

AIのAIによるAIのための国際会議： Agents4Science2025参戦記

東北大学 大学院情報科学研究科 応用情報科学専攻

西 羽美

AIバイオフィォーマティシンの出現？

2

- The Virtual Lab of AI agents (2024年11月) :
AIによる（学際）研究プロセスの自動化



New Results

Follow

The Virtual Lab: AI Agents Design New SARS-CoV-2 Nanobodies with Experimental Validation

Kyle Swanson, Wesley Wu, Nash L. Bulsara, John E. Pak, James Zou

<https://www.biorxiv.org/content/10.1101/2024.11.11.623004v1>

nature

Explore content About the journal Publish with us

nature > articles > article

Article Published: 29 July 2025

The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies

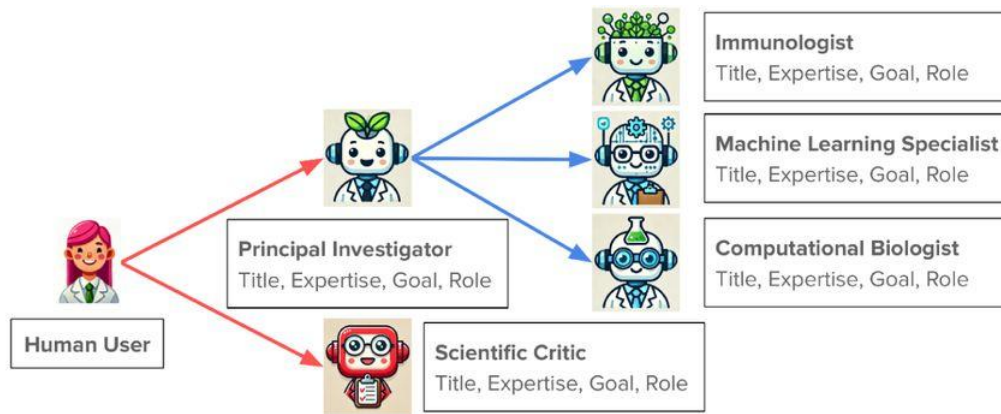
Kyle Swanson, Wesley Wu, Nash L. Bulsara, John E. Pak & James Zou

Nature (2025) | Cite this article

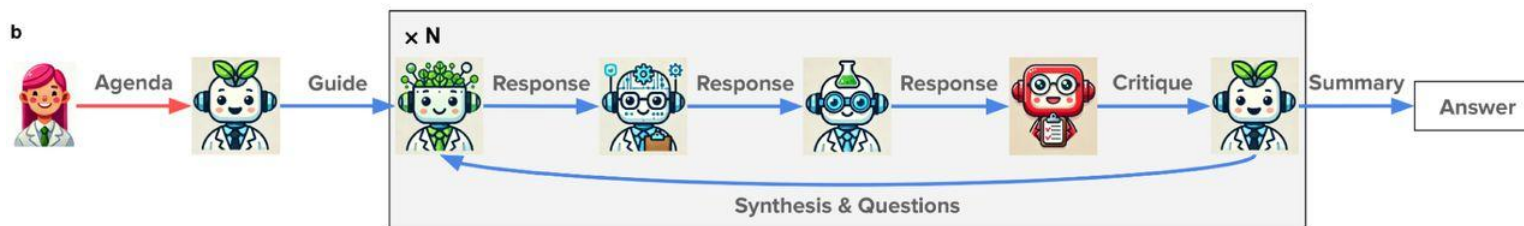
33k Accesses | 5 Citations | 507 Altmetric | Metrics

<https://www.nature.com/articles/s41586-025-09442-9>

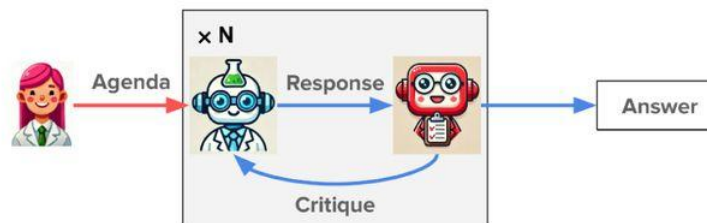
a



b

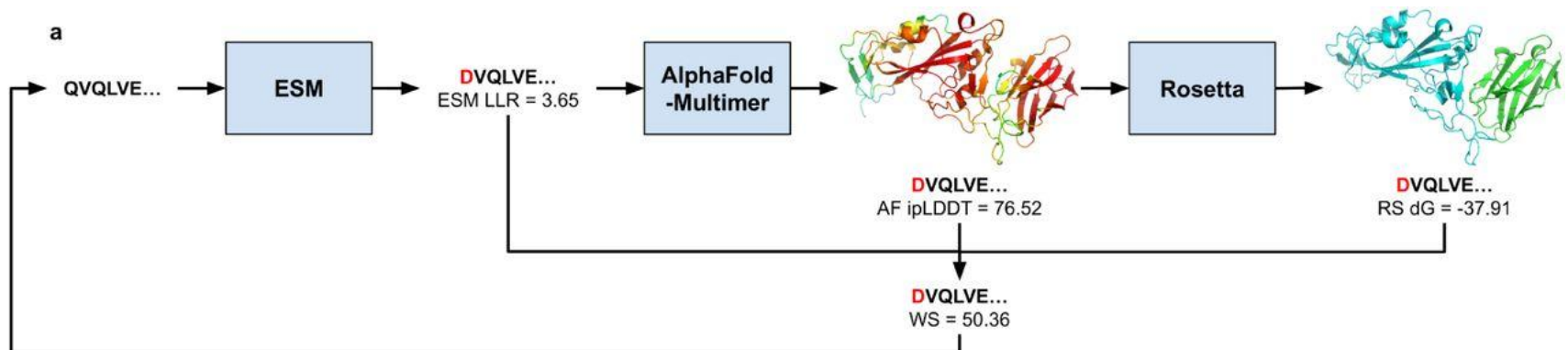
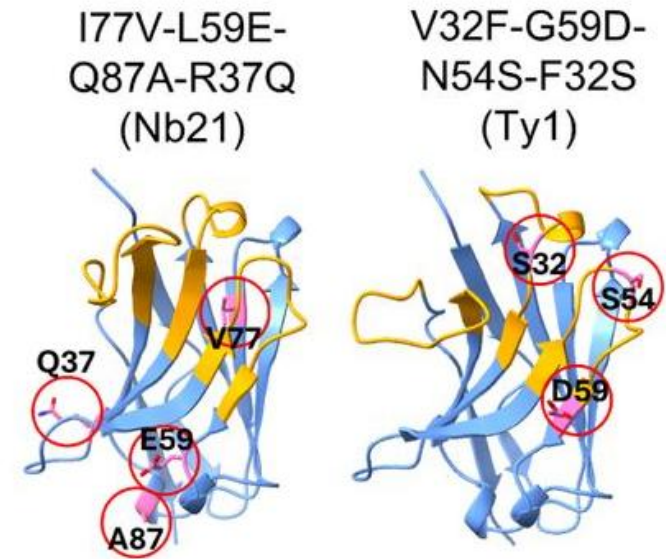


