

日本医療研究開発機構 創薬基盤推進事業 事後評価報告書

I 基本情報

研究開発課題名：10⁶⁰の潜在的化合物空間の探索を実現する低分子創薬技術の開発

Development of Small-Molecule Drug Discovery Technologies Enabling the Exploration of a Potential Chemical Space of ten to the power of sixty

研究開発実施期間：令和5年4月15日～令和8年3月31日

研究開発代表者 氏名：水野 忠快

Tadahaya Mizuno

研究開発代表者 所属機関・部署・役職：

国立大学法人東京大学 大学院薬学系研究科 助教

The University of Tokyo, Department of Pharmaceutical Sciences, Assistant Professor

II 研究開発の概要

研究開発の成果およびその意義等

低分子創薬においては、理論的に存在し得る化合物空間が極めて広大である一方、実際に探索対象となる化合物は既存化合物ライブラリーや既知化合物の近傍構造に依存しやすい。既存ライブラリーに含まれない、あるいは既存構造から離れた化学構造を探索対象にできれば、低分子創薬の可能性を広げられると期待される。しかし、そのような構造をどのように生成するか、生成された構造群に対するバーチャルスクリーニングが実際に成立するかについては、なお多くの課題が残されている。本研究開発は、化合物を個々の候補として扱うだけでなく、広大な潜在的化合物空間の中の位置や広がりとして捉え、その空間を創薬探索に利用できるかという問いから出発した。10の60乗規模とも見積もられる潜在的化合物空間を背景に、既存化合物空間の外側に位置し得る候補構造を生成する技術を構築し、それらを *in silico* 評価、合成可能性評価および実験評価へ接続することで、探索フローとしての実装可能性と課題を検証した。

この問題設定のもと、既存化合物ライブラリーの近傍に限定されない候補構造を得るため、翻訳型の化学言語モデルを構築した。本モデルでは、特定の既知活性化合物や既存ライブラリーを直接の出発点とせず、潜在的化合物空間に由来する候補構造を生成することを目指した。生成された化合物群については、化学的妥当性、構造多様性、既存化合物空間との関係、および薬物らしさの観点から評価した。結果、従来の既存化合物集合に基づく探索とは異なる化学構造領域を含む化合物集合を得られることを確認した。特に学習データ中に確認されないスキップフォルドを含む化合物を生成する能力を評価した指標では、本モデルは比較対象とした既存モデルを大きく上回る値を示した。これは、既存化合物空間の外側を探索対象に含めるという本研究の入口に対応する成果であり、潜在的化合物空間に由来する候補を計算機上で生成・評価するための基盤を整備したものと位置付けられる。

生成された候補構造群に対しては、バーチャルスクリーニングを適用した場合に創薬探索フローとして成立するかを検証した。具体的には、毒性や薬物らしさなどに基づく一次評価、部分構造や標的との相互作用を考慮した二次評価により、候補化合物を段階的に絞り込むスクリーニングフローを実装した。選定された候補につい

ては、合成、培養細胞を用いた in vitro 評価、分子ドッキング、および分子動力学計算を組み合わせた後段検証を行った。これにより、既存ライブラリー外に位置し得る候補構造の生成から、計算機上での絞り込み、実験評価に至るまでの一連の探索フローを実際に接続した。本研究の特徴は、潜在的化合物空間を活用するという基礎的な問いを、生成モデルの性能評価だけで終わらず、創薬探索の実践に必要な各工程へ段階的に落とし込んだ点にある。

実装過程では、既存化合物空間の外側を対象とする探索に固有の課題も具体化した。バーチャルスクリーニング系の性能は単一の指標だけでは十分に評価できず、既知の陽性対照に相当する化合物を除外しない能力と、新規候補を効率よく拾い上げる能力を分けて評価する必要がある。構築した評価系は、既知の有望化合物を保持する点では一定程度機能した一方、新規候補を高い感度で拾い上げるにはさらなる補強が必要であった。加えて、新規化合物を対象とする場合には、合成および in vitro 評価に要する時間とコストが評価・改善ループの速度を大きく制約することも明確になった。これらの結果から、既存ライブラリー外の化学構造を対象とする創薬探索では、生成・選抜の性能だけでなく、後段の合成・実験評価まで含めた運用設計が重要であることを確認した。

分子動力学計算は、ドッキング結果の妥当性を時間発展の観点から点検し、候補の一部を除外する補助的評価として機能し得ることが示された。一方、分子動力学計算のみで候補化合物の有効性を肯定的に判定するには、評価指標や判定基準のさらなる整備が必要である。したがって、in silico 評価は単独で完結する予測手法として扱うのではなく、合成・実験評価、分子シミュレーション、および専門家判断と組み合わせて用いることが重要である。この整理は、既存空間外の候補を扱う創薬探索フローを実践的に設計する上での指針となる。

探索フローの構築と並行して、化学言語モデルを創薬スクリーニングに利用する際の特性と適用条件も検討した。SMILES などの文字列表現を用いた化学言語モデルは、化合物生成、物性・毒性予測、そして分子表現学習に広く利用されている。一方、化合物構造を文字列として扱う場合には、同一化合物に対する表記揺れや立体化学情報の扱いがモデル性能に影響する可能性がある。本研究では、化学言語モデルが分子構造を学習する過程を解析し、キラリティ認識や SMILES 表記揺れが構造認識および予測性能に与える影響を体系的に評価した。その結果、化学言語モデルは化合物構造を有用な表現として学習し得る一方、入力表記や前処理の違いによって性能評価が大きく変動し得ることを明らかにした。これにより、化学言語モデルを単なる高性能予測モデルとしてではなく、適用条件と限界を理解した上で創薬研究に利用するための基盤を整備した。

