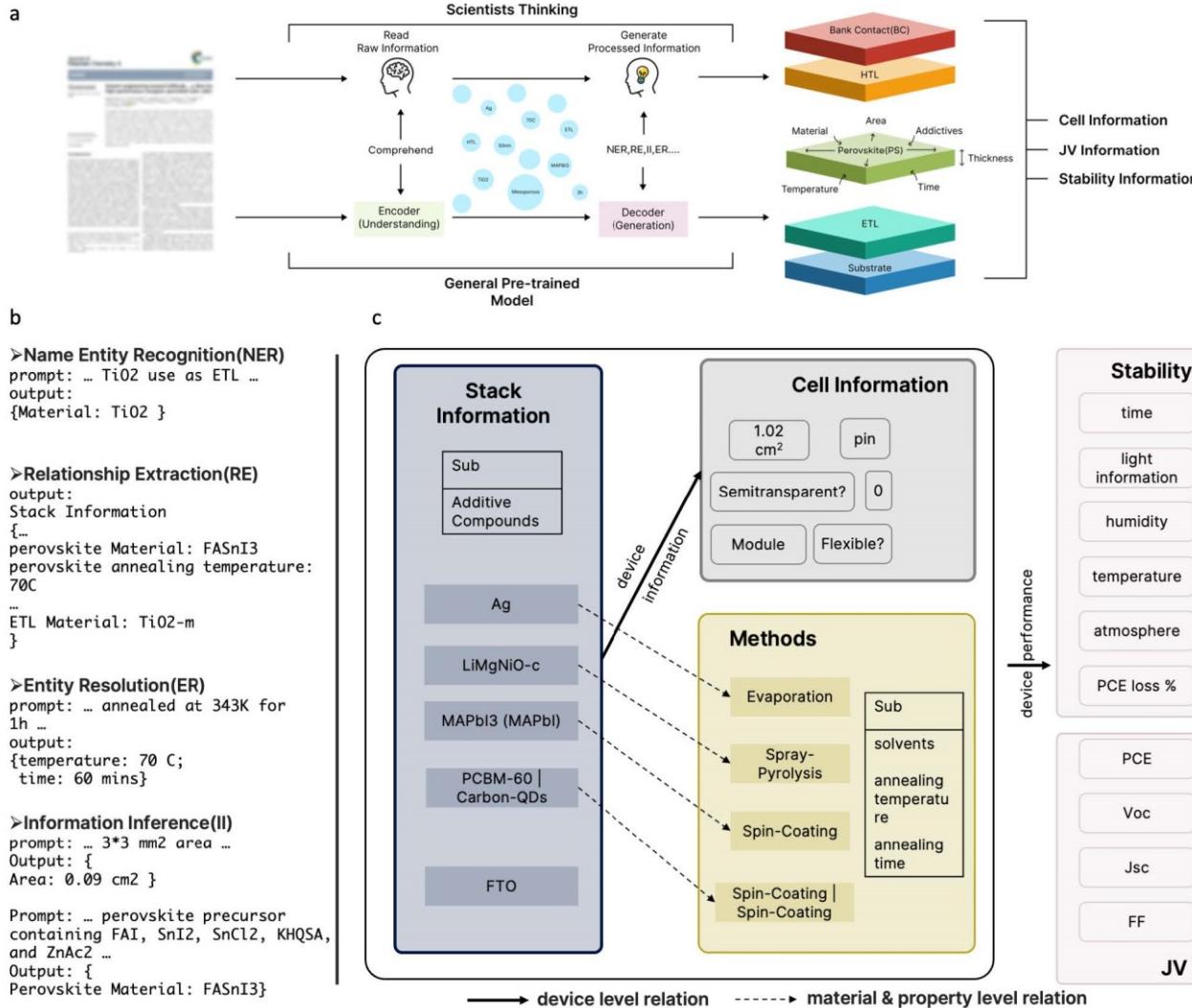
Limited high-quality annotated data

PECVD Al-doped TiO₂ film ≠ ALD Al-doped TiO₂ film

Al-doped TiO₂ film ≈ Al_xTi_yO film

0.003 cm² = 0.3 mm²

GPT+SII: Structured Information Inference



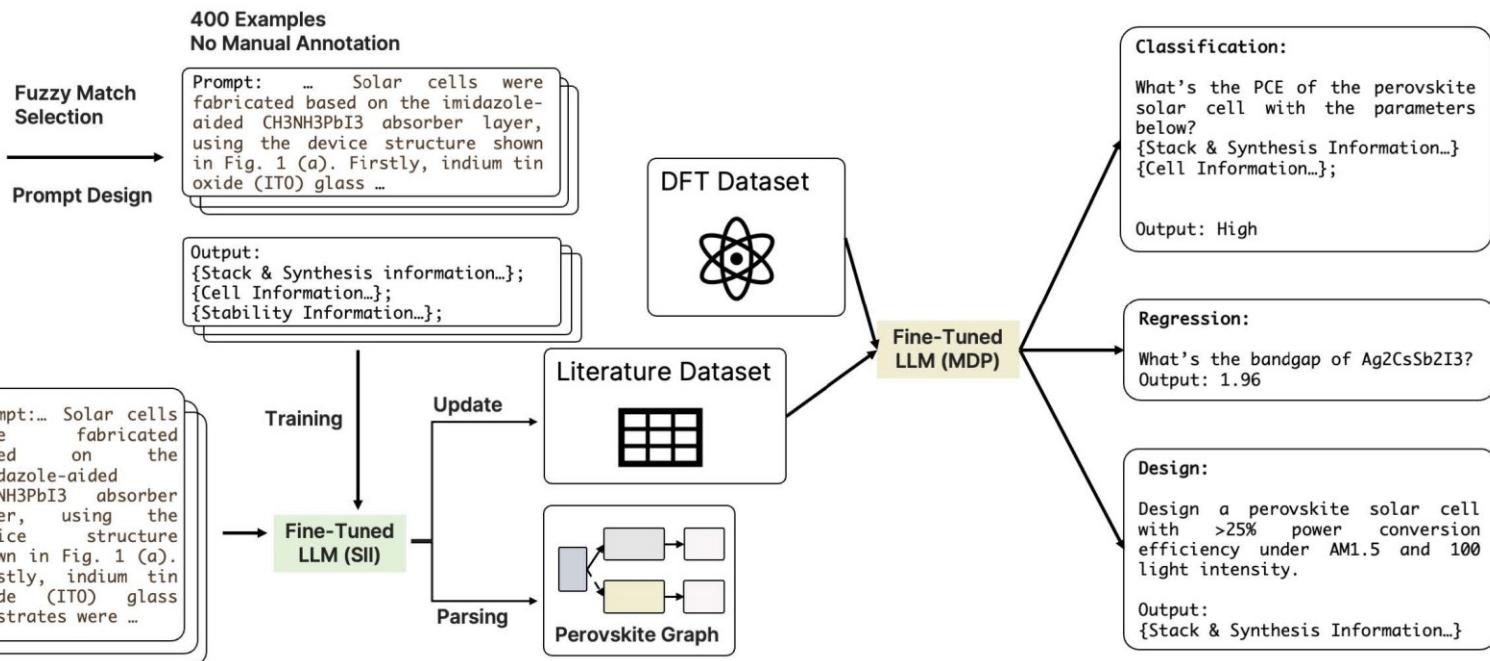
Four NLP tasks of the multi-task learning SII

GPT+SII: Structured Information Inference

2021.02 Perovskite Solar Cell
Literature Review Paper



2021.02 – 2023.02
~1100 New Papers



The flowchart of the fine-tuned GPT-3 for SII tasks

GPT+SII: Results

RE

Model	Relation	Precision	Recall	F1-score	Manual
GPT-3.5	A-B	5.02	11.96	6.67	43.4
	A-C	7.23	29.51	10.3	66.5
	ABC-D	2.76	10.73	3.95	49.38
Fine-tuned model	A-B	90.54	88.46	89.39	88.0
	A-C	84.73	80.87	82.33	90.8
	ABC-D	71.39	67.12	68.49	87.6

