2.1.3 AI創薬

（1）研究開発領域の定義

創薬研究の各段階（例えば、標的探索や医薬品候補分子の最適化など）の効率化を目的として、広義のAI（機械学習のみでなく、従来人間が行っていた高度な判断をコンピュータによって代替する広範な技術や研究分野）を適用する技術の確立、またそれら要素技術の統合を通して、創薬研究のあり方そのものを変革する試みまでを指す領域である。

（2）キーワード

人工知能、機械学習、深層学習、バイオインフォマティクス、ケモインフォマティクス、シミュレーション、数理モデリング、分子動力学計算（Molecular Dynamics: MD）、データベース

（3）研究開発領域の概要

［本領域の意義］

近年、製薬産業を取り巻く状況が大きく変化している。低分子を中心とした創薬ターゲットが枯渇傾向にあると認識されており、既存の研究手法の延長線上での新薬創出の成功確率が低下していることは大きな課題である。開発コストも年々上昇し、製薬産業の高コスト体質を悪化させている。

創薬研究には様々なボトルネックが存在し、上記の課題にも単純な解決策を見出すことは難しい。究極的には、これまでの創薬の概念を大きく覆す変革が必要と考えられる。そこでは、病態を制御する化学物質（伝統的には低分子化合物）を設計して投与するという形にとって変わって、病気の予防から予後のQuality of Lifeの向上までの「ペーシャントジャーニー」を広くサポートするという考え方が中心になってくると考えられる。その中心にいる「患者」を理解するためには、これまで以上に幅広いデータを統合的に活用する必要があり、AIによる解析・可視化は必須と言える。例えば、電子化された診療データや個人ゲノムの情報だけでなく、ウェアラブル機器などから得られる各種ヘルスデータの蓄積も飛躍的に進むものと想定され、コンピュータを用いた大規模データ解析の意義はますます高まる。

その上で、伝統的な創薬研究の各段階においても様々な技術革新が進行中であり、創薬DXとでも言うべき変革への流れが築かれつつある。合理的な医薬品設計のためには、何らかの形でターゲット（標的）分子を選定して特徴付ける必要がある。ターゲットの選定ミスが、後の臨床試験で薬効が出ずに開発中止に繋がるという深刻な問題は10年以上前から広く認識されているが、未だに決め手になる解決策は見出されておらず、AIへの期待は大きい。ターゲットに物理的に作用して薬効を及ぼす医薬品候補分子の探索と最適化（ここでは広く分子設計と呼ぶ）については、データの蓄積や予測結果の評価が比較的容易であり、AIを用いた手法は既に成果を上げている（例えば、AIを用いて短期間で臨床候補化合物の創出に成功など）。もちろん、薬物動態や毒性、活性の全てを考慮して分子を改変していく試みは、AIが熟練の創薬化学者を代替できるところまでには程遠いが、現在のAI創薬の重要な挑戦課題に位置付けられる段階には来ている。また、低分子以外の新規モダリティ（抗体、核酸、細胞など）については、機械学習を主体とする狭義のAIだけでは不十分で、分子シミュレーションや数理モデリングなどを含めた複合的なアプローチが有効と考えられる。さらに、上記の探索研究段階のみでなく、有効な治験のデザインなど開発研究においても様々なAIの活用が進んでいる。

［研究開発の動向］

計算科学技術を医薬品設計に応用する試みの歴史は古く、1979年にワシントン大学のスピンアウトとして設立されたTripos社のレガシーは、現在もCertara社に受け継がれている。 Dassault Systèmes社、Schrödinger社、CCG社などの販売する医薬品設計支援統合ツールは現在も幅広く使われているが、最近

はアカデミアからAI・機械学習を中心とする多数の新たな手法が提案されると共に、AI創薬を標榜するスタートアップも国内外に多数現れるようになってきた。

上記の通り、分子設計の分野におけるAIの活用はかなり一般的になり、2020年には英国Exscientia社と大日本住友製薬社との協業により、従来よりも飛躍的に短期間で臨床候補化合物の創出に成功した事例などが報告されている。国内のスタートップでもやはり2020年にElix社とアステラス製薬社との共同研究開始などが公表されている。

低分子化合物については、*in vitro*あるいは細胞レベルでの測定が可能な活性や薬物動態パラメータを大量に取得して学習データとし、化学構造のみから活性を予測する機械学習モデルを構築する手法が確立されてきている。もちろん、化学構造の記述法（伝統的な記述子か、或いはグラフ表現を用いるのかなど）や予測モデルの適用範囲をどのように評価するかなど、技術的な課題は存在する一方で、学習データの重要性についての認識は一層高まっている。データの質と量は共に重要であるが、例えば公共データベースなどから取得可能なデータは、単位や実験条件などが整っていないことが多い。手作業によるデータの取捨選択や編集作業（マニュアルキュレーション）が、予測モデルの精度向上に重要となる[1)]。製薬企業内部では、実験条件が統一されquality controlのしっかりしたデータが取得されているが、社内データのみでは量が不十分な場合が多い。しかし、他企業のデータを利用することはこれまで不可能であった。

AMEDの創薬支援インフォマティクスシステム構築プロジェクト（2015–2020年）において、国内の7つの製薬企業とアカデミア機関との企業連携を確立したことは、国際的にも社内データ共有事例の先駆けといえる。化学構造と薬物動態パラメータとの組み合わせに加えて、化学構造に戻ることのできない形で構造から計算された記述子のみを中立的なアカデミア機関に提供することで、秘匿性を保ったデータ共有を実現した[2)]。また複数企業のデータを共有することで、より大きな化合物空間を扱えるようになり、個社のデータのみを使うよりも、より有用なモデルの構築が可能であることが示された[3)]。

