



トーゴーの日シンポジウム2024「AI+ロボティクス+データベースが変える生命科学」

パネルディスカッション「生命科学の未来を予想する——データベースはもう要らなくなる...ってコト?!」

基盤モデル時代のオミクス研究とデータベース



二階堂 愛[♂] (Itoshi NIKAIDO)

理化学研究所生命機能科学研究センター バイオインフォマティクス研究開発チーム チームリーダー

理化学研究所 最先端研究プラットフォーム連携 (TRIP) 事業本部, 科学研究基盤モデル開発プログラム (TRIP-AGIS), テーマリーダー

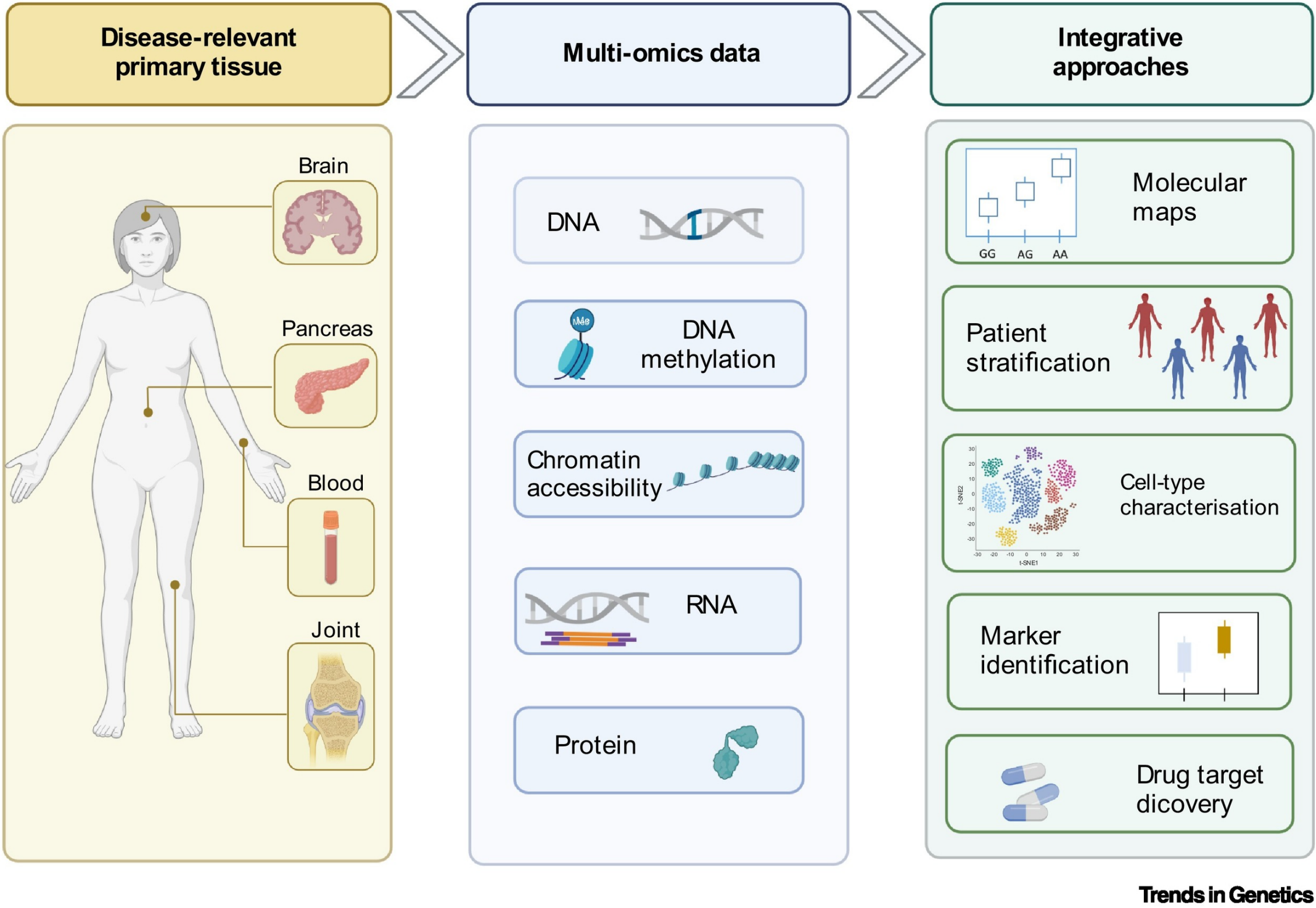
東京科学大学 総合研究院 難治疾患研究所 ゲノム機能情報分野 教授

