生命科学・創薬研究における人工知能活用データ基盤の構築に向けて

技術解説

水口賢司*

Towards the building of an AI-utility data platform in biological and drug discovery research

Key Words：machine learning, database, curation, pharmacokinetics

はじめに

生命科学や創薬研究においても、人工知能（artificial intelligence, AI）の活用は急速に進んでいる。原子・分子レベルでのタンパク質立体構造予測や、細胞レベルでの1細胞RNA-seqの大規模言語モデル解析などは、その代表的な例と言えるだろう。特に前者は、AlphaFold[1]開発者の2024年度ノーベル賞受賞も相まって、幅広い関連分野で注目を集めている。

生物学でのビッグデータというとゲノム情報が思い浮かぶが、実際には、遺伝子の核酸配列情報よりも先に、タンパク質の立体構造情報がデータベースとして蓄積され、研究コミュニティーでの利用の枠組みが整えられた。タンパク質立体構造の国際リポジトリであるProtein Data Bank（PDB）[2]は、1971年設立で50年以上の歴史を誇る。米欧日の連携により、共通のデータが公共資産として維持管理され、阪大蛋白研が運営するPDBjapanもその一翼を担っている。このような、質を担保した大規模データが無償で利用可能であったことが、AlphaFoldなどの立体構造予測AIモデルの成功に繋がったという点は、見逃してはいけない。

一方で、生物学の他の多くの問題については、必ずしも同様の恵まれた状況が存在しておらず、適切なデータが不足している、あるいは存在しないことが、AI活用のボトルネックになっている。本稿では、この問題意識の基に、どのようにしてAI活用のためのデータ基盤を整えるのかについて、特に筆者が関わっている創薬研究分野の幾つかの問題を例に取って議論したい。

* Kenji MIZUGUCHI
1967年5月生まれ
京都大学大学院　理学研究科　化学専攻
博士後期課程（1995年）
現在、大阪大学　蛋白質研究所　教授
博士（理学）
TEL：06-6879-4743
E-mail：kenji@protein.osaka-u.ac.jp

教師あり学習

現代のAIの中心にある機械学習は、データから規則や知識を抽出する技術・手法と説明される。多くの生物学的な問題に応用可能な代表的な機械学習の枠組みとして、ここでは教師あり学習に絞って話を進める。図1には、化合物の活性を化学構造だけから予測するモデルの構築を例として示す。まず、注目している活性の有無（例えば、特定の標的タンパク質の機能を阻害するかどうか）が既知である化合物の情報を収集する。図1左の表の例では、化合物1は活性あり、化合物2は活性なし、とわかっている。文献やデータベースを検索する、あるいは新規に実験を行うなどして、このような例をできるだけ多数集める。これが基礎のデータとなる。

次に、コンピュータ解析を可能にするために、化合物の構造情報を何らかの形で数値化する必要がある。図1の例では、化合物1に、287.3, -2.1, 0, 0, 1, …といった一群の数値を付与しているのがそれに当たる。これを化学構造に対する「表現」と呼ぶ。数値が定まれば、各化合物を点として配置し、各点に活性の有無を示す属性（図1では「ラベル」と呼び、赤色か黄色かで区別している）を付与できる。多数の点からパターンを導き出す作業が、学習と呼ばれる（図1では、直線あるいは曲線のどちら側に位置するかで点の色を判別する）。学習済みのモデルは、数値化された化学構造を与えられれば（入力）、活性の有無を判定できる（出力）。