さらに、構造や記述子など、データそのものは共有せず、モデルのみを共有する連合学習（Federated Learning）の試みが広がってきた。欧州のMachine Learning Ledger Orchestration for Drug Discovery（MELLODDY）コンソーシアムはその一つであり、多数の欧米のビッグファーマとテクノロジー企業が参画している。連合学習では、携帯端末上にデータを分散させたままクラウド上で機械学習モデルを共有する仕組みが既に2017年にGoogle社により提案されており、様々な実装が試みられているが、創薬分野における具体的な成果については、上記MELLODDYからの情報公開が待たれる。国内では、AMEDの産学連携による次世代創薬AI開発プロジェクト（DAIIA; 2020年–）において、日本製薬工業協会の主導により、上述の「創薬支援インフォマティクスシステム構築」よりも拡大された製薬企業18社の参画のもとで、新規化合物創出が進められている。

一方、分子設計以外にターゲット探索についてもAIの活用が進展している。上述の創薬ターゲットの選定ミス（あるいは、確度の高いターゲットの枯渇）問題に対して、ブレークスルーとなることが期待されている。実際、複数のAIスタートアップにより、炎症性腸疾患（IBD）や筋萎縮性側索硬化症（ALS）[4)]、慢性腎臓病（CKD）や特発性肺線維症（IPF）などの疾患に対して新たなターゲットが提案されて、新薬開発に向けたプロジェクトが進行している。これらの試みは、従来の細胞や動物モデルの実験に基づいて仮説を組み立てていくのではなく、ヒト（患者）由来のビッグデータをデータ駆動的に解析することで、新たなターゲットを見出す点が大きな特色になっている。イメージングやシーケンシング技術の進歩は著しく、一人の患者から、ゲノムだけでなくトランクリプトームなどを含む多層的なデータを取得することが容易になってきており、これら分子レベルのデータと診療情報などの個人レベルのデータとを統合的に解析するために、AIの利用は必須と考えられる。国内においては、アカデミア主体で同様のコンセプトによる研究開発が進行中であり、官民研究開発投資拡大プログラム（PRISM）「新薬創出を加速する人工知能の開発」により、IPFと肺がんにおける新規創薬ターゲット探索が行われている。また、フロンテオ社（自然言語処理技術）と東工大（細胞分析技術）によるターゲット探索に向けた共同研究が発表されている。

広義のAI創薬に関わる技術として、何らかの基本原理に基づくモデリング手法の活用も進んでいる。モデリングに関連する事例は歴史が古く、物理学の原理に基づく分子動力学（molecular dynamics: MD）や、1970年代に行われたタンパク質のダイナミクス解析にシミュレーション手法を応用する試みにまで遡ることができる。その後、現在の富嶽に至るスーパーコンピュターやMD専用計算機による計算機能力の向上と力場パラメータの更新により、適用可能な分子サイズや時間スケールが着実に進化してきた。最近は、機械学習と計算化学との融合により、例えば、分子力場という簡易的な方法で、量子化学計算に基づく精密な相互作用エネルギーを計算できるパラメータをAIで決定する試みや、計算コストの高いMDによって得た結果を学習データとしてAIモデルを構築してより高速に解析結果を得るなど、幅広い試みがなされている[5)]。さらに、物理学の分野ではAIを用いて基礎方程式に基づくシミュレーションの解を得る試みが始まっているが、分子系などの複雑なシステムへの応用可能性については今後を待たねばいけないだろう。

細胞レベルのシグナル伝達、あるいはより高次の生命現象のモデル化のためには、より抽象化した構成要素（タンパク質など）の間のネットワークに基づく数理モデリングが広く用いられている[6)]。数理モデリングについては、非専門家が簡単に使えるツールにまで成熟しているとは言い難いが、多数の数理モデルを格納したデータベースBioModelsが既に存在し、Python言語によるBioMASSライブラリーなど、数理モデルの利用拡大を目指したリソースが現れてきている。

このように、AI・機械学習を中心とするデータ駆動的なモデリングと分子から個体レベルに至る広義の数理モデリングとを組み合わせて創薬に応用する試みはまだ緒についたばかりであり、今後の一層の展開が期待される。

（4）注目動向

［新展開・技術トピックス］

• 深層学習によるタンパク質の立体構造予測

ここ1、2年の科学技術一般で最も注目の大きかった話題の一つと言える。メタゲノムデータを含む多重配列アラインメントの利用など、長年のバイオインフォマティクス研究の成果に立脚する一方で、従来はなかったタンパク質のアミノ酸配列を入力として原子座標を直接出力するend–to–endの学習モデルの与えた衝撃は大きく、一般のメディアを含めて既に多数紹介されている。深層学習を利用した予測モデルとしては、BakerらのRoseTTAFold[7)]の他、より最近はMultiple Sequence Alignment（MSA）を用いない方法が提案されているが、DeepMind社のAlphaFold[8)]が、この技術のほぼ同義語として広く使われている。