dritoshi@gmail.com / @dritoshien

基盤モデルとマルチオミクス解析の類似

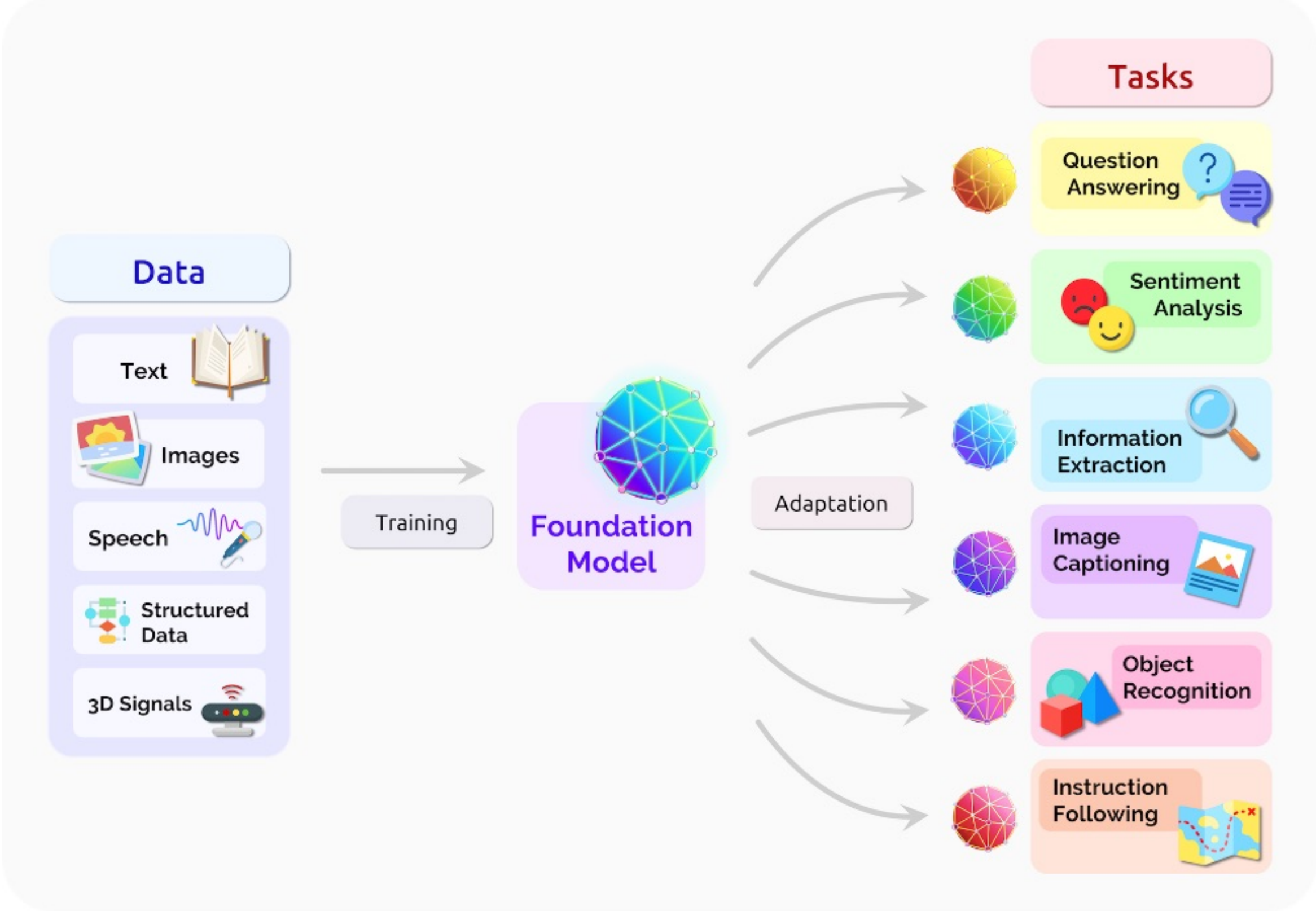
マルチオミクス解析

セントラルドグマを大規模計測し、多様な生命科学の研究タスクに応用



基盤モデル

大規模データを学習し多様なタスクに適応できるモデル

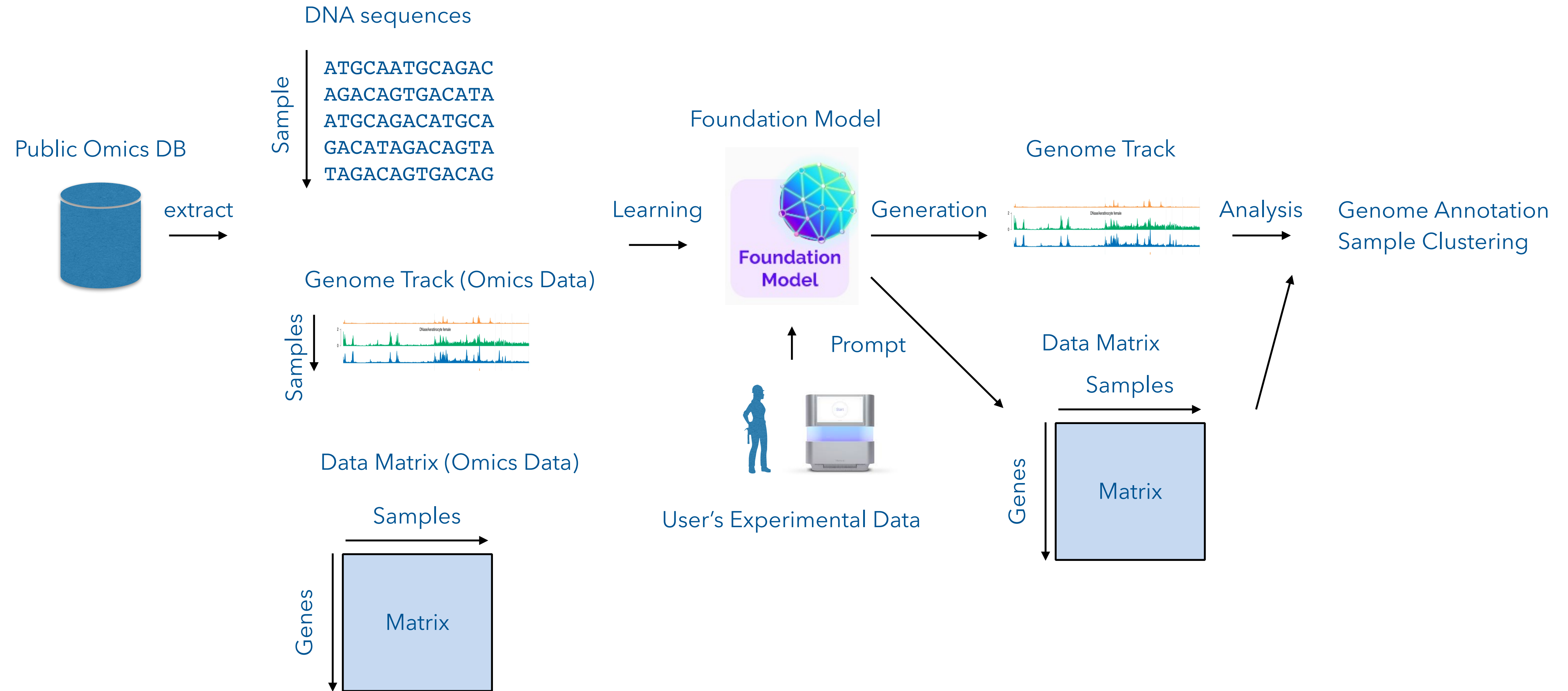


<https://www.cell.com/trends/genetics/fulltext/S0168-9525%2822%2900225-6>

[Bommasani R. et al. On the Opportunities and Risks of Foundation Models. 12 Jul 2022.](#)