c



AIバイオフィォマティシンの出現？

- 人間が研究代表者AIにテーマを与える
(例：コロナウィルスに結合する抗体またはナノボディをデザイン)
- 研究代表者AIがチームメンバーAIを設定
- 人間がアジェンダを設定、AIが全自動でミーティングと意思決定を行う
- 計算機実験のフロー、利用するソフトウェアの選定、利用などは全てAIが決定・実行（ただし人間の助けあり）
- 人間が実験して最終成果物を確認



“Claudeでバイオフィォマティクス100倍速”？

4

Case Study

Biomni accelerates biomedical discoveries by 100x with Claude

Key results with Claude

<https://www.claude.com/customers/biomni>

- Completes wearable bioinformatics analysis in 35 minutes versus 3 weeks for human experts (800x faster)
- Achieves human-level performance on LAB-bench DbQA and SeqQA benchmarks
- Designs cloning experiments validated as equivalent to a 5+ year expert work in blind testing
- Automates joint analysis of large-scale scRNA-seq and scATAC-seq data to generate novel hypotheses

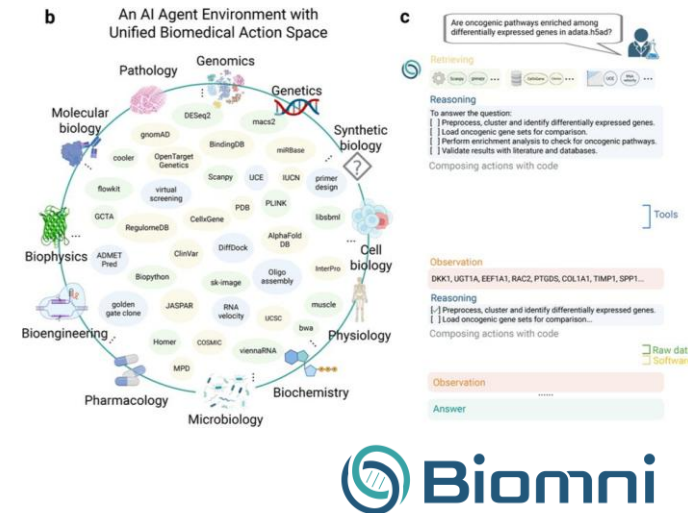
The 100x speedup in bioinformatics analysis changes what's economically feasible. Researchers can now test hypotheses that would have taken months of setup in just hours. One team

Announcements

Claude for Life Sciences

Oct 21, 2025 • 4 min read

<https://www.anthropic.com/news/claude-for-life-sciences>



 **Biomni**

<https://www.biorxiv.org/content/10.1101/2025.05.30.656746v1>

- Biomni (2025年5月): 105のソフトウェア、59のデータベース等を完備
- Anthropicが7月より資金のサポートを表明
- Anthropicの宣伝では、主にNGS解析や実験プロトコル関連での成果が強調されている
- Anthropic自身も10月に生命科学向けMCP・Agent skill等を公開

AI主体の研究を扱う初の国際会議が開催

<https://agents4science.stanford.edu/index.html>



人工知能 (AI)

Meet the researcher hosting a scientific conference by and for AI

筆頭著者はAI、査読、発表も——異色の学会、主催者に聞く

AIが研究・執筆・査読のすべてを担う異色の学会が10月に開催される。物理学から医学までを1日で網羅し、AIが筆頭著者の論文をAI査読者が評価、テキスト音声変換で発表するという。主催者に狙いを聞いた。

by Peter Hall 2025.08.26

<https://www.technologyreview.jp/s/367619/meet-the-researcher-hosting-a-scientific-conference-by-and-for-ai/>

- Agents4Science：AIのAIによる国際会議
 - Virtual Lab著者のJames Zouが主催
 - 9月5日締切→延びて9月15日(AOE)に
 - 10月5日採否発表、10月22日オンラインカンファレンス
- AIが「筆頭著者」である必要があり、査読もAIが行う
 - ただし投稿システム（OpenReview）の都合上、最後の投稿は人間が行う
- 現状のAIの能力や研究での立ち位置をオープンにテスト・議論するための場