AlphaFoldの改良版のAlphaFold2は、マルチプルアライメントやニューラルネットワークを組み合わせて、これまでにない精度で立体構造を予測することに成功し、Critical Assessment of Structure Prediction（CASP）14で優勝した。CASPにはテンプレートモデルとテンプレートフリーモデリングの2つの部門があるが、AlphaFold2は、テンプレートフリーモデリング部門で正解構造との差異を表すglobal distance test–total score（GDT–TS）で、92.4の中央値スコアを達成した。スコア90を超えることはX線結晶構造解析やクライオ電子顕微鏡法などの実験手法と同等であることを表している。これらの高精度で予測された立体構造を利用することで、これまで立体構造がわかっていなかったタンパク質に対しても化合物の結合親和性などを評価することが可能になる。

本技術はあくまでタンパク質立体構造の予測であり、創薬応用への成果が出てくるのははまだこれからだと考えられるが、AlphaFoldは創薬研究、あるいは生物学研究一般のスタイルに既に大きな影響を及ぼしている。その姿は、DeepMind社とEMBL–EBIとの連携によるAlphaFold Protein Structure Databaseに見ることができる。このデータベースには、既に2億個以上のタンパク質の予測構造が収められており、オープンにアクセスすることができる。従来は、標的候補タンパク質を評価する際、或いは副作用の予測をする際に、関係するタンパク質の立体構造は不明でアミノ酸配列などの限られた情報のみを用いた議論が行われていた。AlphaFold Protein Structure Databaseの登場により、現実的なプロジェクトにおいて、関連するタンパク

質のほとんどについて何らかの立体構造情報を利用できるようになった。部分的、あるいは精度が限られた情報であったとしても、タンパク質の機能や相互作用部位の推定など、タンパク質立体構造が重要な示唆を与える事例は多い。 AlphaFold2を利用したタンパク質の立体構造解析やドッキングシミュレーションの研究成果も報告され始めている。

一方で、これらのモデルは従来の構造ベースの医薬品設計で用いられてきた立体構造データをそのまま置き換えられるものではない。 AlphaFoldが扱うことのできない問題は多数あるが、例えば、医薬品などの分子が結合した場合や外的環境の変化（pH、温度など）、一箇所だけアミノ酸残基が変化した或いはリン酸化などの修飾を受けた際の構造の変化は、いずれの場合も（基本的に）予測できない。これらは全て、現実の医薬品設計において重要な役割を果たす要素になる場合が多い。これらの要素を考慮して立体構造データを利用する場合には、従来のホモロジーモデリング法を用いるか、あるいは実験による構造決定が必要となる。これらの問題が扱えないのは、そもそも適切な学習データが存在しないからであり、機械学習の本質的な限界に由来している。しかし、いくつかの要素については、現状のAlphaFold2の実行パラメータの運用や、統合データベースを用いたデータセットの構築と再学習による対応が可能と考えられる。

• 医薬品の標的となる疾患原因分子（主にタンパク質）の探索技術

疾患サンプルと正常サンプルの分子レベルや分子ネットワークレベルの違いを計算によって解析し、疾患発症・進行の分子メカニズムの解明と原因分子を同定することを目的とするものである。これまで実施されてきた研究は、疾患の分子メカニズムの解明などの医学研究に付随する形で創薬ターゲット分子の探索研究が含まれることが多い。

ゲノムワイド関連解析（GWAS）やオミクス解析によって患者と健常者の分子プロファイルの比較を行い、疾患感受性遺伝子や疾患特異的発現分子などを同定する研究が行われてきたが、見出された分子が必ずしも治療標的になるとは限らず候補が非常に多いという問題がある。近年、疾患とタンパク質の治療標的の関係性を予測する機械学習手法も提案されている。タンパク質をコードする遺伝子に摂動を導入（ノックダウン・過剰発現など）した際のトランスクリプトーム情報を利用して、疾患に対する治療標的の可能性を予測する機械学習手法が提案されている[9)]。また、様々なデータベースにおけるタンパク質の間の機能的な関係を利用して、そのタンパク質が治療標的となりうる疾患を予測する手法なども提案されている[10)]。

• テンソル分解アルゴリズムを用いた予測

疾患や医薬品に関するオミクスデータの構造が複雑化し、行列形式のデータだけでなく、テンソル構造のデータが創生されている。シングルセルレベルでのオミクスデータも得られるようになり、データ構造が複雑化している。これまでの統計手法や機械学習手法は行列データが主に対象であったが、テンソル構造のデータを扱うための機械学習手法が提案されている。例えば、テンソル分解を用いた特徴抽出、次元削減や欠損値・未観測値の補間などが提案されている[11-13)]。補完した薬物応答遺伝子発現データの解析によって、治療薬探索の予測精度を向上できることが報告されている[14)]。

• 医薬品候補化合物の構造生成

目標の特性（薬効など）を持つ化合物を、大量の化合物のスクリーニングから見つけるのではなく、目標の特性を持つ化学構造を新しく予測する逆構造活性相関解析（通常の構造活性相関解析とは逆方向のアプローチ）も最近研究が進められている。化合物の構造をSimplified Molecular Input Line Entry System（SMILES）という文字列で表記し、確率的言語モデルを用いて文字列のパターンを学習する研究が、創薬やマテリアルインフォマティクスの分野で出現している。深層学習を応用したニューラルネットワークモデルによる医薬品候補化合物の設計手法が、近年特に注目されている[15)]。 SMILESに基づく構造生成の先行研究事例として、N–gramという言語モデルや変分自己符号器や再帰的ニューラルネットワークなどディープラー