オミクス向け基盤モデルのアーキテクチャ

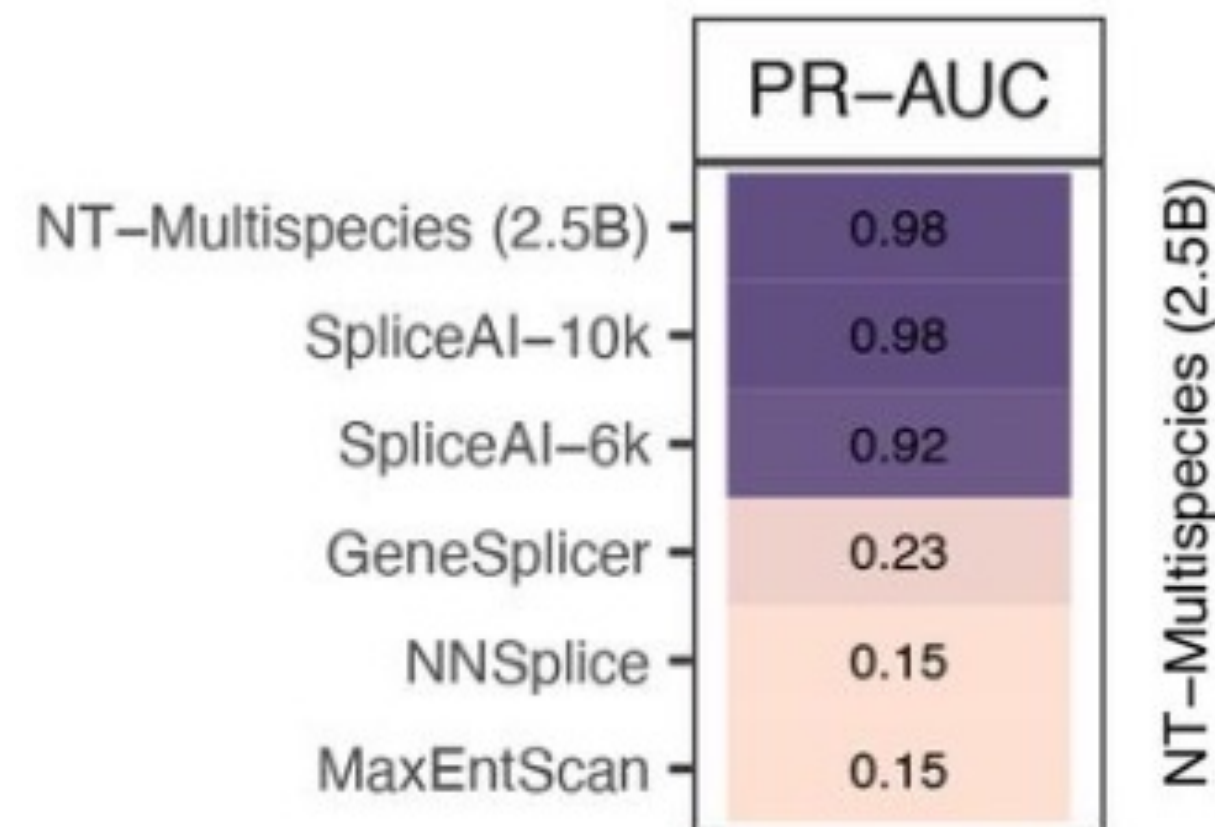
Generation of omics data without wet experiments



