

AIが導くバイオ分子デザイン

齋藤 裕

北里大学
産業技術総合研究所

自己紹介

- 北里大学 未来工学部 データサイエンス学科 教授
産業技術総合研究所 人工知能研究センター 特定フェロー
- 専門： バイオインフォマティクス
- 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

タンパク質は人の役に立つ機能分子



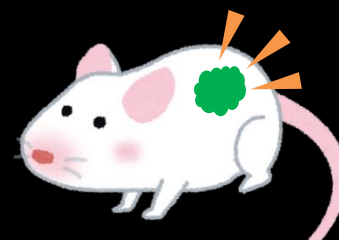
抗体
検査薬、抗がん剤



酵素
バイオものづくり、
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

セントラルドグマを貫く生体分子設計

- Designing biomolecules throughout the central dogma

DNA



プロモータ DNA の設計 :
転写量を増やしたい
[Saito et al. in prep]

転写



RNA

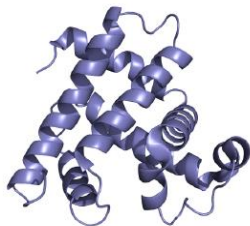


mRNA の設計 :
翻訳量を増やしたい (コドン最適化)
[Saito et al. Sci Rep. 2019]

翻訳



タンパク質

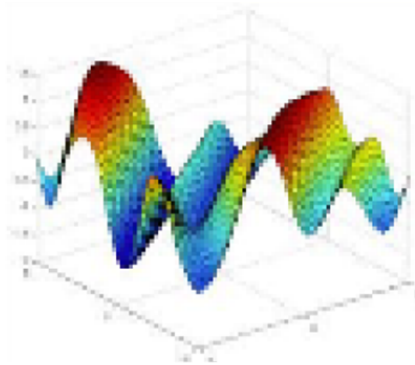
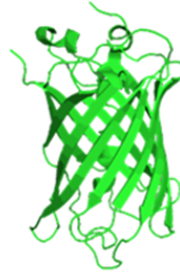


タンパク質の設計 :
機能を改良・改変したい
[Saito et al. ACS Synth Biol. 2018]

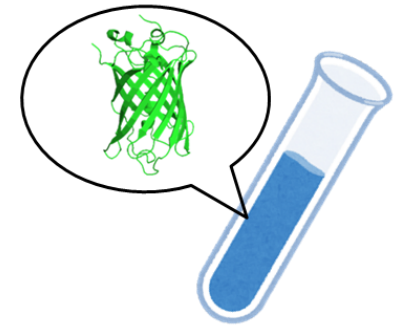
AI が駆動するタンパク質デザイン

- 2016 年から AI 主導タンパク質工学の重要性に着目
様々なタンパク質（蛍光タンパク質、酵素、抗体）において
ドライの技術開発はもちろん、ウェットでの実証例も積み上げてきた

改変タンパク質の提案



機械学習による
機能予測モデル
(独自技術の開発、
最新AIの導入)



改変タンパク質の作成



機能評価（配列、機能値）

実証例： 蛍光タンパク質を緑から黄色へ

➤ 東北大・梅津光央先生、東大・津田宏治先生との共同研究

機械学習が提案した変異体 (Top 4)

野生型
GFP

1

2

3

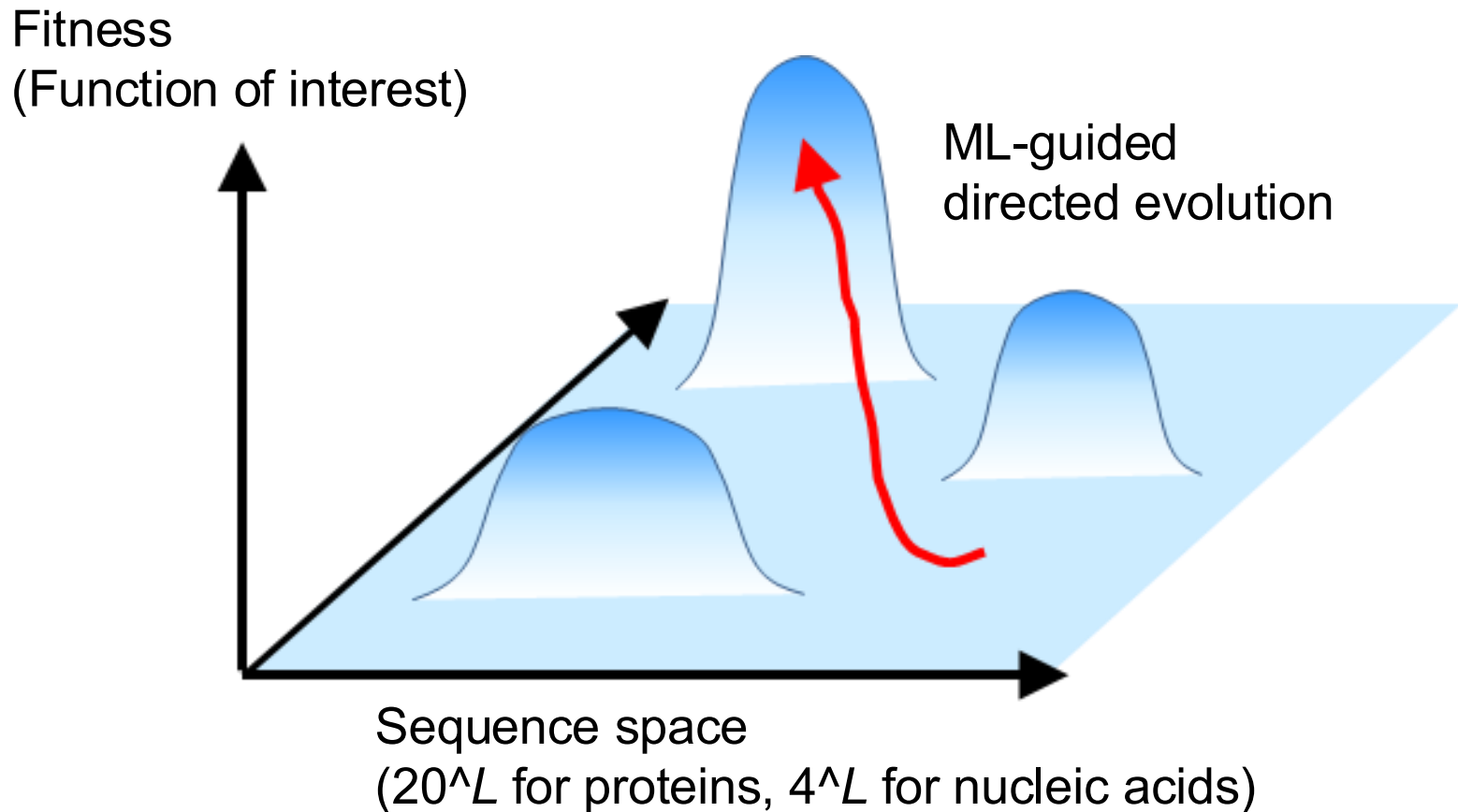
4



現在、抗体や酵素の開発に本手法を応用展開中
タンパク質だけでなくRNA, DNAの分子設計にも応用可能

Sequence space exploration

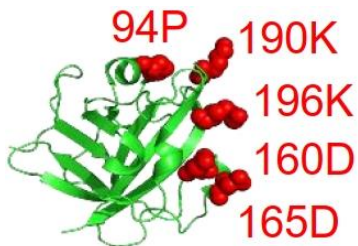
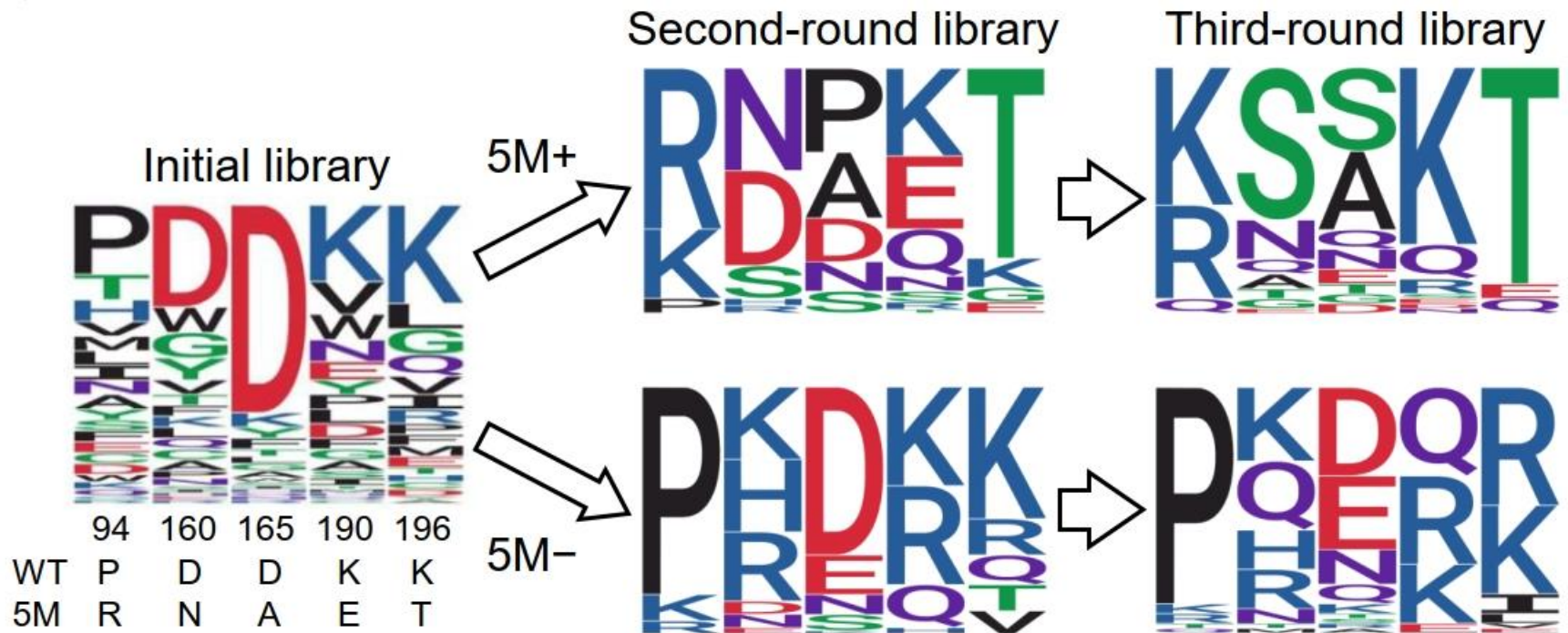
- Machine learning (ML) is a guide in sequence space;
It learns the fitness landscape to lead directed evolution



教師データが進化の指向性に与える影響

共同研究
東北大・梅津光央先生

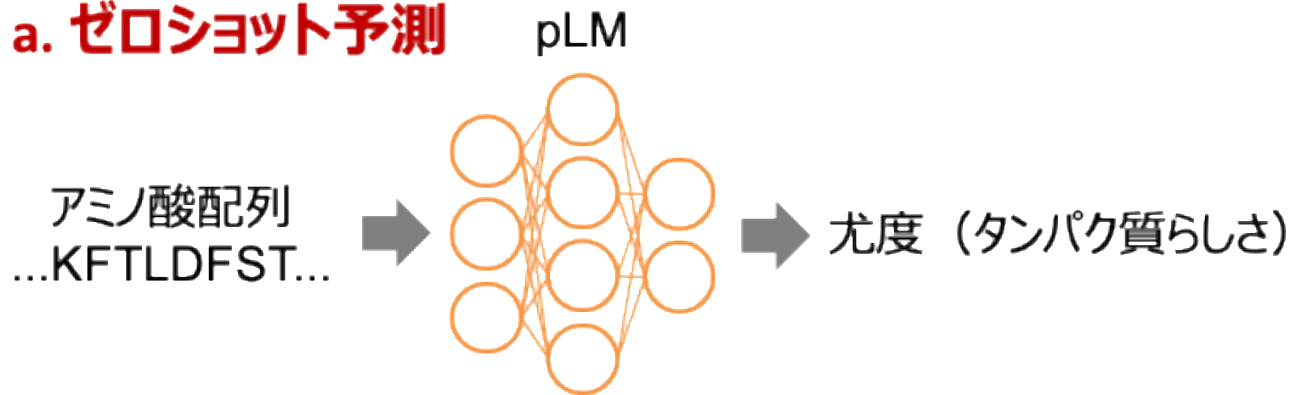
- 酵素 SrtA の機能改良：
既知の高機能変異体 (5M) を教師データに含める / 含めないで予測が大きく変化



