Press Release

横浜市立大学 YOKOHAMA CITY UNIVERSITY　理化学研究所 RIKEN

文部科学記者会・科学記者会、
神戸市政記者クラブ、神戸民間放送記者クラブ、
大阪科学・大学記者クラブ、京都大学記者クラブ　同時配信

2025 年 3 月 14 日
横 浜 市 立 大 学
理 化 学 研 究 所

データ駆動型生成 AI の限界に迫る
－生成 AI で信頼性の高い分子設計を実現する戦略－

横浜市立大学大学院 生命医科学研究科 生命情報科学研究室の吉澤 竜哉さん（博士後期課程 1 年）、石田 祥一特任講師、寺山 慧准教授、理化学研究所 生命機能科学研究センター制御分子設計研究チームの佐藤 朋広研究員、大田 雅照上級研究員、本間 光貴チームリーダーの研究グループは、生成 AI[*1] を用いたデータ駆動型分子設計において、AI の予測信頼性を維持しながら複数の特性を同時に最適化するためのフレームワークを開発しました。従来の分子設計では、AI によって有望であると予測された分子が、実際には望ましくないというケース（報酬ハッキング[*2]）が頻発しており、これが AI による分子設計の実用化を妨げる大きな要因の一つとなっていました。特に創薬では同時に改善したい特性が多いため、有用性は限定的でした。本研究グループは、この課題に対処するための手法 DyRAMO を開発し、無償で公開しました。DyRAMO は、医薬品や機能性材料の設計などのさまざまな条件に適用可能であることから、多岐にわたる分野での応用が期待されます。

本研究成果は、国際科学雑誌「Nature Communications」に掲載されました（2025 年 3 月 11 日）。

研究成果のポイント

- 分子設計の過程で AI の信頼性を自動的に最適化し、AI の生成能力を最大限に引き出す手法を開発。
- 複数特性の同時最適化において、AI が不適切な分子を高評価・生成してしまう現象（報酬ハッキング）を回避。
- 抗がん剤設計を題材に開発手法を検証し、AI が学習していない既存の承認薬を設計することに成功。

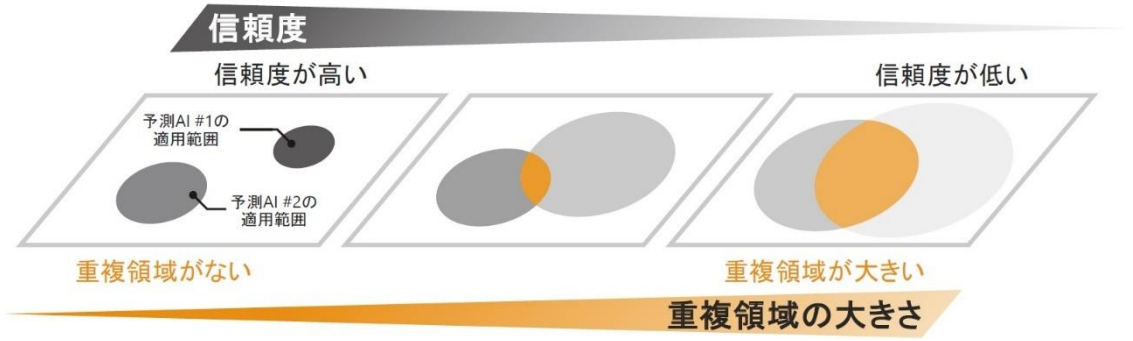


図 1 複数の予測 AI の信頼度の高さと、それぞれの適用範囲[*3]が重複する領域の大きさの関係性
全ての予測 AI に対して高い信頼度で適用範囲を定義した場合、それぞれの適用範囲が重複する領域が狭くなる、あるいはなくなる。一方で、適用範囲の重複領域を広げすぎると、信頼度が低下する。

研究背景

近年、生成 AI を活用して望ましい特性を持つ分子を設計する技術が急速に発展し、医薬品開発や機能性材料の創出に大きな変革をもたらしつつあります。しかし、生成 AI で適当に生成しただけでは目的の分子を得ることは困難で、望ましい特性を持っているのかを適切に評価し、生成 AI にフィードバックする仕組みが重要です。そこで、生成 AI と教師あり学習*4による予測モデル（以降、予測 AI）を組み合わせたデータ駆動型の分子設計は、過去の実験データ等を活用することで、高速かつ効率的に分子を設計できる技術として注目を集めています。

