

人工知能でタンパク質をデザイン

齋藤 裕

産業技術総合研究所

人工知能研究センター

<https://staff.aist.go.jp/yutaka.saito/>

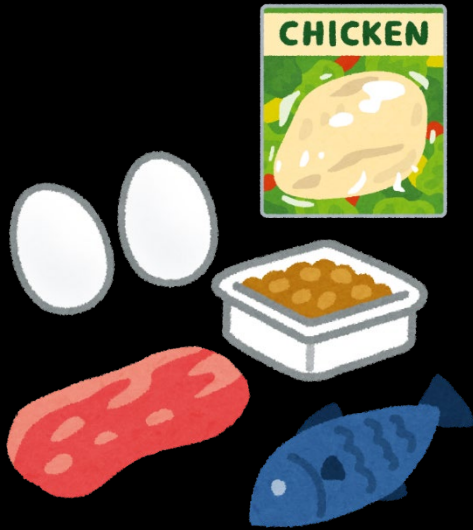
自己紹介

- 産業技術総合研究所 人工知能研究センター 主任研究員
専門：バイオインフォマティクス
情報技術をもちいて
生命科学データ（DNA、RNA、タンパク質など）を解析する研究分野
- 研究モットー： Bioinformatics for Bioengineering

Reverse bioengineering
大量データ解析から生命のルールを読み解く

Forward bioengineering
ルールを用いて生体分子を設計・改良する
- 齋藤研究室（東大院・メディカル情報生命、北里大・データサイエンス）
<https://staff.aist.go.jp/yutaka.saito/>
Twitter @ytkasai

タンパク質と聞いて何を思い浮かべますか？



食べ物？



筋トレ？

タンパク質は人の役に立つ機能分子



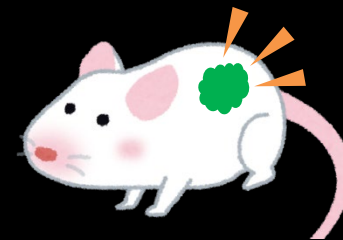
抗体
検査薬、抗がん剤



酵素
洗剤、胃薬



ワクチン
感染症の予防



病変部位を
光らせる

蛍光タンパク質
生体イメージング

もっと高機能なタンパク質を創りたい

人工知能でタンパク質をデザイン！

タンパク質は文字列データ

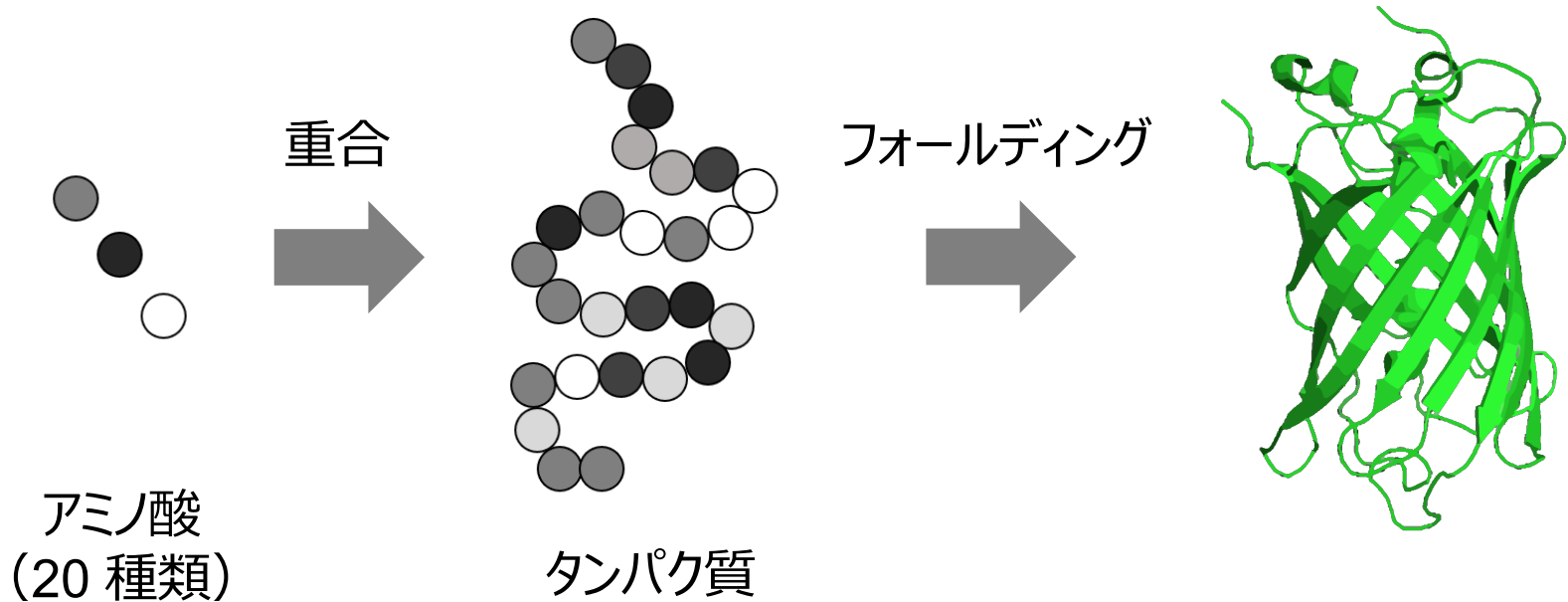
- タンパク質は 20 種類のアミノ酸が重合したポリマー
各アミノ酸を 1 文字で表すと、タンパク質は文字列として表現できる