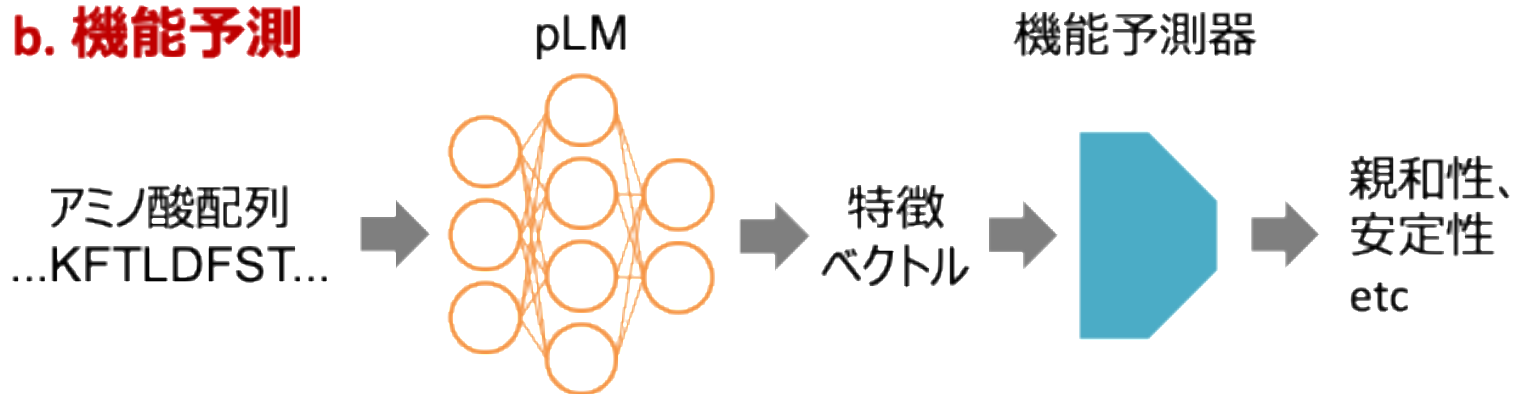
SrtA :
2つのペプチドを連結する酵素
タンパク質の修飾などに有用

タンパク質言語モデル（pLM）の使用方法

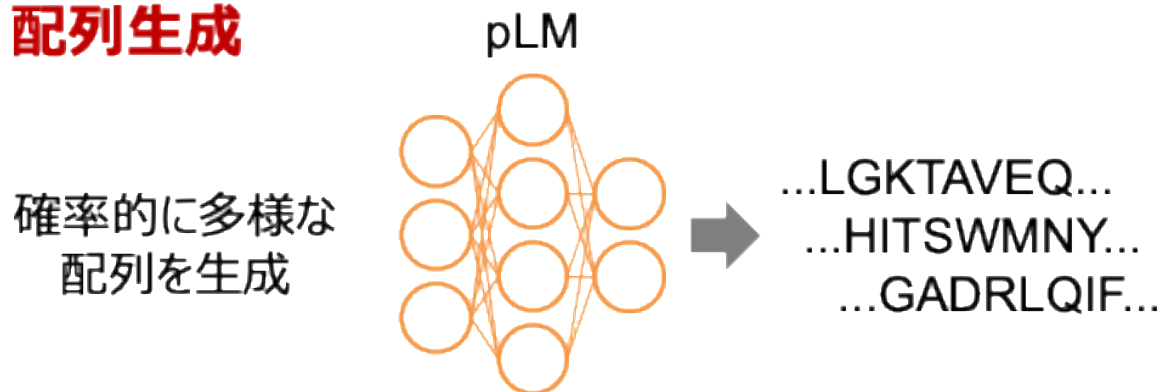
a. ゼロショット予測



b. 機能予測

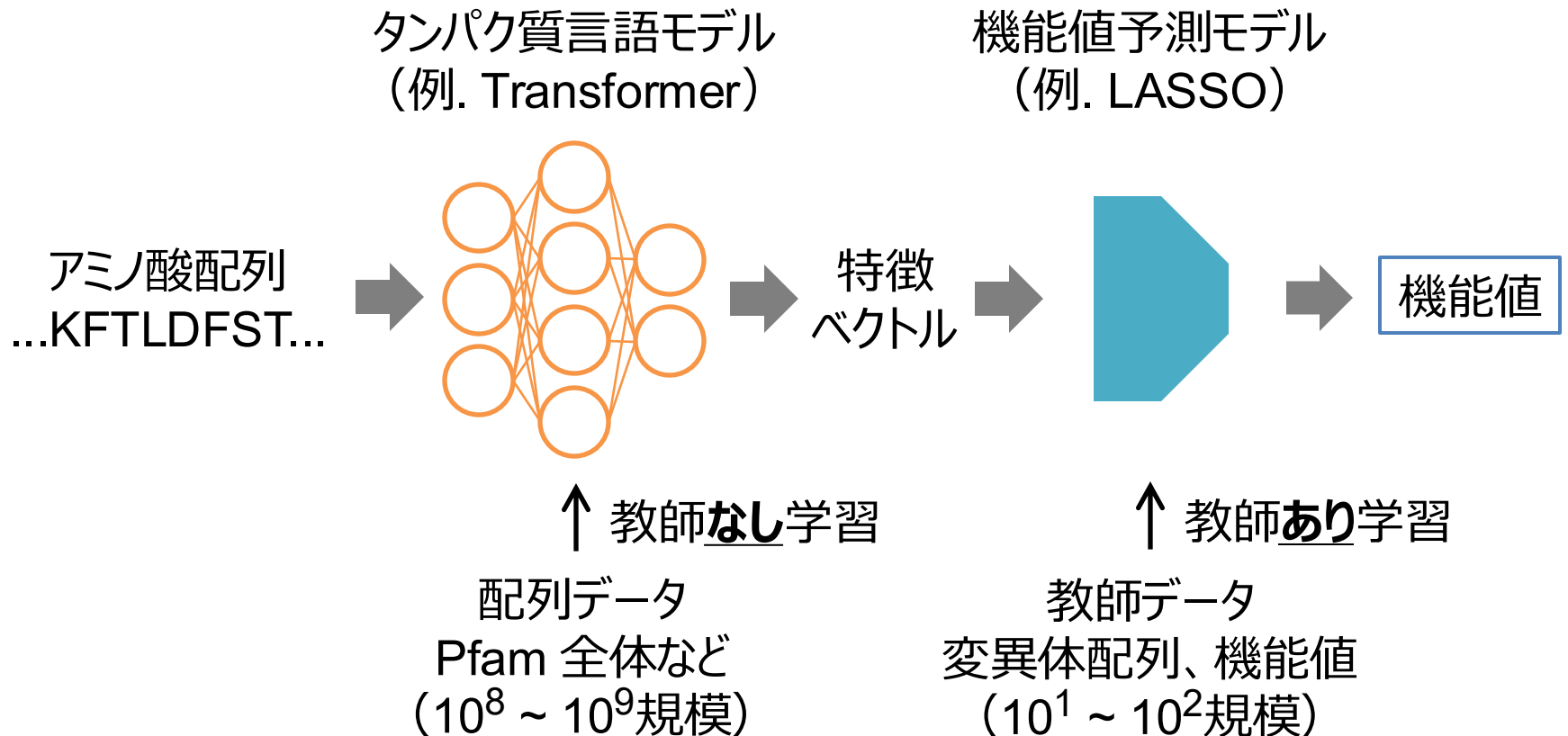


c. 配列生成



タンパク質言語モデルによる表現学習

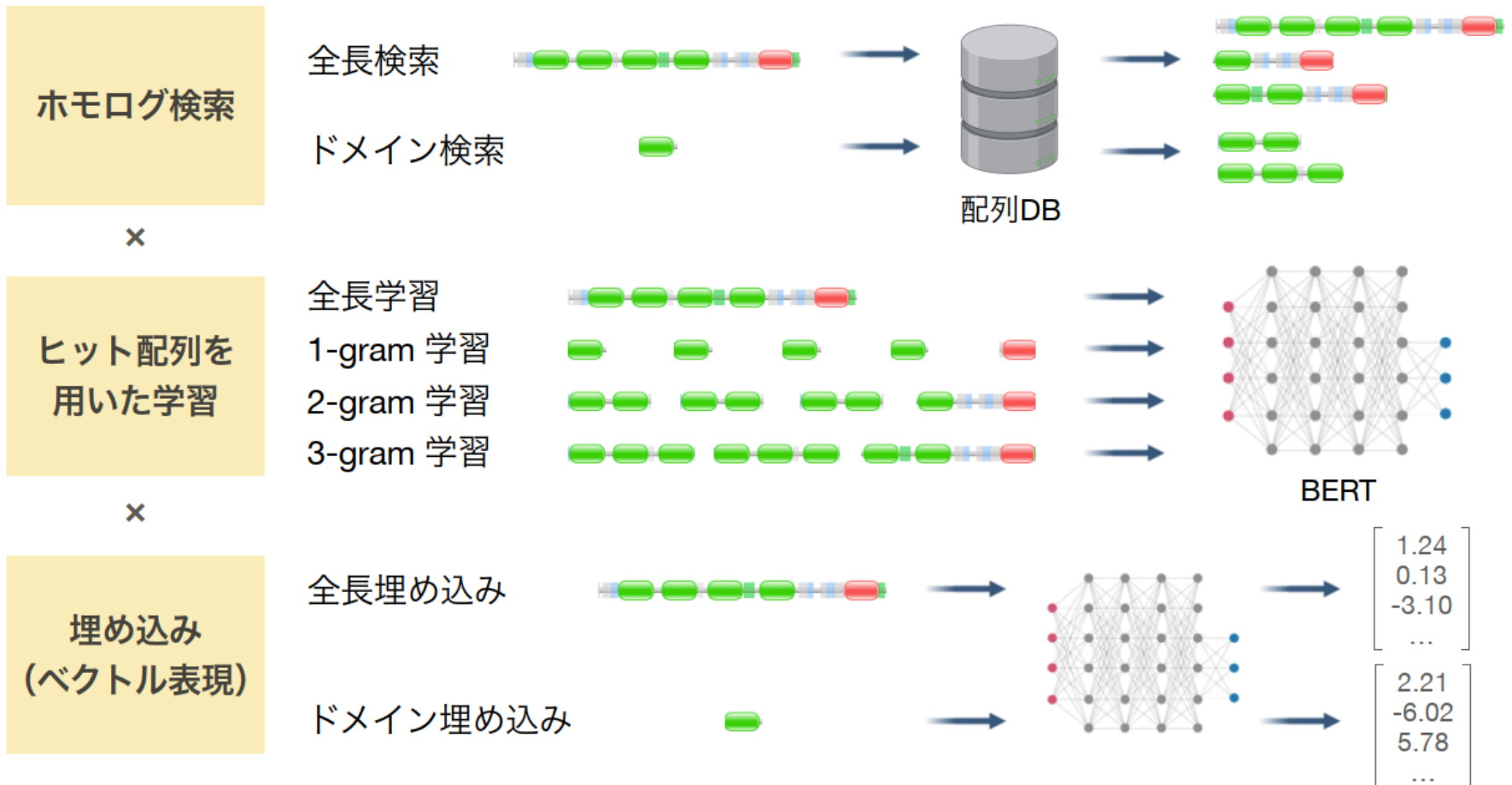
- 生体分子設計における教師データの不足を補うため
機能未知の大量配列データから特徴表現を学習
- 自然言語処理分野の深層学習モデル Transformer を利用