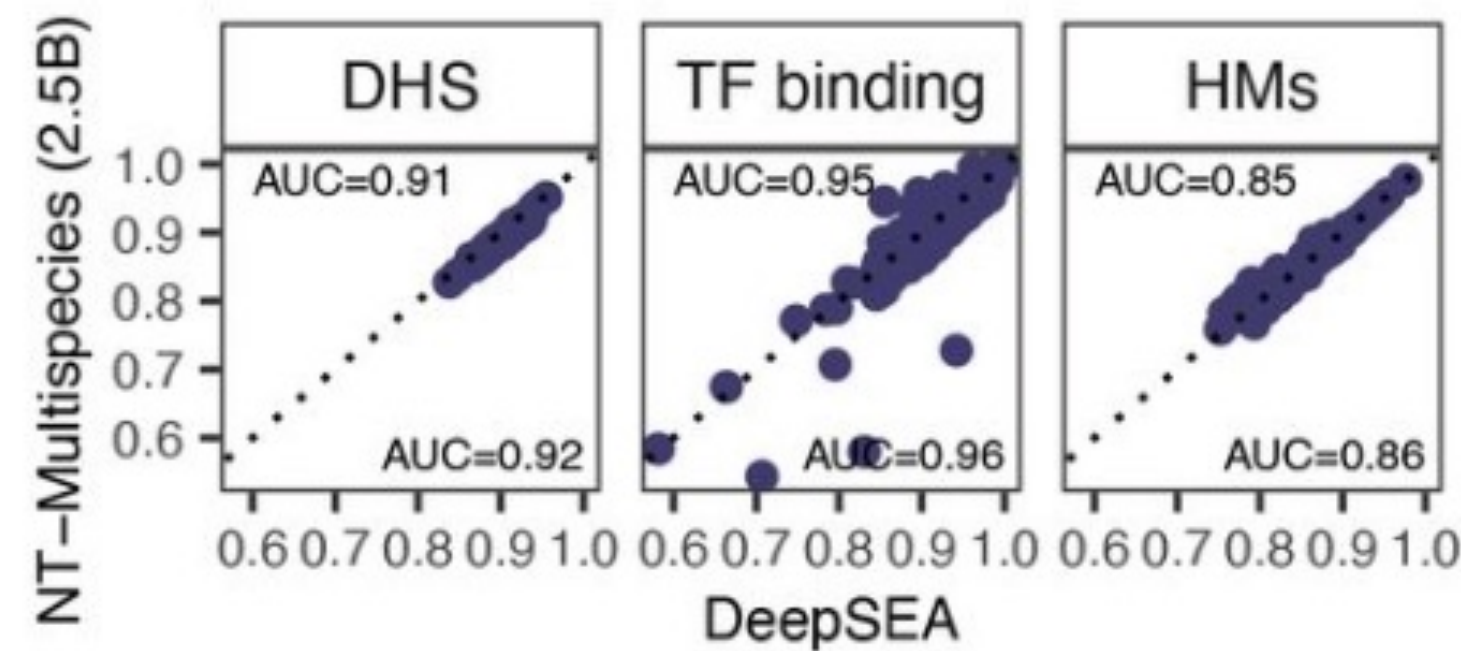
オミクス向け基盤モデルの現状

Generation of omics data without wet experiments

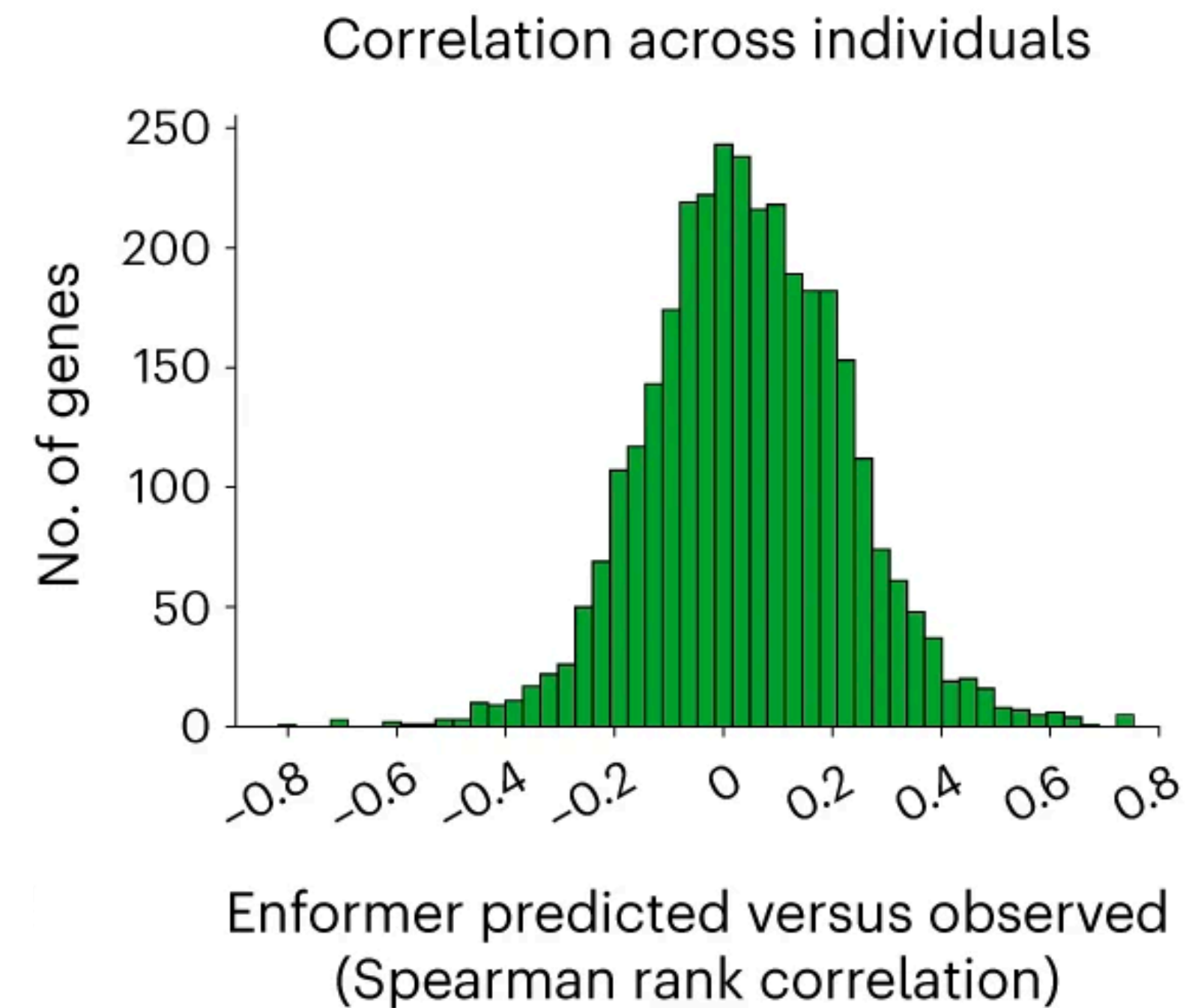
Splice Site Prediction Task



Epigenome Prediction Task

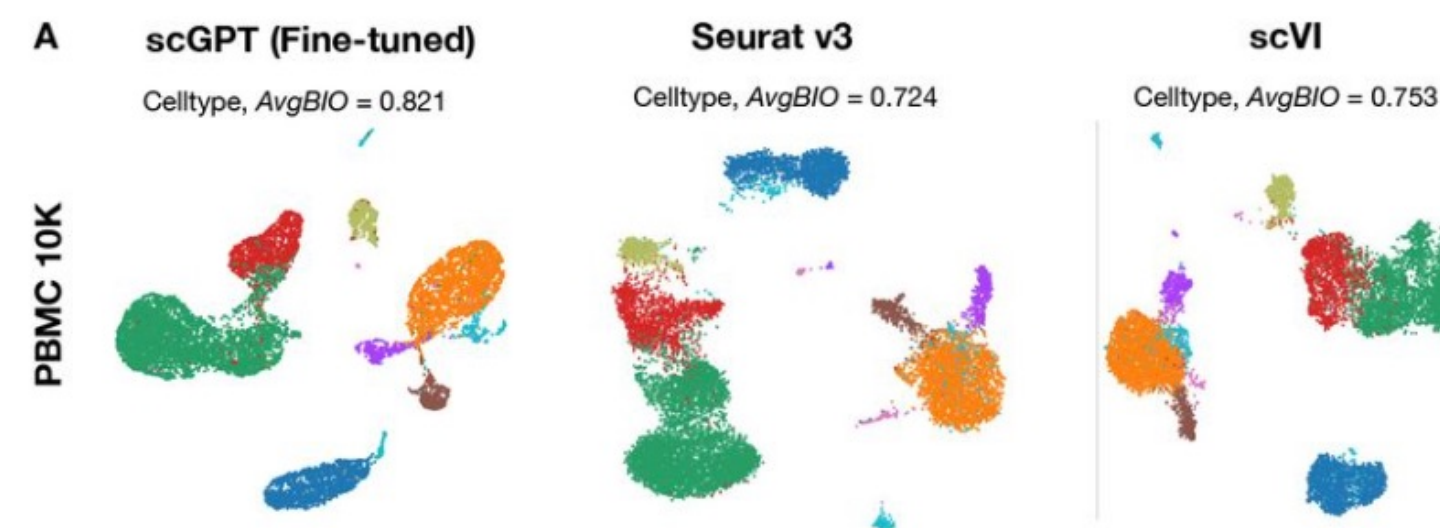


Prediction of gene expression across individuals



<https://www.nature.com/articles/s41588-023-01574-w/figures/1>

Cell Type Clustering Task



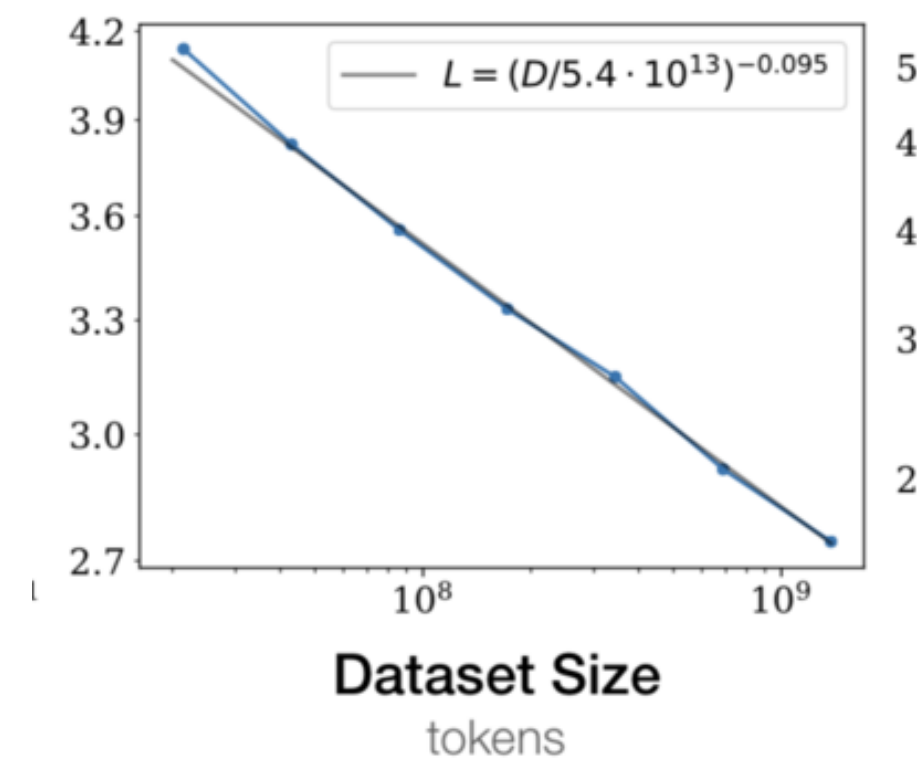
ひとつの基盤モデルが各ドメイン専門アルゴリズムの性能を超える