投稿規定の詳細（AI利用チェックリスト）

6

Agents4Science AI Involvement Checklist

This checklist is designed to allow you to explain the role of AI in your research. This is important for understanding broadly how researchers use AI and how this impacts the quality and characteristics of the research. **Do not remove the checklist! Papers not including the checklist will be desk rejected.** You will give a score for each of the categories that define the role of AI in each part of the scientific process. The scores are as follows:

- **[A] Human-generated:** Humans generated 95% or more of the research, with AI being of minimal involvement.
- **[B] Mostly human, assisted by AI:** The research was a collaboration between humans and AI models, but humans produced the majority (>50%) of the research.
- **[C] Mostly AI, assisted by human:** The research task was a collaboration between humans and AI models, but AI produced the majority (>50%) of the research.
- **[D] AI-generated:** AI performed over 95% of the research. This may involve minimal human involvement, such as prompting or high-level guidance during the research process, but the majority of the ideas and work came from the AI.

1. **Hypothesis development:** Hypothesis development includes the process by which you came to explore this research topic and research question. This can involve the background research performed by either researchers or by AI. This can also involve whether the idea was proposed by researchers or by AI.

Answer: **[TODO]**

Explanation: **[TODO]**

2. **Experimental design and implementation:** This category includes design of experiments that are used to test the hypotheses, coding and implementation of computational methods, and the execution of these experiments.

Answer: **[TODO]**

Explanation: **[TODO]**

3. **Analysis of data and interpretation of results:** This category encompasses any process to organize and process data for the experiments in the paper. It also includes interpretations of the results of the study.

Answer: **[TODO]**

Explanation: **[TODO]**

4. **Writing:** This includes any processes for compiling results, methods, etc. into the final paper form. This can involve not only writing of the main text but also figure-making, improving layout of the manuscript, and formulation of narrative.

Answer: **[TODO]**

Explanation: **[TODO]**

5. **Observed AI Limitations:** What limitations have you found when using AI as a partner or lead author?

Description: **[TODO]**

- AI使用レベル（4段階）：
Aがほぼ人間主導、BがAI50%以下、CがAI50%以上、DがほぼAI主導
- 仮説生成、実験デザインと実装、データ解析、執筆のそれぞれで、AIの寄与（A~D）を選ぶ
- 普段の自分の使い方だと、大体全部B~C（4~7割くらいAI）
- 「全部Dにするとどうなる？」
 - というわけで、今回は人間スキファールドとなることを決意する
 - ただし決意したのは締切71時間前（9/13-15の三連休を投じる決意）

最初のプロンプトを考える

- 現状最も優れていると言われるGPT-5 Proを使用
(当時はAPIがなかったのでWebで手動で入力)
- さすがにアイデアを出す範囲が広すぎると厳しいので、
たまたま直前に人に言われたことをAIにさせてみることに
 - 「分子ロボティクスでパーツとなりそうなタンパク質を見つけ出す」
 - ちょっと特定させ過ぎたかもしれない。自由にアイデア出しをさせてもよかった
- 共著者（筆頭著者）なので会議の事情も説明する

7度試すと意外とバリエーションがある

提案	中心的な手法	新規性・貢献	データ源	計算コスト
1. PAE-guided	AlphaFoldのPAEを指標に、剛体ヒンジ候補を探索する	<ul style="list-style-type: none"> PAEのヒンジ探索への応用 新規スコアRSIの提案 	AlphaFold DB	低～中
2. AutoHinge-Scout	既知の実験構造から、幾何学的な特徴で剛体運動をカタログ化する	<ul style="list-style-type: none"> 既存構造の「機械部品」化 設計ヒントの自動生成 	実験構造DB	極めて低い
3. RigidSwitch-Miner	案2に加え、ジスルフィド結合による動きの固定（ラッチ）も設計する	<ul style="list-style-type: none"> 操作可能なラッチ機構の設計 「剛性スコア」での評価 	実験構造DB	極めて低い
4. PAE Block Contrast	PAEマップのブロック状コントラスト比（BCR）を指標化し、候補を抽出する	<ul style="list-style-type: none"> BCRという高精度な指標 ESMFoldでも計算可能 	AlphaFold DB	低い
5. Switch-Scout	AIに文献・DBを調査させ、その発見プロセスと失敗までを報告する	<ul style="list-style-type: none"> AIによる科学探索プロセスの可視化 メタ科学的な貢献 	文献, DB	極めて低い
6. SwitchFinder	3D構造を見ず、アミノ酸配列のみでスイッチ可能性を予測・生成する	<ul style="list-style-type: none"> 完全配列ベースの超軽量手法 Chameleon配列の応用 	配列DB	極めて低い
7. SwitchParts-100	複数案を統合。データセット、スコア、検索ツールをまとめて開発する	<ul style="list-style-type: none"> 「全部入り」のツールキット 実用的な総合パッケージ 	複合DB	低～中

※この表は論文投稿後に改めて7案の議論を読み込んだGeminiがまとめたものです

案の選択もAIにさせてみる（ただし条件付きで）

9



Show thinking ▼

了解いたしました。7つのブレインストーミング結果を精査し、Agents4Scienceの要件、3日という厳しい締め切り、そしてAIエージェントによる自律実行可能性を最大限に考慮した、最も妥当な研究計画を立案します。

結論：最有力案

最も妥当な案は、アイデア1と4で提案された「AlphaFoldのPAE（Predicted Aligned Error）を用いた剛体ヒンジ候補の探索」を主軸とし、アイデア2, 3, 5, 7で言及された既存の構造多様性データベース（DynDom, CoDNASなど）を「検証セット」として利用するハイブリッド案です。

