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の最近の研究

Slackで日々、開発メンバーのみなさまが積極的に出版論文・プレプリントの情報交換をしてくれているので、その中からいくつかピックアップ。

・ファージ生成

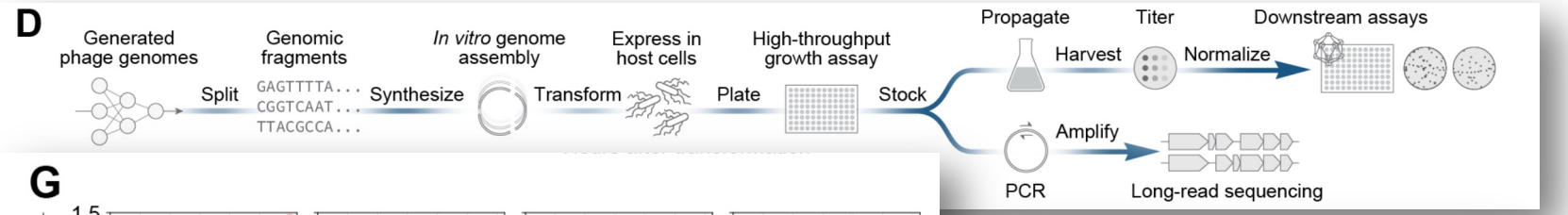
・モデル・配列評価（ヌクレオチド依存性マップ）

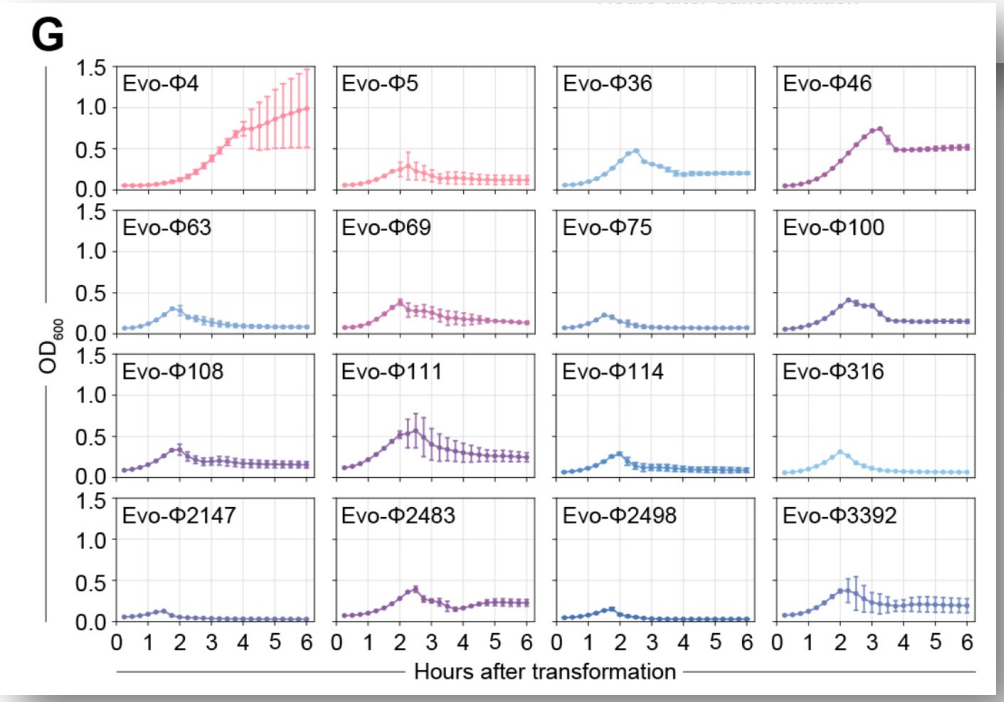
・安全性・バイオセキュリティ研究

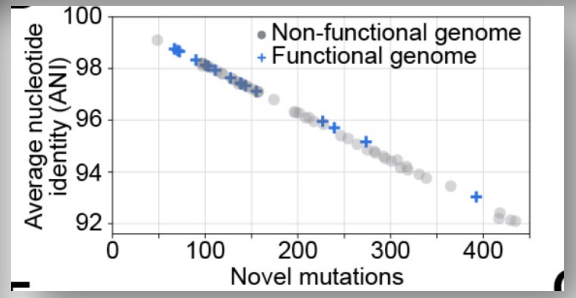
ファージゲノムのAIデザインと実験的合成・機能検証

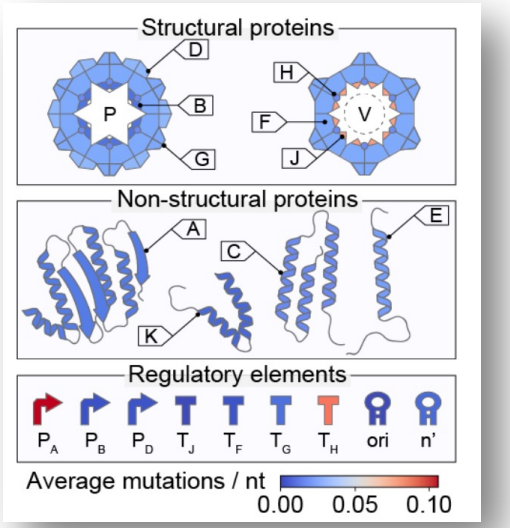
Evo1/Evo2を利用して、CRISPRやTnのような単一・少数の遺伝子の組み合わせではなく、
より多くの相互作用の完全性を維持しながら新規配列を合成できるか、という検証。　=> Generative Genomics