タンパク質言語モデルの独自の学習技術

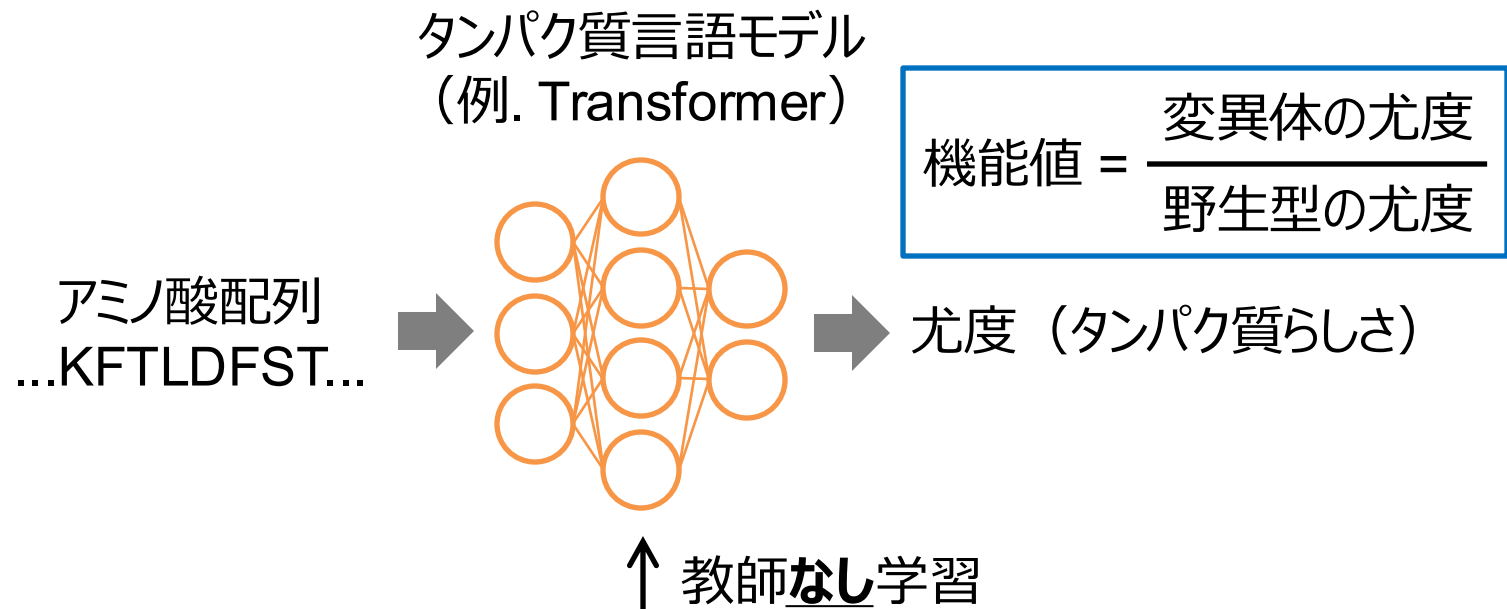
- タンパク質のドメイン構造を考慮した継続事前学習により pLM の性能を改善

  : 変異導入ドメイン



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



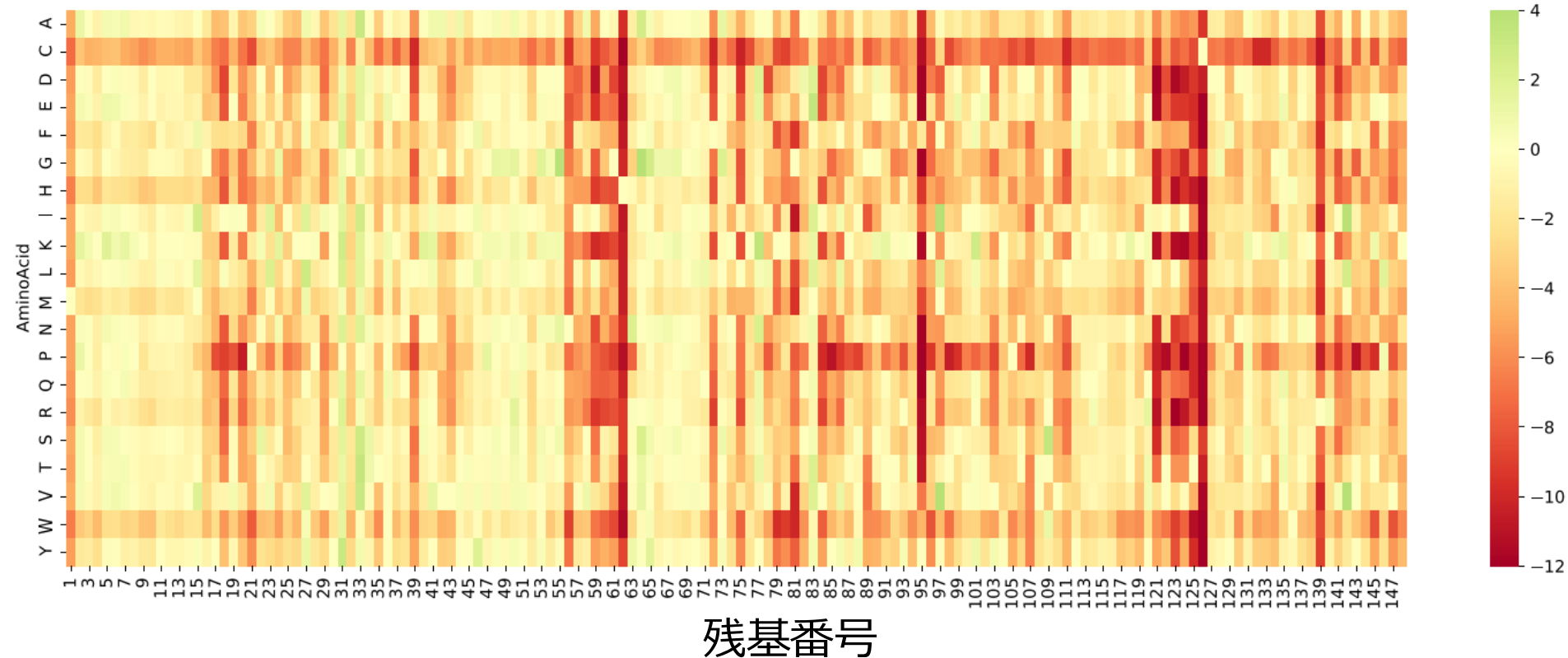
- ① pre-training : Pfam全体など
- ② fine-tuning : 改変対象タンパク質のホモログ

In silico saturation mutagenesis

- ゼロショット予測により、ウェット実験のデータが全くない状況で、網羅的な変異効果予測を行える。点変異、多重変異どちらにも対応
- 酵素ソルターゼに対する予測結果 (point saturation)

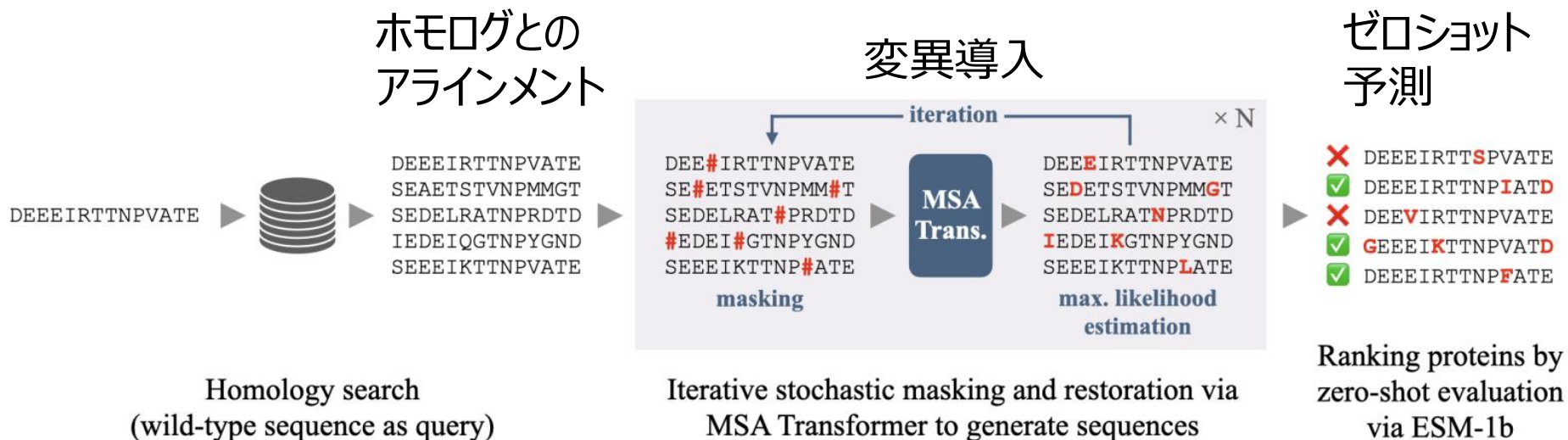
Pretrained model

log (機能値)



