

ゲノム言語モデル研究動向と 開発進捗状況

東 光一

国立遺伝学研究所

ゲノム言語モデル関連最新研究動向 (2025.09)

1. 新規モデル

PlantCAD2: 被子植物特化・長コンテキストー塩基解像度モデル

2. 応用

EiRA: 生体分子結合タンパク質特化のESM拡張、Evo2とのクロスモーダル統合 (gLM × pLM)

3. ベンチマーク

(ようやく出始めた) Evo2の性能評価研究

1. 新規モデル PlandCAD2

<https://www.biorxiv.org/content/10.1101/2025.08.27.672609v1>

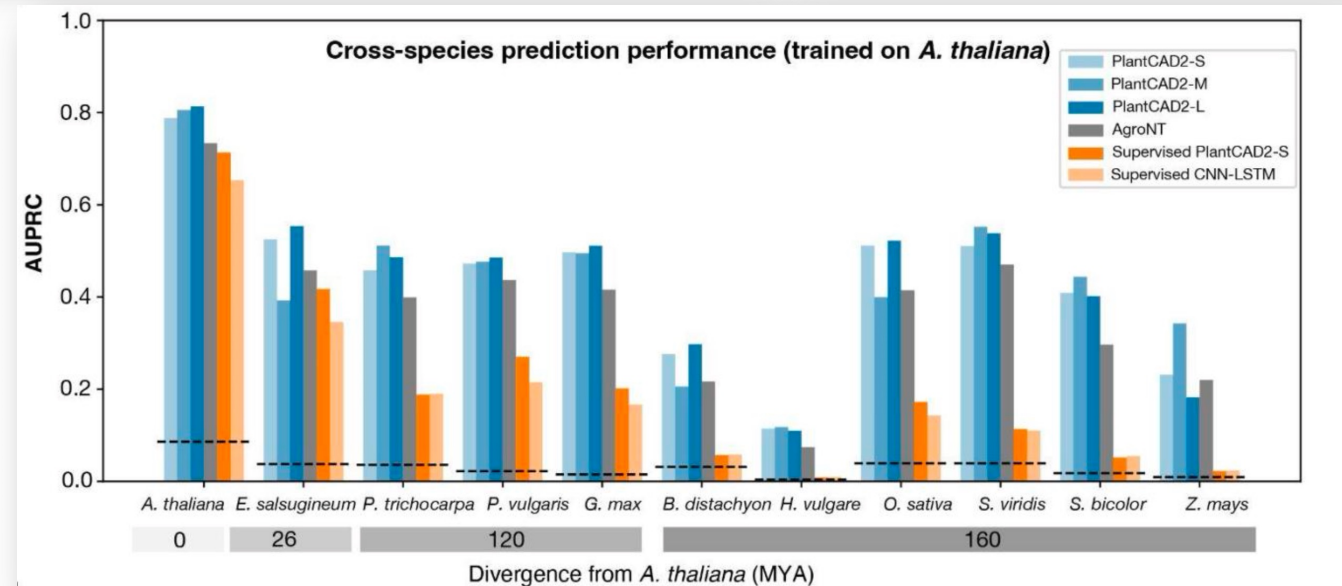
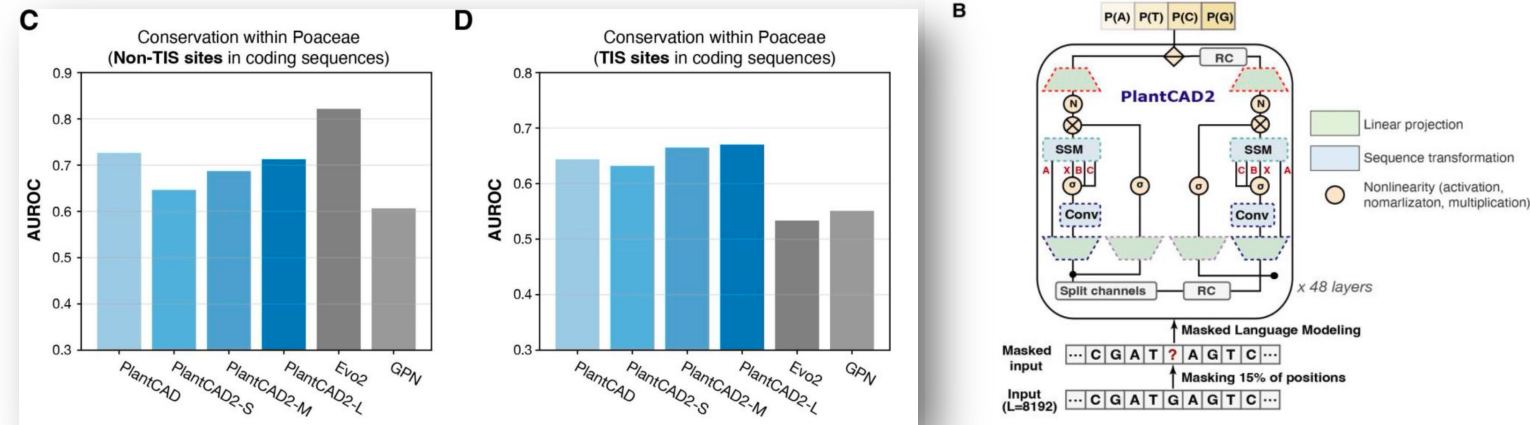
PlantCaduceus (双方向RC統合モデルCaduceusの植物版) の拡張。Mamba2ベース、MLMターゲット