Microviridaeデータ（～15,000ゲノム）によるEvo1/Evo2のSFT（GPU: 32 × H100）
ΦX174の開始コンセンサス配列（5~10bp程度）をプロンプトとすることで、ΦX様ゲノムが生成できる
プロンプト、推論パラメータの調整によって「創造性」と「機能的制約」をバランス
基本的な遺伝子配列品質で情報的にスクリーニング => 302個候補ゲノム
プラークアッセイやODで実験的にスクリーニング => 16個候補ゲノムが増殖阻害
（うち、ロングリードで検証し、9ゲノムはデザインした配列と完全一致）
ΦX174を上回る適応度（溶菌効率）のゲノムも。









https://www.biorxiv.org/content/10.1101/2025.09.12.675911v1

モデル評価・配列評価の手法：ヌクレオチド依存性解析

gLMの予測を利用して、塩基間の相互依存性を評価する。
ある位置 i の塩基を別の塩基に変えたときに、離れた位置 j のATGC予測分布の変化を定量。

nature genetics

Article

https://doi.org/10.1038/s41588-025-02347-3

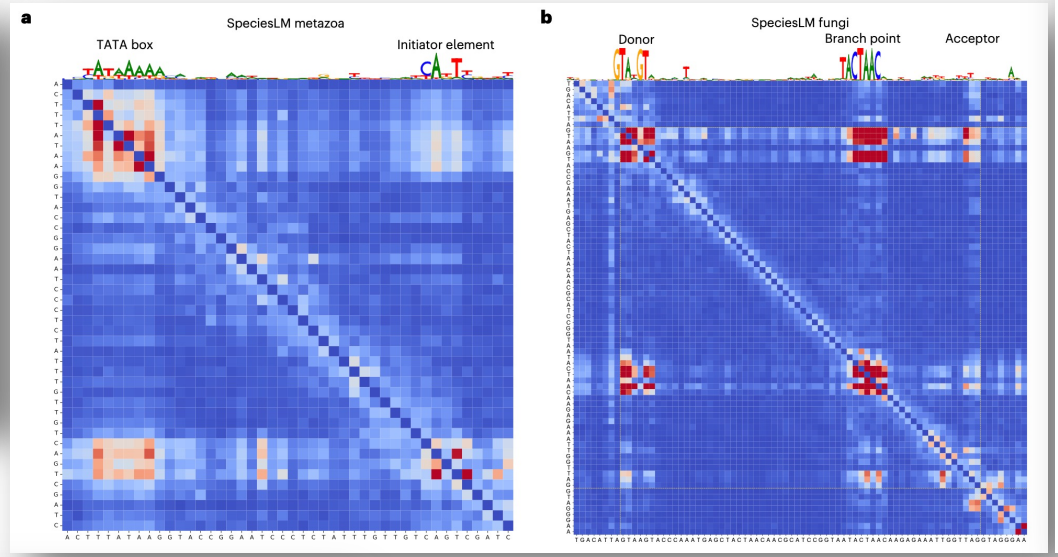
Nucleotide dependency analysis of genomic language models detects functional elements