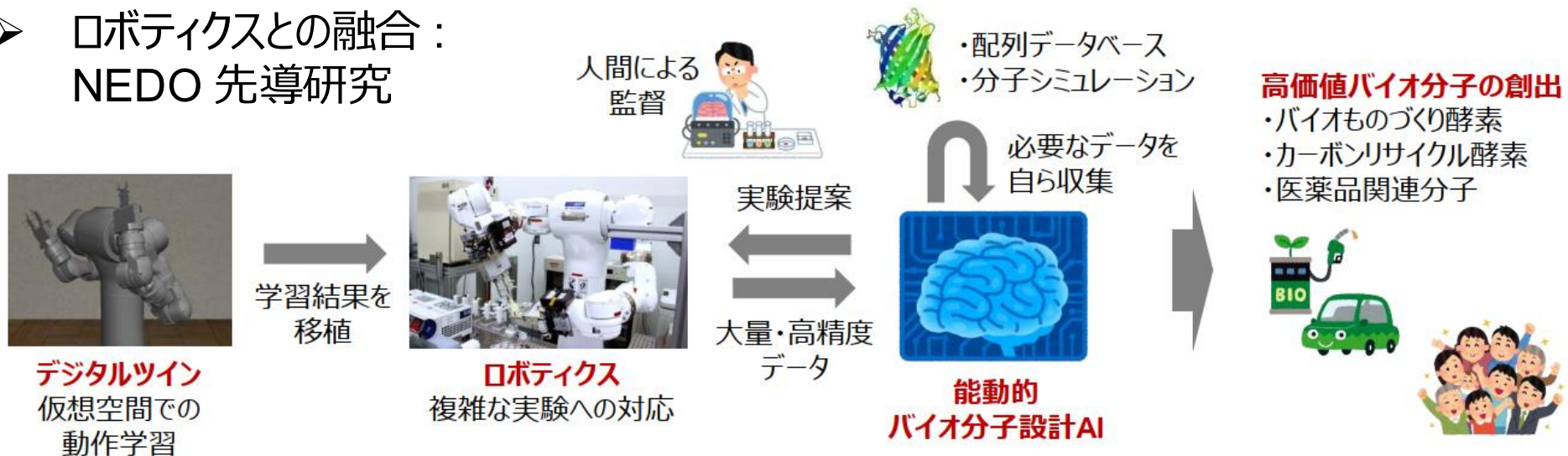
EvoOpt : ゼロショット予測による配列最適化

- 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



その後の展開

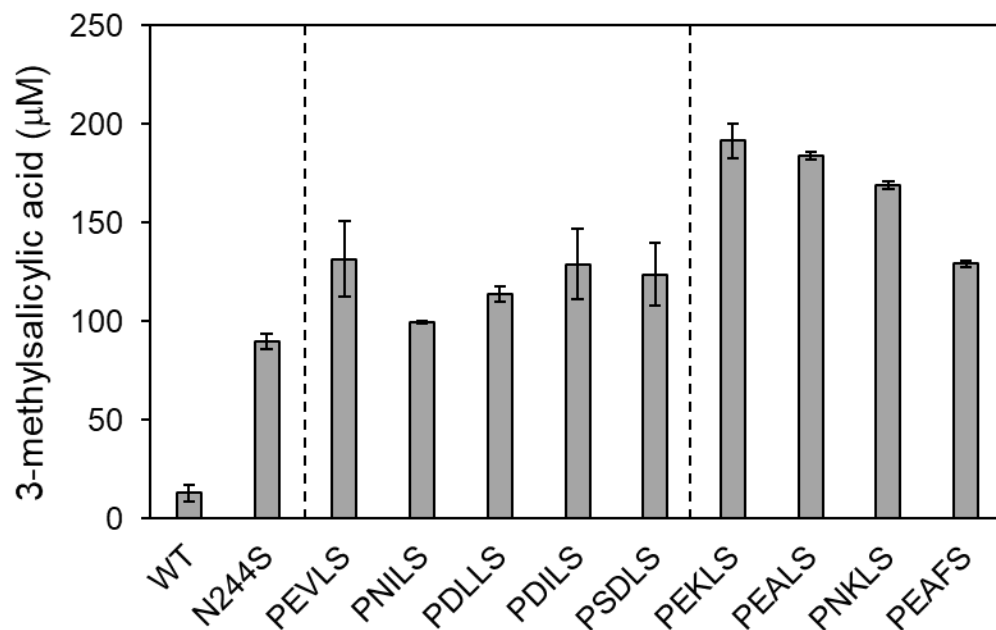
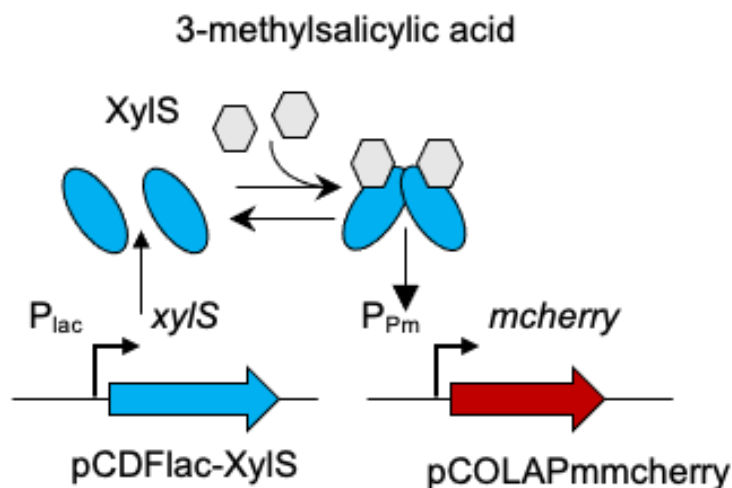
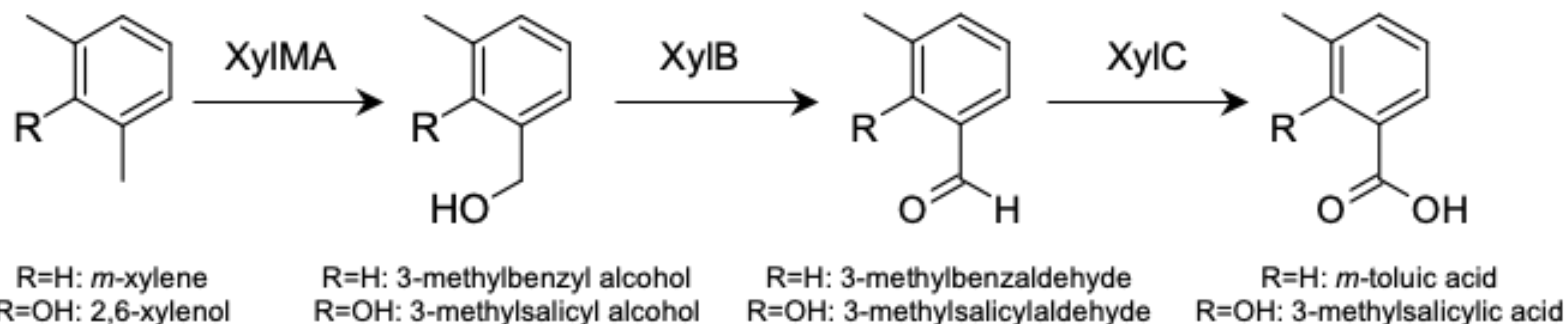
- バイオものづくり：
酵素の基質特異性改変
[Ogawa et al. ACS Synth Biol. 2023] [Matsushita et al. ACS Catal. 2024]
- バイオ医薬品：
抗体模倣分子の結合親和性向上 [Ito et al. mAbs. 2023]
- 量子 + AI による探索空間の拡大：
抗体の標的結合部位全体の設計 [Andrejs et al. Sci Rep. 2024]
- ロボティクスとの融合：
NEDO 先導研究



酵素 XylIM の基質特性改変

共同研究
東大・勝山陽平先生

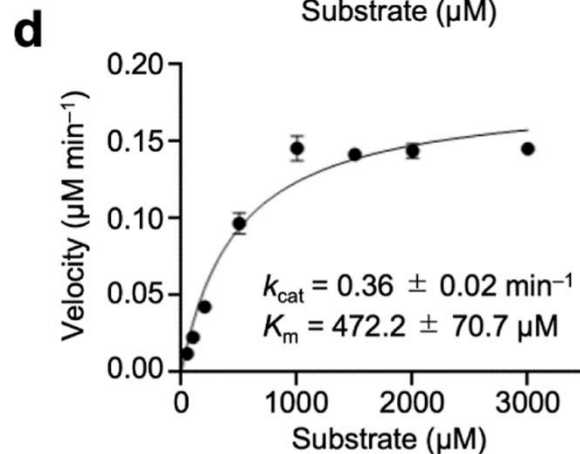
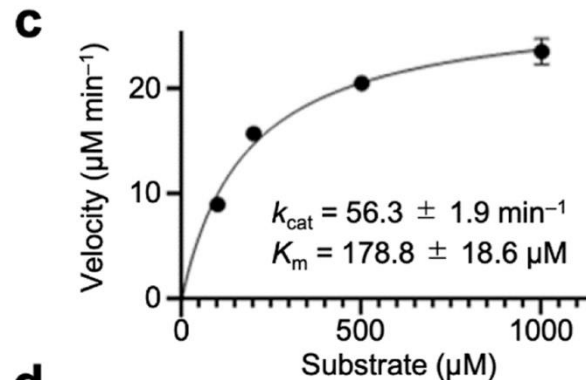
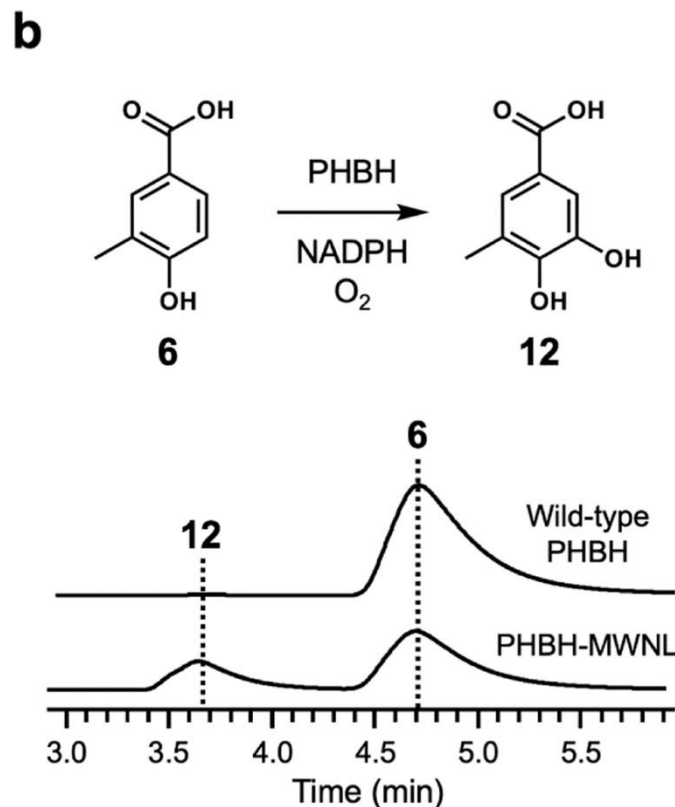
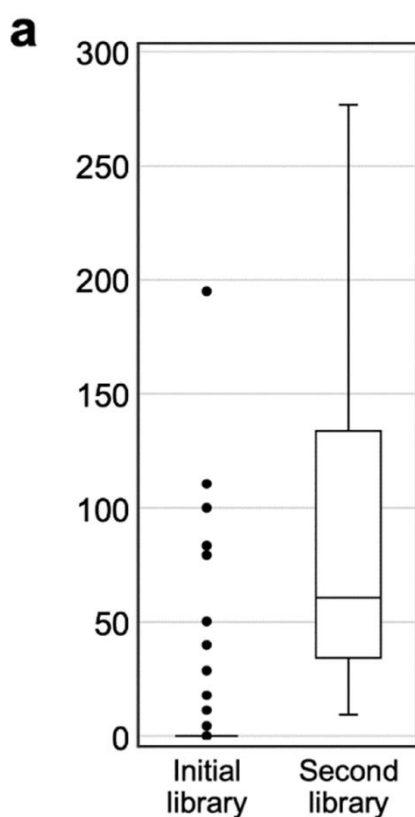
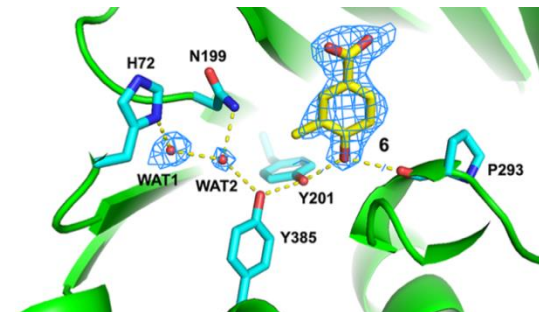
- 非天然基質 2,6-xyleneol への活性を 2 ラウンドの機械学習で 15 倍向上
多段階反応の教師データを効率よく取得するため新規アッセイ系を利用



酵素 FMO の基質特性改変

共同研究
静岡県大・渡辺賢二先生

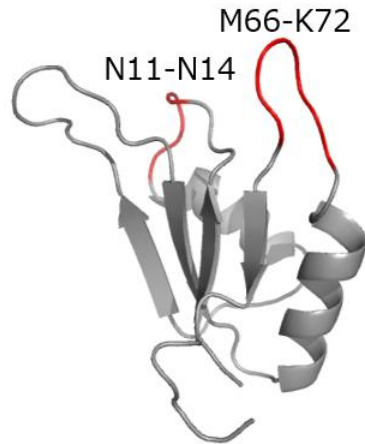
- *p*-hydroxybenzoate hydroxylase (PHBH) を題材として
野生型では活性を示さない基質への反応性を付与
変異体の立体構造から反応機構を考察



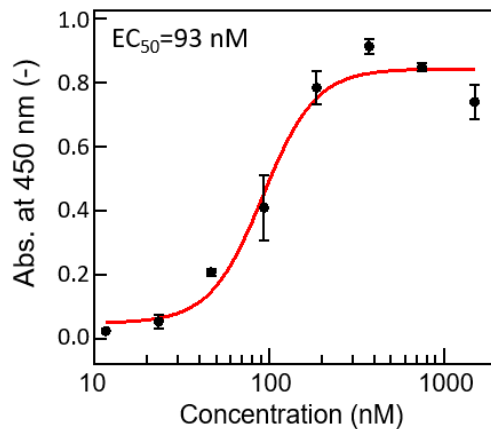
抗体模倣分子の結合親和性向上

共同研究
東北大・梅津光央先生

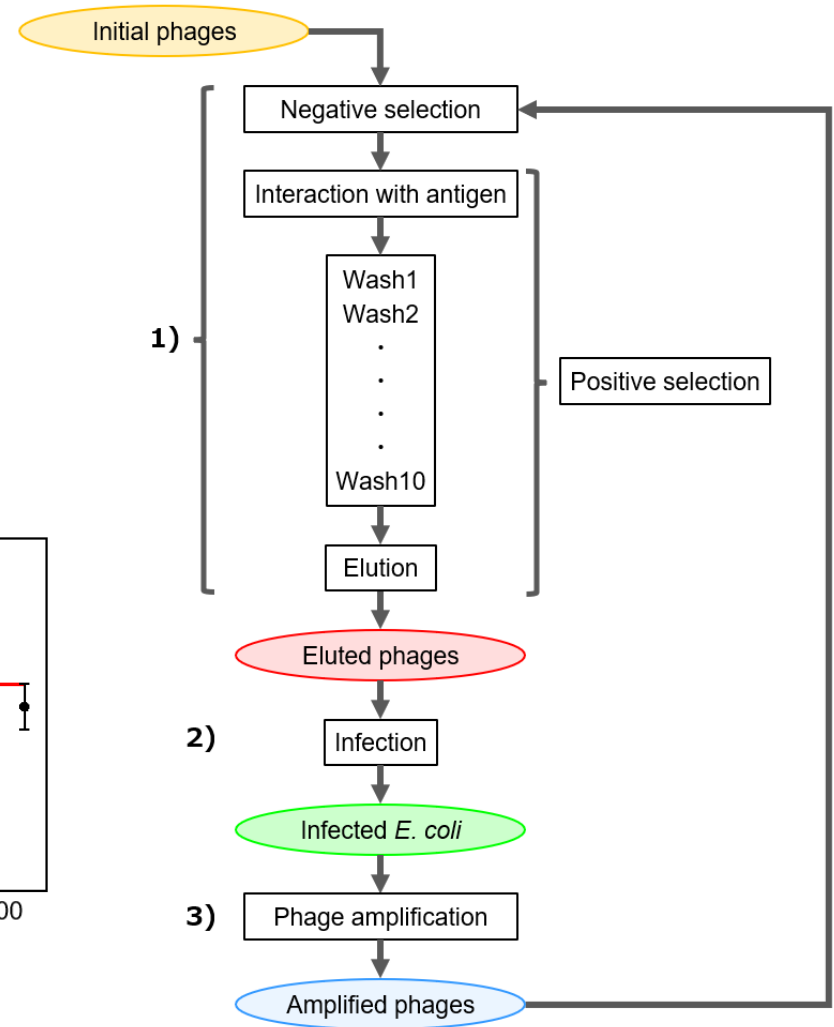
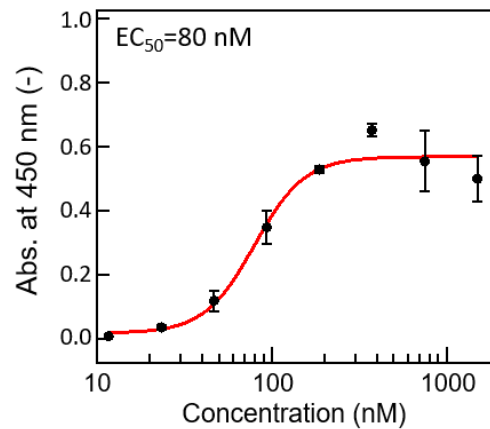
- 非抗体骨格のループ領域を改変して標的 galectin-3 への結合親和性を付与
ファージディスプレイライブラリの NGS データを教師データとして利用