ニングによる言語認識・生成を用いた分子設計手法、モンテカルロ木探索と再帰的ニューラルネットワークを組み合わせた構造発生手法ChemTSなども提案されている。深層学習を応用した構造生成の先行研究として、変分オートエンコーダ（Variational Autoencoder：VAE）[16] や敵対的生成ネットワーク（Generative Adversarial Network：GAN）などを応用したChemical VAEや Grammer VAE、DruGANなどの手法が報告されている[17–20]。再帰的ニューラルネットワーク、強化学習、GANを組み合わせたObjective–Reinforced Generative Adversarial Networks（ORGAN）が提案されている[21]。ドイツ製薬企業Bayer社の研究グループから、化学構造や物性情報だけでなく、遺伝子発現プロファイルなどのオミクス情報を使う化合物生成モデルなどもGANを基盤として提案されている[22]。 VAEの潜在空間上で遺伝子摂動応答遺伝子発現プロファイルと化合物応答遺伝子発現プロファイルの相関解析を行いヒット化合物候補の構造生成方法も提案されている[23]。合成可能性を考慮し、複数の化合物から一つの化合物を生成するための手法（Molecular Chef）なども提案されている[24]。しかしながら、化学的にはあり得ない構造を出力する手法も多く、提案される化学構造が実際に合成可能なものかどうかは保証が無いのが現状である。近年のトレンドとして、自然言語処理分野で注目を浴びているトランスフォーマー[25] を活用した構造生成の方法も提案されてきている。トランスフォーマーと変分オートエンコーダを組み合わせたTransVAE、トランスフォーマーと強化学習、敵対的生成ネットワークを組み合わせたTransORGAN[26] などが挙げられる。化合物の構造の文字列での表記法として、SMILESがよく利用されているが、他の文字列表現も提案されており、SELFIESやDeepSMILESなどが挙げられる。通常のSMILESだと出力された文字列が化学的に妥当な構造でない場合があるが、SELFIESだと出力された文字列が常に何らかの化学構造に対応するので、全く化学構造が出ない事態を回避できることが特長である[27]。 DeepSMILESは、ニューラルネットワークでの学習に特化したSMILES表記の拡張版として提案されている[28]。

構造生成の性能は化合物の構造の表現方法に依存するので、化合物の構造を文字列ではなく、グラフとして捉えて構造生成を行う手法の研究開発も進んでいる。JT–VAE[29] やGraph–AF[30] などが挙げられる。合成可能性を考慮した化合物の構造生成手法として、所望の物性を持つ化合物の構造を予測するだけでなく、その化合物を合成するための化学反応経路も提示する手法（casVAE）が提案されている[31]。

• 医薬品候補化合物の構造最適化

化合物の構造最適化は創薬における重要課題の一つであり、合成可能性、体内動態、毒性など複数の項目も同時に考慮して最適化された構造を出力できるように、多目的最適化に特化したアルゴリズムが必要である。リード最適化の過程では、現時点で最良の化合物の一部を構造変換した化合物群を合成して活性の変化を調べるが、それを効率化する情報技術としてMatched Molecular Pairs（MMP）解析がある。MMP解析では、指定した構造変換に対応する化合物ペアを検索し、構造変換による特性値の変化を確認することによって、構造の一部が異なる化合物ペアにおける置換基効果を調べることができる。最近、活性化合物の最適化に繋がるSAR Matrixと呼ばれる情報技術がドイツのボン大学のBajorathらにより提案されている[32]。化合物間の大規模な組み合わせに対して、化学構造の部分構造変換パターンと生物活性情報の対応を地図として視覚化するようなソフトウエアも日本の民間企業である理論創薬研究所によって開発されている[33]。また、トランスフォーマーとMMPを融合したアプローチも提案されている[34]。さらに、医薬品候補化合物の構造全体を生成するのではなく、構造の一部であるスキャフォールド（基本骨格）は固定した状態で構造生成を行う研究も行われている。 AstraZaneca社を中心にSMILES文字列の一部だけを変換する構造生成器[35]、Sanofi社を中心にSMILES文字列の一部だけを変換するように制約を加えた構造生成器[36]、VAEによる深層学習モデルとビルディングブロック型モデルを組み合わせた構造生成器などが提案されている[37]。

［注目すべき国内外のプロジェクト］

• Machine Learning for Pharmaceutical Discovery and Synthesis（MLPDS）Consortium

マサチューセッツ工科大学が中心となり、2018年に立ち上がった医薬品発見と合成のための機械学習コンソーシアムである。Pfizer社、Novartis社、Eli Lilly社など大手製薬会社も参画している。同コンソーシアムは、低分子の発見と合成に役立つソフトウエアのデザイン促進を目的としている。

• Machine Learning Ledger Orchestration for Drug Discovery（MELLODDY）project

欧州を中心とした製薬企業 10 社とアカデミア、IT 企業が参画して 2019 年から開始された。 機械学習による有望な化合物予測プラットフォームの構築を目指している。各企業は連合学習を用いることで、データの秘匿性を保持しながら 1,000 万を超える低分子化合物のデータを共同利用できる枠組みを構築し、世界最大のコレクションを活用した、より正確な予測モデルの確立、創薬の効率化が期待されている。

• Life Intelligence Consortium（LINC）

2016年11月に日本で設立された産学官連携コンソーシアムLINCでは、ライフサイエンス分野の産業競争力強化を目的として、アカデミア、製薬・ライフサイエンス企業、IT 企業など 100 以上の機関が参画し、10組のワーキンググループがシームレスな AI 創薬プラットフォームの構築を目指した技術開発を進めてきた。この成果を受け、2021年4月からは、一般社団法人ライフインテリジェンスコンソーシアムとして活動を行っている。第1期の個別AIモデルの研究開発に加えて、広くライフサイエンス分野のデジタルトランスフォーメーションを目指す活動（AI/データ基盤の構築、シンクタンク機能の確立や人材育成など）を進めている。また、活動領域は製薬だけでなく、ヘルスケア、化学、食品、農業などに拡大している。