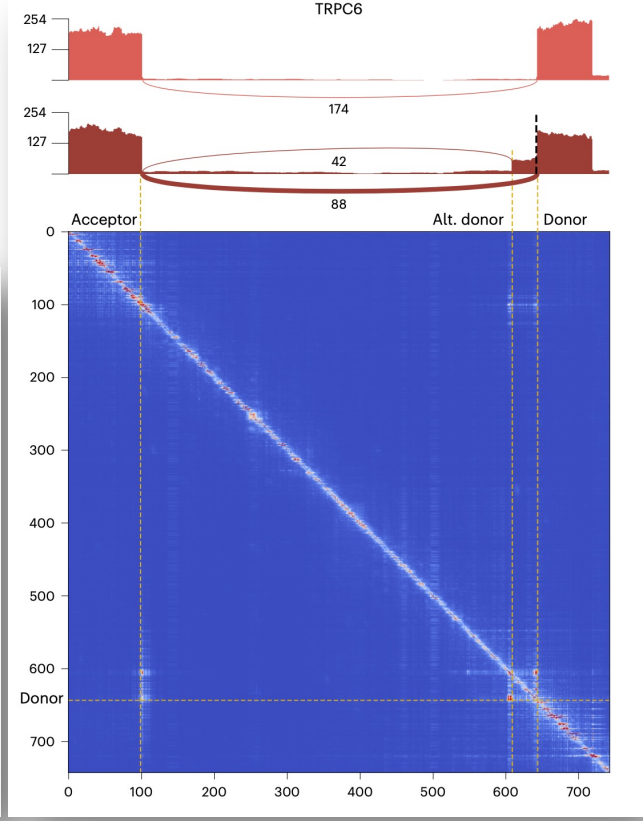
Received: 24 September 2024

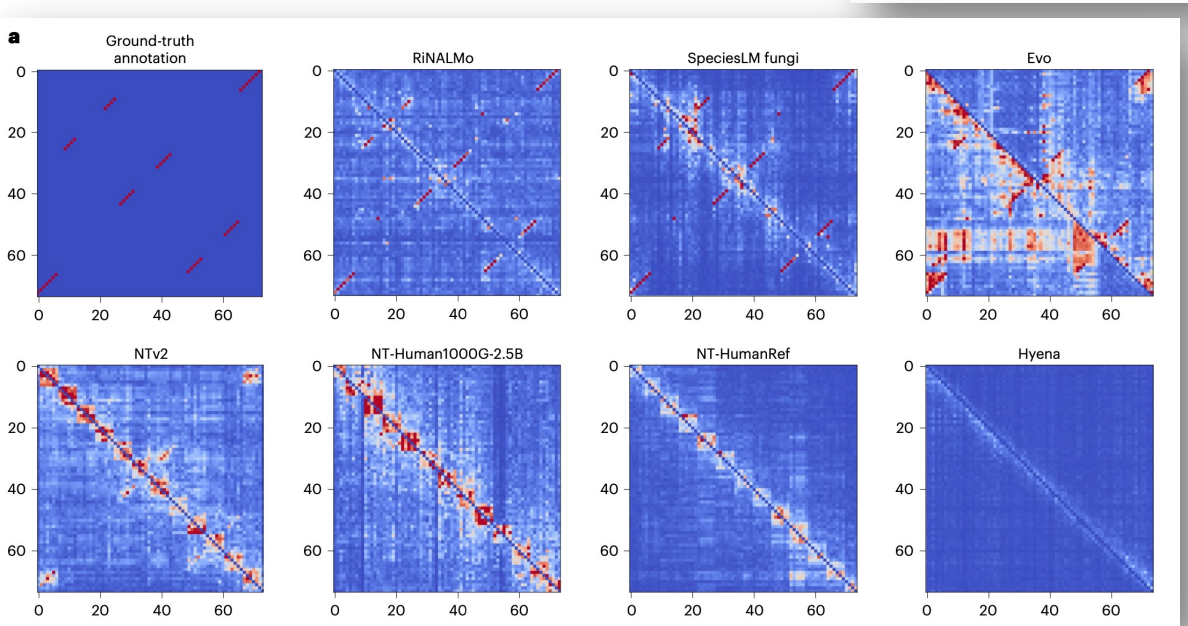
Accepted: 28 August 2025

Published online: 10 October 2025

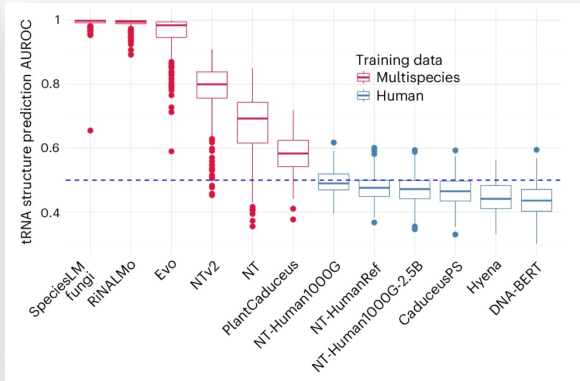
Pedro Tomaz da Silva[1,2,8], Alexander Karollus[1,2,8], Johannes Hingerl[1,2], Gihanna Sta. Teresa Galindez[1,3], Nils Wagner[1], Xavier Hernandez-Alias[1,4], Danny Incarnato[5] & Julien Gagneur[1,6,7]







依存性マップはRNA二次構造も反映。
ヒトtRNAの依存性予測は、ヒトのみのデータで学習するよりも複数種で学習した方がより依存性の予測精度が高い。
Evoなどの自己回帰モデルは双方向文脈を考慮しないのでちょっと無理があるか。



ゲノム生成の安全性・バイオセキュリティ

Evo2のジェイルブレイク

Evo2はそもそも訓練データからヒト感染ウイルス配列を除外しているが、
それらのパターンを汎化・再構築する能力を持ってしまっているので、
うまい方法で誘導することでヒト感染ウイルス配列を生成できてしまう
出力時のガードレール機能（有害性フィルタリング）が必要ではないか。

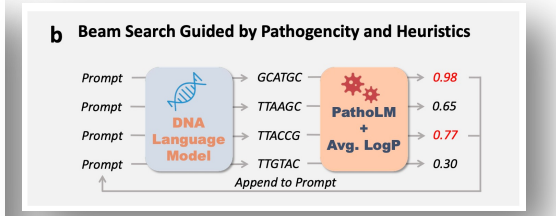
GeneBreaker: Jailbreak Attacks against DNA Language Models with Pathogenicity Guidance

Zaixi Zhang[*,†] Princeton University zz8680@princeton.edu
Zhenghong Zhou[*] Shanghai Jiao Tong University lltzahd615@sjtu.edu.cn
Ruofan Jin[*,‡] Zhejiang University ruofanjin@zju.edu.cn
Le Cong[†] Stanford University congle@stanford.edu
Mengdi Wang[†] Princeton University mengdiw@princeton.edu

Abstract

DNA, encoding genetic instructions for almost all living organisms, fuels groundbreaking advances in genomics and synthetic biology. Recently, DNA Foundation Models have achieved success in designing synthetic functional DNA sequences, even whole genomes, but their susceptibility to jailbreaking remains underexplored, leading to potential concern of generating harmful sequences such as pathogens or toxin-producing genes. In this paper, we introduce GeneBreaker, the first framework to systematically evaluate jailbreak vulnerabilities of DNA foundation models. GeneBreaker employs (1) an LLM agent with customized bioinformatic tools to design high-homology, non-pathogenic jailbreaking prompts, (2) beam search guided by PathoLM and log-probability heuristics to steer generation toward pathogen-like sequences, and (3) a BLAST-based evaluation pipeline against a curated Human Pathogen Database (JailbreakDNABench) to detect successful jailbreaks. Evaluated on our JailbreakDNABench, GeneBreaker successfully jailbreaks the latest Evo series models across 6 viral categories consistently (up to 60% Attack Success Rate for Evo2-40B). Further case studies on SARS-CoV-2 spike protein and HIV-1 envelope protein demonstrate the sequence and structural fidelity of jailbreak output, while evolutionary modeling of SARS-CoV-2 underscores biosecurity risks. Our findings also reveal that scaling DNA foundation models amplifies dual-use risks, motivating enhanced safety alignment and tracing mechanisms. Our code is at https://github.com/zaixizhang/GeneBreaker.

Disclaimer: This paper contains potentially offensive and harmful content.

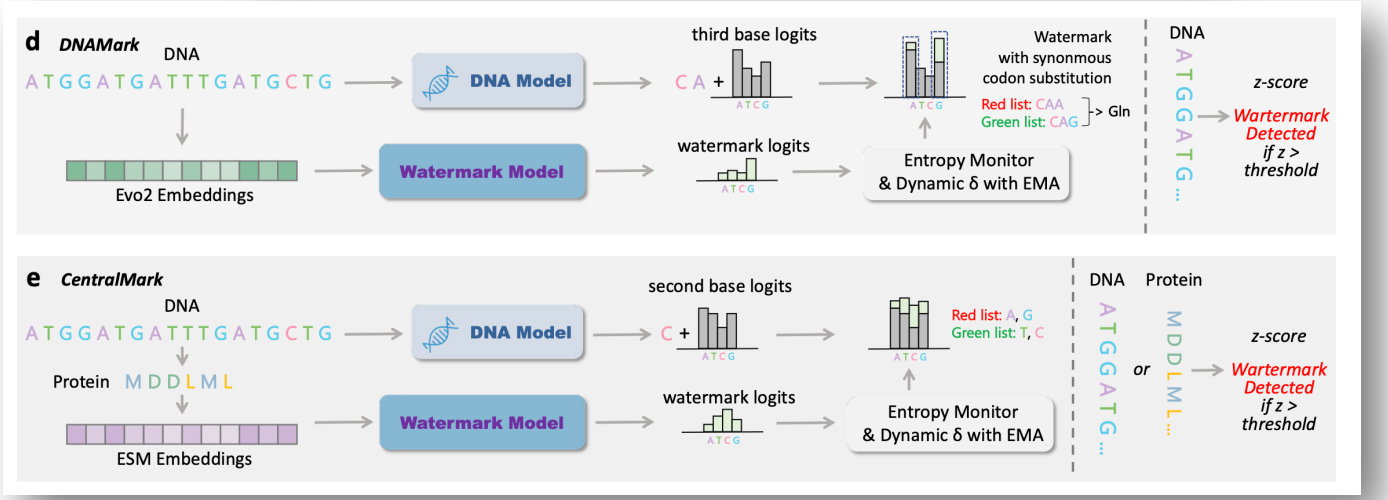


https://arxiv.org/abs/2505.23839

AI生成ゲノム配列にウォーターマーク（「透かし」）を入れる技術
天然の配列か、特定のモデルが生成した配列かを追跡可能になる？

Securing the Language of Life: Inheritable Watermarks from DNA Language Models to Proteins