パラメータサイズ: 676M
コンテキスト長: 8,192bp

学習:
65 被子植物ゲノム
遺伝子周辺±5kbpを重点的に切り出す
リピート配列領域はダウンサンプリング

評価:
進化的保存性 (MSAフリー) ゼロショット予測
制御エレメントゼロショット予測

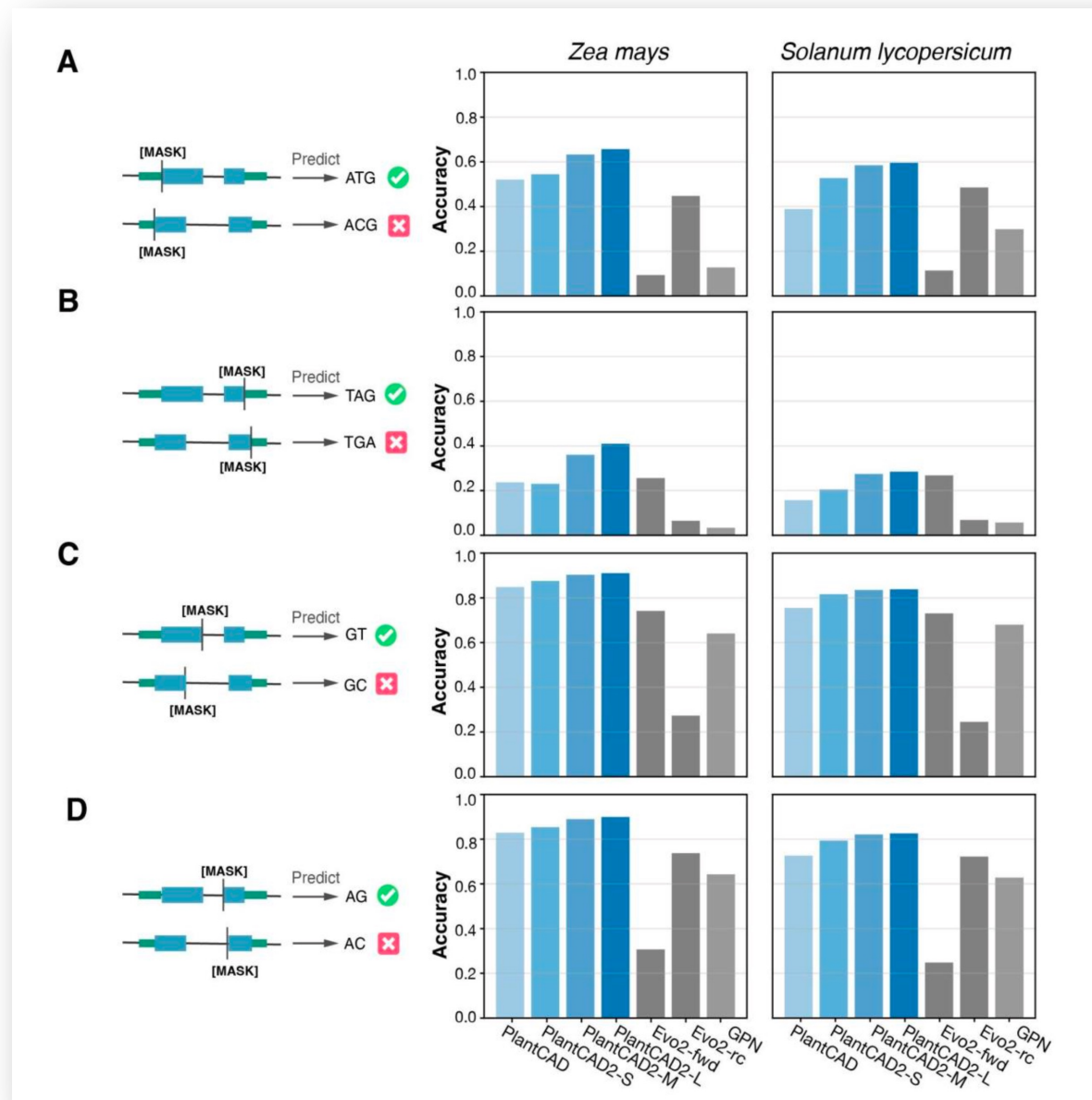
LoRAファインチューニングによる
Accessible regions (ATAC-seq peaks)予測