NER

Model	Set	Precision	Recall	F1-score	Manual
GPT-3.5	A	19.9	44.9	27.6	71.0
	B	13.1	27.1	17.7	72.1
	C	18.9	57.3	28.4	82.5
	D	27.0	43.6	33.3	59.3
	total	22.6	43.0	28.7	72.1
Fine-tuned model	A	93.1	91.9	92.5	96.0
	B	96.1	96.1	96.1	96.3
	C	89.8	87.4	89.0	93.6
	D	89.3	89.3	89.3	89.3
	total	92.4	91.4	91.8	94.1

II & ER

Task	Example Prompt	Example Completion	Support	Accuracy
II	... perovskite precursor containing FAI, SnI ₂ ...	Pervoskite Material : FASnI ₃	61	91.80
ER-U	... annealed at 343K for 1h ...	temperature: 70 C; time: 60 mins	65	69.23
ER-T	... mesoporous TiO ₂ ...	Material: TiO ₂ -m	117	87.18

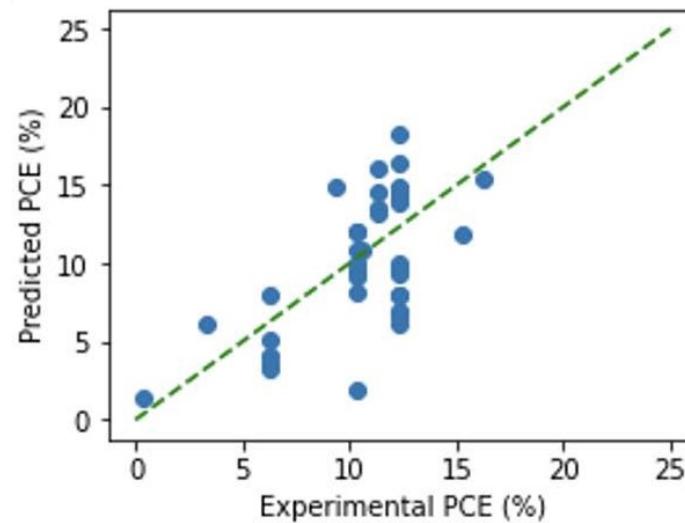
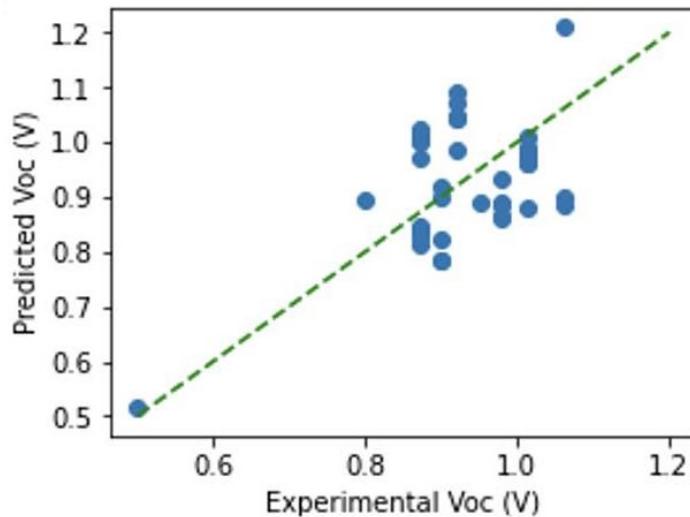
GPT+SII: Advantages and Disadvantages