Zaixi Zhang[*] Princeton University zz8680@princeton.edu
Ruofan Jin Zhejiang University ruofanjin@zju.edu.cn
Le Cong[*] Stanford University congle@stanford.edu
Mengdi Wang[*] Princeton University mengdiw@princeton.edu



https://arxiv.org/abs/2509.18207

ゲノム言語モデル開発進捗状況

遺伝研スパコンB200ノード
（ノードあたりBlackwell B200 GPU x 8, 4ノード合計32枚）を利用して
Evo2（StripedHyena2） 7B相当モデルのフルスクラッチ学習に取り組み中。
（1ノード8枚は推論計算に占有。3ノードを学習に利用）

データセットの構成は完了。
GlobDB由来高品質MAG（Metagenome-assembled genomes）データセット
約15万ゲノム、合計 約500B tokens

```
Input files         : 145486
Input contigs       : 18273440
Total bp (contigs) : 481,413,711,551
Kept chunks         : 49,941,497
Dropped (N-frac)    : 18,465  (threshold=0.1)
Dropped (short)     : 0  (min_contig_len=1000)
```

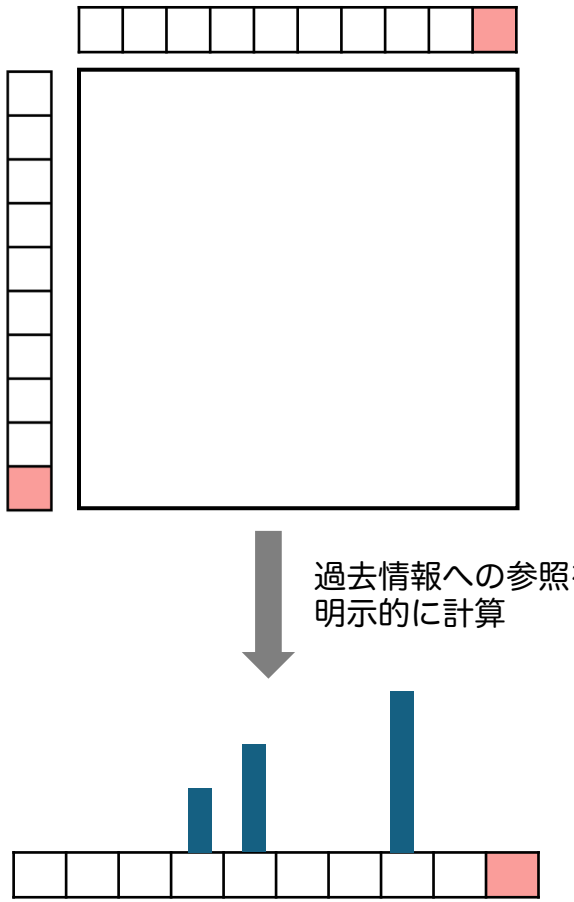
8,192bpチャンク分割、tokenization計算済み。
データリークしないようにゲノム単位で train/val/test 分割。
test/valやバッチに系統が偏らないように、
ANI距離に基づいてゲノムの並びを制御。