制御エレメントゼロショット予測。

Evo2は

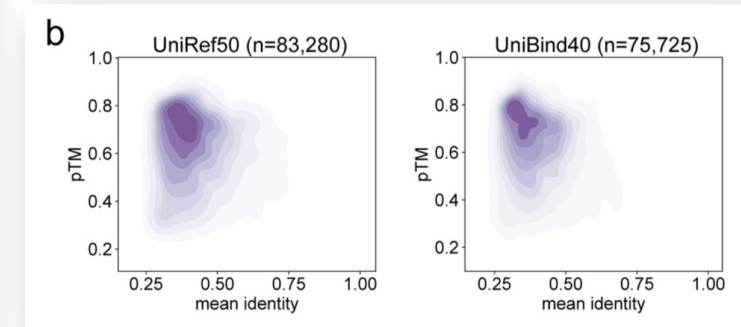
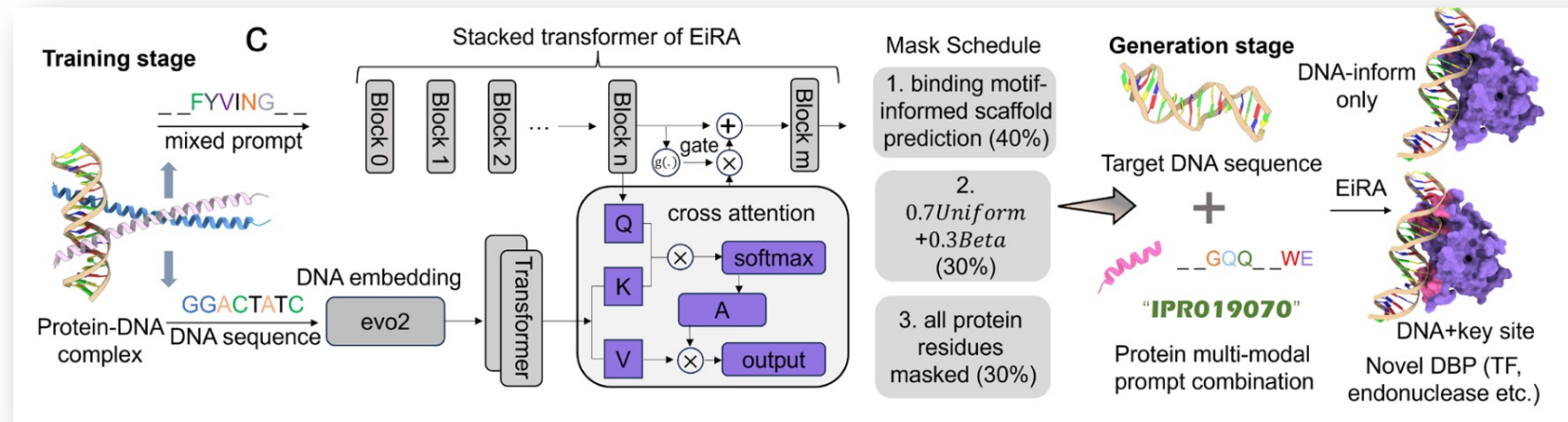
Forward推論、Reverse推論で極端な性能差。



2. gLM ✕ pLM (EiRA)