- 自然言語処理などの文字列データ用の機械学習技術が適用できる
RNN, Transformer,
タンパク質言語モデル (protein language model; pLM)

タンパク質は画像データ

- タンパク質は機能と関連した 3D 構造をとる（フォールディング）
各原子の座標からなる 3D 画像データとして表現できる

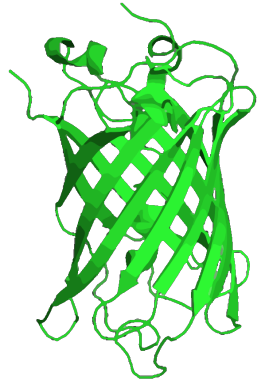


- 画像データ用の機械学習技術が適用できる
3D CNN, SE3 Transformer,
拡散モデル (diffusion model)

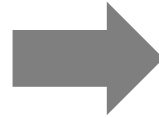
タンパク質の機能改良

➤ 人工的に変異を導入してタンパク質の機能を改変

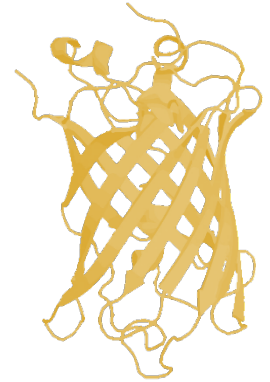
野生型
タンパク質
(低機能)



変異導入



改良型
タンパク質
(高機能)



MSKGEELFTGVVPILVELDG
DVNGHKFSVSGEGEGDATYG
KLTLKFICTTGKLPVPWPTL
VTTF**S**YGVQCF**S**RYPD**H**MKQ
HDFFKSAMPEGYVQERTIFF
KDDGNYKTRAEVKFEGDTLV

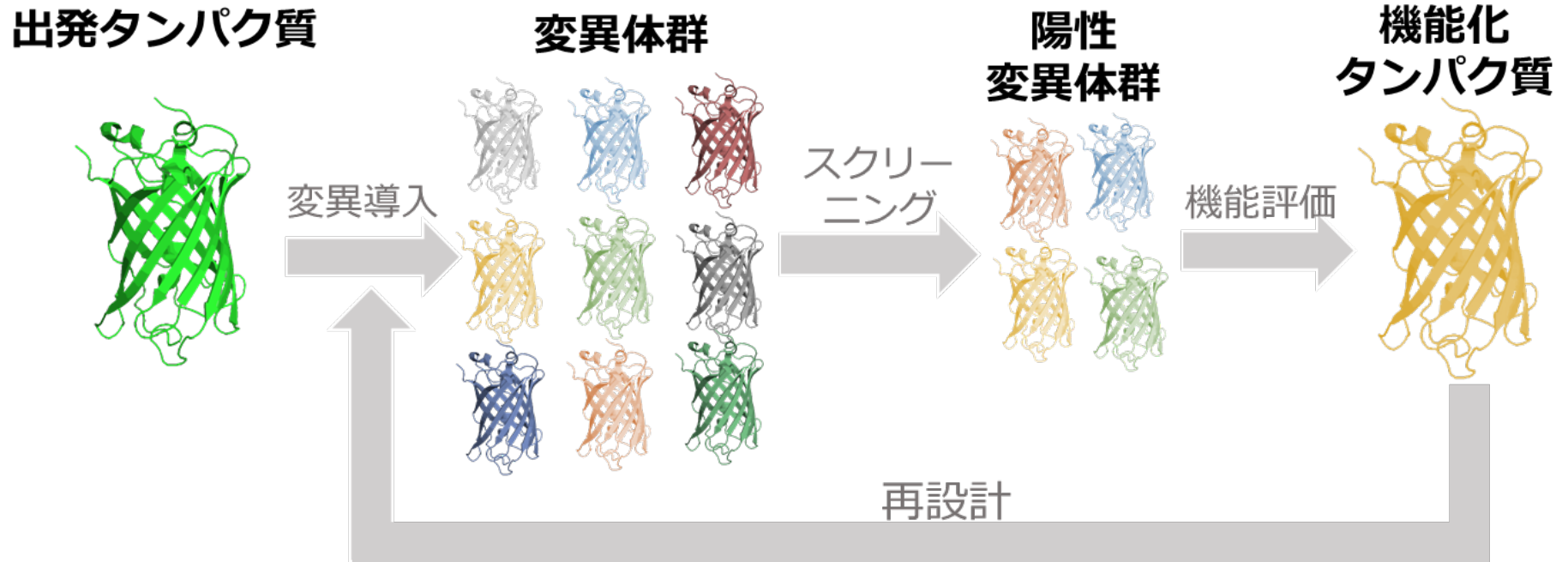
...

MSKGEELFTGVVPILVELDG
DVNGHKFSVSGEGEGDATYG
KLTLKFICTTGKLPVPWPTL
VTTF**G**YGVQCF**A**RYPD**Y**MKQ
HDFFKSAMPEGYVQERTIFF
KDDGNYKTRAEVKFEGDTLV

...