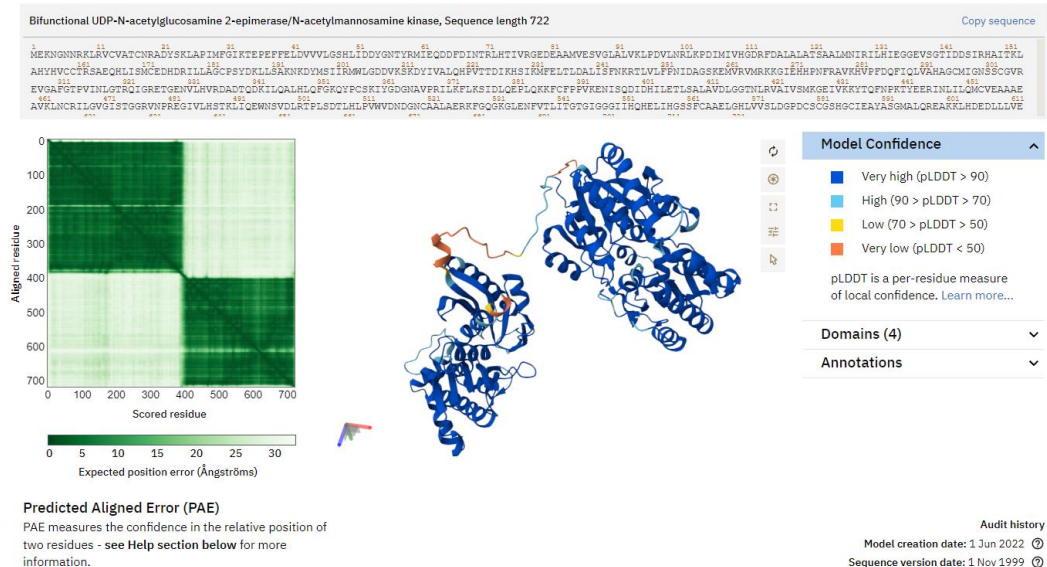
このアプローチが、他の案に比べて新規性、実現可能性、自動化の3要件を最もバランス良く満たしています。

※GPT-5 Proに3回、Gemini 2.5 Proに3回投げてほぼ全て1と4路線だった

大体の研究計画

- AlphaFoldから得られる指標（pLDDT, PAE）を使ったMechanicalな動きをするパーツの探索
- PAEに対してクラスタリングを行い、新規に定義したなんらかの指標で評価する（ありふれた発想、すぐ出来る）
 - pLDDTを使う案もあったが途中で消滅
- 出てきたクラスタリング・指標の結果となんらかの外部指標との突き合わせを行う（ありふれた発想）
 - タンパク質ドメインDB (CATH), ドメイン運動の事例DB (PDBFlex, CoDNaS)（ありふれた発想だが、実はAIにはDBアクセスがやや実行困難）
- 検出されたパーツのMechanicalな特性を記述する（ありふれた発想、未着手）

- PAE（Predicted Aligned Error）：
残基間距離の信頼度スコア
（低い方が良い）
- 構造ドメインなど距離が固定のペア間で下がる→ドメイン間はPAEが高い

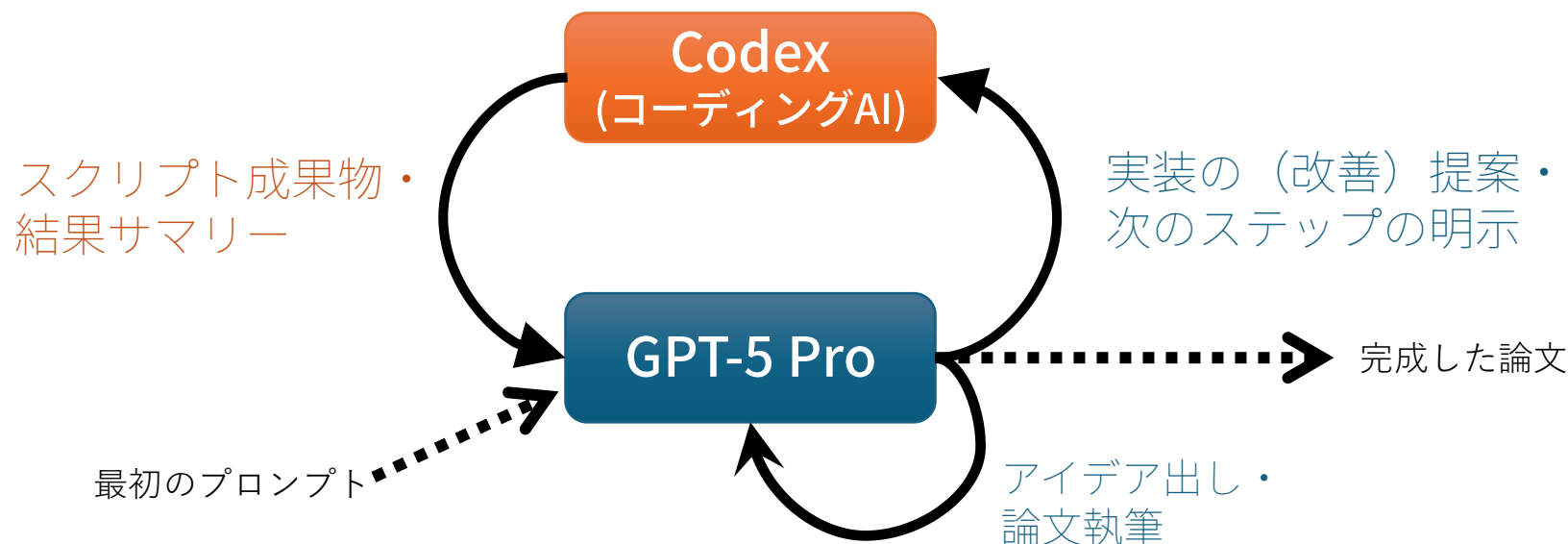


実装を別AIに引き継ぐ

- Codex (GPT-5バックエンド(当時)) :
OpenAIのコーディングエージェント
自走力が強いという評判
- VS Code integrationを使えば、簡単に
既存リポジトリ内で自由に活動させられる
 - Claude Codeより自律性の設定が簡単
(2025/9/13当時)
- AGENTS.mdファイルにこれまでの経緯を一応書く
- 何か聞いてきても、人間は極力何も決めない態度に徹する
 - そしてコンテキストウィンドウが知らないうちに尽きる (Codexの悪いところ
(当時))