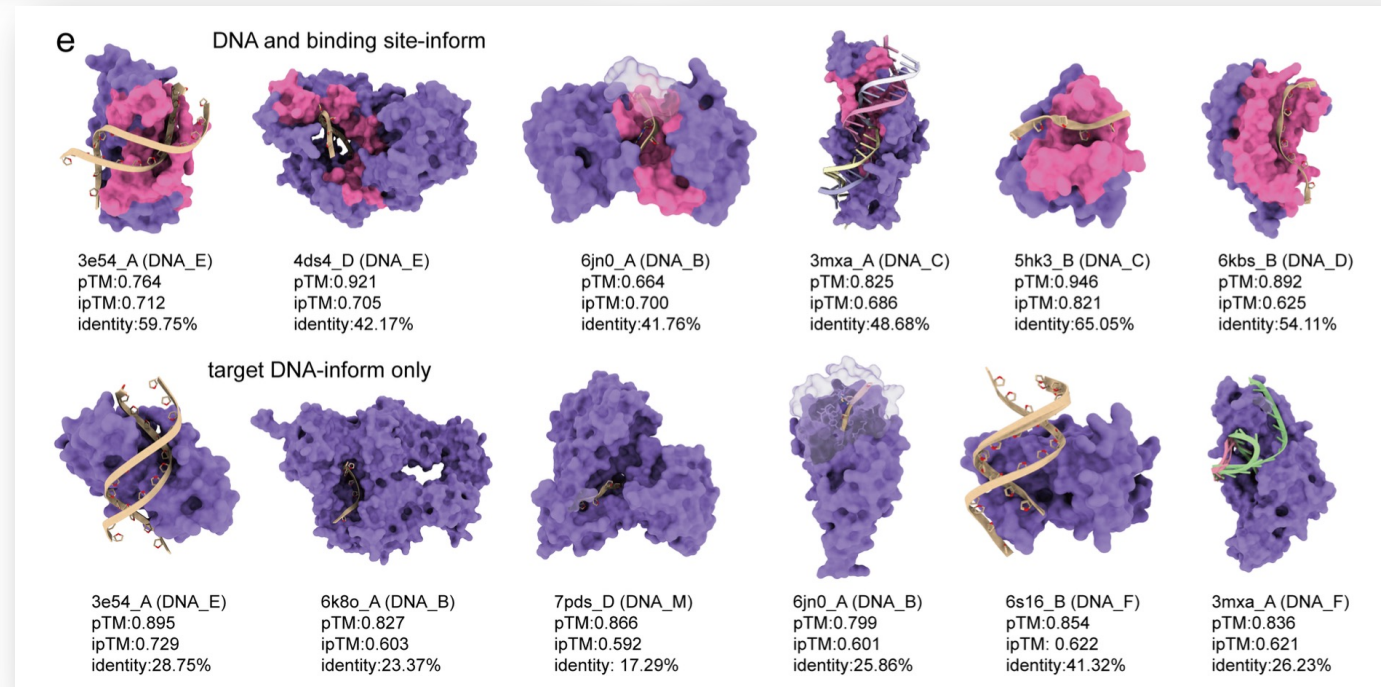
<https://www.biorxiv.org/content/10.1101/2025.09.02.673615v1>

EiRA: 生体分子結合タンパク質 (DNA, RNA, 金属など) でドメイン適応 (DPO) したESM3拡張。



Evo2パラメータを固定してEira最終層近くで Gated cross attention計算。
DNA配列で条件づけて、それと結合親和性の高いタンパク質を生成できる。

UniRef50/UniBind40と相同性が低いが、pTMが高い、De novo DNA結合タンパク質が多数生成された。



3. ベンチマーク関連

Research highlights

Bioinformatics

<https://doi.org/10.1038/s41592-025-02829-6>

Benchmarking genomic language models

Supervised deep learning models have a regulation," notes Koo. "In contrast, most

依然として、「生成」を評価するベンチマーク・応用タスク性能評価が不足。

→の論文では、NT, Hyena, DNABERT2などの埋め込みを利用したエピゲノム予測のためのSFTが、one-hotエンコーディングの教師あり学習と性能において大差ないことを明らかにした。

また、そもそもEvo2などの大規模モデルは（推論動かしだけでも大変なので）あまり検証されていない。

Tang et al. *Genome Biology* (2025) 26:203
<https://doi.org/10.1186/s13059-025-03674-8>

Genome Biology

RESEARCH

Open Access

Evaluating the representational power of pre-trained DNA language models for regulatory genomics

Ziqi Tang¹, Nirali Somia¹, Yiyang Yu² and Peter K. Koo^{1*}



3. ベンチマーク関連

Confirmatory Results

[Follow this preprint](#) [Previous](#)

A Benchmark of Evo2 Genomic AI Models for Efficient and Practical Deployment

Posted September 12, 2025.