抗体、酵素、蛍光タンパク質の性能向上など
バイオ産業の研究開発において広く行われている

従来の機能改良法：指向性進化



問題点：

ランダムに変異を導入するため作成した変異体のほとんどは機能を失う
多数の変異体をスクリーニングするには多額の費用が必要



本研究の目標：

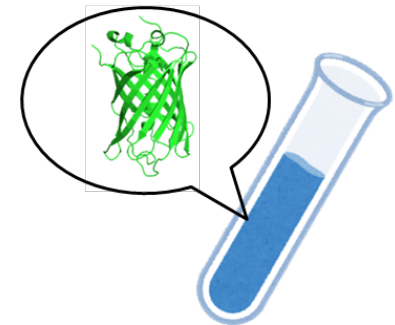
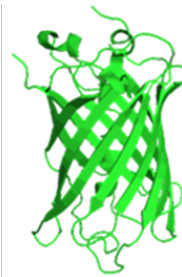
機械学習により機能を向上すると予測された変異のみを導入
高機能な変異体を豊富に含む小規模な変異体群を作成

提案手法：機械学習が導く指向性進化

- 機械学習とウェット実験を組み合わせたタンパク質機能改良プラットフォームを開発

ウェット実験：
東北大 梅津光央 教授

改変タンパク質の提案

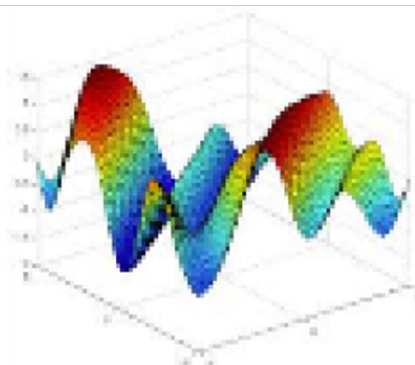


改変タンパク質の作成

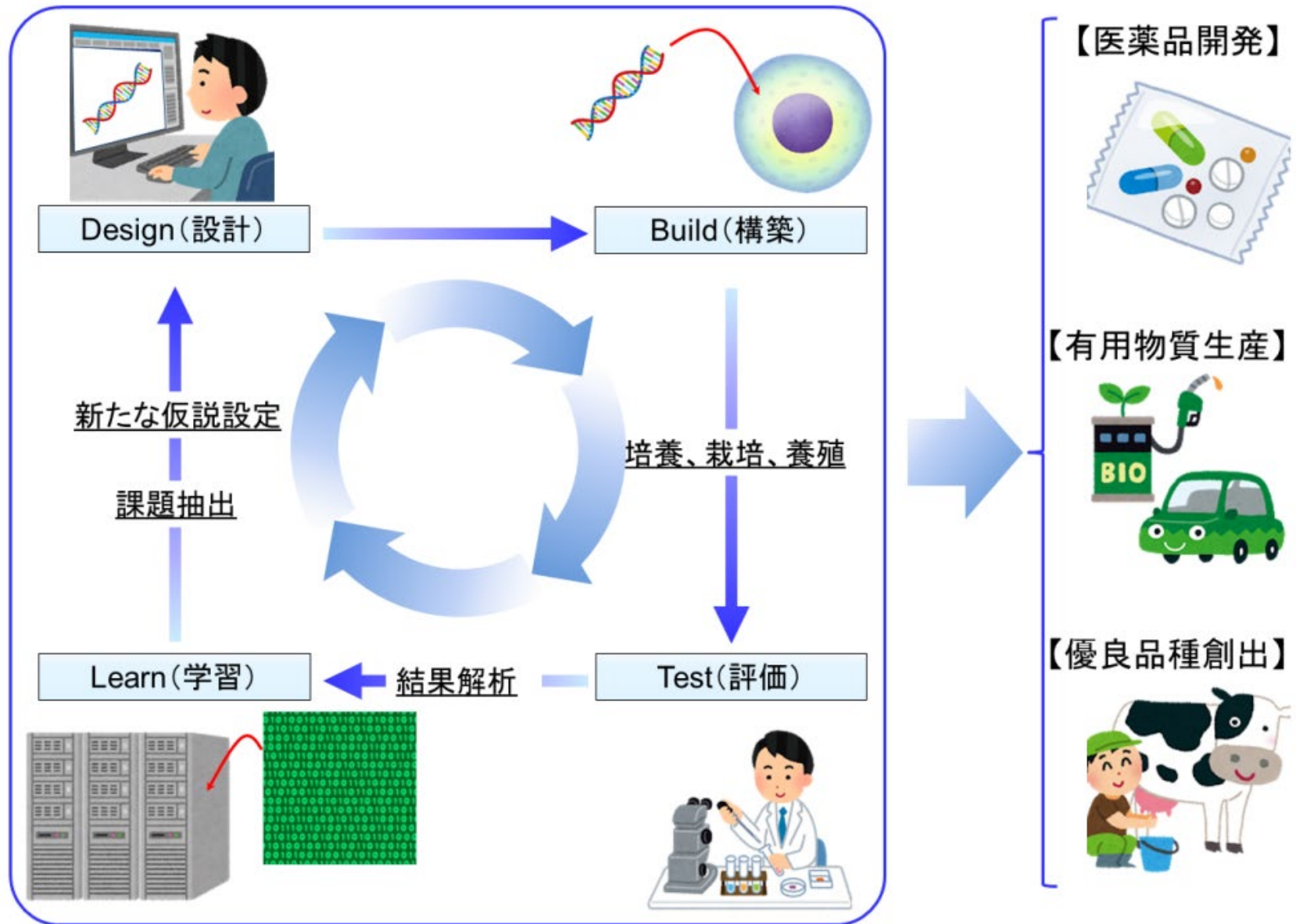
機能評価（配列、機能値）



機械学習による
機能予測モデル
（ベイズ最適化）



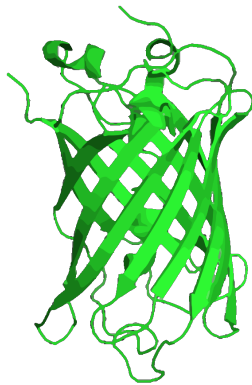
cf. バイオものづくりの DBTL サイクル