• 官民研究開発投資拡大プログラム（PRISM）「新薬創出を加速する人工知能の開発」

「創薬標的の枯渇」問題を克服するための取り組みとして、医薬基盤・健康・栄養研究所、理化学研究所、科学技術振興機構など17の産学官研究機関が参画して2018年より開始されている。特発性肺線維症（IPF）と肺がんを対象疾患とし、それぞれの疾患の臨床情報、オミクス情報、医療データや既存知識を収集した、世界初あるいは世界最大規模の疾患統合データベースが構築されている。そこから新規創薬ターゲットを同定するため、新規解析プログラム、医療テキストや学術論文から医学・生物学分野の専門用語を自動抽出する自然言語処理プログラム、分子データを解析するための機械学習アルゴリズムなどのAIの開発が行われている。最終年度である2022年度には、オープンプラットフォーム「峰」という形での事業成果の提供が計画されている。

• AMED創薬支援推進事業「産学連携による次世代創薬AI開発（DAIIA）」

2020年度から開始されたプロジェクトであり、化合物―生体分子親和性予測AI、化合物構造発生AI及びオミクス情報に基づく標的予測AIの開発を目的としている。 AIの性能は学習データの化合物のケミカルスペースに大きく依存する。日本製薬工業協会の協力を得て、日本国内の製薬企業18社が参画し、公共データベースの化合物データだけでなく、企業が保有する大規模で多面的な化合物情報の提供を受けて、緊密な産学連携のもとにケミカルスペースの拡大を図っている。連合学習を利用し、企業間で化合物そのものを共有するのではなく、モデルのみ共有して学習することによって予測精度を高める試みが計画されている。真に実用的な統合創薬AIプラットフォームを構築し、開発されたシステムは事業化などを通して当該プロジェクト終了後も継続して活用することを目指している。

• AI–based Substances Hazardous Integrated Prediction System（AI–SHIPS）

毒性関連ビッグデータを用いた人工知能による次世代型安全性予測を目指した経済産業省のプロジェクト

である。基本的には一般化合物の毒性が予測対象となっており、動態的アプローチ、代謝的アプローチ、AI的アプローチ、トキシコゲノミクス的アプローチを組み合わせ、統合的予測システムの構築が進められた。従来の毒性予測の情報技術は、化学構造から機械学習によって毒性を予測する手法がほとんどであり、予測過程はブラックボックスであるため、なぜ毒性が予測できたのか解釈するのは困難である。本システムは、毒性の予測結果だけでなく毒性の発現メカニズムの情報も示唆できる点が特長である。同じ手法は医薬品開発における毒性研究にも利用可能なので、創薬応用の視点からも研究成果の応用が期待されている。

• 全ゲノム解析等実行計画

政府方針として「臨床情報と全ゲノム解析の結果等の情報を連携させ搭載する情報基盤を構築し、その利活用に係る環境を早急に整備する」（経済財政運営と改革の基本方針；2022年6月閣議決定）などが示され、がん・難病についての全ゲノム解析等のプロジェクトが計画されている。詳細はまだ明らかでないが、データ基盤の構築、解析手法の開発、医薬品開発への応用などの面で、AI創薬にも大きな関係をもつプロジェクトと考えられる。

（5）科学技術的課題

• 医薬品の標的となる疾患原因分子（主にタンパク質）の探索技術

上で述べた通り、ヒトデータのAIによる解析が進展しているが、診療情報だけではターゲット分子に到達するのは難しいので、ヒト検体と紐づいた分子情報（最も一般的にはオミクス解析データ）も同時に扱う必要がある。技術的には、診療情報を如何にAI解析可能な形に整形するかという大きな課題があり、データ生成時から構造化を保証する試み（電子カルテの規格の統一、オントロジーの整備など）と、非構造化データからの情報の自動抽出の試み（自然言語処理の活用など）の両方を推進するのが現実的なアプローチと考えられる。

分子情報については、オミクスデータ、特に最近一般的になっている一細胞オミクス解析に関わる実験コストの問題が大きいが、解析手法についても一層の進展が望まれる。ゲノム、トランスクリプトーム、プロテオーム、エピゲノム、メタボロームなど、多階層のオミクスデータの取得が可能になってきているが、現状は各階層のオミクスデータを個別に解析してパスウェイなどを抽出した結果を後から組み合わせるといったアプローチが主流で、文字通りの意味でマルチオミクスデータの統合解析を実行するアルゴリズムはほとんど存在しないか、少なくとも一般的に浸透していない。

また、診療情報と分子情報を組み合わせるAIモデルは、ほとんどの場合病態に関与する遺伝子あるいはパスウェイの候補をリストとして提案するのにとどまる。通常はその出力結果を現在知られている様々な知見と照らし合わせ、確からしいメカニズムと共に特定のターゲット候補を絞り込むという専門家による作業を必要としている。この作業までを自動化することが、本来の意味でのターゲット探索AIの究極の目的であり、今後の重要な技術的課題と言える。その実現のためにはまず、「現在知られている様々な知見」をデータベース化する必要がある。そのようなデータベースは知識ベースと呼ばれ、タンパク質、遺伝子、疾患など、各分野での整備が進みつつある。しかし、創薬ターゲットとしての妥当性を判断するために、個別分野の知識を超えた統合的な判断や推論が必要となる。一般に、用語や概念は分野毎に整理されることが多く、分野間でそれらを統合するのは困難であることが多い。その解決策として、既存知識を知識グラフという形でグラフ表現にすることで、比較的簡単にデータを繋げて拡大していうことができるのではないかと期待されている。