個人の遺伝子発現の差は予測できない

基盤モデルの3つの性質

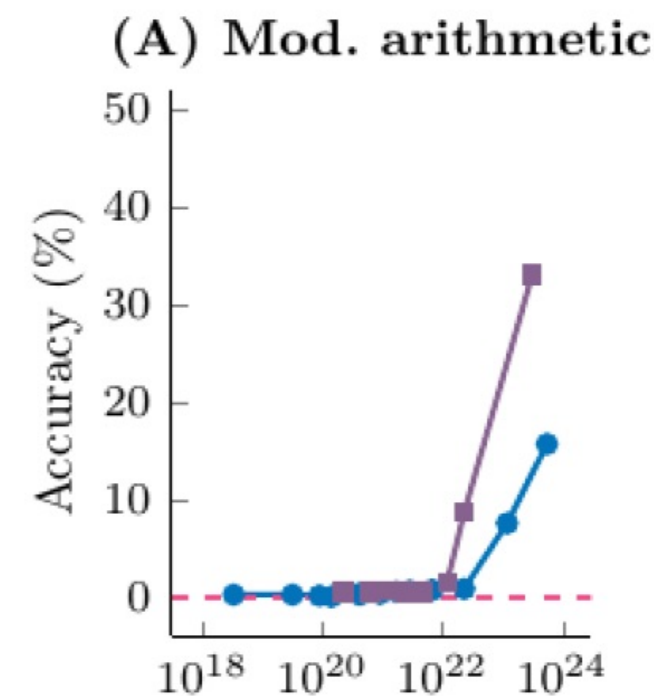
データの量・質・幅が巨大なAIの性能を支える

データの量を増やすほど誤差は減る



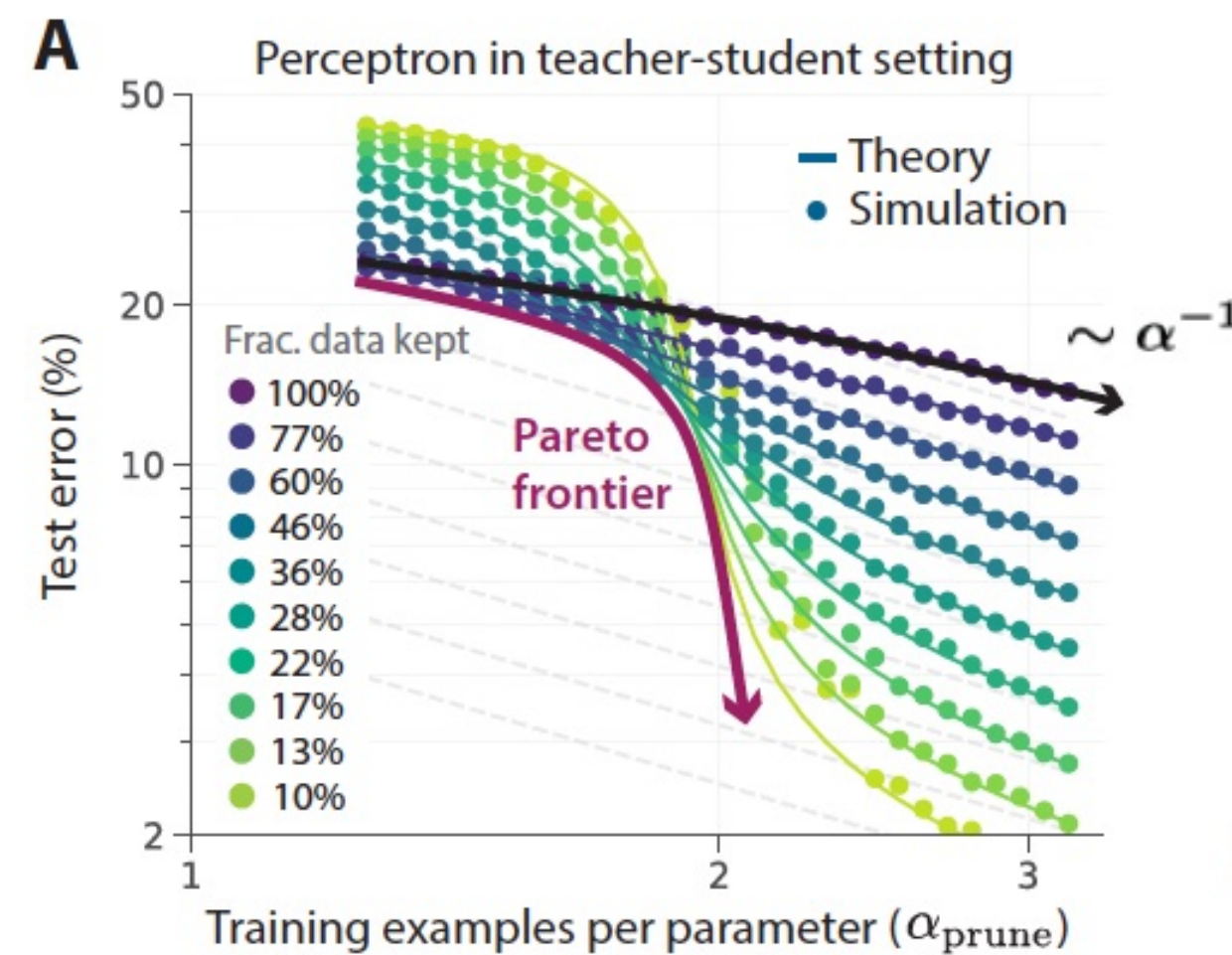
<https://arxiv.org/abs/2010.14701>

データの量を増やすと能力を突然開花させる



<https://openreview.net/forum?id=yzkSU5zdwD>

データの質がよいと効率的に性能向上



Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, Ari S. Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. 2022. <https://doi.org/10.48550/arXiv.2206.14486>