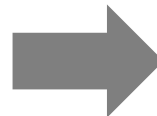
実証： 蛍光タンパク質を緑から黄色へ

➤ GFP を YFP に改変できるか？

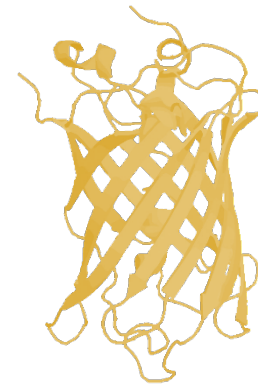
Green fluorescent
protein (GFP)



タンパク質機能改良
プラットフォーム



Yellow fluorescent
protein (YFP)



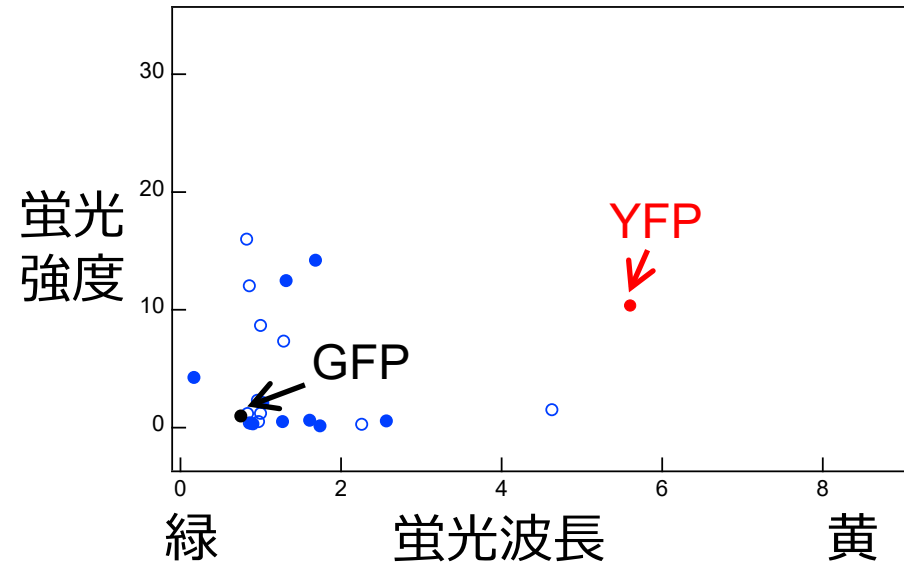
MSKGEELFTGVVPILVELDG
DVNGHKFSVSGEGEGDATYG
KLTLKFICTTGKLPVPWPTL
VTTF**S**YGVQCF**S**RYPD**H**MKQ
HDFFKSAMPEGYVQERTIFF
KDDGNYKTRAEVKFEGDTLV

...

MSKGEELFTGVVPILVELDG
DVNGHKFSVSGEGEGDATYG
KLTLKFICTTGKLPVPWPTL
VTTF**G**YGVQCF**A**RYPD**Y**MKQ
HDFFKSAMPEGYVQERTIFF
KDDGNYKTRAEVKFEGDTLV

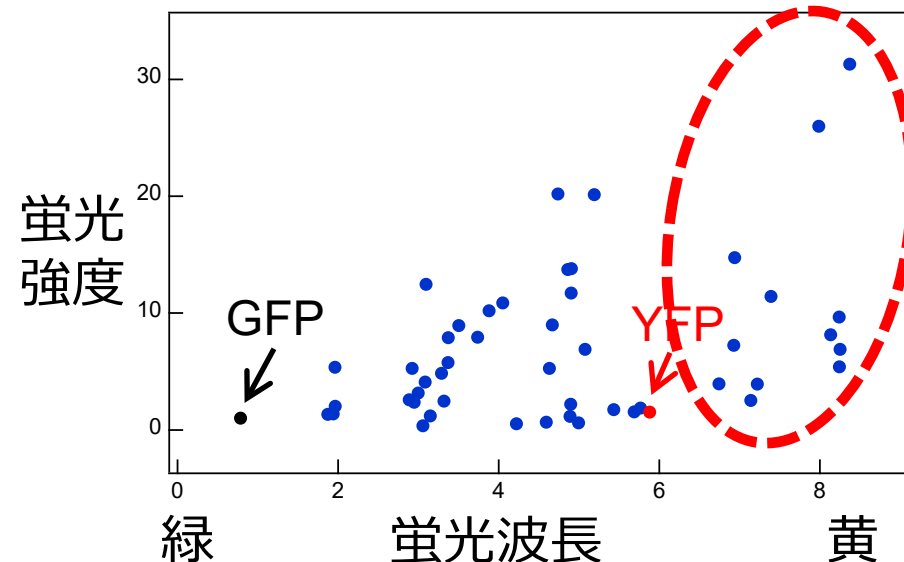
...