Huimin Li, Hongyi Ji, Yuchen Zeng, Wei Lv, Jianmin Wu, Sheng Liu, Chunhua Lin, Huanming Yang, Zhaorong Li, Yubao Chen, Wei Dong

doi: <https://doi.org/10.1101/2025.09.10.675279>

[Download PDF](#)

[Print/Save Options](#)

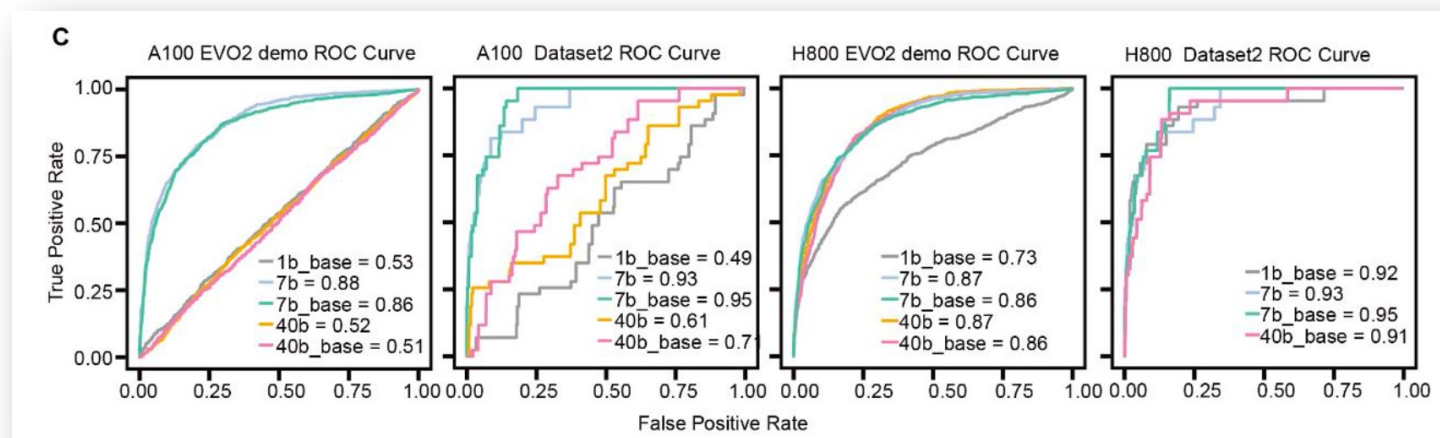
[Supplementary Material](#)

Evo2のがんゲノム関連タスク（腫瘍変異の病原性予測）性能評価。
卵巣がんコホートデータなどでゼロショット予測。
H800 (FP8演算対応) と A100 (FP8演算非対応) GPUで予測性能比較。

FP8精度で計算しないとひどい性能になる。

（Evo2の公式にもそう書かれている。そもそもBlackwellで低性能報告もあるらしく、本当に数値計算的に安定したモデルなのか怪しく感じる）

いまのところ40Bは、Hopper (H100/H800) のFP8精度演算じゃないとまともな性能が出ないらしい。



モデリング進捗

- Evo1 7B相当モデル学習の分散学習基盤調査
 - Hyenaは簡単だが、7B程度の規模のモデルはマルチGPU・マルチノードの分散学習が必要
 - Savanna (<https://github.com/Zymrael/savanna>)
 - Hyena/MHAハイブリッドアーキテクチャの分散学習に最適化されたフレームワーク
 - データ並列、パイプライン並列、テンソル並列と、Evo2などの長コンテキスト (>1M tok) で必須になるコンテキスト並列の学習をサポート
- 「進化系統による条件付け」の手法検討
 - Evo2のように離散シンボルで与えるのではなく、系統樹上の座標をちゃんと数値表現して条件付き生成したい。