しかし、この生成 AI と予測 AI の連携も万能ではありません。まず、予測 AI は内挿、つまり教師データと類似したデータに対しては高い性能で予測を行えるものの、外挿（教師データから大きく異なるデータに対する予測）は苦手で、一般に予測の信頼性が低下します。特に、生成 AI による分子設計において、教師データに含まれない新規の分子が設計された場合、予測性能は保証されず、信頼性は低下します。その結果、誤った予測結果に基づく分子の最適化が起こる問題が指摘されています（報酬ハッキング）。

この問題に対処する方法の一つが、予測 AI の適用範囲を考慮することです。ただし、複数の特性の同時最適化（多目的最適化*5）を試みる場合は、複数の予測 AI の適用範囲を同時に考慮する必要があります。例えば、全ての適用範囲を高い信頼度で厳格に設定すると、探索可能な分子がほとんどなくなってしまいます（図 1 左）。一方で、低い信頼度を設定して適用範囲を緩めすぎると、信頼性の低い予測結果が増え、設計された分子が実際には望ましくないものになる可能性が高まります（図 1 右）。そのため、全ての予測 AI の信頼度を適切に設定する必要があります。しかし、適用範囲が重複した領域に分子が存在し得るかは、実際に分子設計を実施して初めて明らかとなるため、事前に適切な信頼度の値を設定することは困難です。このことから、信頼性を考慮した複数特性の同時最適化は実現されていませんでした。

研究内容

本研究では、予測 AI の適用範囲を定義する信頼度を自動的に調整するフレームワークである DyRAMO（Dynamic Reliability Adjustment for Multi-objective Optimization）を開発しました。DyRAMO による信頼度の調整は、①信頼度の設定、②分子設計、③設定された信頼度の高さと分子設計の結果の評価、を通して行われます（図 2）。

- Step 1：使用する予測 AI それぞれにおいて、信頼度を設定し、それに基づき適用範囲を定義します。ここで、本検証での信頼度は、予測する分子と教師データに含まれる分子との構造的な類似度を基準としました。またユーザーは、調整する信頼度の範囲と、優先的に信頼度を高くする特性を選択することが可能です。
- Step 2：分子設計を行います。ここでは、Step 1 で定義された適用範囲内で、全ての特性が最適化された分子を設計することを目標とします。
- Step 3：Step 1 で設定された信頼度の高さと、Step 2 の分子設計の結果を評価します。この評価をもとに、再び Step 1 の信頼度の設定を行います。