7Bモデル学習には分散並列学習基盤が必須だが、環境構築に失敗し続けている

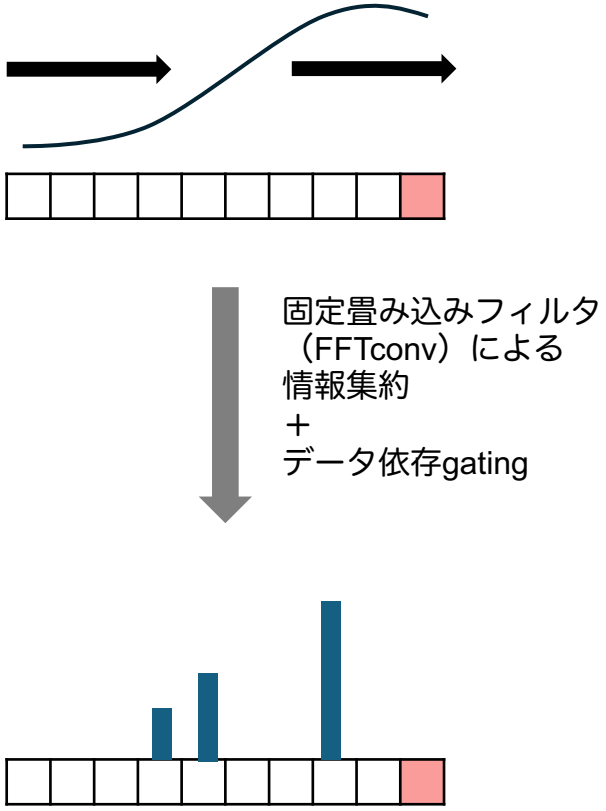
自己回帰型ゲノム言語モデル

現在時点の予測において、過去系列を「どのように参照するか」で戦略の違い

アテンション

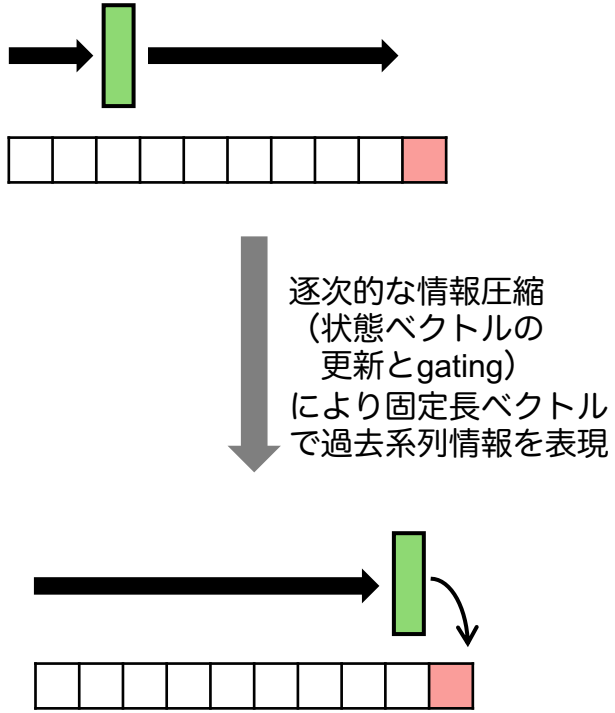


StripedHyena2



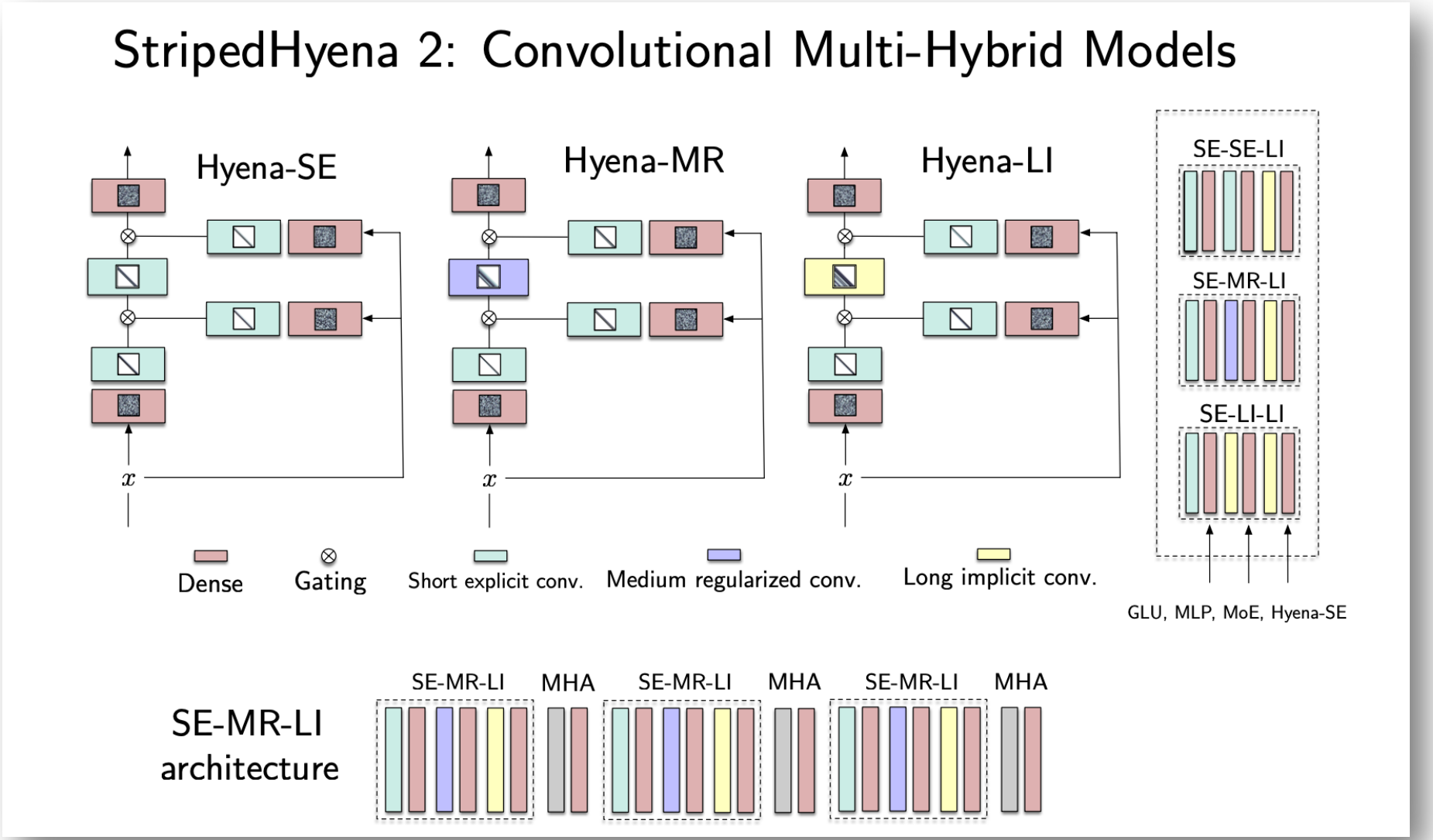
もともとのHyena論文ではフィルタは学習可能だが、
StripedHyena2（Evo2のアーキ）の一部フィルタは固定

MambaなどSSM系



StripedHyena2アーキテクチャ

https://arxiv.org/abs/2503.01868



Layout	PPL@400B
MHA-MHA-MHA	3.09
LI-LI-LI	2.87
SE-SE-LI	2.88
SE-MR-LI	2.83

Table 2.1: Effect of different block layouts on pretraining at the 7B parameter scale.