• 医薬品候補分子の活性や各種パラメータを予測する技術

低分子化合物については、化学構造と活性あるいは薬物動態パラメータなどを結びつける機械学習モデルを構築する技術は既に幅広く利用されている。しかし新しい創薬モダリティである中分子、ペプチド、核酸などについては類似のデータが十分に蓄積されていないため、同様のアプローチを適用することが難しい。そこ

で、これらの物質についても各種パラメータを網羅的に取得する実験を行ってデータの蓄積を図ることが重要であると考えられる。同様に、mRNAを用いた創薬が注目されているが、mRNAの物性やキャリアとの相互作用、それらと薬理活性との関係などについては、一般に利用できるデータがほとんど存在していない。これらについても、系統的に実験データを取得することで、機械学習を適用できる可能性が広がると考えられる。

より本質的な問題として、中分子、ペプチド、核酸など分子量の大きな分子については、分子構造の揺らぎが無視できない。これについては、低分子化合物の機械学習モデルの単純な延長で取り扱うことは難しいので、MDなどを取り入れた何らかの別のアプローチが必要となる。物理化学原理に基づくアプローチはデータの量に依存せず適用範囲が広いという利点があるが、一般に大きな計算機資源を必要とし、ハイスループットの解析には適さない。そこで、シミュレーションと機械学習を組み合わせるアプローチなどが新規モデリティ創薬には重要になると考えられる。

（6）その他の課題

・本分野で人材育成が課題であることは広く認識されているが、各現場で要求されるスキルなどをより明確化する必要がある。例えば、データサイエンティストという言葉は、人材募集などでもよく登場するが、実際には、ビジネスとデータ分析/解析チームをつなぐビジネストランスレータ、課題に応じたデータ収集や解析方法の選定を行うデータアナリスト、機械学習モデルを実装するAIエンジニア、といった異なる役割が存在する。 AIエンジニアには、プログラミング技術を含む相応の情報科学のバックグラウンドが必須である一方、創薬研究においてはむしろデータアナリストなどの果たす役割が大きい。そのタスクについては、例えば医薬系学部やウェット生物学の出身者が十分に対応可能であり、そのような領域からのリクルートを進めることが可能だと考えられる。その実現に向けて、求められる職種を広くアピールすると共に、大学などで必要なトレーニングの機会を提供する必要がある。

・上記で強調した患者由来情報の利用について、現在の日本の個人情報保護行政は、必ずしも情報の有効な利活用の促進に繋がっていない点が懸念される。例えば、個人情報保護法では匿名加工による第三者への情報提供が示されているが、実際の運用上は様々な障壁があり、必ずしも幅広いデータ利活用が進んでいないように見受けられる。これは、法律だけの問題だけでなく、データを提供する個人や、データ取得に関わる医療機関などのこれまでの慣習や意識にも関わる課題であり、データを公開することにより得られる利益を各コミュニティーに浸透させ、意識改革を促すという側面も重要である。

・個人情報以外のデータについても、これまで共有が難しかった製薬企業社内の化合物情報などの共有と利活用の動きについて上で取り上げた。こちらの課題は、法規制よりむしろ慣習や意識に関わる部分が大きい。個人情報や知的財産権に必ずしも直結しないデータであっても、データ産生者がデータを「囲い込む」事例は多い。 AlphaFoldの成功は、Worldwide Protein Data Bankというタンパク質立体構造のオープンデータがなければ生まれ得なかった。 AI技術を活用した新たなブレークスルーの実現に向けては、如何にオープンデータあるいはオープンサイエンスの文化を醸成していくかが課題だと思われる。

（7）国際比較：

国・地域	フェーズ	現状	トレンド	各国の状況、評価の際に参考にした根拠など
日本	基礎研究	○	↗	・複数の国家プロジェクトが進行し、政府の骨太の方針などでも、AI等の技術の創薬への有効活用が明記されている。
	応用研究・開発	○	↗	・複数のAI創薬スタートアップが現れ、アカデミアや製薬企業との共同研究が進展している。

米国	基礎研究	◎	→	・NIH–NCATS, Havard–MIT や西海岸の創薬研究の多くが、AlphaFoldなどを含むAIの利用を前提としている。
	応用研究・開発	◎	↗	・Illumina社とAstraZeneca社は、ヒトオミクス解析を通した創薬標発見の加速を目指した提携を発表している。 ・アカデミアの成果などを基に、Atomwise社、InveniAI社、Aria Pharmaceuticals社、Genesis Therapeutics社などAI創薬ベンチャーが数多く立ち上がっている。 ・MELLODYコンソーシアムには、米国のビッグファーマやIT企業も参画し、プラットフォーム構築を進めている。
欧州	基礎研究	◎	↗	・Google社傘下のDeepMind社によるAlphaFold及びAlphaFold Protein Structure Databaseの利用が進み、各分野に大きな影響を与えている。 ・Genomics EnglandやFinnGen programなどの大規模バイオバンクがAI創薬の基盤としての地位を確立しつつある。
	応用研究・開発	◎	↗	・BenevolentAI社、Relation Therapeutics社、Exscientia社などのスタートアップが投資を集め、新規ターゲットの導出などに成功している。
中国	基礎研究	○	↗	・北京大学・中国科学院上海薬物研究所など複数の拠点で、創薬応用に向けた研究が数多く行われている。
	応用研究・開発	○	↗	・国内の製薬・バイオ医薬品企業による新薬の開発や香港に拠点を置くInsilico Medicine社などによるAI創薬が進展している。
韓国	基礎研究	△	→	・KAIST, KIAS, Ewha Womans Universityなどのアカデミア機関で、AIや計算技術を活用した創薬研究が行われている。
	応用研究・開発	△	→	・Standigm社などのAIスタートアップが大手製薬企業との協業を開始している。
台湾	基礎研究	–	–	–
	応用研究・開発	–	–	・台湾工業技術研究院（日本の産総研に相当）では、バイオ分野を含むAI技術の先端技術研究を推進し、特定の対象疾患についての新薬探索を進めている。