機械学習の効果



初期変異体群：
従来のランダム変異導入で作成

YFP より良い変異体は取得不可
155 変異体中 23 個が蛍光陽性
(15%)



機械学習が提案した変異体群

YFP より黄色く蛍光も強い
変異体を 12 個取得することに成功
63 変異体中 48 個が蛍光陽性
(76%)

機械学習による指向性進化の
有効性を実証した世界初の結果
[Saito et al. ACS Synth Biol. 2018]

機械学習が提案した変異体 (Top 4)

野生型
GFP

1

2

3

4



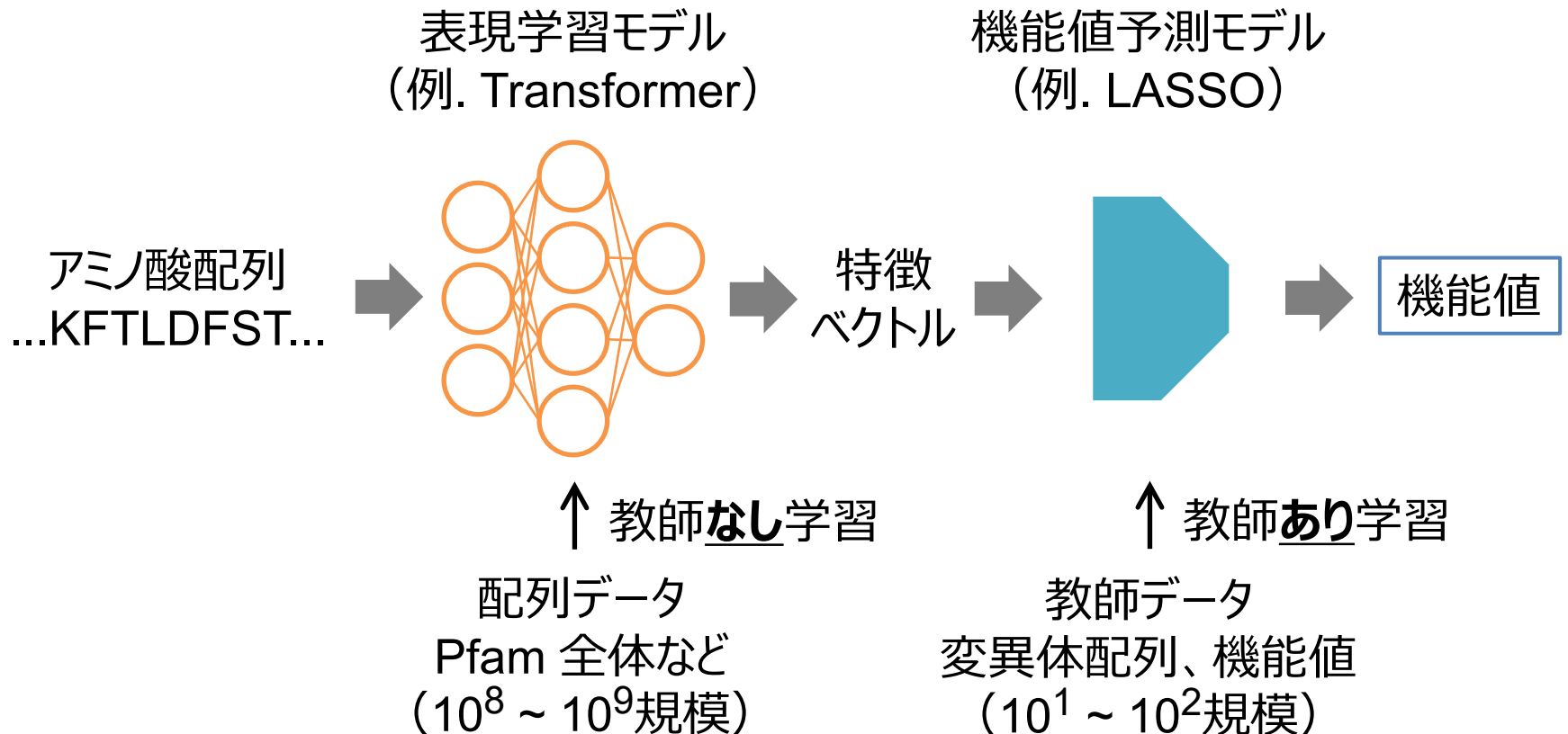
その後の展開

- バイオものづくり：
酵素 SrtA の機能改良 [Saito et al. ACS Catal. 2021]
酵素 XylM の機能改良 [Ogawa and Saito et al. ACS Synth Biol. 2023]
- バイオ医薬品：
抗体様標的結合タンパク質の設計 [Ito and Nguyen et al. mAbs. 2023]
- NEDO「AI×ロボティクスによるバイオ分子設計デジタルラボの研究開発」
教師データ取得のための実験をロボティクスで効率化