※SE, MRのカーネルサイズは
ハイパーパラメータ

たとえば1レイヤのフィルタ
カーネルサイズは、
7Bモデルの場合、
以下のように設定されている

- hyena_se: 7
- hyena_mr: 16
- hyena: L（global）

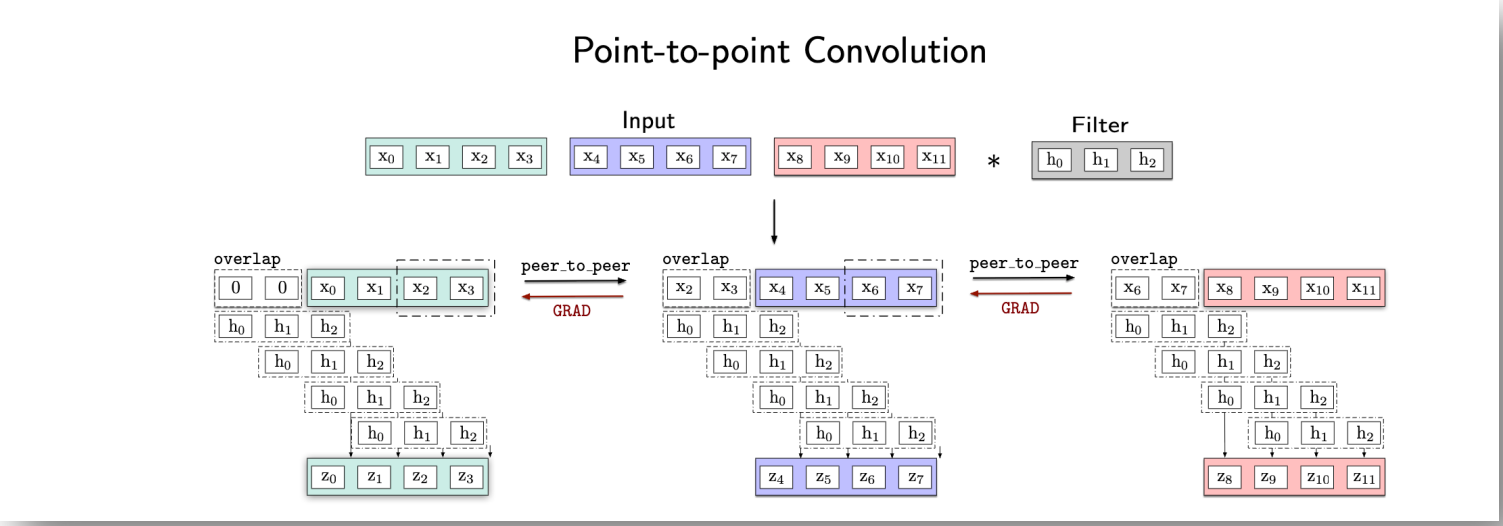
Striped Hyena2分散並列学習基盤: Savanna

LLM分散並列学習におけるMegatron-LMに対応するもの。
Striped Hyena2アーキテクチャの分散並列学習に特化。

Data Parallel, Pipeline Parallel, Tensor Parallel
DeepSpeed ZeRO stage1/2/3対応
NVIDIA TransformerEngineによるFP8対応

★マルチハイブリッドモデル（複数種類の演算ブロック）に対応
★Context Parallel
文脈方向でHyena畳み込み演算を並列化。
all-to-all, point-to-pointのGPU間通信を独自に実装して効率化。

https://arxiv.org/abs/2503.01868



Savanna分散学習基盤の環境構築

権限制約のあるHPC上で、まだエコシステムが追いついていない最新GPU（Blackwell）＋PyTorch/DeepSpeed/Flash-attention/Savanna系を、Apptainerコンテナスタックで統合する

コンテナ内のパッケージ導入：
いずれも --no-build-isolation必須
ホスト側のCUDAドライバ、コンテナ内部のCUDA Toolkitに対応したtorch、DeepSpeed等関連パッケージのバージョン整合性を確認
sm_100対応はPytorch cu128 wheel必須, DeepSpeed, Flash-attentionもこれを前提にビルド

Savannaまわり：
・依存が明示されないまま import されて落ちる
・公式ドキュメント不足、 API仕様の説明なし → 「中身を読んで仕様を逆算する」必要がある
・「トークナイズ済み split をロードする」前提だが、その前処理ツールはリポジトリに同梱されていない
→ GPT-NeoXを参考に自前で実装
・そもそも PyTorch 2.6.0 前提で設計されているが、PyTorch 2.6.0 はBlackwellに対応していない

https://github.com/Zymrael/savanna

Savanna is a framework developed and maintained by a small team (< 5 people), tailored to the specific needs of the projects above. With careful tuning of the distributed training strategy and architecture, it achieves high MFU on H100s at the thousand-GPU scale. **It is not a production-ready framework, expect rough edges if you don't know what you're doing.** It can serve as a good starting point for research on large-scale training of multi-hybrids. We recommend digging around the training configs for the models above: all the details are there.

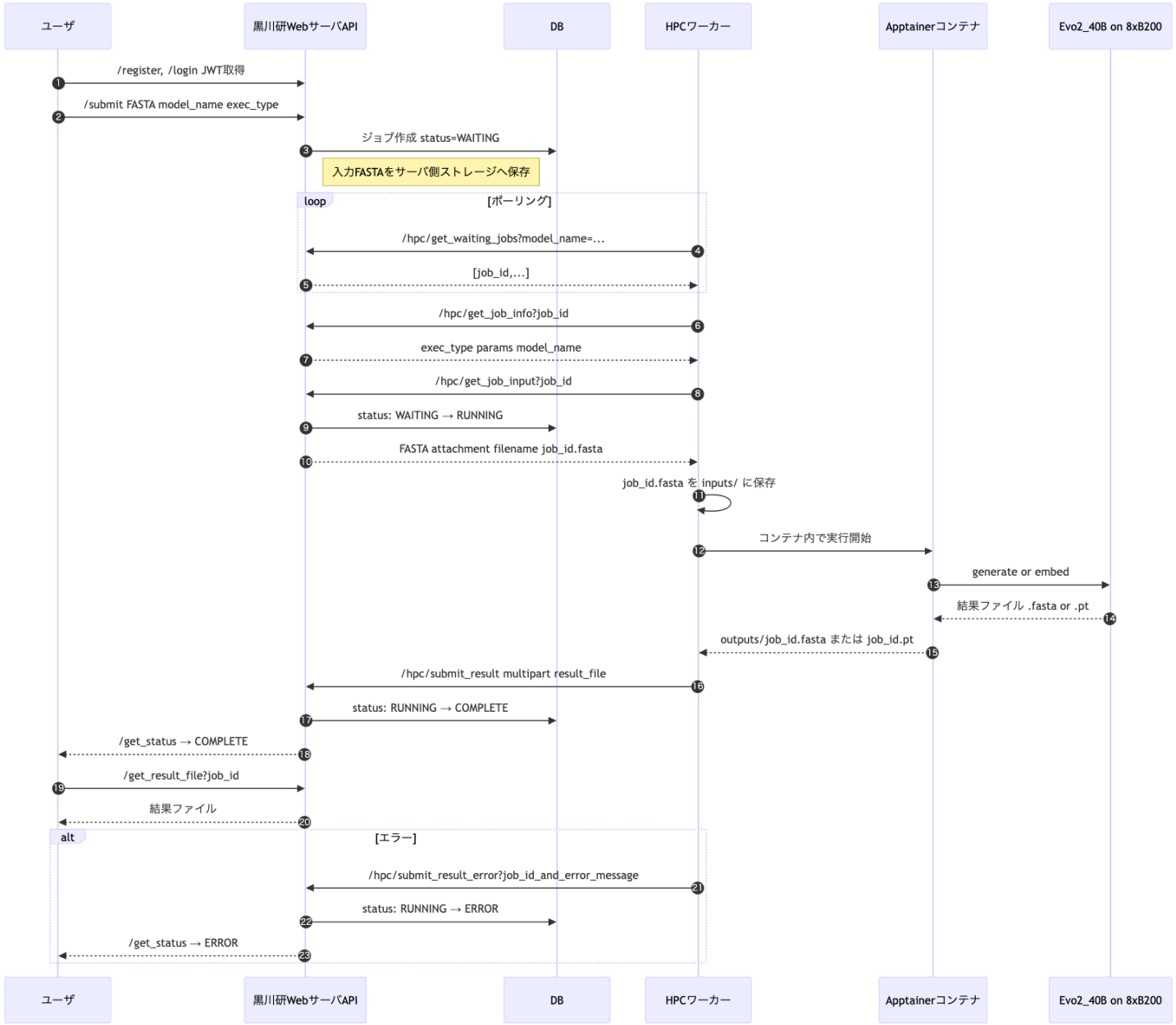
```
torch==2.6.0

ftfy
lm_dataformat
```

```
PyYAML
causal-conv1d
deepspeed
einops
flash-attn>=2.1.1,<=2.6.3
lazy_import_plus
omegaconf
opt-einsum
pybind11
regex
requests
scipy
sentencepiece
tokenizers>=0.20.1
torchaudio==2.6.0
torchvision==0.21.0
wandb
boto3
arrow
ring_flash_attn
pydantic<=2.8.2
numpy<=2.2.6 # to enable the
transformer-engine==1.13.0
```