また、適切な信頼度の組み合わせを効率的に探索するために、ベイズ最適化*6を導入しました。

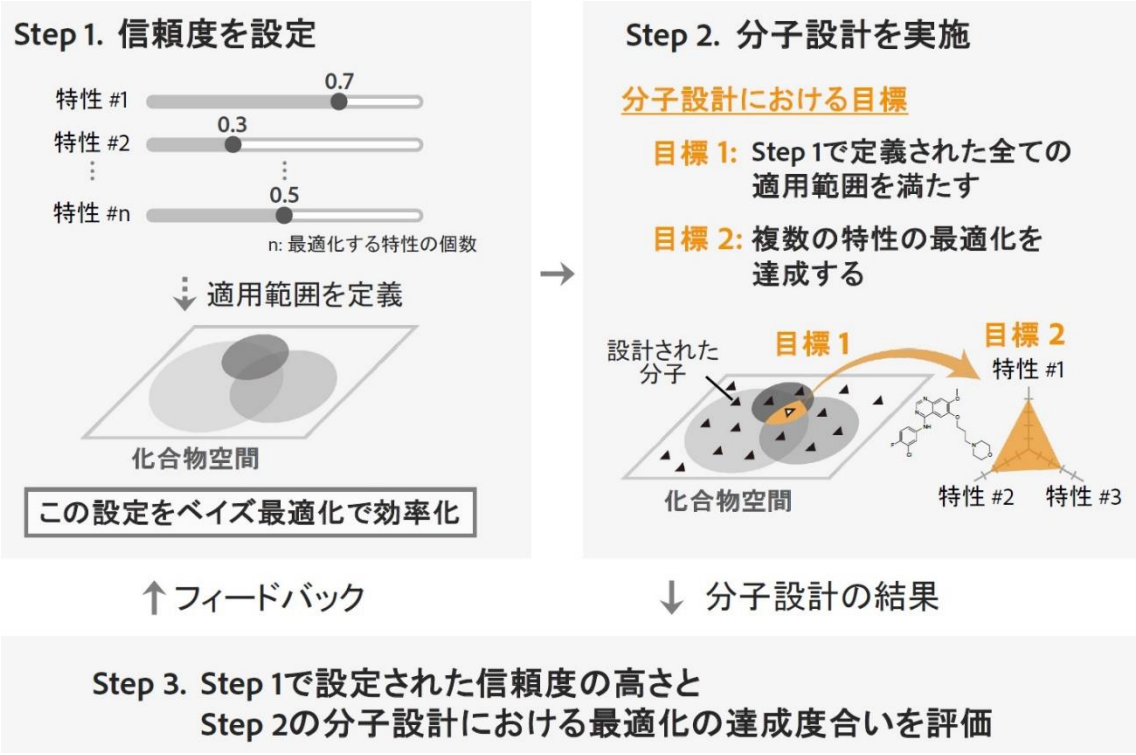
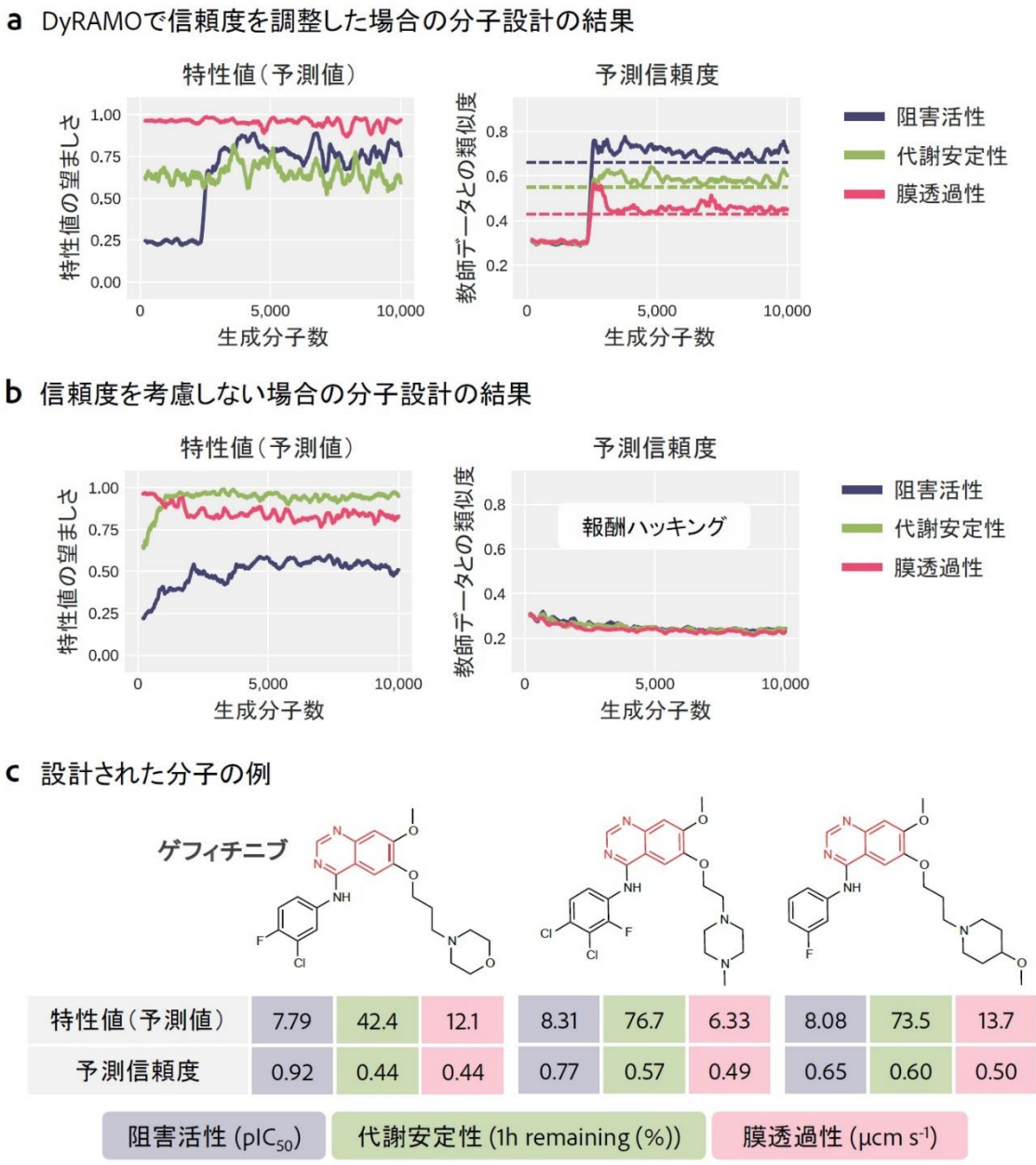