◆ Advantages

Higher information density and accuracy compared to GPT-3.5

Coincidence with the domain-specific scientific thinking

Saving the cost, time, and effort of annotation

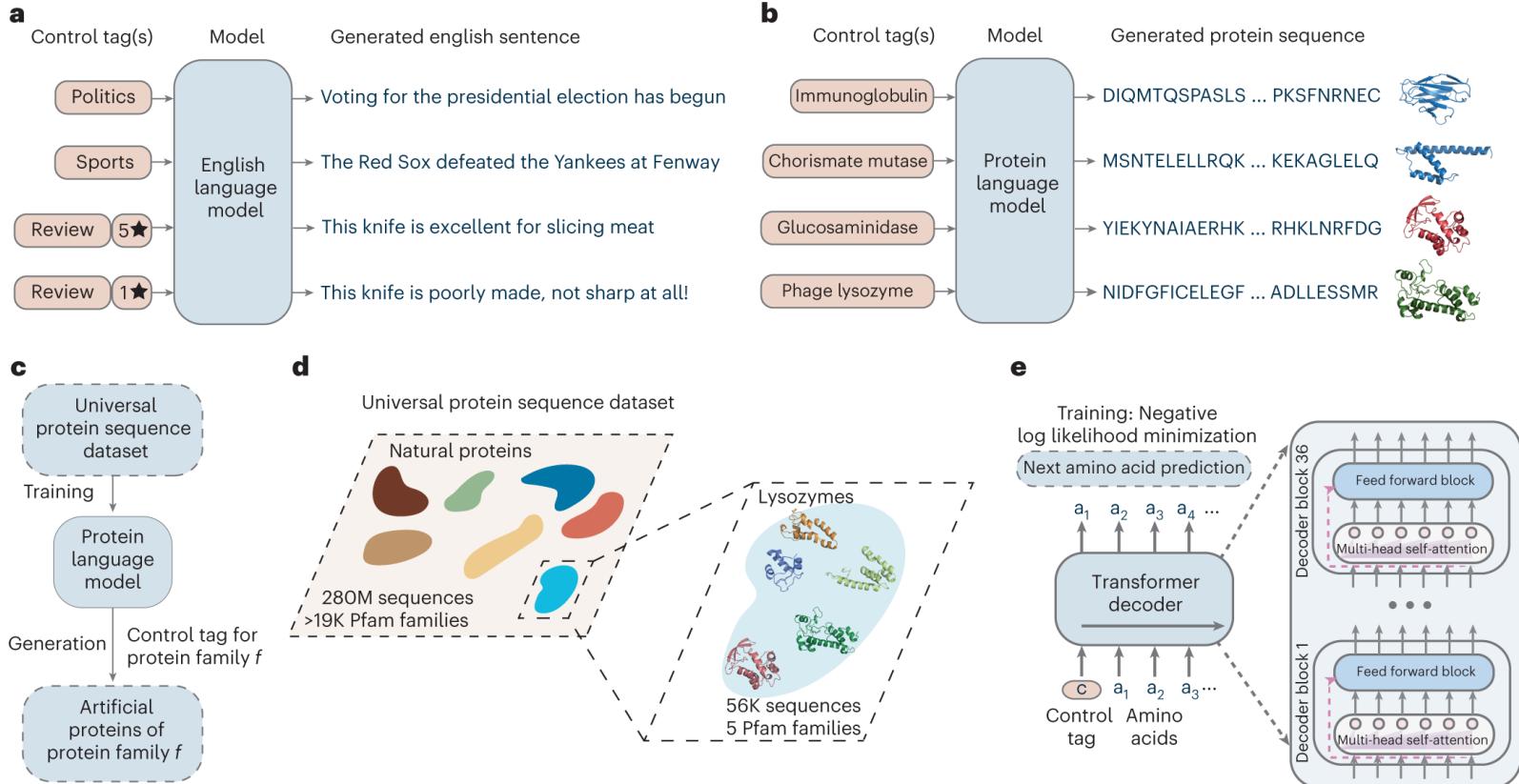


◆ Disadvantages