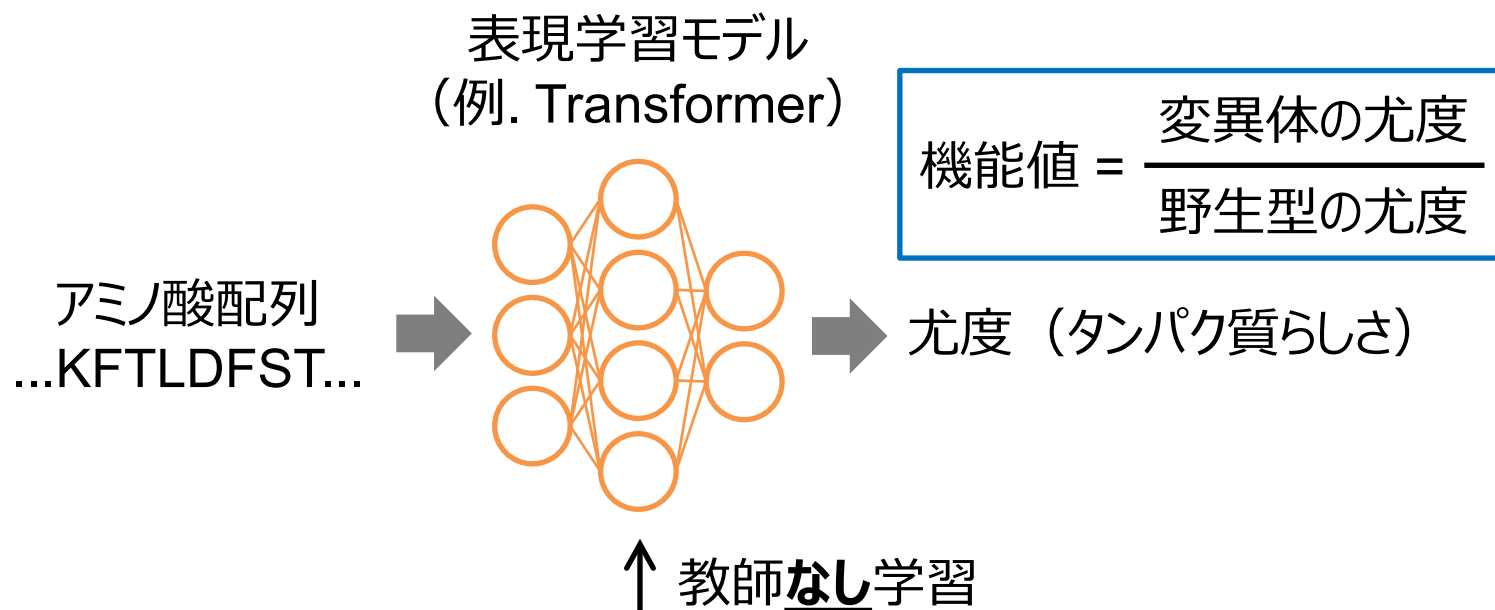
Transformer による表現学習

- 生体分子設計における教師データの不足を補うため
機能未知の大量配列データから特徴表現を学習
- 自然言語処理分野の深層学習モデル Transformer を利用



ゼロショット予測

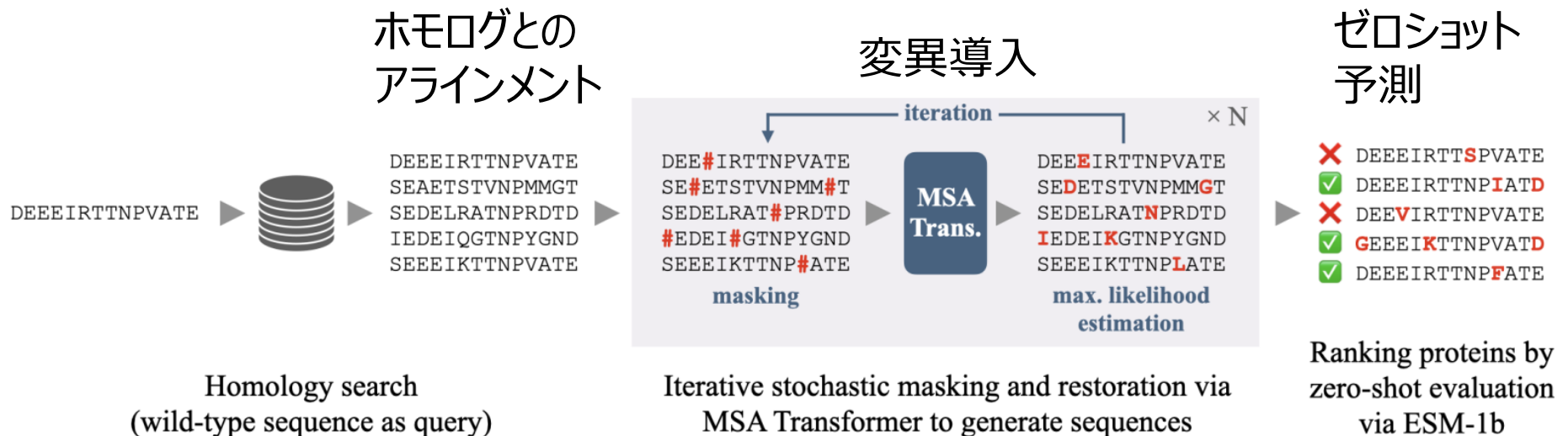
- Transformer のネットワーク構造を少し書き換えると、
入力配列の尤度（タンパク質らしさ）を評価できるようになる
- 実験で測定した機能値のデータが全くない状況でも予測を行える