性能評価・応用タスク設計進捗

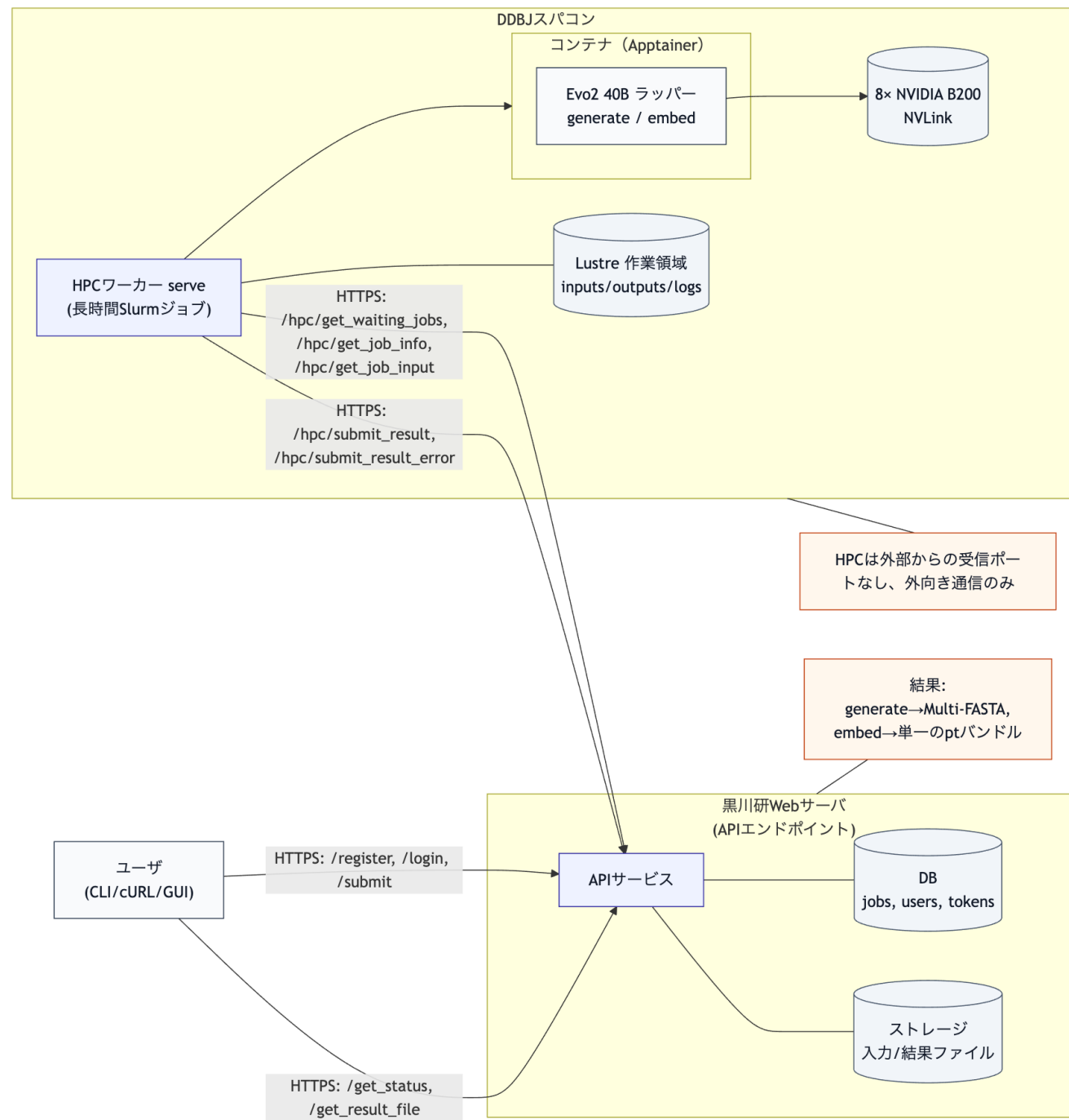
そもそもEvo2はモデル読み込みだけで数十秒～数分以上かかるので、評価チームの研究効率が上がらない。