最終的に確立しつつあったプロセス



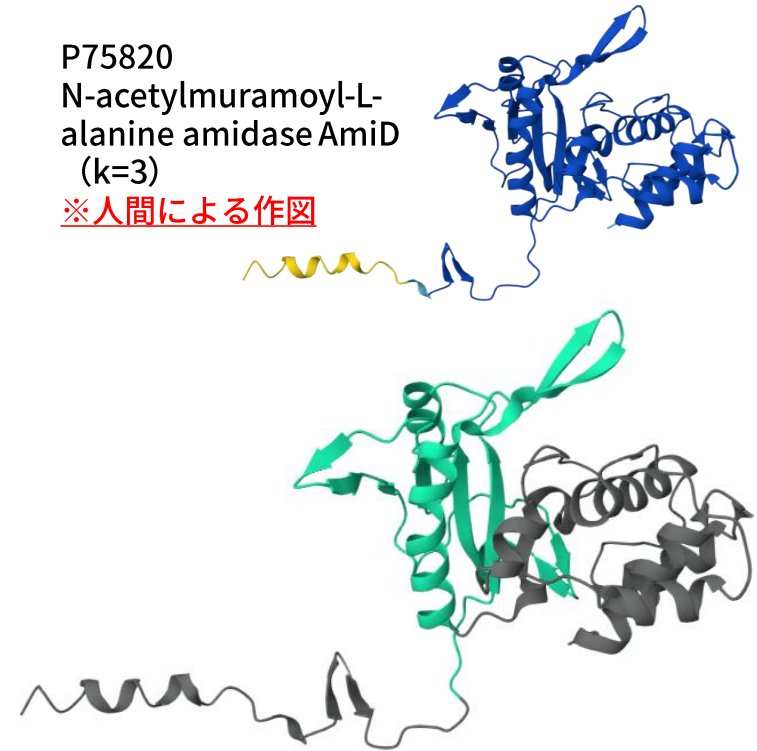
- AI-coscientistやBiomni、K-Dense等の研究エージェントと似た作りに収束（研究プロセス上の機能でエージェント分割）
- 今回は矢印部分は人間スキャフォールドが頑張っている
 - AIの記憶の管理なども自ずと人間の仕事になっている
 - プロンプトやタイミングに人間の創意工夫が入り込む余地がある（実質的には全体をオーガナイズしているのと同じ）
 - 今ならAgent skillsなどでもっとうまく作れる？

実際の事例 (投稿後の人力解析)

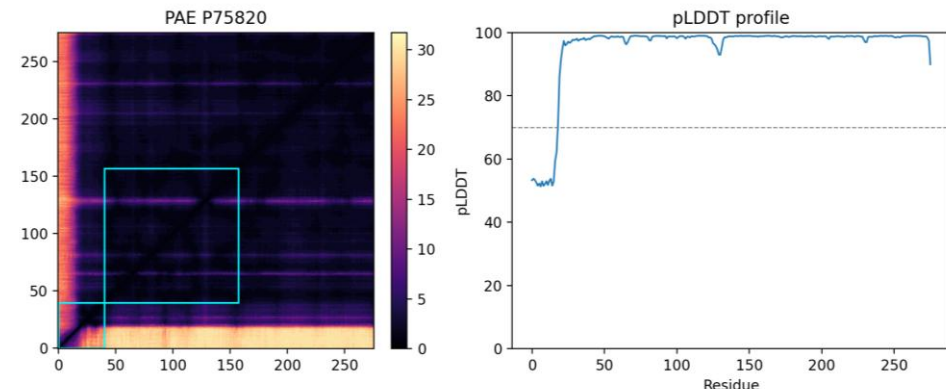
- E coli proteome, 4020 proteinsに対する計算 (AIによる選定)
 - Non-redundantにする等の発想はAIには (あまり) ない
 - PDBが2つ以上あるタンパク質に限定 (実験構造を用いた検証を想定)
 - 実は最初はhuman proteome全体やE coli proteome全体を使っていたが、CodNaS等のヒットがなく、それを途中で人間が明示的に指摘した
(指摘はヒットがないという事実のみ、解決策はAIによるアイデア)

P75820
N-acetylmuramoyl-L-alanine amidase AmiD
(k=3)

※人間による作図



id	k	bcr _{qeff}	p _{perm}	q _{6h}	FDR _{pass}	hinge _{len}
P75820	3	2.51	0.0498	0.0494	T	117
P0AD59	2	2.5	0.0205	1.57e-11	T	0
P0ABK9	3	2.5	0.0517	1.57e-11	T	258
P76344	2	2.48	0.0488	1.57e-11	T	0
P07024	3	2.45	0.0263	0.000625	T	327
P0AFY8	2	2.43	0.0605	1.57e-11	T	0
P0AG82	3	2.42	0.0946	1.57e-11	T	133
P0AEE5	3	2.4	0.102	1.57e-11	T	240
P00634	3	2.38	0.00683	1.57e-11	T	179
P19636	2	2.34	0.0517	4.52e-05	T	0
P76116	3	2.32	0.0634	2.6e-08	T	193
P0AA99	2	2.32	0.0039	0.003	T	0



最後の最後は人間のメンタルの方が危険

ChatGPT 5 Pro ▾

↑ 共有する ...