- ① pre-training : Pfam全体など
- ② fine-tuning : 改変対象タンパク質のホモログ

EvoOpt : ゼロショット予測による配列最適化

- ゼロショット予測による機能値が高くなるように、
計算機上で変異導入を繰り返して配列を人工進化させる
- 複数のタンパク質言語モデルを組み合わせたパイプライン



宣伝：日本語総説

創立100周年記念特別企画～バイオテクノロジーが拓く未来社会～ 特集

深層学習によるタンパク質の機能予測と設計

齋藤 裕

 著者情報

ジャーナル

フリー

2022 年 100 巻 11 号 p. 589-592

DOI https://doi.org/10.34565/seibutsukogaku.100.11_589

[齋藤. 生物工学会誌. 2022]

Review Article

タンパク質の言語モデル

山口 秀輝 , 齋藤 裕 

 著者情報

ジャーナル

オープンアクセス

HTML

2023 年 4 巻 1 号 p. 52-67

DOI <https://doi.org/10.11234/jsbibr.2023.1>

[山口ら. JSBi Bioinformatics Review. 2023]

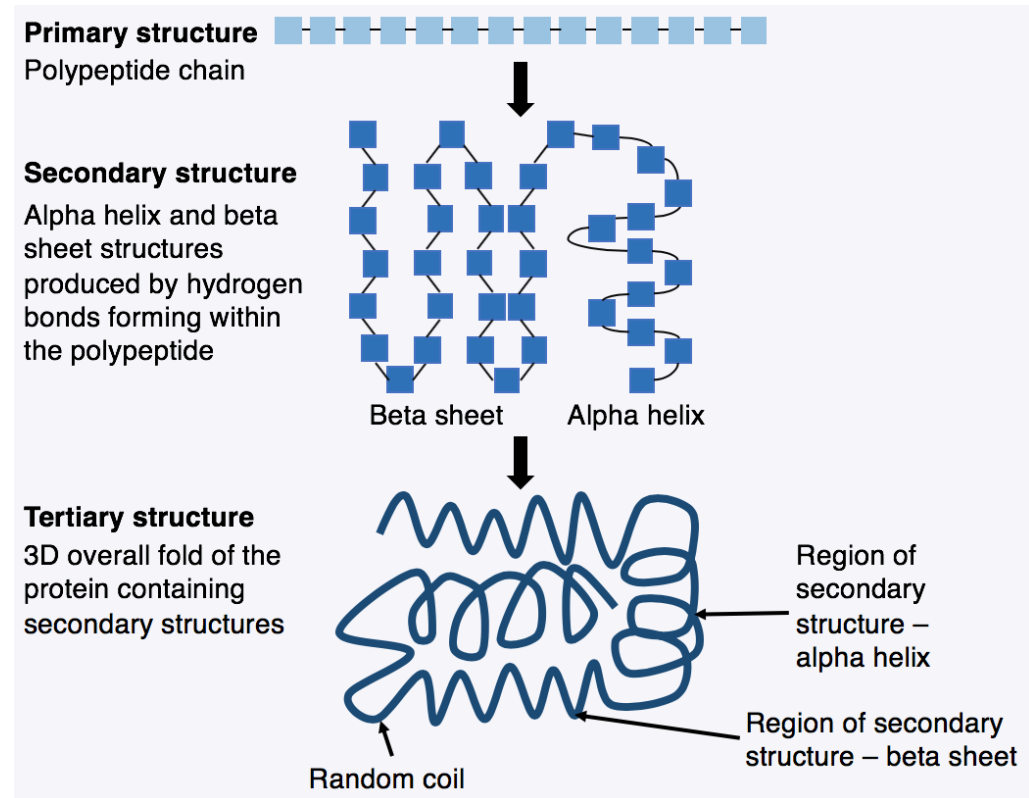
そして世界は戦国時代へ

- GAFA を筆頭に巨大 IT 企業が続々と参入
- 有名な例：AlphaFold

Google DeepMind が開発した
タンパク質の 3D 構造予測手法

3D 構造予測の国際コンペ
CASP14 を驚異的な成績で圧勝

中身はアミノ酸配列や
3D 構造を扱うための
各種 Transformer の組み合わせ



タンパク質言語モデルの大規模化

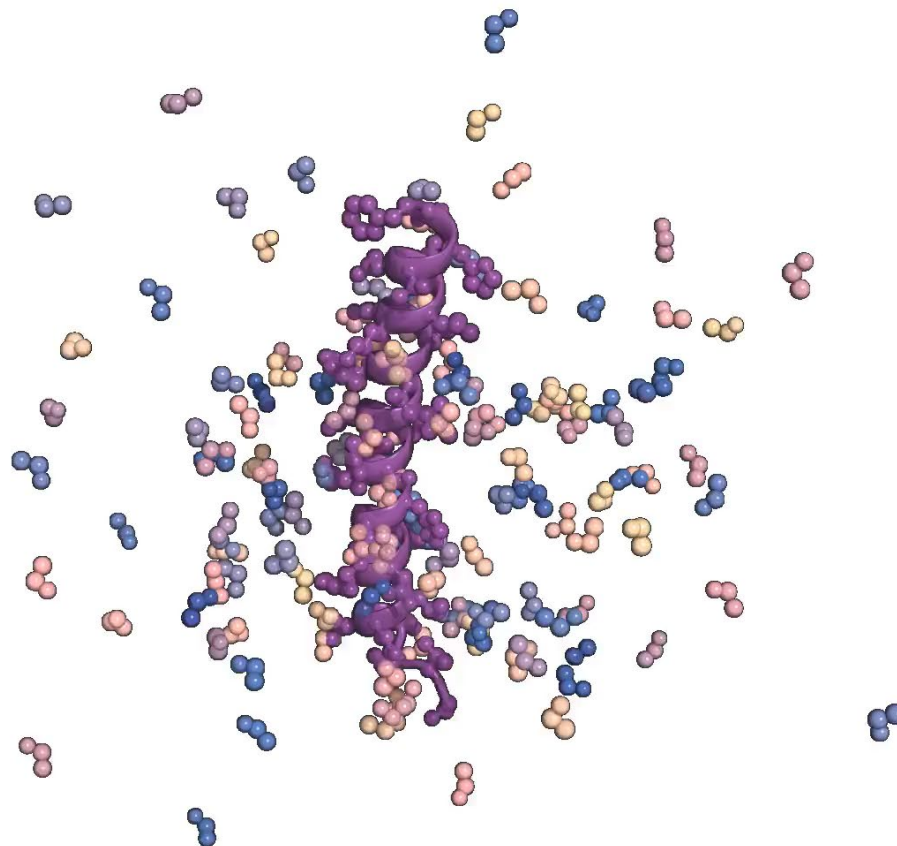
- TAPE [Rao et al. NeurIPS 2019]