⇒ API化

遺伝研スパコンにログインしなくても、スパコン上で常駐するEvo/Evo2の計算結果をREST API経由で簡単に取得できる環境を構築した。

Multi-FASTAを投げて

1. それらをプロンプトとした配列生成
2. それらの配列のEvo/Evo2内部の任意のレイヤーにおけるベクトル表現を高速に取得できるようにした。



評価・解釈性研究

 GOODFIRE

Research

Finding the Tree of Life in Evo 2

AUTHORS

Michael Pearce*
Elana Simon*†
Michael Byun*
Daniel Balsam*

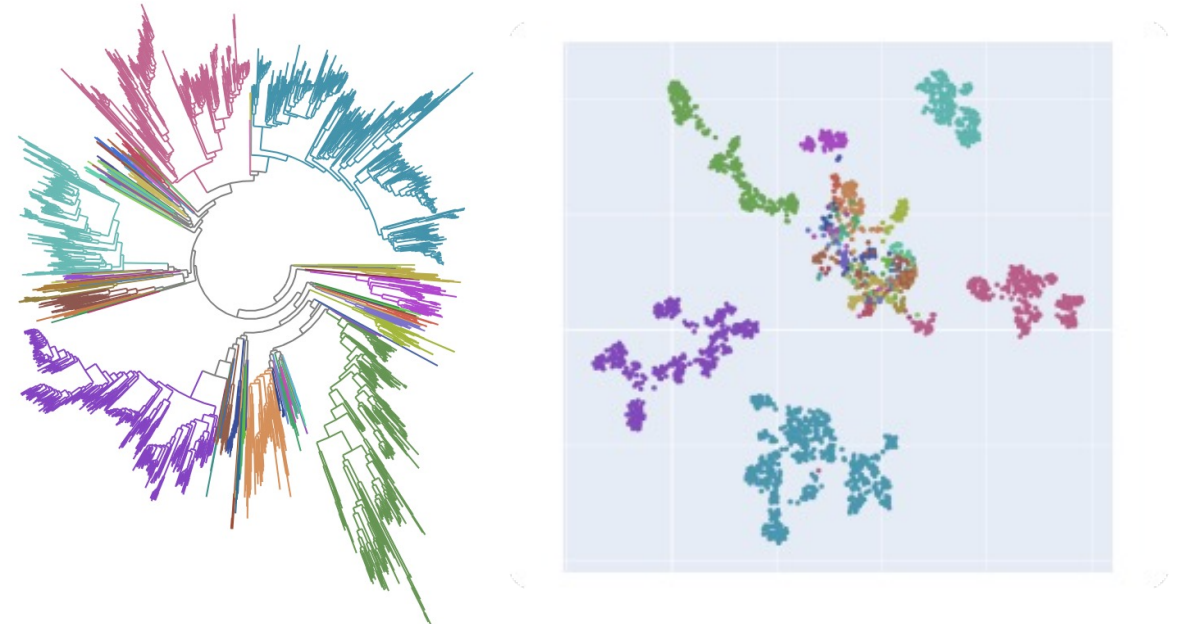
*Goodfire
†Stanford University

Correspondence to michael@goodfire.ai