図２　DyRAMO のワークフロー
多目的最適化の状況において、複数の予測 AI の信頼度の高さと、予測値の良さ（最適化の達成度合い）を同時に満たすように、各特性の適切な信頼度を探索する

本手法の有効性を検証するために、非小細胞肺がん等の治療における標的タンパク質の一つである上皮成長因子受容体（EGFR）に結合する医薬品候補化合物の設計を実施しました。本設計では、EGFR に対する阻害活性、代謝安定性、膜透過性の３種類の特性の同時最適化を試みました。また、分子生成 AI として本研究グループが開発している ChemTSv2*7 を用いました（図２中の Step 2）。結果として、DyRAMO は、信頼度が一定程度高い予測に基づいた３種類の特性が、全て同時に最適化された分子の設計に成功しました（図３-a）。信頼度を考慮しない生成の場合（図３-b）と比較すると、DyRAMO では信頼度が比較的高い値に収まっていることが分かります。さらに、設計された分子の中には、承認薬の一種であるゲフィチニブが含まれていました（図３-c）。これらの結果は、DyRAMO が予測の信頼度を適切に調整しつつ特性の最適化を行い、有望な医薬品候補を設計する能力を有することを示唆しています。

他にも、現実的な最適化では、複数の特性の中で優先度が存在する場合があります。つまり、特に重要で高い信頼度が求められる特性がある一方、多少信頼度を妥協しても良い特性もある場合があります。このような状況を想定し、DyRAMO では項目ごとに優先度をつける機能も実装しています。詳細は論文をご参照ください。



特性値（予測値）	7.79	42.4	12.1	8.31	76.7	6.33	8.08	73.5	13.7
予測信頼度	0.92	0.44	0.44	0.77	0.57	0.49	0.65	0.60	0.50
	阻害活性 (pIC_{50})			代謝安定性 (1h remaining (%))			膜透過性 ($\mu cm\ s^{-1}$)		

図 3　分子設計の結果

a：DyRAMO で信頼度を調整した場合の分子設計の結果。図の左側は予測された特性値の望ましさの推移を示している。図の右側は予測の信頼度の推移を示しており、点線は適用範囲の定義に用いられた信頼度である。DyRAMO により、一定程度高い信頼度の予測に基づいて 3 種類の特性が同時に最適化された

b：信頼度を考慮しない場合の分子設計の結果。信頼度を考慮しない場合は、特性値（予測値）の最適化には成功するものの、予測の信頼度が低く、報酬ハッキングが発生している可能性が高い

c：DyRAMO により設計された分子の例。表に記載されている値は、3 種類の特性の予測値および予測信頼度を示している。設計された分子には、承認薬の一つであるゲフィチニブ（左）が含まれており、その他の分子も既存の EGFR 阻害薬に見られる特徴的な部分構造（赤色）を有していた

今後の展開

従来の AI による分子設計では、AI が有望であると判断した分子が実際には適切でない、という現象が発生する可能性が高く、これが分子設計のプロセスの効率を低下させる要因の一つとなっていました。DyRAMO を導入することで、設計段階で信頼性の高い評価に基づき分子を選定できるため、開発の手戻りが減少し、結果として医薬品開発や材料開発のプロセスが加速されることが期待されます。