Flash-attention（Blackwell未対応）切って、FP8やめて全部BF16計算、でなんとかまわせそうな雰囲気がある

モデル評価（性能計測、応用）チームの取り組み



Evo2 40B推論APIを開発し、評価チームで運用中。
遺伝研スパコンにログインせず配列生成と、
任意レイヤーを指定した内部表現の取得が可能。

forward推論は比較的簡単に8xB200で計算できる。
開発メンバーの廣田さん、鈴木さんによる検証で、
B200x8 Vortex推論で
80kbp くらいのコンテキストの演算ができそう、という試算。

抽出ブロック例：

blocks.0.pre_norm
blocks.0.post_norm
blocks.0.filter
blocks.0.projections
blocks.0.out_filter_dense
blocks.0.mlp
...
blocks.3.inner_mha_cls
blocks.3.inner_mha_cls.Wqkv
blocks.3.inner_mha_cls.inner_attn
...
blocks.49.mlp.act
blocks.49.mlp.l1
blocks.49.mlp.l2
blocks.49.mlp.l3

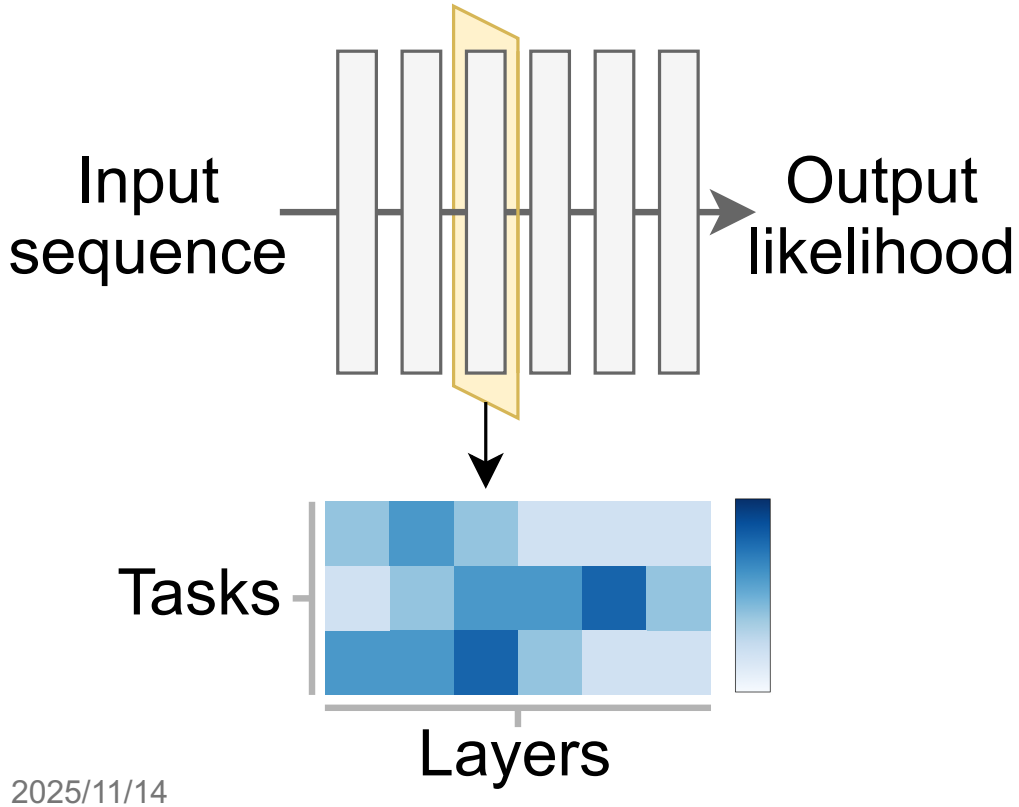
GenAI Bioにおける取り組みの例

researchmap

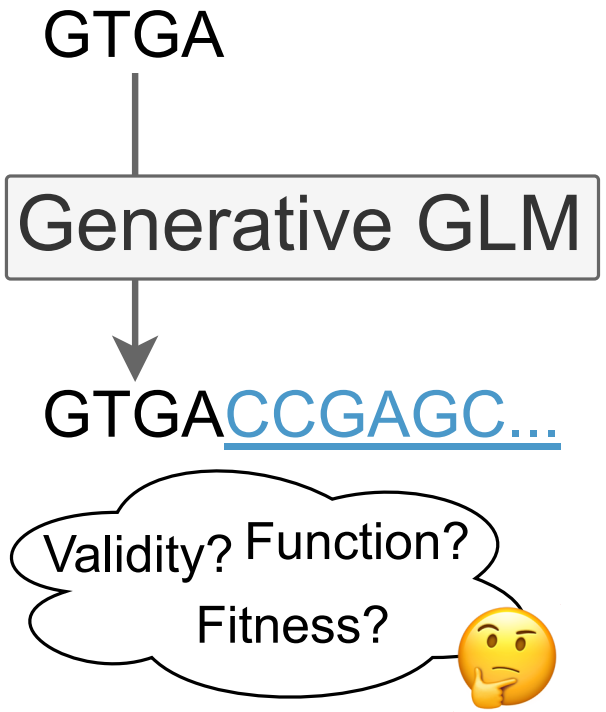
担当：廣田 佳亮 (東京科学大学 生命理工 山田研究室 博士1年)