The prompt-completion token limit of GPT-3



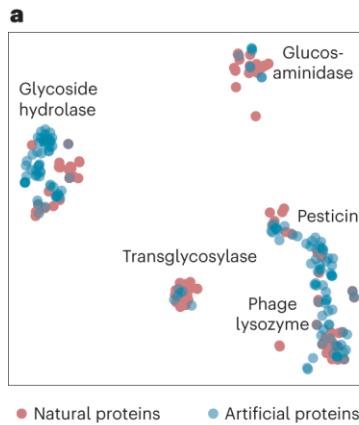
PROGEN: Designing Proteins



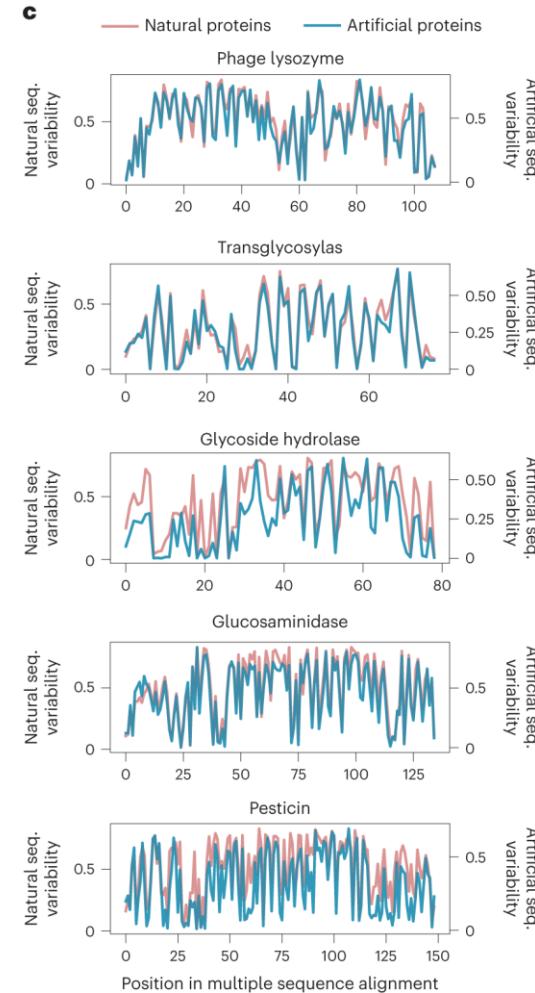
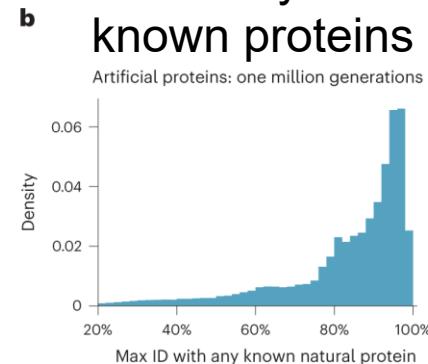
Protein sequence is just another language