生命科学や創薬研究では、他にも多様な対象を扱

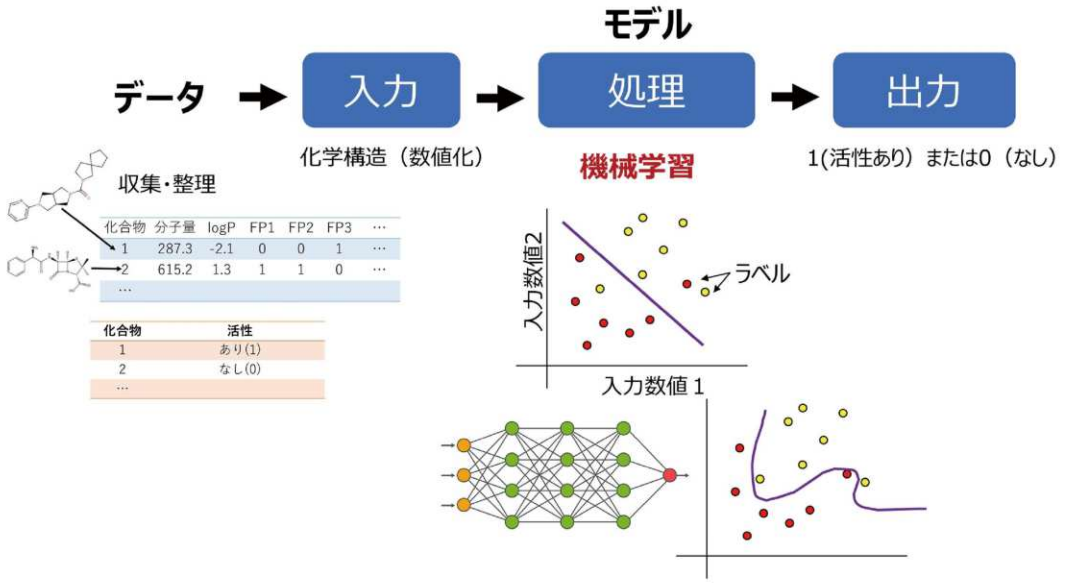


図1　化合物の活性を化学構造だけから予測するモデルの構築

う。例えば、化合物の活性ではなく、ウイルスが作るタンパク質の各種変異体と病原性との関係をAIで予測したい場合、あるいは遺伝子発現データから細胞の属性を予測したい場合、それぞれタンパク質変異体の表現、遺伝子発現データの表現を用意する必要がある。データの表現によって、AIモデルの性能が大きく変わることが知られているため、問題の本質を捉えた適切な表現を選ぶことが重要になる。近年は、大規模言語モデルなど、元々自然言語処理の分野で開発された表現が生命科学の分野でも幅広く利用できることが見出されているが[3]、一般的には適切な表現を見出すことは、対象についての知識を必要とする重要な研究課題である。そのような解析を可能にする前提として、コンピュータフレンドリーな形でデータが整理されていることが必須となる。

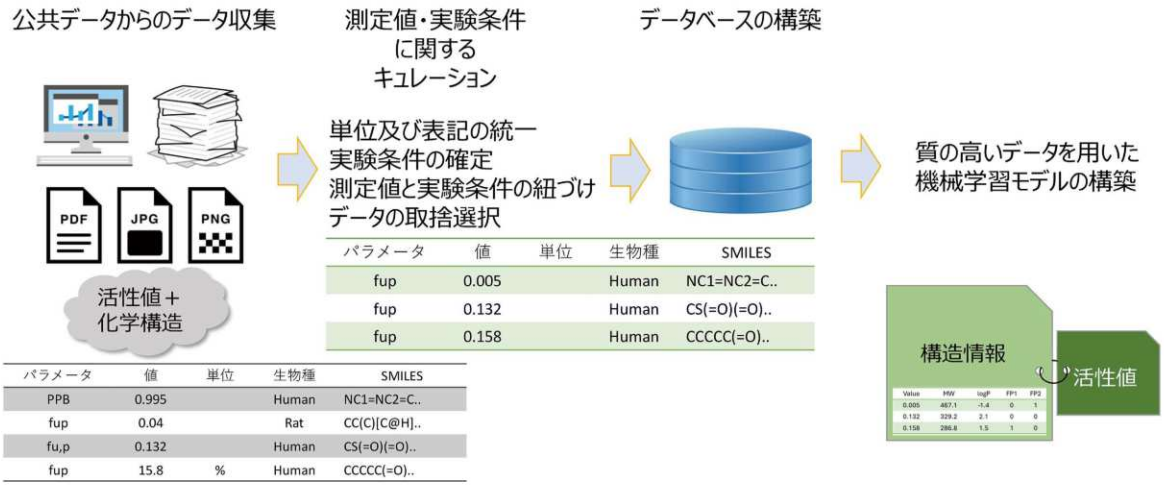
キュレーション

データの選別、整形、統合などの作業のことをキュレーションと呼んでいる。これは、AIモデル構築の基礎であり、また整理・選別の度合い（データの質）が予測モデルの性能に影響を与えることが、我々及び他のグループの研究で具体的に示されている[4]。化合物の活性予測を例とした一般的なキュレーションの流れを図2に示す。

データの出所としては、データベース、論文、インハウス実験など様々なものが考えられるが、実験条件や単位などが統一されていない場合が多い。表記の異なるデータについて実験条件を精査し、統合が可能かどうかを判定する。また、同じ条件で複数の測定値が存在する場合にどう対応するかを含め、モデル構築の目的に合わせてデータを取捨選択する必要がある。単位の変換については、単純に係数をかけるだけでなく、実験手法に応じて、より複雑な変換を施さねばいけない場合がある。