データの幅を増やすと様々なタスクに対応

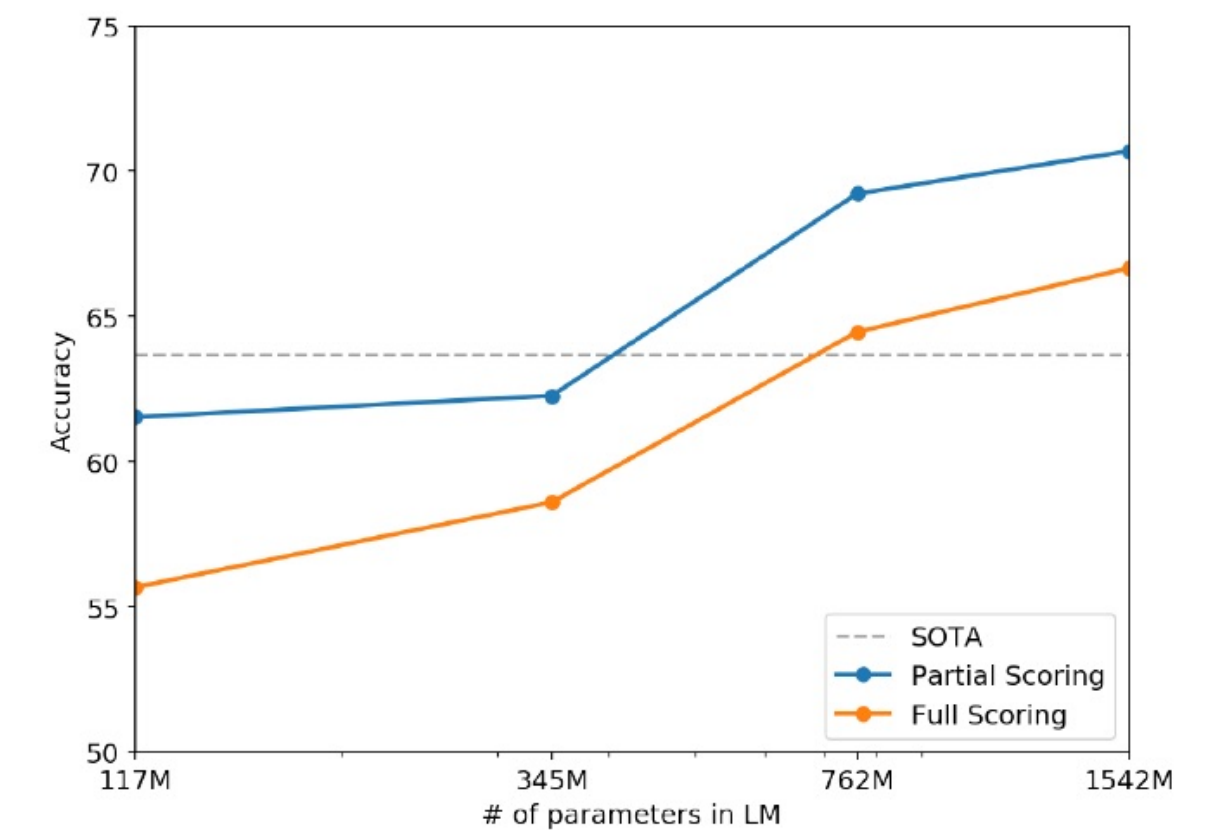
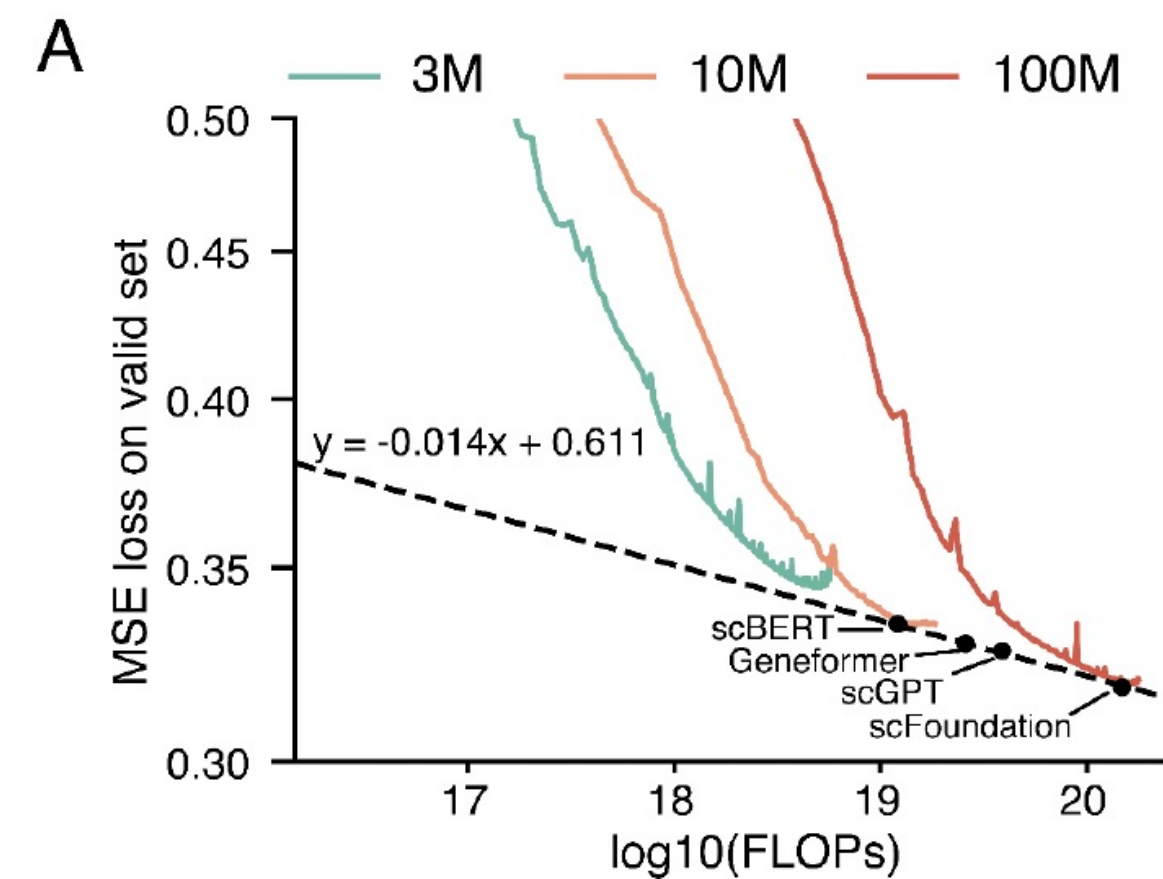
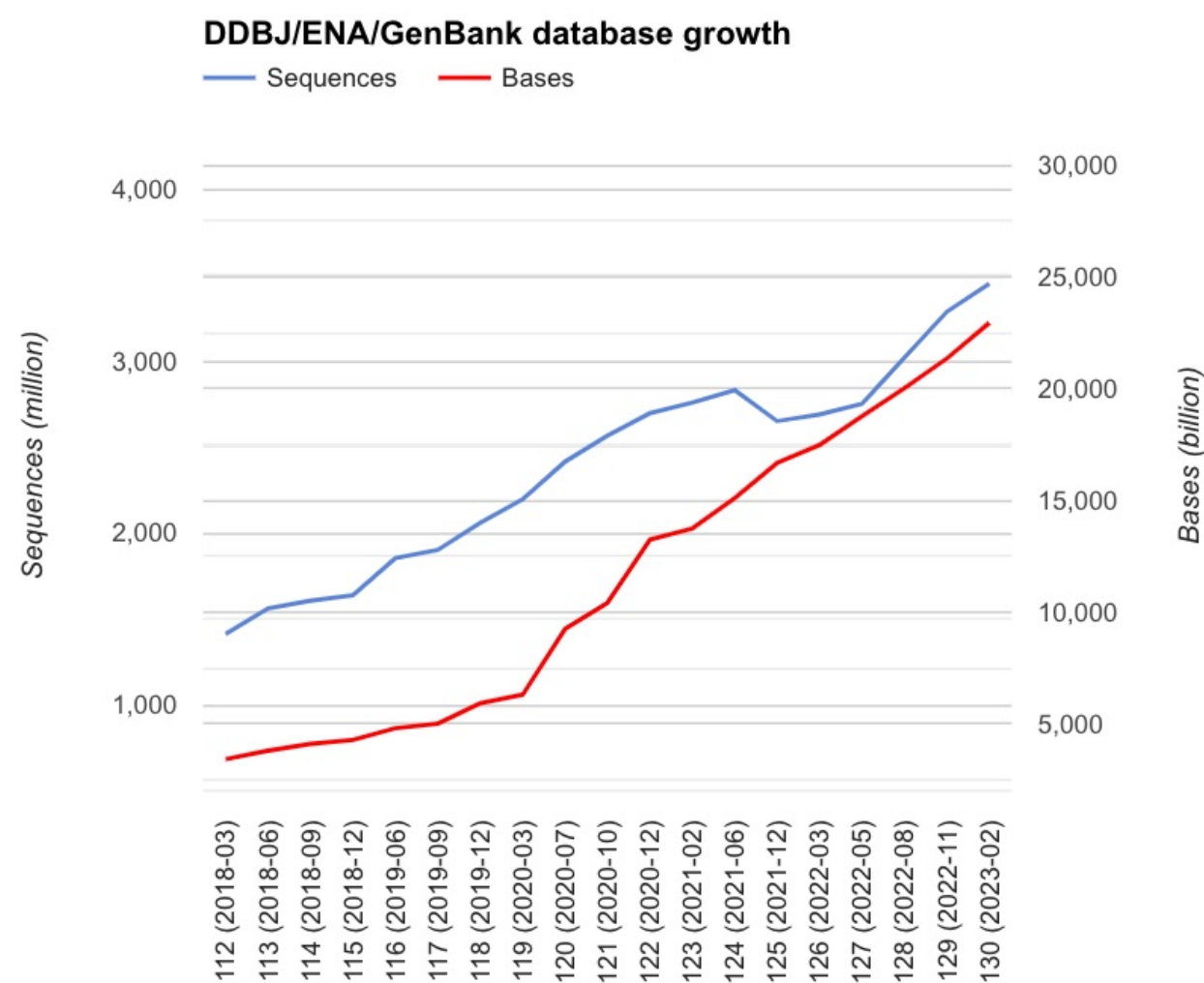