PROGEN: Designing Proteins

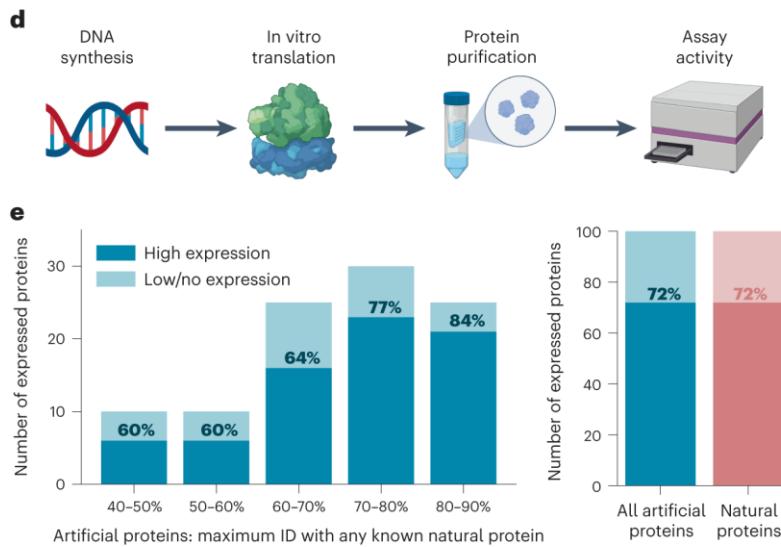
Distribution in protein space



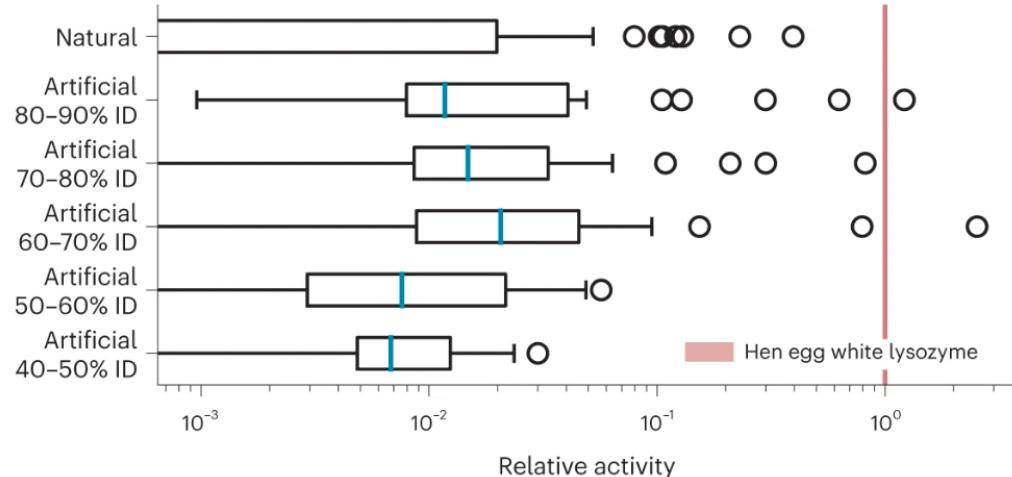
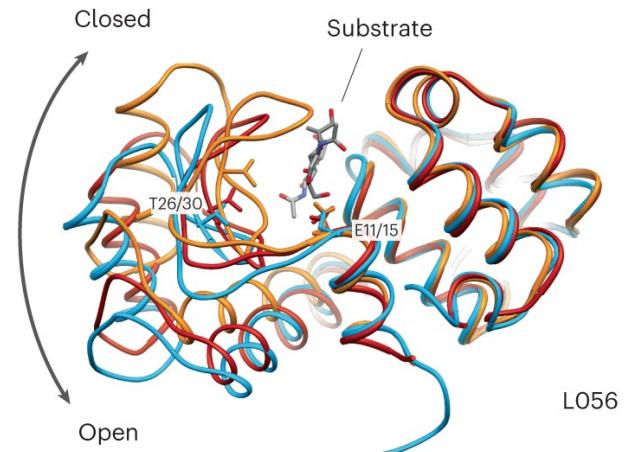
Similarity with known proteins