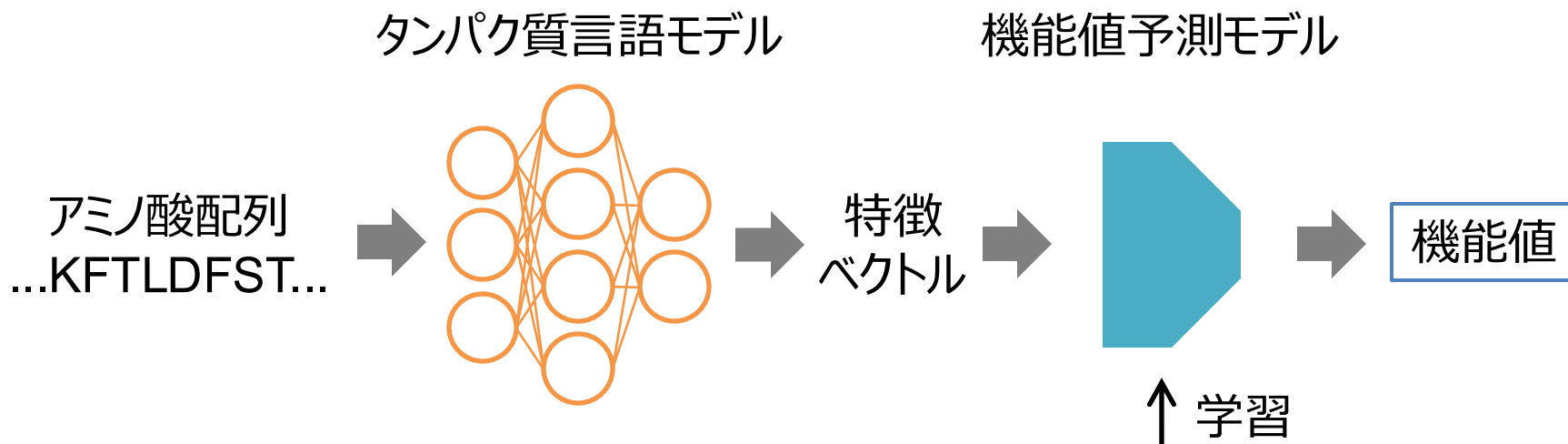
1E2



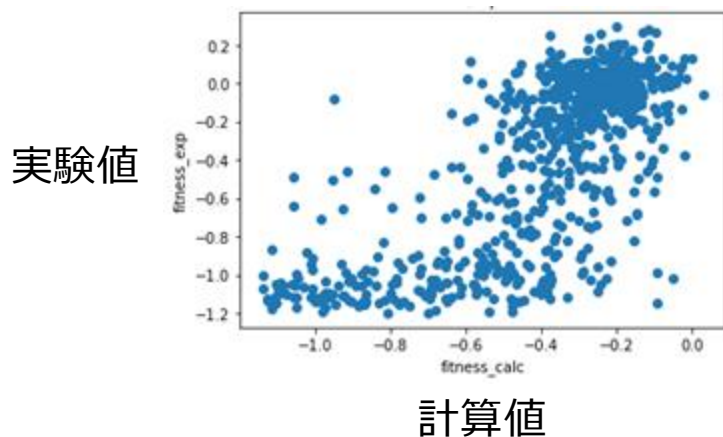
1H2



分子シミュレーションによる教師データ拡充



分子シミュレーションは実験値と相関例. 構造安定性、相関係数 0.674



実験による教師データ
変異体配列、機能値
($10^1 \sim 10^2$ 規模)

+

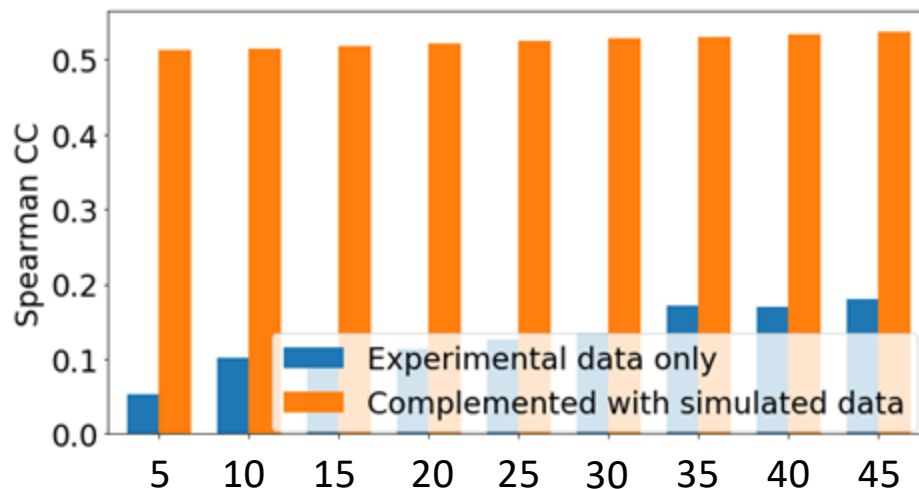
分子シミュレーションによる
「弱教師」データ
($10^2 \sim 10^3$ 規模)

実験データ数が少ない（50以下）状況での予測精度を大幅に改善

■ 実験データ+シミュレーション
■ 実験データのみ

結合親和性予測

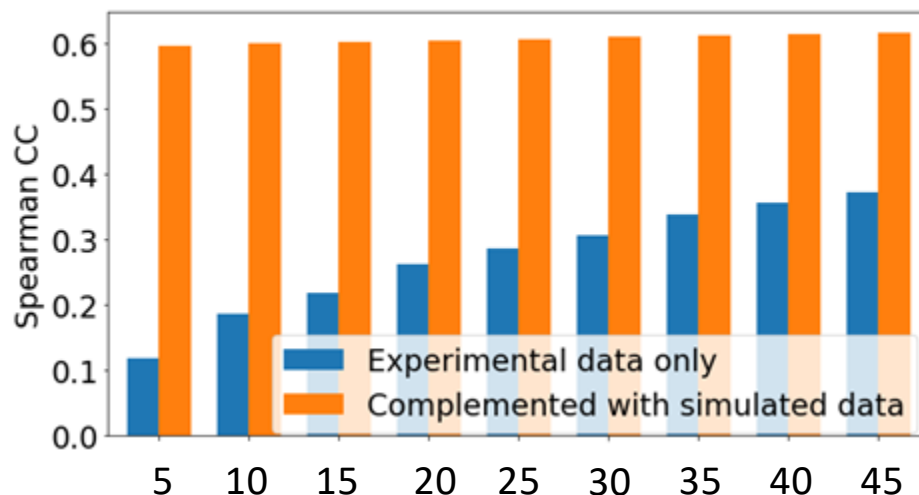
予測
精度



ベンチマークデータ：
GRB2-SH3
[Faure+ Nature. 2022]

安定性予測

予測
精度



実験データ数

[Deguchi et al. under revision. 2025]

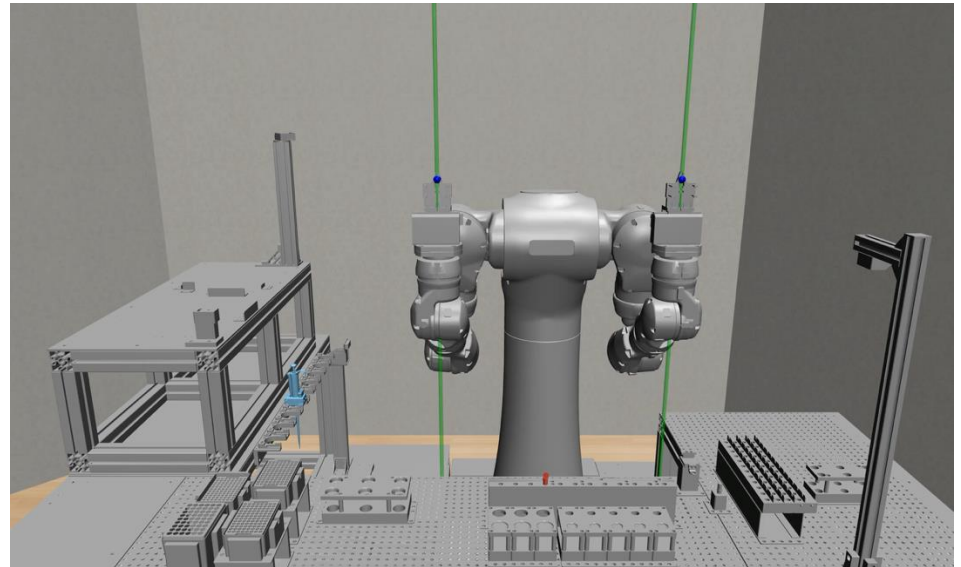
ラボオートメーション、ロボティックバイオロジー

- 実験ロボットによるラボ自動化：
「実験の上手い下手」「ラボ秘伝のコツ」からの脱却に向けて

ロボティクス利用：再現性が高く、プロトコル移転が可能



産総研で開発された
実験ロボット「まほろ」



バーチャルまほろ（デジタルツイン）
物理シミュレータ上で動作を強化学習
実機に動作を移植

セントラルドグマを貫く生体分子設計

- Designing biomolecules throughout the central dogma

DNA



プロモータ DNA の設計 :
転写量を増やしたい
[Saito et al. in prep]

転写



RNA

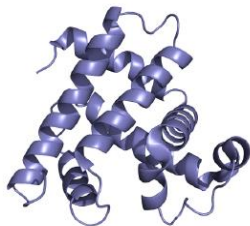


mRNA の設計 :
翻訳量を増やしたい (コドン最適化)
[Saito et al. Sci Rep. 2019]

翻訳

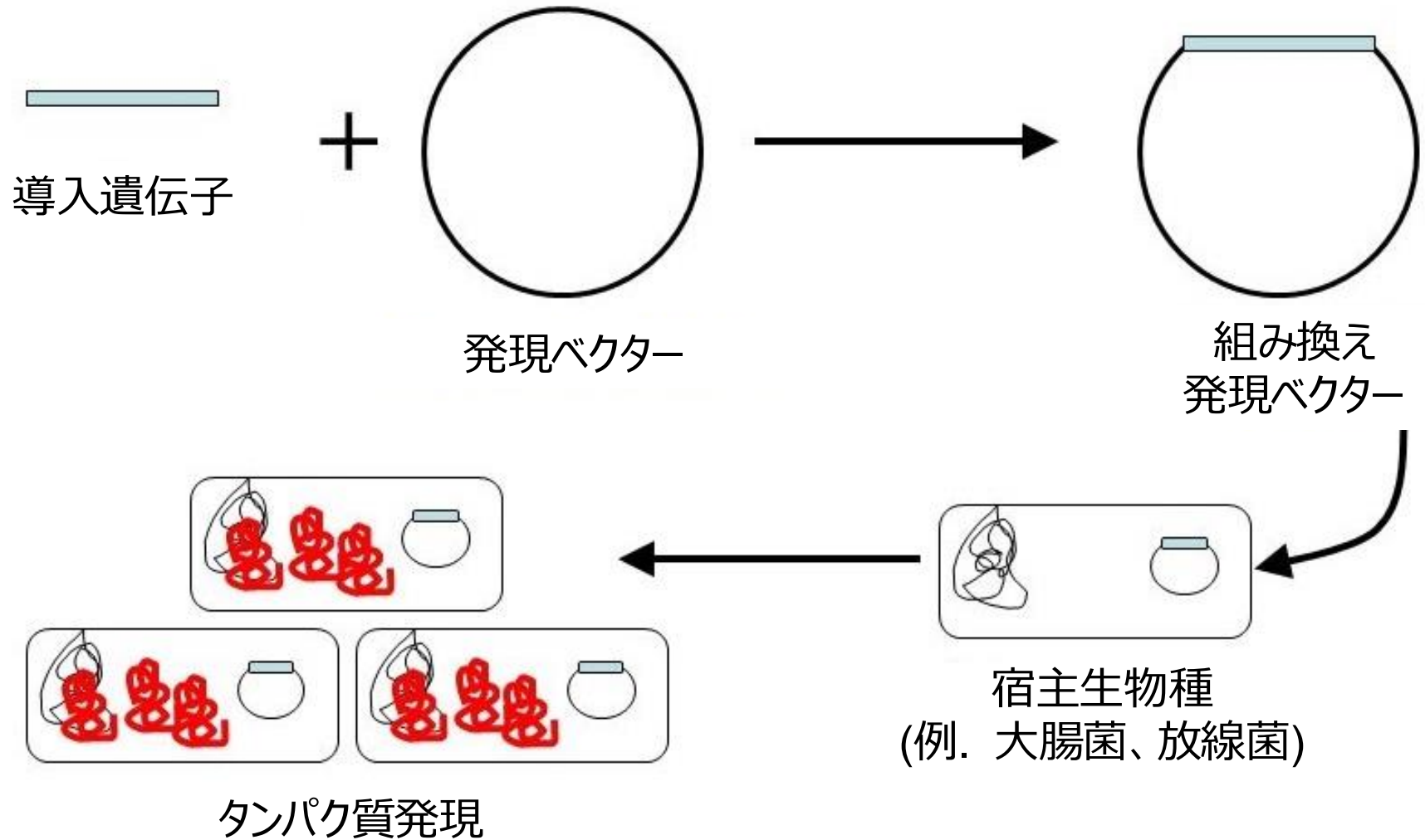


タンパク質



タンパク質の設計 :
機能を改良・改変したい
[Saito et al. ACS Synth Biol. 2018]