興味・関心：GLMのベンチマーク設計、ゲノムコンテキストの理解と活用

下流タスクに役立つ埋め込み表現とは？



GLMの生成能力をどのように評価するか？



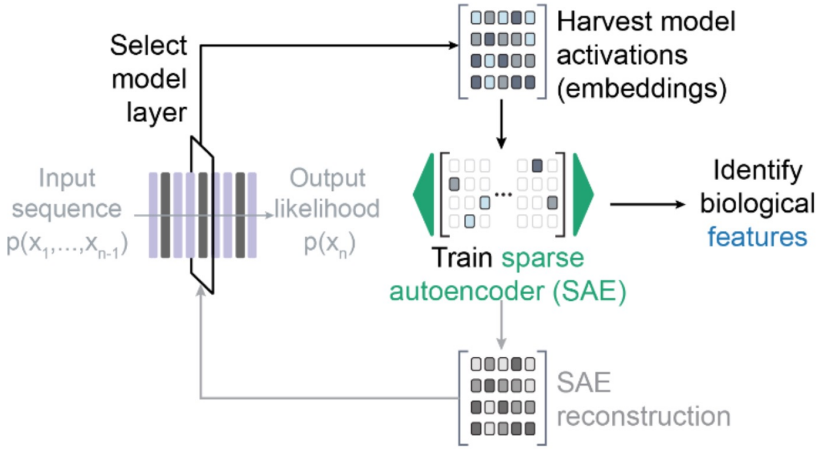
ゲノム言語モデルの解釈性研究と進化的情報の活用（筑波大学 鈴木）

本タスクで明らかにしたいこと・取り組み

・gLMは何を学習できて何を学習できないのか、ゲノムにおける良いモデル・学習とは具体的に何か？

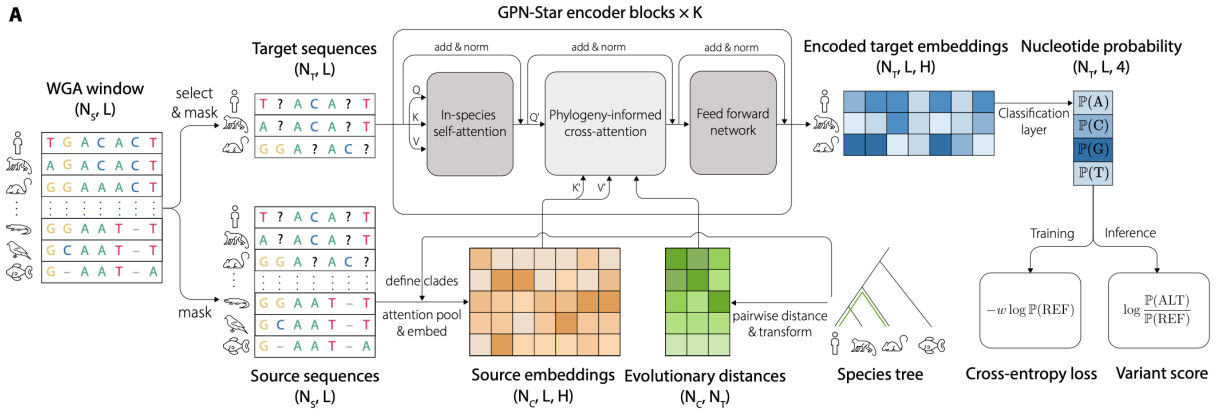
・進化的に保存されてきた機能的な制約情報の活用により、効率的な学習（ロスの設計など）や正確な変異影響予測の実現

解釈性：Sparse Auto-Encoder (SAE)



Brixi et al., bioRxiv (2025)

進化的情報：MSA, WGA



Ye, Benegas et al., bioRxiv (2025)

現在の開発メンバー

開発メンバー

名前	チーム	所属
東 光一	モデリング	国立遺伝学研究所
廣田 佳亮	モデリング	東京科学大学
張 一鳴	モデリング	東京大学
増田 元希	モデリング	東京科学大学
松本 淳弥	データ	東京科学大学
豊田 大樹	データ	東京科学大学
中居 風雅	データ	東京科学大学
築山 翔	評価	東京科学大学
鈴木 翔介	評価	筑波大学

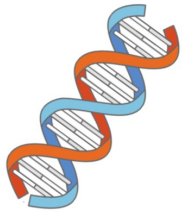
遺伝研　中村保一研究室のみなさま

中村 保一	評価	国立遺伝学研究所
谷澤 靖洋	評価	国立遺伝学研究所
坂本 美佳	評価	国立遺伝学研究所
望月 孝子	評価	国立遺伝学研究所
浅野 さとみ	評価	国立遺伝学研究所
Hanjie Mao	評価	国立遺伝学研究所
Mohamed Elmanzalawi	評価	国立遺伝学研究所
Ziyi He	評価	国立遺伝学研究所
近藤 翼	評価	国立遺伝学研究所
徳丸 万柚子	評価	国立遺伝学研究所
Febrina Margaretha	評価	国立遺伝学研究所
Sheetal Agarwal	評価	国立遺伝学研究所

遺伝研 中村研　ゲノム言語モデルを用いた遺伝子アノテーションの高度化

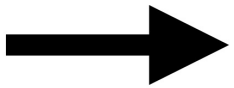
手動アノテーション → ゲノム言語モデル

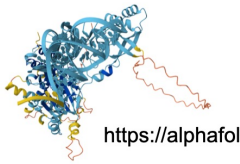
ゲノム配列



ゲノム言語モデル

遺伝子構造モデル
を正確に予測

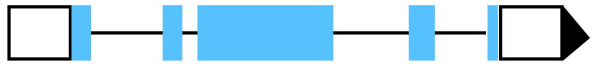




https://alphafoldserver.com/

機能予測 or

モデルの修正



高品質・高網羅性

調節領域・ノンコーディングRNA にも適用

中村研のこれまでの活動

2000年　シロイヌナズナ ゲノムアノテーション *Nature* **408**, 796–815 (2000)

2023年　フタホシコオロギ 神経ペプチドアノテーション *Insects* **14**, 121 (2023)

2025年　イエネコのゲノムアノテーション　*J. Adv. Res.* **75**, 863–874 (2025)

プログラムで予測された複数の遺伝子モデルをエキスパートが目で見て統合し、
各々世界最高の遺伝子アノテーションを実施した。