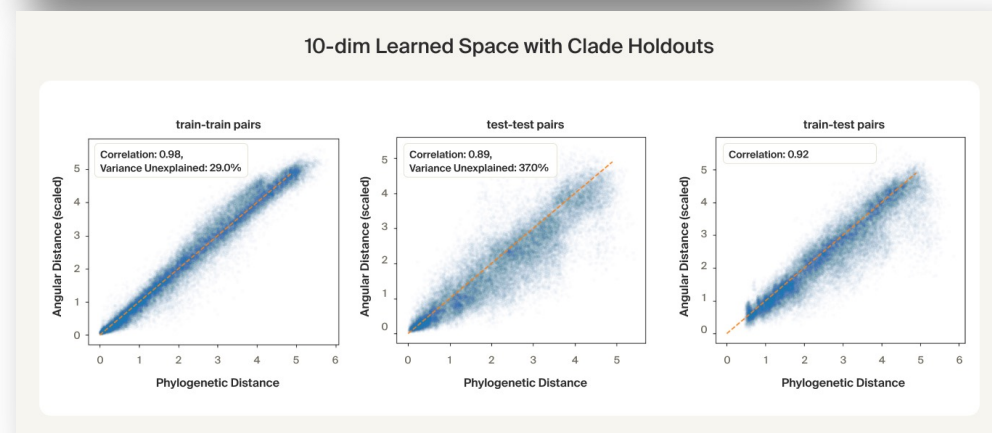
PUBLISHED

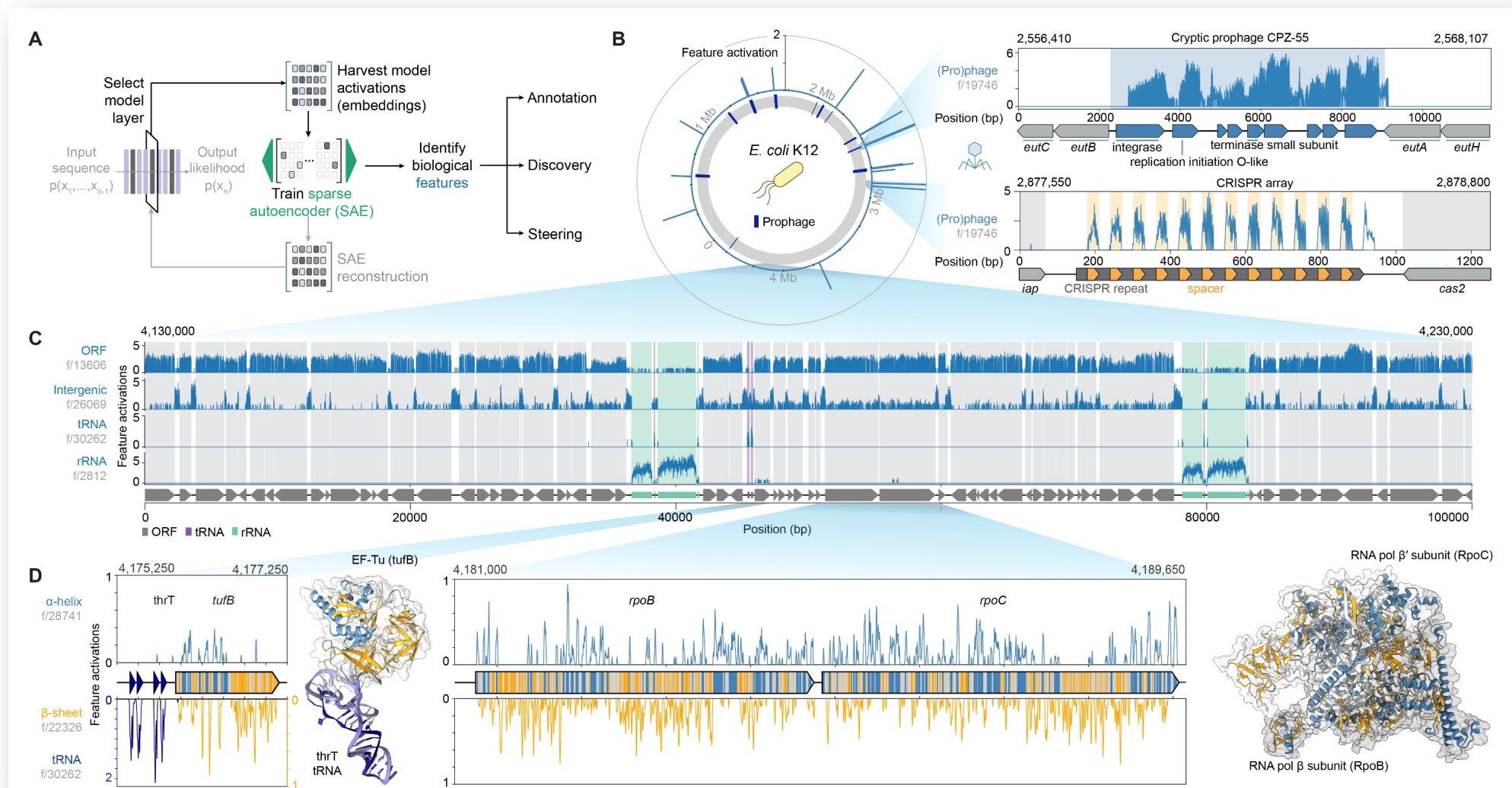
August 27, 2025

<https://www.goodfire.ai/research/phylogeny-manifold>



ゲノムごとにランダムに切り出した領域のForward演算（全32レイヤー中）レイヤー24の内部ベクトルを取得、平均化して「ゲノムベクトル」を構成、UMAP可視化距離学習によって、内部ベクトル表現だけからほぼ正確に進化距離を推定できる。





ゲノムを入力して埋め込まれた潜在変数を、スパースオートエンコーダで評価。
ゲノム上の領域ごとに、どのニューロンが発火するかを詳細に見ることができる。
多くのニューロンが既知のなんらかのゲノム上のパターンを表現している。
個別の特徴に特化して手作りのHMMなどのモデルはもはや不要？