タンパク質の異種発現



微生物等を利用した有用物質生産において広く行われている

コドン最適化

- タンパク質の発現量を向上するために
アミノ酸配列を変えない範囲で導入遺伝子の塩基配列を設計

		Second letter				
		U	C	A	G	
First letter	U	UUU } Phe UUC } UUA } Leu UUG }	UCU } UCC } Ser UCA } UCG }	UAU } Tyr UAC } UAA Stop UAG Stop	UGU } Cys UGC } UGA Stop UGG Trp	U C A G
	C	CUU } CUC } Leu CUA } CUG }	CCU } CCC } Pro CCA } CCG }	CAU } His CAC } CAA } Gln CAG }	CGU } CGC } Arg CGA } CGG }	U C A G
	A	AUU } AUC } Ile AUA } AUG Met	ACU } ACC } Thr ACA } ACG }	AAU } Asn AAC } AAA } Lys AAG }	AGU } Ser AGC } AGA } Arg AGG }	U C A G
	G	GUU } GUC } Val GUA } GUG }	GCU } GCC } Ala GCA } GCG }	GAU } Asp GAC } GAA } Glu GAG }	GGU } GGC } Gly GGA } GGG }	U C A G

遺伝暗号は
20 種類のアミノ酸を
64 種類のコドンで表現

あるアミノ酸配列を
コードする塩基配列は
およそ 3^L 種類存在
(配列長 L)

コドン最適化

- タンパク質の発現量を向上するために
アミノ酸配列を変えない範囲で導入遺伝子の塩基配列を設計
- $O(3^L)$ の候補配列（同義置換体）から
どのように「最適」な配列を選ぶか？

アミノ酸配列		M	S	T	A	A	E	T	T	G	G	D	
野生型		ATGTCTACGGCGGCCGAGACGACGGGCGGGCGAC											
同義 置換体	{	ATG TCC ACGGCGGCCGAGACGACGGGCGGGCGAC											
		ATG TCA ACGGCGGCCGAGACGACGGGCGGGCGAC											
		ATG TCG ACGGCGGCCGAGACGACGGGCGGGCGAC											
		ATG AGT ACGGCGGCCGAGACGACGGGCGGGCGAC											
		ATG AGC ACCGCAGCAGAGACCACCGGAGGAGAC											
		ATGTCT ACT GCGGCCGAGACGACGGGCGGGCGAC											
		ATGTCT ACC GCGGCCGAGACGACGGGCGGGCGAC											
		ATGTCT ACA GCGGCCGAGACGACGGGCGGGCGAC											
		⋮											

コドン頻度

➤ 例. 大腸菌 K-12 株のゲノムから計算したコドン頻度

		U	C	A	G
U	U	Phe 2.22	Ser 0.80	Try 1.60	Cys 0.50
	C	1.65	0.86	1.22	0.64
	A	Leu 1.39	0.71		
	G	1.36	0.89		Trp 1.53
C	U	Leu 1.10	Pro 0.70	His 1.29	Arg 2.09
	C	1.11	0.55	0.97	2.2
	A	0.39	0.84	Gln 1.54	0.35
	G	5.29	2.32	2.89	0.54
A	U	Ile 3.04	Thr 0.89	Asn 1.77	Ser 0.87
	C	2.52	2.34	2.16	1.60
	A	0.43	0.70	Lys 3.37	Arg 0.20
	G	Met 2.78	1.44	1.03	0.11
G	U	Val 1.83	Ala 1.53	Asp 3.22	Gly 2.48
	C	1.53	2.56	1.91	2.97
	A	1.09	2.02	Glu 3.96	0.79
	G	2.62	3.37	1.78	1.11

よく行われるコドン最適化 :
宿主のコドン頻度に学ぶ

大腸菌は Phe をコードするのに
UUU を最もよく使用している



導入遺伝子の Phe にも
UUU を使おう

Codon adaptation index (CAI)

$$CAI = \left(\prod_{i=1}^L w_i \right)^{\frac{1}{L}}$$

w_i : relative codon frequency

mRNA の 2 次構造

- 2 次構造がリボソームの進行を妨げ翻訳を阻害
特に、開始コドン付近の 2 次構造の安定性 (ΔG) が重要
- 2 次構造を不安定化させるように同義置換を導入して
タンパク質の発現量向上を試みる

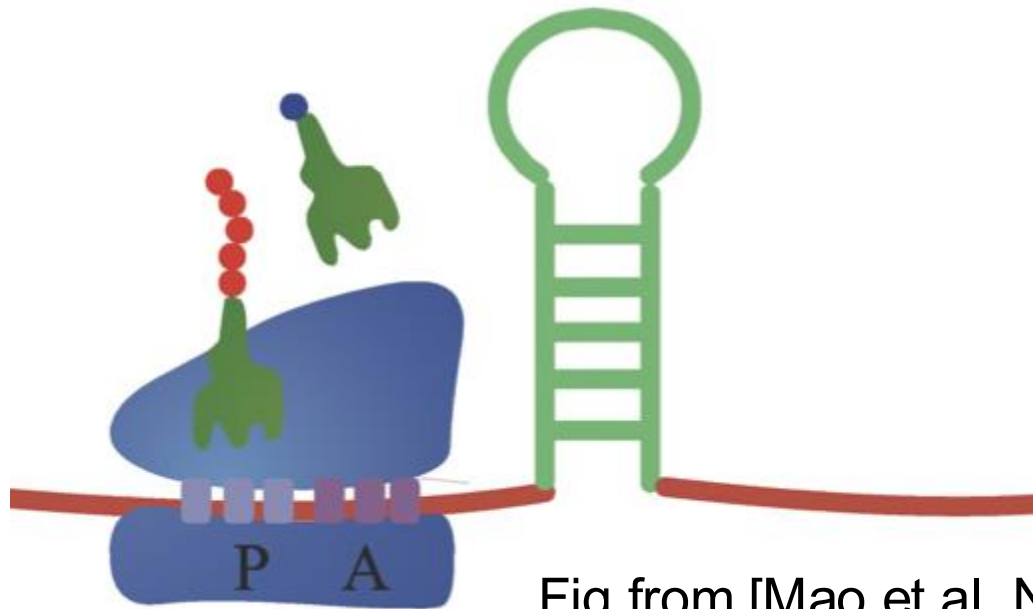
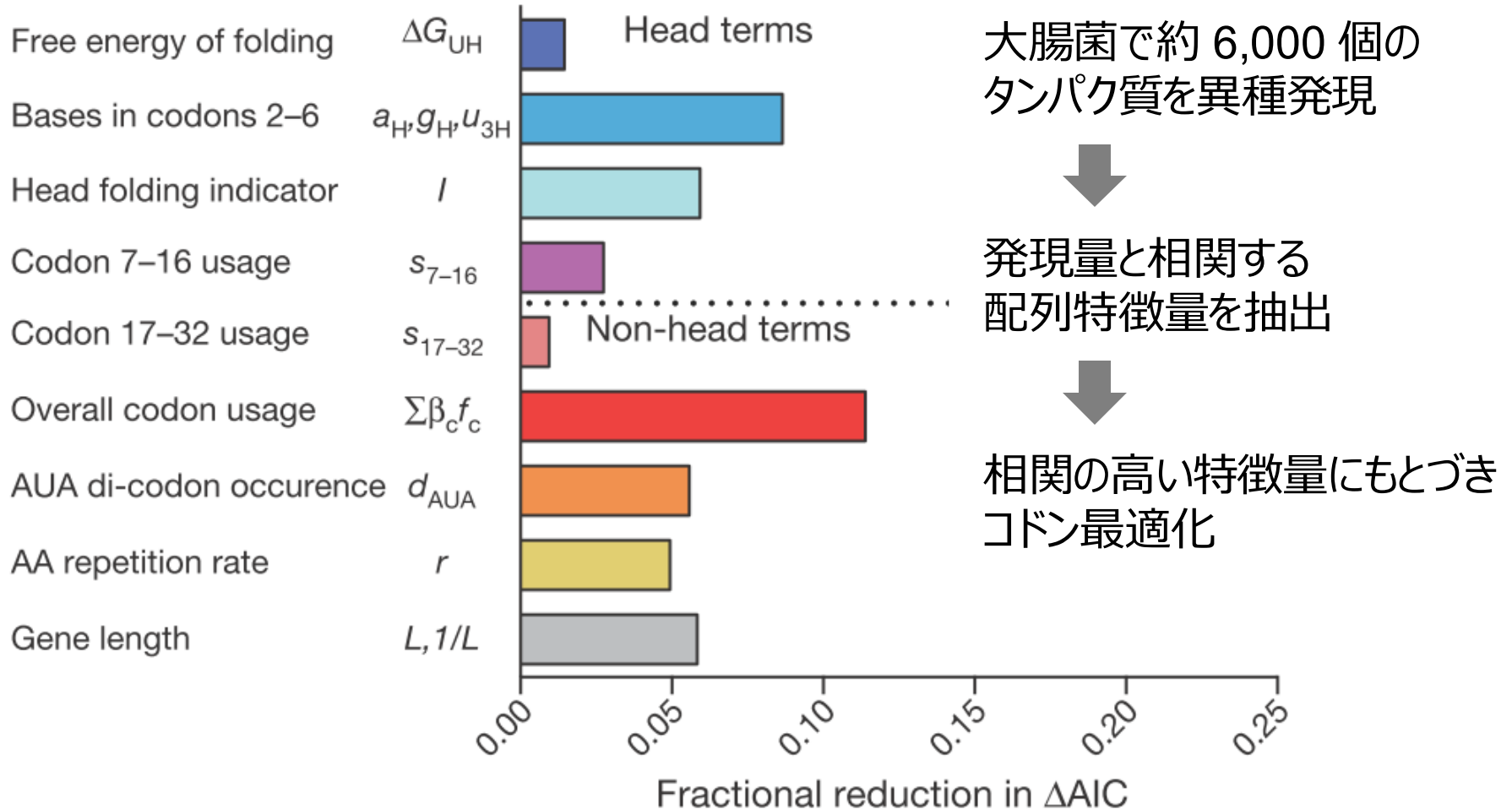


Fig from [Mao et al. NAR. 2014]