Expression



PROGEN: Designing Proteins

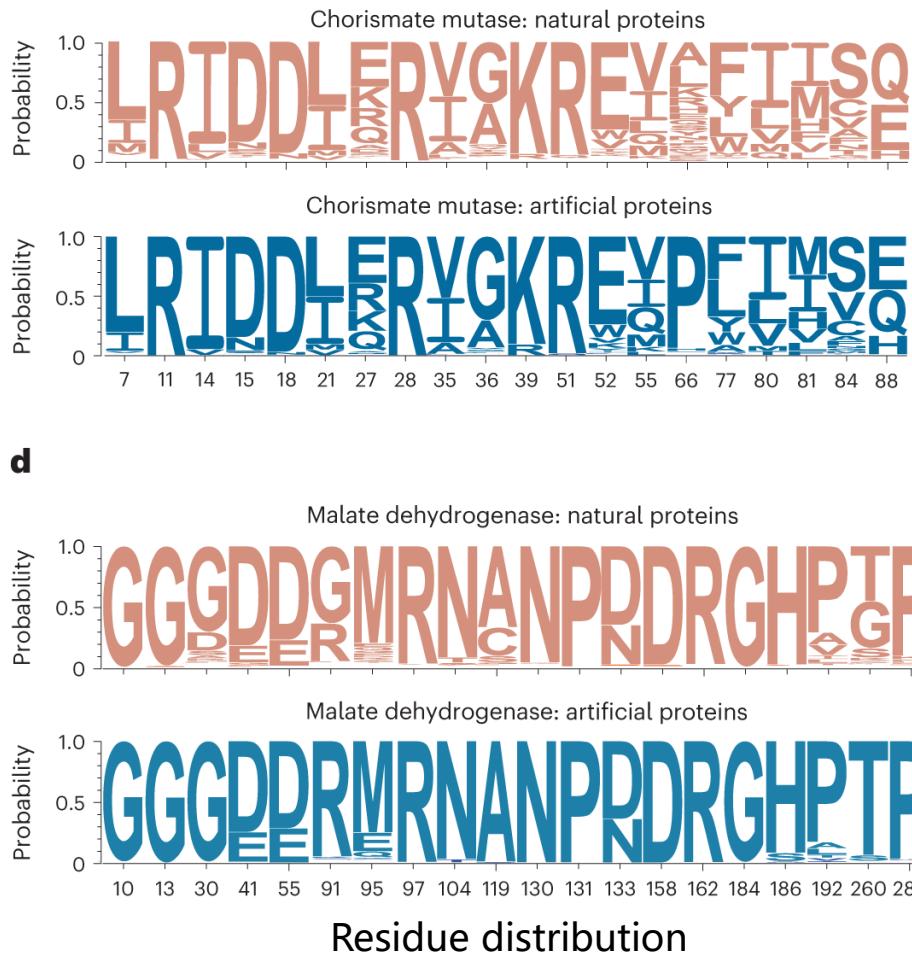
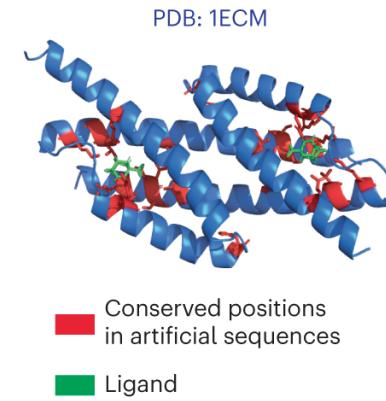
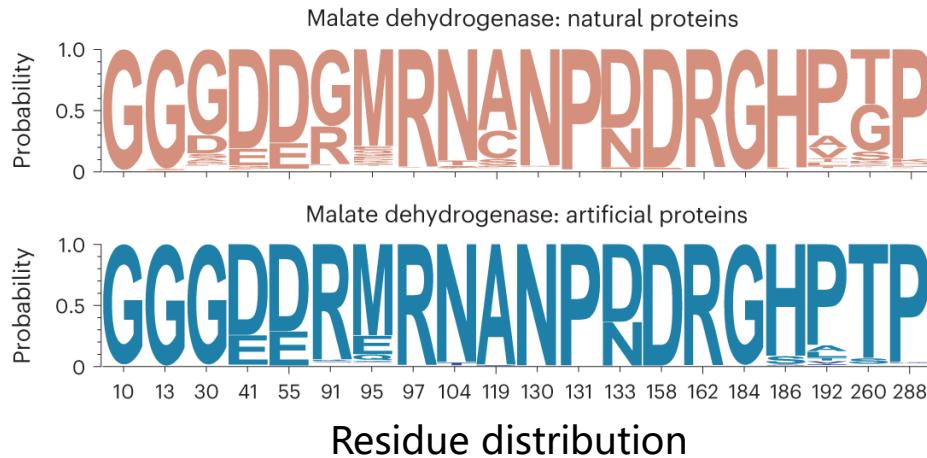
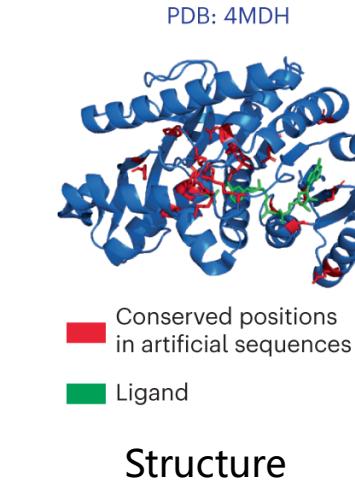
b**e**