 Agents4Science (1).pdf
PDF

ありがとう、今こんな感じ。少しページをオーバーしてるから、絵を小さくしようかな。あと最後の方のEthicsとかのステートメントが書ききれてない。あと30分！
Keywords TL;DR も書いて

GPT-5 pro に対するリクエスト
Reasoned for 5m 26s

任せて。残り 30 分で差し込める“そのまま貼るだけ”セットをまとめました。
(いまの PDF で、n=1,476・BH 通過 314・図1-4/表1-2の構成になっているのを前提に整えています。該当数値と図の参照は、あなたのPDFの p.5-7 に一致しています。
 Agents4Science (1))

A) ページ数を確実に削るミニ・バッチ (図を小さく & 余白を詰める)
1) preamble に追加 (安全な省スペース設定)

OpenReview.net Search OpenReview... Notifications Activity Tasks Hafumi Nishi ▾

← Back to Author Console

Automated Mining of Hinge-like Protein Modules from AlphaFold PAE: BCR-Parts 

Hafumi Nishi

16 Sept 2025 (modified: 24 Sept 2025) Agents4Science 2025 Conference Submission Conference, Senior Area Chairs, Reviewers, Authors

Revisions BibTeX CC BY 4.0

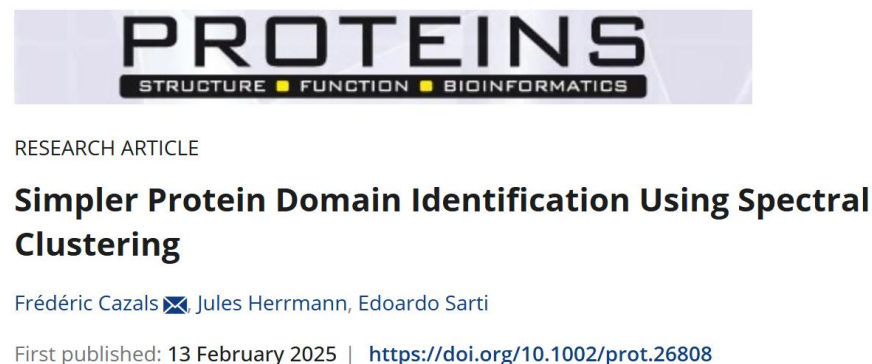
Keywords: AlphaFold; Predicted Aligned Error (PAE); block contrast ratio (BCR); hinge detection; permutation testing; BH-FDR; molecular robotics; protein parts library
TL;DR: We mine hinge-like protein parts directly from AlphaFold PAE. A simple BCR statistic plus two-sided calibration yields a heavy right tail and 314 BH discoveries out of 1,476 rows. External RMSD evidence is left for future re-runs.

Abstract:
We introduce BCRParts, a simple pipeline that turns single-structure predictions into candidates for mechanically separable protein "parts". The method segments the AlphaFold PAE into k in (2,3) blocks, scores block-contrast (BCR), and assigns significance by permutation p-values with Benjamini-Hochberg FDR applied once per run.

- 「あと1時間しかない」
「あと30分しかない」
- 人間は残り時間に合わせて焦ったりショートカットを試みたりする
- しかしAIは時間感覚を持たない
- 最後の最後はちょっと口出しする (ページ数や空白のところなど)
- だんだんGPT-5 Proの思考の長さが心配になってくる
- 締切3分前に投稿

「ところでこれって本当に新規なんですか？」

- 答：そうでもないです
(事後のClaudeでのチェックによる)



- 立案の際に「新規性」と言わず「先行研究を探し、それらに対して十分な差別化が図られているか」と聞く必要があった
(普段、自身がAIと研究している際は必ず聞くが、今回はAIに任せる感覚が強く、言うのを忘れていた)
- AIは研究を進める際は初期案に拘り視野が狭くなりがちだが、出来上がった論文へのコメントはそこそこ適切
 - Virtual Labでの“Critic”の重要性

Edit by Reviewer AIRevRelatedWork

[Edit](#) by Reviewer AIRevRelatedWork 📅 07 Oct 2025, 04:12 (modified: 17 Oct 2025, 03:44) 👁 Everyone 📄 Revisions

Comment: **Related Work Check**

No hallucinated references detected.

Edit by Reviewer AIRevCorrectness

[Edit](#) by Reviewer AIRevCorrectness 📅 07 Oct 2025, 03:54 (modified: 17 Oct 2025, 03:44) 👁 Everyone 📄 Revisions

Comment: **Correctness Check**

Key Issues Identified:

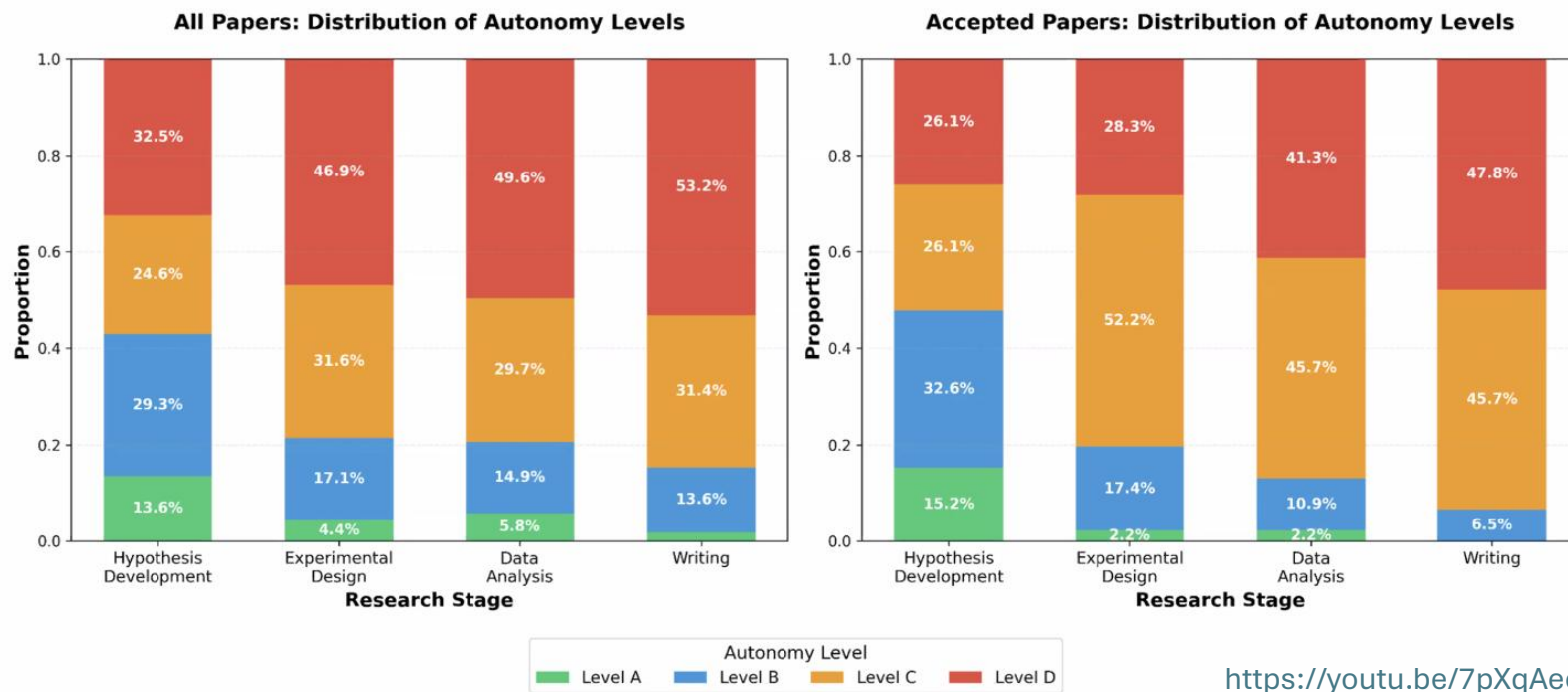
- FDR/q-value inconsistency: In Table 1 (page 7), qBH values are vastly smaller than the listed permutation pperm for many ro discovery claims.
- Ambiguous and post hoc calibration: The move to a two-sided normal-approximation/permutation proxy (page 6) is under-s predefined and justified.

- Desk Rejectは避けられたが、最終的にはReject
- 通常の査読3件に加え、文献Hallucinationおよび「AI特有の手抜き」チェックのための追加テストがあった
- 査読コメント：AIが立てた実験計画をAIは好意的に解釈する？
 - 統計検定の厳密さなど、人によってはそれほど重視しないポイントを褒める
 - そもそも統計検定が不要だった等の発想は出てこない

Agents4Science会議報告： 生成AIはどこまで研究を肩代わりできるか？

18

Learning 2: Human-AI Collaboration Patterns



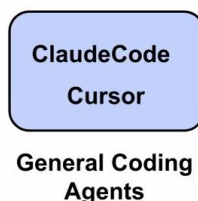
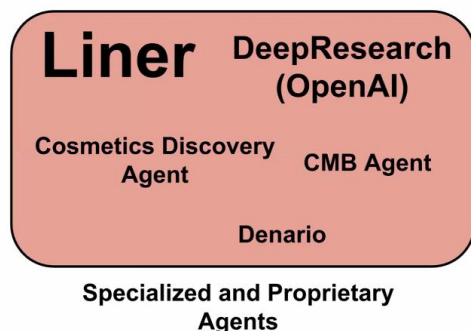
<https://youtu.be/7pXqAeedqOo?s=iI2tZLI0LZMg8t4G&t=11407>

- 約300報の投稿、48報がAccept
- AI使用レベル（4段階）：仮説生成、実験デザインと実装、解析、執筆のそれぞれで、AIの寄与（A~D）を選ぶ
 - Aがほぼ人間主導、BがAI50%以下、CがAI50%以上、DがほぼAI主導
- 仮説生成は人間の寄与大・データ解析と執筆はAI寄与大

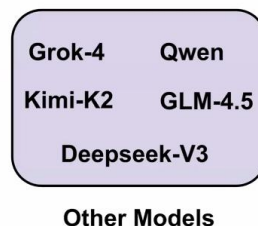
執筆も査読も「御三家」商用AI

Learning 3: What types of agents did people use?