（註1）フェーズ

基礎研究：大学・国研などでの基礎研究の範囲

応用研究・開発：技術開発（プロトタイプの開発含む）の範囲

（註2）現状　※日本の現状を基準にした評価ではなく、CRDS の調査・見解による評価

◎：特に顕著な活動・成果が見えている　　○：顕著な活動・成果が見えている

△：顕著な活動・成果が見えていない　　×：特筆すべき活動・成果が見えていない

（註3）トレンド　※ここ1～2年の研究開発水準の変化

↗：上昇傾向、→：現状維持、↘：下降傾向

関連する他の研究開発領域

・AIソフトウェア工学（システム・情報分野　2.1.4）
・AI・データ駆動型問題解決（システム・情報分野　2.1.6）

参考文献

1）Tsuyoshi Esaki, et al., “Data Curation can Improve the Prediction Accuracy of Metabolic Intrinsic Clearance,” *Molecular Information* 38, no. 1-2（2019）: 1800086., https://doi.org/10.1002/minf.201800086.

2）Hiroshi Komura, et al., “A public-private partnership to enrich the development of in silico

predictive models for pharmacokinetic and cardiotoxic properties," *Drug Discovery Today* 26, no. 5（2021）: 1275-1283., https://doi.org/10.1016/j.drudis.2021.01.024.
3）Masataka Kuroda, et al., "Utilizing public and private sector data to build better machine learning models for the prediction of pharmacokinetic parameters," *Drug Discovery Today* 27, no. 11（2022）: 103339., https://doi.org/10.1016/j.drudis.2022.103339.
4）Michael Eisenstein, "Machine learning powers biobank-driven drug discovery," *Nature Biotechnology* 40, no. 9（2022）: 1303-1305., https://doi.org/10.1038/s41587-022-01457-1.
5）大田雅照, 池口満徳「タンパク質立体構造に基づく創薬における人工知能技術の応用」『MEDCHEM NEWS』28 巻 4 号（2018）: 175-180., https://doi.org/10.14894/medchem.28.4_175.
6）Hiroaki Imoto, Sawa Yamashiro and Mariko Okada, "A text-based computational framework for patient-specific modeling for classification of cancers," *iScience* 25, no. 3（2022）: 103944., http://doi.org/10.1016/j.isci.2022.103944.
7）Minkyung Baek, et al., "Accurate prediction of protein structures and interactions using a three-track neural network," *Science* 373, no. 6557（2021）: 871-876., https://doi.org/10.1126/science.abj8754.
8）John Jumper, et al., "Highly accurate protein structure prediction with AlphaFold," *Nature* 596, no. 7873（2021）: 583-589., https://doi.org/10.1038/s41586-021-03819-2.
9）Satoko Namba, Michio Iwata and Yoshihiro Yamanishi, "From drug repositioning to target repositioning: prediction of therapeutic targets using genetically perturbed transcriptomic signature," *Bioinformatics* 38, Suppl 1（2022）: i68-i76., https://doi.org/10.1093/bioinformatics/btac240.
10）Yingnan Han, et al., "Empowering the discovery of novel target-disease associations via machine learning approaches in the open targets platform," *BMC Bioinformatics* 23（2022）: 232., https://doi.org/10.1186/s12859-022-04753-4.
11）Victoria Hore, et al., "Tensor decomposition for multiple-tissue gene expression experiments," *Nature Genetics* 48, no. 9（2016）: 1094-1100., https://doi.org/10.1038/ng.3624.
12）Jianwen Fang, "Tightly integrated genomic and epigenomic data mining using tensor decomposition," *Bioinformatics* 35, no. 1（2019）: 112-118., https://doi.org/10.1093/bioinformatics/bty513.
13）Yoshihiro Taguchi and Turki Turki, "Tensor-Decomposition-Based Unsupervised Feature Extraction in Single-Cell Multiomics Data Analysis," *Genes* 12, no. 9（2021）: 1442., https://doi.org/10.3390/genes12091442.
14）Michio Iwata, et al., "Predicting drug-induced transcriptome responses of a wide range of human cell lines by a novel tensor-train decomposition algorithm," *Bioinformatics* 35, no. 14（2019）: i191-i199., https://doi.org/10.1093/bioinformatics/btz313.
15）Petra Schneider, et al., "Rethinking drug design in the artificial intelligence era," *Nature Reviews Drug Discovery* 19, no. 5（2020）: 353-364., https://doi.org/10.1038/s41573-019-0050-3.
16）G. E. Hinton and R. R. Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks," *Science* 313, no. 5786（2006）: 504-507., https://doi.org/10.1126/science.1127647.
17）Rafael Gómez-Bombarelli, et al., "Automatic Chemical Design Using a Data-Driven