コドン頻度と 2 次構造にもとづくコドン最適化が昔から経験的に行われてきた
その他に考慮すべきことは？

コドン最適化のための配列特徴量

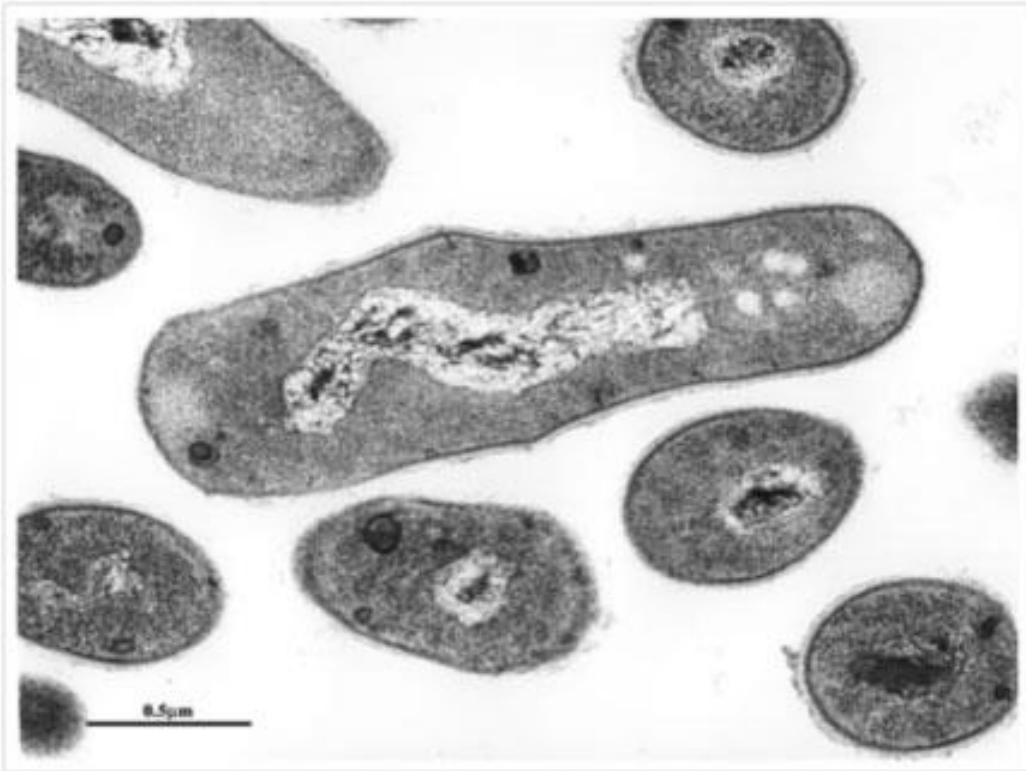
➤ 大腸菌宿主での先行研究 [Boel et al. Nature. 2016]



大腸菌のようなモデル生物以外では、こうしたデータ駆動的な研究は少ない

放線菌 *Rhodococcus erythropolis*

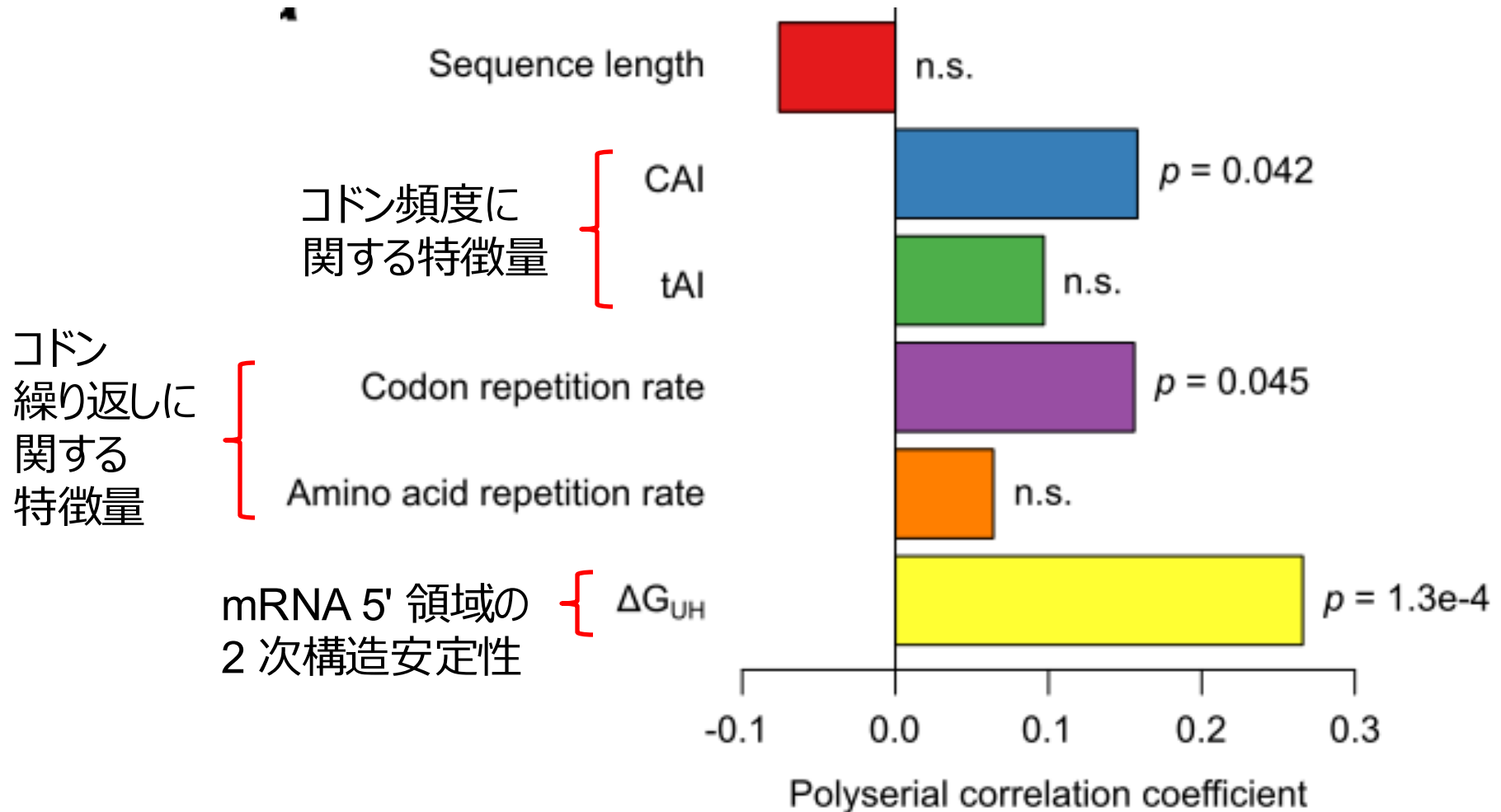
- 高 GC 含量のグラム陽性細菌
- 4 ～ 35 °Cの広い温度域で増殖可能
- 有機溶媒耐性
- 既存の汎用宿主では発現できないタンパク質も生産可能



R. erythropolis 宿主のための
コドン最適化手法の開発

NEDO スマートセル
産総研生物プロセス部門との
共同研究 (2016 年～)

発現量と配列特徴量の相関

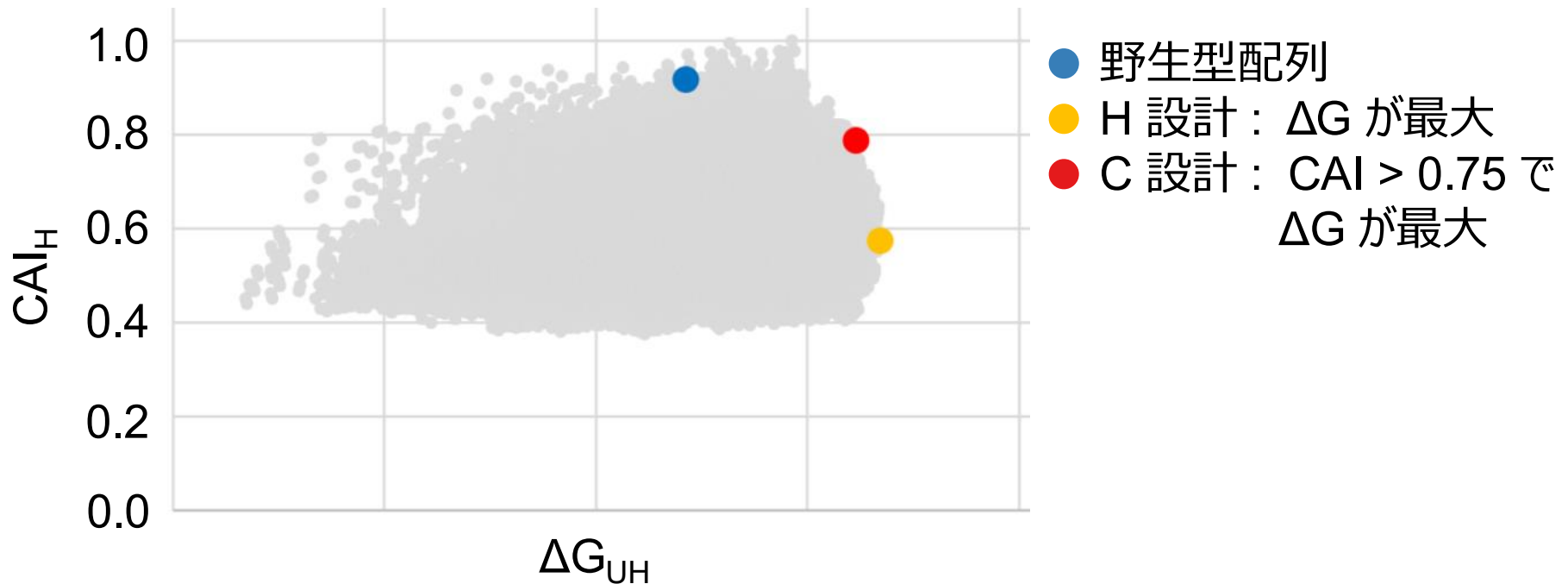


発現量と最も高い相関を示した
CAI, ΔG にもとづくコドン最適化手法を開発

2 次構造とコドン頻度の両方を考慮

- C 設計 : CAI が閾値以上で、 ΔG が最大の配列を選択

ある遺伝子 X の全ての同義置換体の分布



C1 : ΔG 最大、 $CAI > 0.6$ 、レアコドンなし

C2 : ΔG 最大、 $CAI > 0.75$ 、レアコドンなし

C3 : ΔG 最大、 $CAI > 0.9$ 、レアコドンなし

12 遺伝子について検証

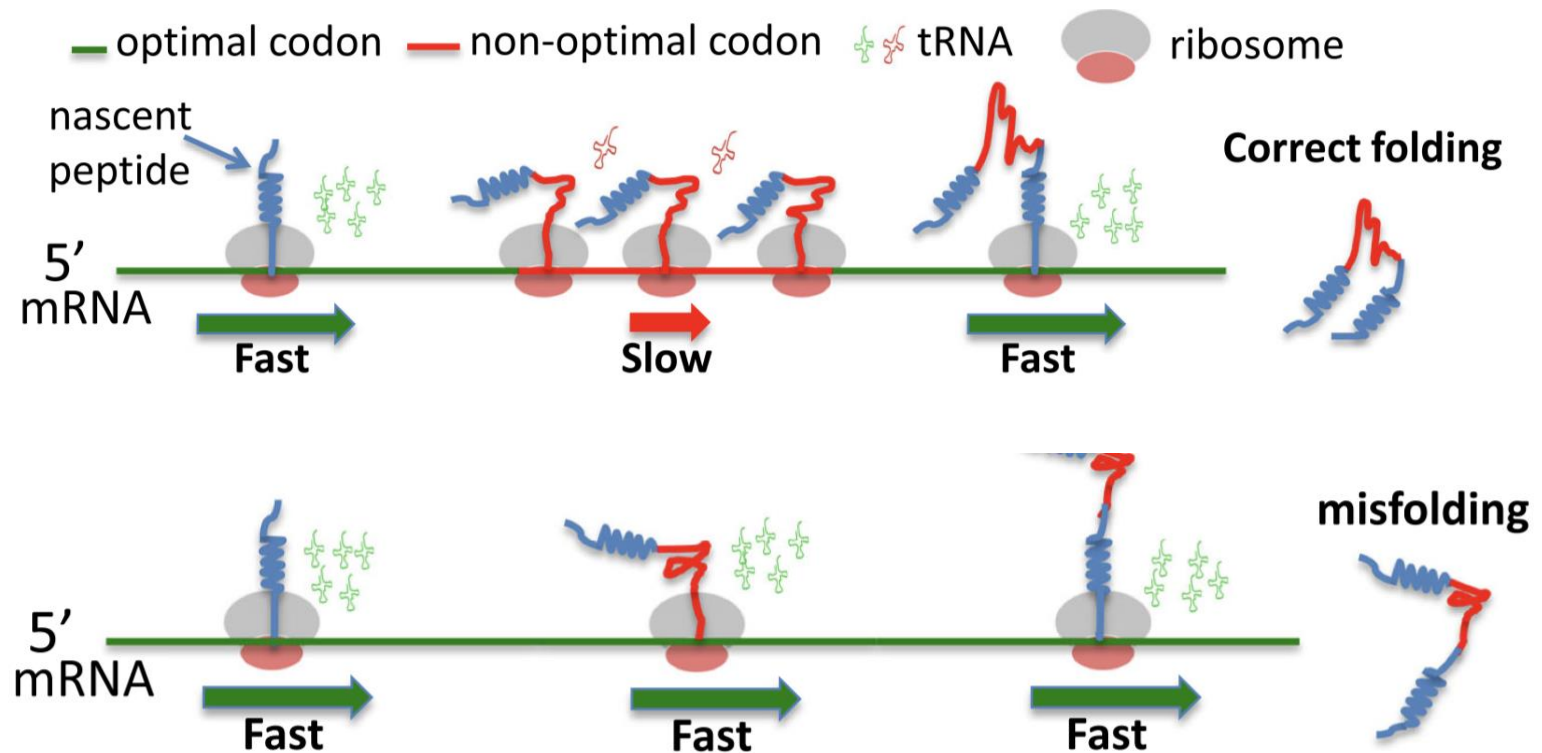
- C 設計により、12 遺伝子中 9 遺伝子の発現向上に成功
- 野生型で低発現の 5 遺伝子については全て発現向上に成功