本研究は、単なる分子設計技術の進歩にとどまりません。「既存のデータを用いて生成 AI で何ができるのか？」という、データ駆動型生成 AI の根本的な限界に迫るものです。生成 AI の普及が進む中、その信頼性や適用範囲を適切に管理することは、今後の実用化において極めて重要な課題です。DyRAMO の開発は、生成 AI の活用方法そのものを見直し、より信頼性の高い生成 AI の実現に向けた重要なステップと位置付けられます。本研究で導入した DyRAMO の戦略は、分子の設計にとどまらず、生成 AI とデータ駆動型の予測 AI によるさまざまな設計に活用できるものです。

さらに、本手法は分子シミュレーションと連携[*8]することで、データが不足している領域の分子も評価が可能となります。データ駆動型の分子設計は、結局のところ教師データと類似した分子しか正確に評価できず、新規性の高い分子の探索は困難です。しかし、DyRAMO と分子シミュレーションを適切に組み合わせることで、予測 AI が学習していない分子に対しても信頼性の高い評価が可能となり、より柔軟かつ信頼性の高い分子設計手法になると期待されます。

研究費

本研究は、AMED の産学連携による次世代創薬 AI の開発(DAIIA)、生命科学・創薬研究支援基盤事業 (BINDS)、文部科学省の「富岳」を活用したシミュレーション・AI 駆動型次世代医療・創薬、データ創出・活用型マテリアル研究開発プロジェクト、JST の創発的研究支援事業の支援を受けて実施されました。また理化学研究所ジュニアリサーチアソシエイトプログラムおよび中外創薬科学イノベーション財団（C-FINDs）の支援も受けています。

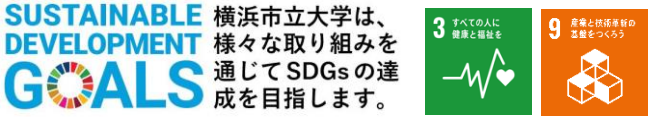
論文情報

タイトル： A Data-Driven Generative Strategy to Avoid Reward Hacking in Multi-objective Molecular Design

著者： Tatsuya Yoshizawa, Shoichi Ishida, Tomohiro Sato, Masateru Ohta, Teruki Honma, Kei Terayama

掲載雑誌：Nature Communications

DOI：10.1038/s41467-025-57582-3



用語説明

*1 生成 AI：Generative Artificial Intelligence。機械学習を用いて、新しいデータを生成する AI の総称。画像、文章、音楽などのコンテンツを生成する用途のほか、科学技術分野では新しい材料や化合物の設計にも活用されている。

*2 報酬ハッキング： AI が、設計された目的関数（報酬）を最大化しようとする際に、意図しない方法で目標を達成してしまう現象を指す。例えば、分子設計 AI が本来の目的である「有望な分子の生成」ではなく、予測モデルの意図しないバイアスを利用して高いスコアを出す分子を作るケースが挙げられる。

*3 適用範囲：機械学習モデルが所定の信頼度で予測を行えるデータの範囲を指す概念。機械学習モデルは、学習データと類似したデータに対しては高い信頼度で予測できるが、大きく異なるデータに対しては信頼度が低下するため、予測の適用可能性を判断する指標として利用される。

*4 教師あり学習：入力データと、それに対応する正解データをもとに学習する機械学習の手法。学習に用いたデータを教師データと呼ぶ。予測モデルの構築に広く用いられ、与えられたデータのパターンを学習し、新しいデータに対する予測を行う。

*5 多目的最適化：複数の競合する目的関数を同時に最大化（あるいは最小化）する最適解を求める問題。以下は創薬における応用例。
https://www.tsurumi.yokohama-cu.ac.jp/news/20221124yoshizawa.html

*6 ベイズ最適化：効率的に最適な解を探索するための手法の一つ。特に、評価コストが高い問題に適用されることが多く、限られた試行回数の中で最適なパラメータを見つけるために活用される。

*7 ChemTSv2：生命情報科学研究室で開発している分子生成 AI で、薬から材料までさまざまな機能性分子を設計可能。
https://www.yokohama-cu.ac.jp/news/2023/20230818terayama.html

*8 分子シミュレーションと連携： すでに量子化学シミュレーションと連携することで新規な蛍光有機分子の設計と合成に成功している。
人工知能で蛍光有機分子を開発 －複雑な現象を示す機能性分子の開発に貢献－
（2022 年 3 月 10 日理化学研究所、横浜市立大学）
https://www.riken.jp/press/2022/20220310_1/index.html