Continuous Representation of Molecules," *ACS Central Science* 4, no. 2（2018）: 268-276., https://doi.org/10.1021/acscentsci.7b00572.
18）Matt J. Kusner, Brooks Paige and José Miguel Hernández-Lobato, "Grammar Variational Autoencoder," arXiv, https://doi.org/10.48550/arXiv.1703.0192,（2023年2月2日アクセス）.
19）Xiufeng Yang, et al., "ChemTS: an efficient python library for de novo molecular generation," *Science and Technology of Advanced Materials* 18, no. 1（2017）: 972-976., https://doi.org/10.1080/14686996.2017.1401424.
20）Artur Kadurin, et al., "druGAN: An Advanced Generative Adversarial Autoencoder Model for de Novo Generation of New Molecules with Desired Molecular Properties in Silico," *Molecular Pharmaceutics* 14, no. 9（2017）: 3098-3104., https://doi.org/10.1021/acs.molpharmaceut.7b00346.
21）Gabriel Lima Guimaraes, et al., "Objective-Reinforced Generative Adversarial Networks（ORGAN）for Sequence Generation Models," arXiv, https://doi.org/10.48550/arXiv.1705.10843,（2023年2月2日アクセス）.
22）Oscar Méndaz-Lucio, et al., "De novo generation of hit-like molecules from gene expression signatures using artificial intelligence," *Nature Communications* 11（2020）: 10., https://doi.org/10.1038/s41467-019-13807-w.
23）Kazuma Kaitoh and Yoshihiro Yamanishi, "TRIOMPHE: Transcriptome-Based Inference and Generation of Molecules with Desired Phenotypes by Machine Learning," *Journal of Chemical Information and Modeling* 61, no. 9（2021）: 4303-4320., https://doi.org/10.1021/acs.jcim.1c00967.
24）John Bradshaw, et al., "Barking up the right tree: an approach to search over molecule synthesis DAGs," in *Advances in Neural Information Processing Systems 33*, eds. H. Larochelle, et al.（NeurIPS, 2020）.
25）Ashish Vaswani, et al., "Attention is All you Need," in *Advances in Neural Information Processing Systems 30*, eds. Isabelle Guyon, et al.（NeurIPS, 2017）, 6000-6010.
26）Chen Li, et al., "Transformer-based Objective-reinforced Generative Adversarial Network to Generate Desired Molecules," in *Proceedings of the 31st International Joint Conference on Artificial Intelligence（IJCAI2022）*, ed. Luc De Raedt（IJCAI, 2022）, 3884-3890., https://doi.org/10.24963/ijcai.2022/539.
27）Mario Krenn, et al., "Self-referencing embedded strings（SELFIES）: A 100% robust molecular string representation," *Machine Learning: Science and Technology* 1, no. 4（2020）: 045024., https://doi.org/10.1088/2632-2153/aba947.
28）Noel O'Boyle and Andrew Dalke, "DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures," ChemRxiv, https://doi.org/10.26434/chemrxiv.7097960.v1,（2023年2月2日アクセス）.
29）Wengong Jin, Regina Barzilay and Tommi Jaakkola, "Junction Tree Variational Autoencoder for Molecular Graph Generation," *Proceedings Machine Learning Research* 80（2018）: 2323-2332.
30）Chence Shi, et al., "GraphAF: a Flow-based Autoregressive Model for Molecular Graph Generation," 8th International Conference on Learning Representations（ICLR 2020）, https://iclr.cc/virtual_2020/poster_S1esMkHYPr.html,（2023年2月7日アクセス）.
31）Dai Hai Nguyen and Koji Tsuda, "Generating reaction trees with cascaded variational

autoencoders," *The Journal of Chemical Physics* 156, no. 4（2022）：044117., https://doi.org/10.1063/5.0076749.
32）Disha Gupta-Ostermann and Jürgen Bajorath, "The 'SAR Matrix' method and its extensions for applications in medicinal chemistry and chemogenomics [version 2; peer review: 2 approved]," *F1000Research* 3（2014）：113., https://doi.org/10.12688/f1000research.4185.2.
33）Atsushi Yoshimori, Toru Tanoue and Jürgen Bajorath, "Integrating the Structure-Activity Relationship Matrix Method with Molecular Grid Maps and Activity Landscape Models for Medicinal Chemistry Applications," *ACS Omega* 4, no. 4（2019）：7061-7069., https://doi.org/10.1021/acsomega.9b00595.
34）Jiazhen He, et al., "Transformer-based molecular optimization beyond matched molecular pairs," *Journal of Cheminformatics* 14（2022）：18., https://doi.org/10.1186/s13321-022-00599-3.
35）Josep Arús-Pous, et al., "SMILES-based deep generative scaffold decorator for de-novo drug design," *Journal of Cheminformatics* 12（2020）：38., https://doi.org/10.1186/s13321-020-00441-8.
36）Maxime Langevin, et al., "Scaffold-Constrained Molecular Generation," *Journal of Chemical Information and Modeling* 60, no. 12（2020）：5637-5646., https://doi.org/10.1021/acs.jcim.0c01015.
37）Kazuma Kaitoh and Yoshihiro Yamanishi, "Scaffold-Retained Structure Generator to Exhaustively Create Molecules in an Arbitrary Chemical Space," *Journal of Chemical Information and Modeling* 62, no. 9（2022）：2212-2225., https://doi.org/10.1021/acs.jcim.1c01130.