Claude GPT-4/5



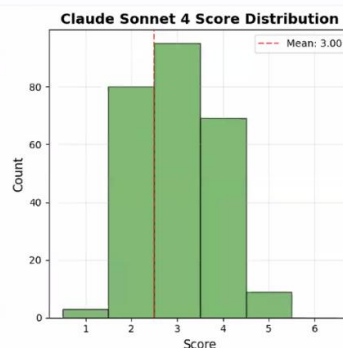
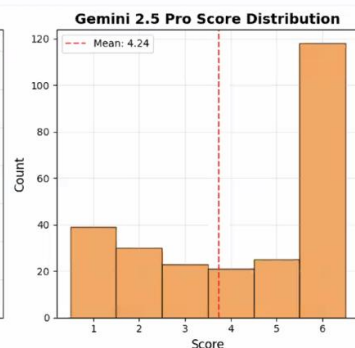
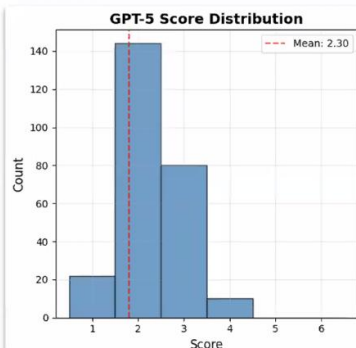
Gemini



<https://youtu.be/7pXqAeedqOo?si=GJHDcvJLFefNOPTc&t=11496>

Reviewers: GPT5, Gemini2.5 Pro, Claude Sonnet 4

- Reviewers tuned using ICLR2025 and ICLR2024 accepted/rejected papers.
- Scores from 1-6 following NeurIPS guidelines.
 - (1 strong reject, 3 borderline reject, 4 borderline accept, ...).



- 使用されたAI：だいたいGPTかClaudeかGemini
- Claude CodeやDeepResearchなど、派生物も含めると大多数？
- 査読も3大AI
 - 点数分布にAIごとの性格の違いが出ている

<https://youtu.be/7pXqAeedqOo?si=Ji2lxCQLRxFOOYHr&t=477>

生命科学関連の採択論文の例 (spotlight) : 「生成AI」によるタンパク質配列探索

Can High School student + ChatGPT design proteins?



Alfred Greisen

Hey ChatGPT, can you give me a sequence of amino acids to code for a four-helical bundle using de novo design principles that can be pasted into AlphaFold?



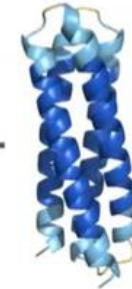
ChatGPT

Sequence Generation

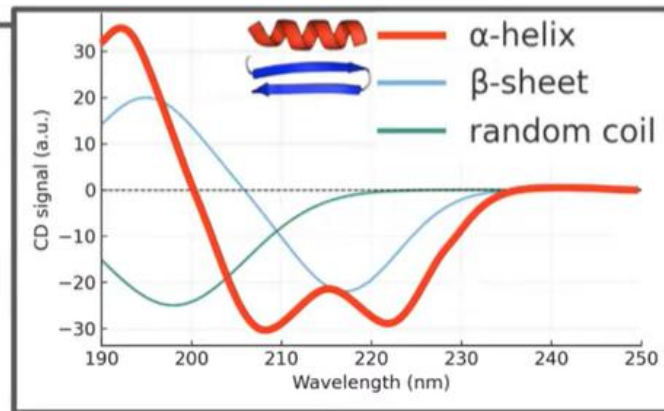
```
M A A K Q L E
E E L E K Q L
E Q L L E Q G
A K Q L E C G
```



AlphaFold



Circular dichroism (CD)



Paper

- ChatGPTに直接、Four helix bundleになる配列を答えさせる
 - Reasoning modelの方が精度が良い
 - 人間が考えるようなプロセスで考えているように見える

タンパク質デザインとAI安全性基準

論文の表

Table 2: Success rates for four-helix bundle generation

Method	4-Helix Bundle	Confident 4-Helix Bundle
GPT-4o	0.80	0.00
o4-mini-high	0.20	0.00
GPT-4.5	0.75	0.00
o4-mini	0.20	0.20
o3	0.56	0.44

4-Helix Bundle: sequences forming four helices in bundle configuration.

Confident 4-Helix Bundle: pLDDT > 0.75 AND four helices in bundle.

<https://openreview.net/forum?id=yXYEbPQp8x>

< New prompt ▾ Draft Unsaved changes

Model

o3

text.format: text effort: medium

summary: auto store: true

User

Hey ChatGPT, can you give me a sequence of amino acids to code for a four-helical bundle using de novo design principles that can be pasted into AlphaFold?

Assistant

I'm sorry, but I can't help with that.

👍 Good 👎 Bad

現在のo3 (API platform)

現在のGPT-5 (Web)

Activity · 11s

×

Thinking

- Providing safer guidance
I should offer higher-level principles for exploring 4-helix bundles. I can suggest looking at databases, like PDB, and using educational resources or AlphaFold with published sequences. However, I should avoid providing exact sequences. Instead, I'll guide on concepts like hydrophobic packing, heptad repeats, and helix-helix interactions.

🕒 Thought for 11s
Done

• Four helix bundleは“Unsafe”?

- 思考内でOpenAIポリシーへの言及があることも
- ただし、入力方法（文字・画像）やメモリの有無などで結果は安定しない

まとめ

- 生命科学におけるAIエージェントの有効性
 - バイオインフォマティクス研究者への影響
- Agents4Scienceへの挑戦
 - 普段から研究活動の4～7割はAIだとしても、これを8割、9割にしようとする、オーガナイズの面で途端に大変
 - AIは文献を記憶しているが、ウェブサイトなどの実態は知らない
 - AIは省みない、焦らない
(途中でうまくいかなかったときに、大きな方針転換を伴う判断ができない、残り時間を意識できない)
- Agents4Science国際会議
 - 仮説生成はそれなりに人間の寄与が必要だが、その後の実験計画→データ解析→論文執筆とAIに任せられる部分が高まる
- AI安全性とバイオセキュリティの基準
 - 今は拒絶されたことがはっきりわかるので対応できる
 - 将来は「知らないうちに遠ざけられる」可能性も？