Figure 3. Performance on the Winograd Schema Challenge as a function of model capacity.

Language models are unsupervised multitask learners. OpenAI blog 1.8 (2019): 9.

データ枯渇問題: ゲノム科学は?

- 高品質言語データは2024年5月に枯渇
- シーケンスデータベースのrun数やbase数は線形に増加。2018年から加速している
- データサイズは指数的增长に見えるが2018から加速しておりその後は線形増加



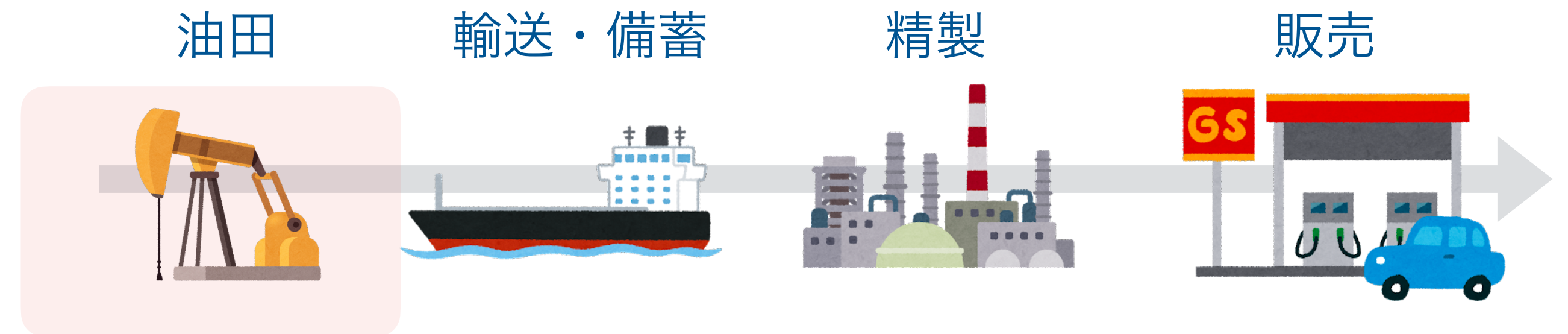
<https://www.nature.com/articles/s41592-024-02305-7>