- ESM [Rives et al. PNAS. 2021]
Meta 社開発、パラメータ数 100M
- ESM2 [Lin et al. Science. 2023]
現在最大のタンパク質言語モデル、パラメータ数 15B
- ESM2 のモデル規模は GPT2 と同じくらい

近い将来、GPT3, 4 相当まで大規模化した際、
自然言語処理分野で見られたような
劇的な性能向上（創発現象）が起こるかもしれない

pLM から pLLM へ

拡散モデルによるタンパク質の生成

- RF diffusion (David Baker Lab, University of Washington)



自然言語プロンプトによるタンパク質生成

- 画像生成 AI みたいにタンパク質を生成できないか？

「AI さん、かわいい犬の画像を描いて！」

「AI さん、肺がんの薬になるタンパク質をデザインして！」

- ProteinDT [Liu et al. arXiv. 2023]

NVIDIA, CalTech 他開発

画像生成 AI で用いられる CLIP という手法を
タンパク質に応用している (Protein CLAP)

自然言語プロンプトによるタンパク質生成

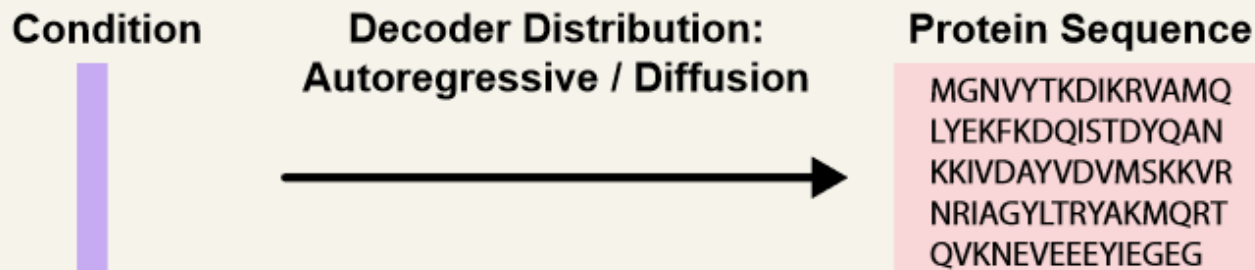
- 教師データとして、タンパク質の配列とその機能を説明した文章のペアを多数用意
両者の特徴ベクトルが近くなるように配列と文章のエンコードを
それぞれ学習する（対照学習）

Stage 1: ProteinCLAP



- さらに特徴空間から配列へのデコードを学習することで、生成が可能に

Stage 3: Decoder



Take-home message

- タンパク質は人の役に立つ機能分子
- 機械学習を利用して高性能なタンパク質を効率よく設計できる
- タンパク質デザインの戦国時代が到来
人工知能分野全体で見ても、今最も「熱い」フィールドになりつつある