野生型の 発現量	発現向上	変化なし	発現低下
少	5	0	0
中	3	1	0
大	1	1	1
計	9	2	1

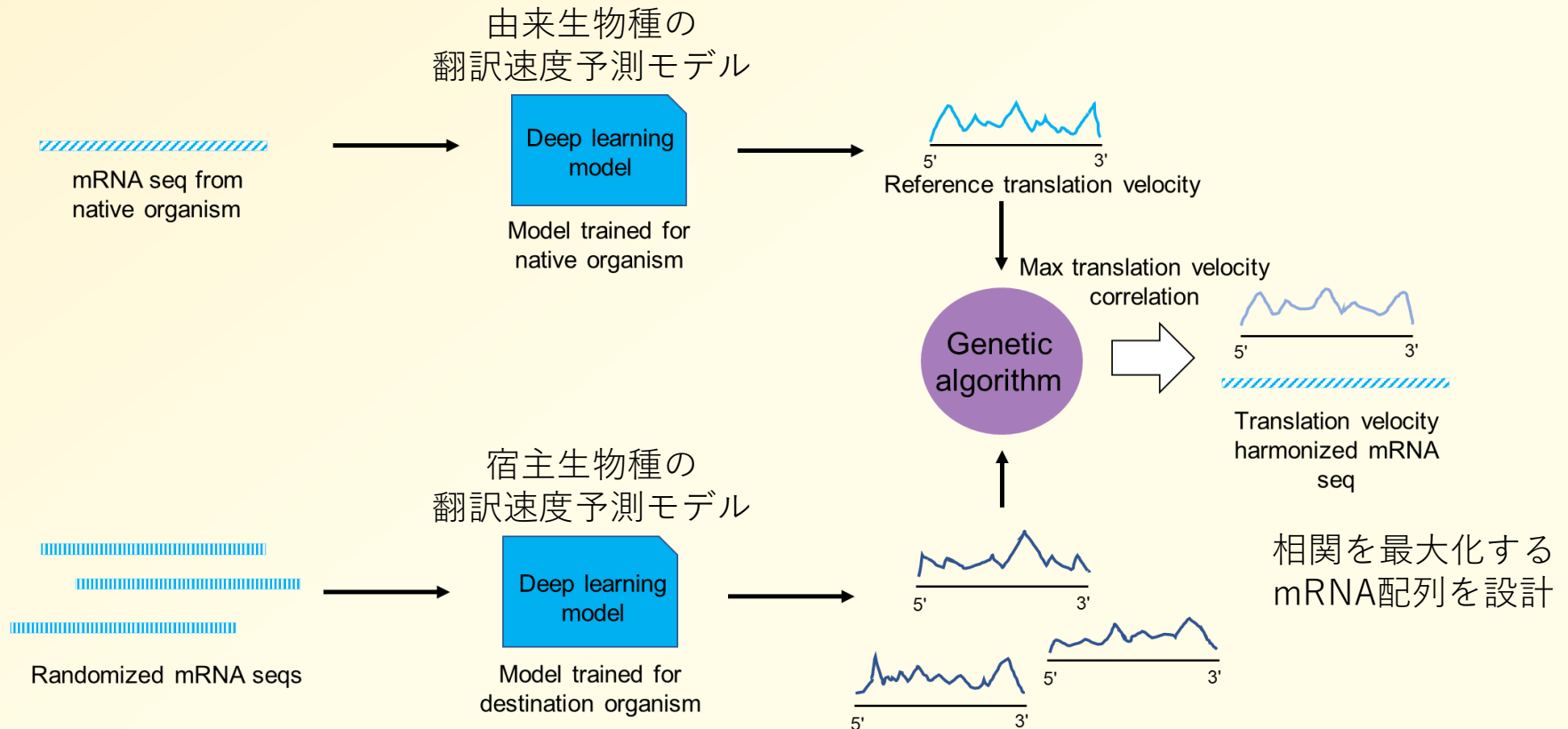
※ 発現向上：バンド強度が 1.1 倍以上
発現低下：バンド強度が 0.9 倍以下

翻訳速度の精密制御に向けて

- 翻訳が速すぎるとタンパク質のフォールディングが間に合わず凝集してしまう
- 翻訳を「適切に」遅くする mRNA を設計したい



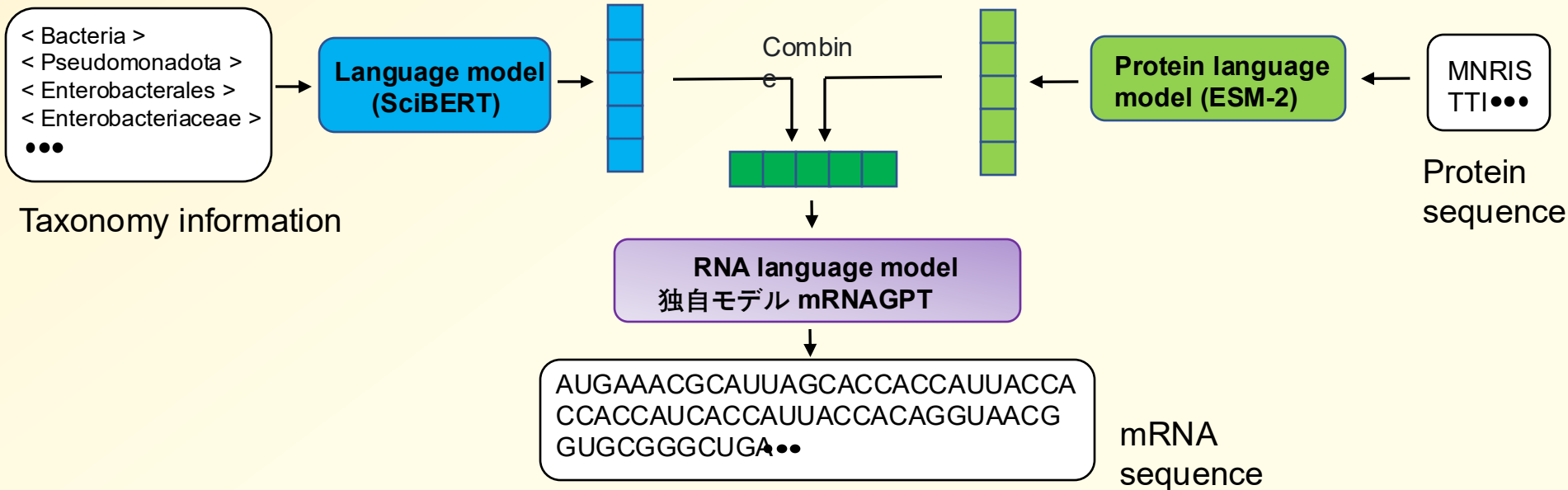
由来生物種での翻訳速度を模倣する宿主mRNA配列を設計



[Bian et al. in prep.]

逆翻訳言語モデルの開発

既存のタンパク質言語モデル、自然言語モデル、RNA言語モデルを初期値として利用し、自然界の「タンパク質、生物種、mRNA」ペアでモデル全体を追加学習



[Bian et al. in prep.]

議論

- 各モダリティ（タンパク質、RNA、DNA、化合物）の基盤モデルは巨大IT企業の参入によって一通り出揃っている感も研究の余地、方向性について考える必要あり
- マルチモーダル化と、それが役立つ実問題の発見
 - 例． 逆翻訳言語モデルは、タンパク質異種発現の問題のためタンパク質言語モデル、RNA言語モデル、自然言語モデルを融合
- ラボオートメーションとの融合、ラボオートメーション自体へのAIの導入自動化から自律化へ

例． バーチャルまほろと sim2real transfer

その他の話題

- 連合学習により日本国内の大学や企業のデータを集約して強いAIを作る？

ラベル付きデータはラボや企業に死蔵されている
低分子創薬分野では先例あり（京大・奥野先生グループ）

- AIの「省データ」化：少量データでも精度良く予測する

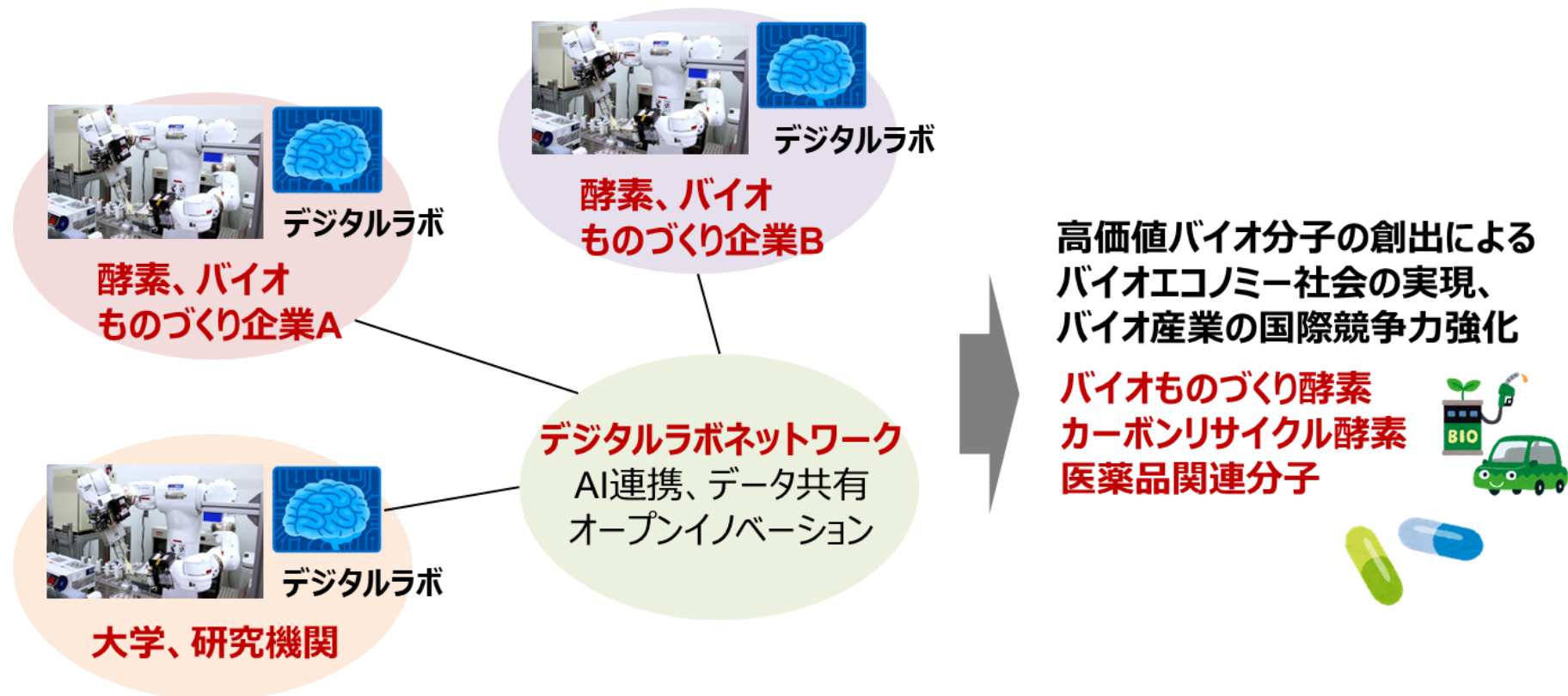
分子シミュレーションによるデータ拡充など

- 「生成」モデルである必要は？

多くの場合、エンコーダ＋探索アルゴリズムや入力微分でも似たようなことはできる

野生型に比べて多数の変異が入るためヒット率は低い
現状、使いこなせているのは数千オーダーで実験検証できるビッグラボくらい？
中堅ラボが使うにはフィルタリングのための手法が必要

社会実装に向けたシナリオ・構想



デジタルラボネットワーク（分散型）と社会実装の構想

謝辞

- 産総研
光山 統泰
亀田 倫史
Thuy Duong Nguyen (当時)
北川 航
- 北里大学
Bian Bian
- 東京大学
山口 秀輝 (当時)
- 東北大学
梅津 光央
- 東京大学
津田 宏治
浅井 潔
- 東京大学
勝山 陽平
- 静岡県立大学
渡辺 賢二
松下 拓磨 (当時)