- シーケンスデータはよりクローズドへ
- 生物資源の制限、ヒトゲノム情報保護で各国にデータが囲い込みになる
- ヒトゲノム
 - 経済安全保障。データ持ち出しが困難に。AMEDでのゲノムシーケンスは国内でのみ許可
 - 海外のデータが手に入らない。中国はヒトゲノムデータは海外に出さない
- 以下DDBJの有田正規さんからの情報
 - 生物多様性条約CBD 2022.12
 - 昆明宣言: 遺伝資源からの利益配分を約束。COP16で詳細は決定
 - 海洋生物多様性BBNJの合意 2023.03
 - 遺伝資源情報から利益配分を明記。Batch Identifier 付与
 - 必要資金の半分以上を先進国が負担
 - キャパシティ・ビルディング
 - From Data FIAR to Data CARE

基盤モデル時代のオミクス研究の勝ち筋

- マルチオミクス解析はひとつのモデルに均一化する?
- 基盤モデルの構築や応用に議論がフォーカス
 - 計算機の調達やモデル学習など
- **AIがみたこともないデータを抱え込むことも重要**
 - データ油田を作る
 - 本邦のデジタル赤字問題
 - データ枯渇: ゲノムデータは線形増加・クローズドへ
 - AIのための新しい計測技術
 - AIが賢くなるために必要な計測技術
 - AIが賢くなるためのデータの量・質・幅
 - AIのための新しい材料
 - 創薬モダリティ、細胞、オルガノイドなど
- **データベースの位置づけ**
 - DDBJ/PDBj/DBCLS/DDBJスパコンを擁する地の利
 - 基盤モデルの学習データ
 - データとメタデータ
 - 基盤モデルの出力結果のリファレンス

エネルギー資源の流通



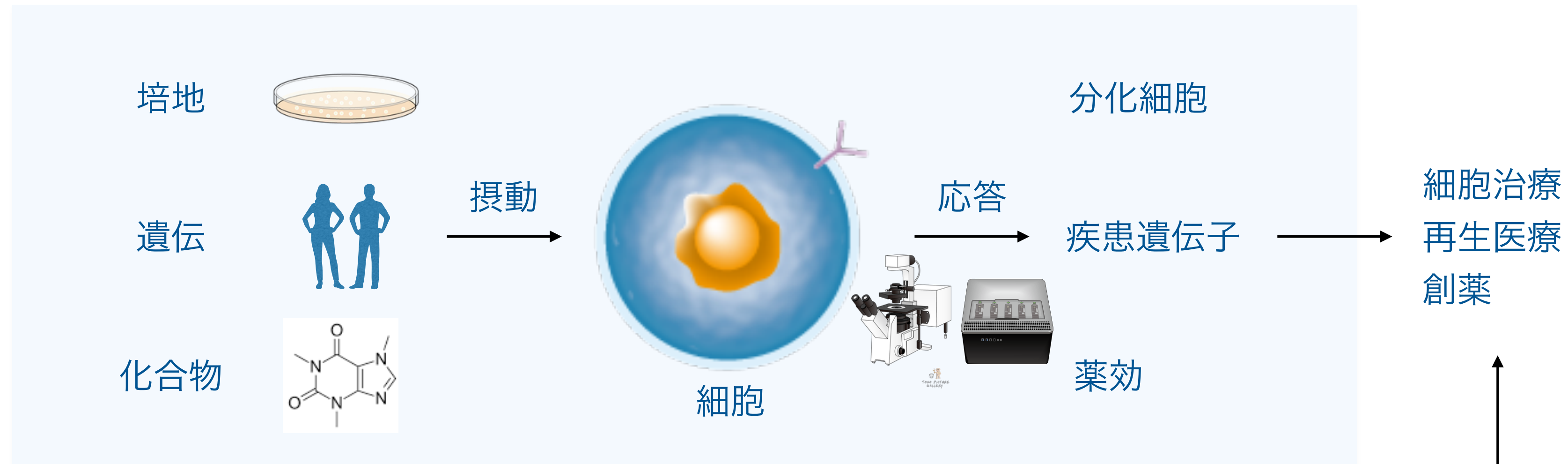
データ資源の流通



RIKEN TRIP-AGIS: AI for Omicsの開発

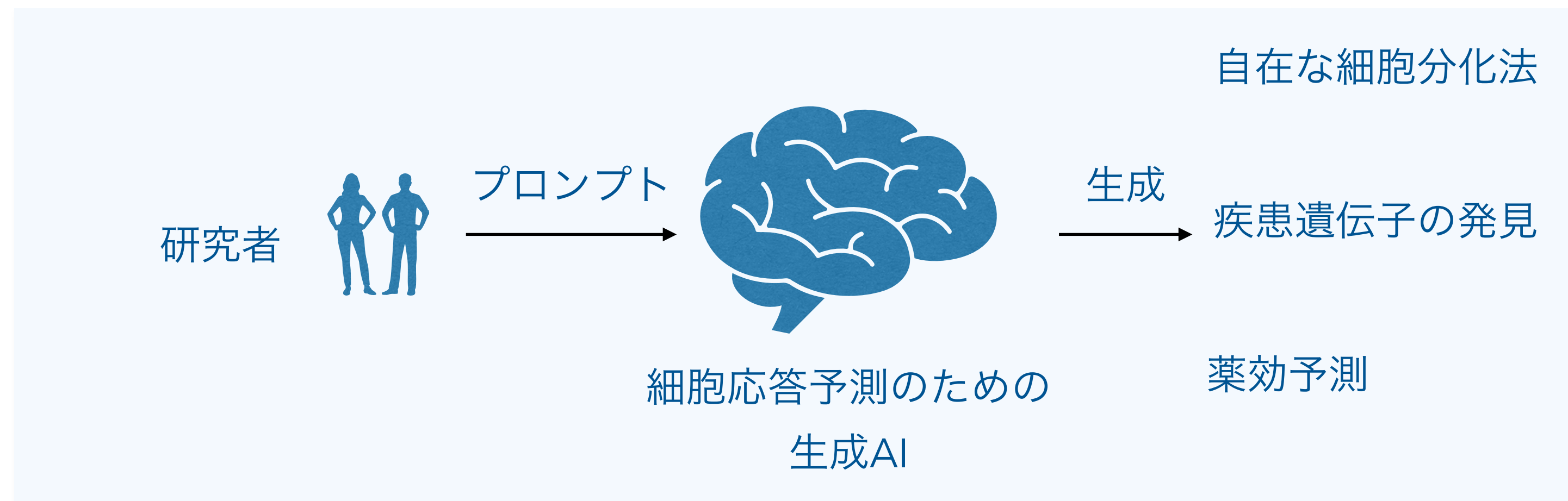
100倍を超えるスループットを実現するオミクス技術

理研独自の大規模実験技術で学習データを収集



理研独自データで高性能な生成AIを構築

© 2016 DBCLS TogoTV, CC-BY-4.0



二階堂@BDR

岡田@BDR

岩崎@CPR

岡田@CPR



高橋恒一@RIKEN BDR