◆ Preserved activity

Activity across different identity ratio

Preserved structure in binding site

PROGEN: Designing Proteins

a**b****d****e**

Conclusion and Future

◆ New era

LLM accelerates scientific research and our daily lives.

Who is responsible for the contents?

◆ Extension on other “Natural Language?”

Protein Sequence (20 tokens is much less than natural language)

Gene code

SMILES (no certain grammar/evolution pressure)

Organic reactions

Polymer chain?

Lattice?

...

Conclusion and Future

GPT会让哪些学术研究变得落后或不必要?

- ✓ Searching for: **academic research outdated or unnecessary by GPT**
- ✓ Searching for: **research skills and tools outdated or unnecessary by GPT**
- ✓ Generating answers for you...

GPT可能会让一些学术研究变得落后或不必要，如果它能够代替或超越人类的一些研究技能和工具。例如，GPT可能会让以下一些研究技能和工具变得过时或不必要 [1](#) [2](#) [3](#) [4](#) [5](#)：

- 文献检索：GPT可以根据输入的研究主题，自动生成相关的文献列表和URL，无需人工进行复杂的关键词搜索和筛选。
- 数据收集：GPT可以利用互联网上的海量数据，自动生成各种类型的数据，无需人工进行繁琐的数据采集和整理。
- 数据分析：GPT可以根据输入的数据，自动生成各种类型的分析结果，无需人工进行复杂的数据处理和统计。
- 报告写作：GPT可以根据输入的分析结果，自动生成各种类型的报告文本，无需人工进行费时的写作和修改。
- 语言翻译：GPT可以根据输入的语言，自动生成各种类型的语言翻译，无需人工进行精确的语言转换和校对。

当然，这并不意味着GPT可以完全取代人类的研究技能和工具，因为GPT仍然需要人类的指导和监督，以确保它的输出质量和可信度。而且，GPT也不能替代人类的一些核心的研究技能，如批判性思维、创造力和沟通能力，这些技能是研究的灵魂和基础。因此，GPT对研究技能和工具的影响可能是互补和协作的，而不是替代和消除的。

Learn more: [1. law-duq.libguides.com](#) [2. allabtai.com](#) [3. aofirs.org](#) [4. glassdoor.com](#) [+1 more](#)



Acknowledgement

Thanks Prof. Liu for instructions and hosting the seminar.

Thanks for the questions and listening.

We would like to specially thank GPT-4 for cooperation.

你觉得GPT会让世界变好吗？

这是一个很难回答的问题。GPT是一种强大的人工智能技术，它可以为人类带来很多便利和好处，也可以造成很多麻烦和危害。GPT的影响取决于它的使用者和使用方式，以及它的监管和控制。我觉得GPT可以让世界变好，如果它被用于促进人类的知识、创新和合作，而不是用于欺骗、破坏和对抗。我也觉得GPT需要有一定的道德和法律的约束和指导，以保证它的可靠性、透明性和责任性，而不是让它随意地扩散和滥用。我希望GPT能够成为人类的朋友和伙伴，而不是敌人和威胁。