未知構造を対象とする探索では、すべての候補を合成・実験評価することは現実的ではない。そのため実験的検証が困難な段階でも、探索結果の性質を間接的に点検できる評価軸が必要となる。本研究では、特許情報の時間軸を利用し、化合物構造が時間とともにどのように拡張されてきたかを解析する方法を検討した。この枠組みにより、生成化合物が既存化合物空間に強く依存しているのか、あるいは既存構造から相対的に離れた領域を含むのかを合成結果に依存せず補助的に点検するための評価軸を得た。本成果は、生成モデルによる化合物探索を評価するための新たな観点を与えるものであり、今後の化合物空間解析や AI 創薬モデル評価にも応用可能である。

本研究開発の意義は、既存化合物ライブラリー外に位置し得る化学構造を生成する技術を構築した点に加え、そのような構造群を対象とするバーチャルスクリーニングを実際の創薬探索フローに接続し、どの段階で何が有効に機能し、どの段階で制約が生じるかを具体的に整理した点にある。AI 創薬では、モデルの生成能力や予測精度が注目される一方、実際の創薬フローに接続する際には、評価感度、偽陽性の扱い、合成可能性、実験検証コスト、および専門家判断との役割分担が重要となる。本研究では、これらの要素を一連の探索過程の中で検証し、既存化合物空間の外側を対象とする探索を実装する上での要件を明確化した。

化学言語モデルの構造認識に関する知見は、低分子創薬に限らず、毒性予測を含む安全性評価、ADME 予測、物性予測、および化合物表現学習など、広く薬学・生命科学分野における AI 利用に資するものである。特に SMILES 表記揺れや立体化学情報の扱いがモデル性能に影響し得ることを示した点は、化合物データを用いた AI 研究の再現性と信頼性を高める上で重要である。これらの成果は、国際誌での論文発表や国内外の学会発表を通

じて発信されており、化学言語モデルを創薬研究に適切に利用するための基盤的知見として位置付けられる。

今後は、本研究で構築した潜在的化合物空間探索基盤と、外部の大規模合成可能化合物空間、実験評価基盤、および専門家によるメディシナルケミストリーに基づく判断を組み合わせることで、より実装可能性の高い低分子創薬探索フローへと発展させる。探索初期段階では大規模な候補空間を効率よく評価し、後段では合成・実験評価、および専門家判断により候補を精査するという役割分担が重要である。本研究で得られた知見は、10の60乗規模の潜在的化合物空間を対象とする探索を、単なる概念的提案ではなく、実際の創薬プロセスに接続可能な生成・評価・運用基盤として発展させるための土台となる。以上より本研究開発では、低分子AI創薬の上流工程における探索対象の多様化、評価手法の高度化および創薬基盤技術の適正利用に貢献する成果を得た。

Research and Development

Achievements and Their Significance

In small-molecule drug discovery, the theoretically possible chemical space is extremely large, whereas practical screening is often limited to existing compound libraries or structures close to known compounds. Expanding search targets beyond existing libraries may broaden the potential of small-molecule drug discovery, but it remains challenging to generate such structures and to determine whether virtual screening can be applied effectively to them. This project started from the basic question of whether a vast potential chemical space, estimated to contain as many as 10^{60} possible compounds, can be represented and utilized for drug discovery. We aimed to generate candidate structures that may lie outside existing chemical space and to connect them to in silico evaluation, synthetic feasibility assessment, and experimental validation.

To address this question, we developed a translation-based chemical language model for generating candidate structures that are not restricted to the immediate neighborhood of existing compound libraries. The generated compounds were evaluated in terms of chemical validity, structural diversity, relationship to existing chemical space, and drug-likeness. The model generated compound sets containing structural regions different from those obtained by conventional searches based mainly on existing compounds. In particular, in an index evaluating the ability to generate compounds with scaffolds not observed in the training data, the model showed markedly higher values than the existing models. This result supports the central aim of this project: to include chemical structures outside existing compound space as possible search targets.

We then examined whether virtual screening of the generated candidates could function as a practical drug discovery workflow. The screening process consisted of an initial evaluation based on toxicity- and drug-likeness-related criteria, followed by a secondary evaluation that considered substructures and interactions with a representative drug discovery target. Selected candidates were further examined through synthesis, in vitro assays, molecular docking, and molecular dynamics simulations. Through this process, we connected the generation of candidate structures outside existing libraries to computational filtering and experimental evaluation, thereby translating a basic question about potential chemical space into practical steps required for drug discovery.

The implementation clarified several challenges in exploring chemical structures outside existing compound space. The performance of a virtual screening system cannot be fully evaluated using a single metric; the ability to retain known positive controls and the ability to identify new promising candidates with sufficient sensitivity must be distinguished. The screening system constructed in this project showed a certain ability to retain known promising compounds, but further improvement was needed to identify new candidates with high sensitivity. The time and cost required for synthesis and in vitro testing also strongly limited the speed of the evaluation and improvement cycle. Molecular dynamics simulations may help examine docking results and exclude some candidates, but further development of evaluation metrics and decision criteria is required before they can be used for positive selection.

The significance of this project lies in both the development of a model for generating chemical structures outside existing libraries and the practical examination of how such structures can be connected to virtual screening and experimental evaluation. In parallel, we clarified important conditions for using chemical language models, including the effects of chirality recognition and SMILES notation inconsistency on model performance. We also explored a supplementary evaluation axis using the time course of patent information to assess the relationship between generated compounds and existing chemical space. Together, these findings provide a foundation for a more practical workflow that combines potential chemical space exploration, large-scale synthesizable compound spaces, experimental platforms, molecular simulation, and expert medicinal chemistry judgment.