このような作業は、当たり前に行われているものと思われるかもしれないが、必ずしも常に実践されているとは限らない。我々は、化合物の体内動態（吸収、分布、代謝、排泄）を規定する重要なパラメータの一つである血漿タンパク非結合率と呼ばれる量を化学構造から予測するモデルを構築して発表した[5]。これに対して、その性能を批判する論文[6]が発表されたが、この論文の著者らが行った解析に用いたデータを精査すると、結合率と非結合率とが混在したものであった。（同じ実験から得られる測定値が、非結合率で表現される場合もあれば、結合率として報告される場合もある。文献6の図5が、現在では大きく修正されている。）キュレーションの重要性を改めて示す例と言えるだろう。

キュレーションを施したデータには、当該分野の専門家の知識が集約されており、データベースなどの形で再利用できることが望ましい。低分子化合物



パラメータ	値	単位	生物種	SMILES
fup	0.005		Human	NC1=NC2=C..
fup	0.132		Human	CS(=O)(=O)..
fup	0.158		Human	CCCCC(=O)..

パラメータ	値	単位	生物種	SMILES
PPB	0.995		Human	NC1=NC2=C..
fup	0.04		Rat	CC(C)[C@H]..
fu,p	0.132		Human	CS(=O)(=O)..
fup	15.8	%	Human	CCCCC(=O)..

図2　キュレーションの流れ

の薬物動態パラメータを集約したDruMAP (https://drumap.nibn.go.jp)[7]や、創薬標的探索のための統合データベースTargetMine (https://targetmine.mizuguchilab.org)[8]は、この意味での貴重なリソースと位置付けられると考えている。

データが不足する場合

冒頭で述べた通り、生物学の多くの問題については、AI・機械学習の活用に適った質を持つデータが十分に存在しない。この解決にもっとも有効な手段は、地道にデータを収集することだが、解析手法の工夫によってある程度の対応は可能になる。

まず、対象とする問題そのものに対するデータは少なくても、関連する問題についてのデータが多く存在する場合には、事前学習や転移学習と呼ばれる手法を適用することで、有効なAIモデルを構築できる可能性がある。図3に、シトクロムP450 (Cytochrome P450, CYP) 阻害の予測を例とした説明を示す。

CYPは主要な薬物代謝酵素であり、基質特異性の異なる多数の分子種が存在する。これらはアミノ酸配列類似度に基づいて分類され、接頭語CYPの後にファミリーを示すアラビア数字、サブファミリーを示すアルファベット、分子種番号を示すアラビア数字の組合せで表される。ある化合物が特定のCYP種の活性を阻害すると、そのCYP種で代謝される医薬品の代謝に影響を与え、薬の飲み合わせの際の副作用などの問題が起こり得る。そこで我々は、深層学習を用いて7つのCYP種（CYP1A2、CYP2B6、CYP2C8、CYP2C9、CYP2C19、CYP2D6、CYP3A4）について、与えられた化合物が各CYP種の活性を阻害するかどうかを予測するインシリコ阻害予測モデルを構築した[9]。

通常のやり方でCYP種毎に活性データを収集してモデルを作ると、CYP2B6とCYP2C8に関しては、実験データが少ないため十分な精度の予測モデルを構築できない（図3左のシングルタスク学習、single task learning）。その場合でも、データ数の多いCYP種（CYP3A4, CYP1A2, CYP2C9, CYP2C19, CYP2D6）とデータ数の少ないCYP種（CYP2B6, CYP2C8）のデータを用いてマルチタスク学習（multitask learning）を行うことができる（図3中央）。さらに、CYP3A4などの大規模なデータセットを用いてモデルを作成し（事前学習）、実際の問題（CYP2B6など）の少数データに適応させるためにモデルのパラメータを微調整するファインチューニングと呼ばれる手法を用いることもできる（図3右）。これらの手法を組み合わせることにより、データ数の少ないCYP2B6およびCYP2C8に対する阻害予測の性能が、単独に構築した予測モデルに比べて統計的に有意に上昇することを示した[9]。

以上とは異なるアプローチとして、何らかの一般原理に基づくシミュレーションと機械学習とを組み合わせることで、データの不足を補う手法が存在する（ハイブリッドモデリングと呼ばれることがある）。気象現象など、一般原理（物理法則）がはっ

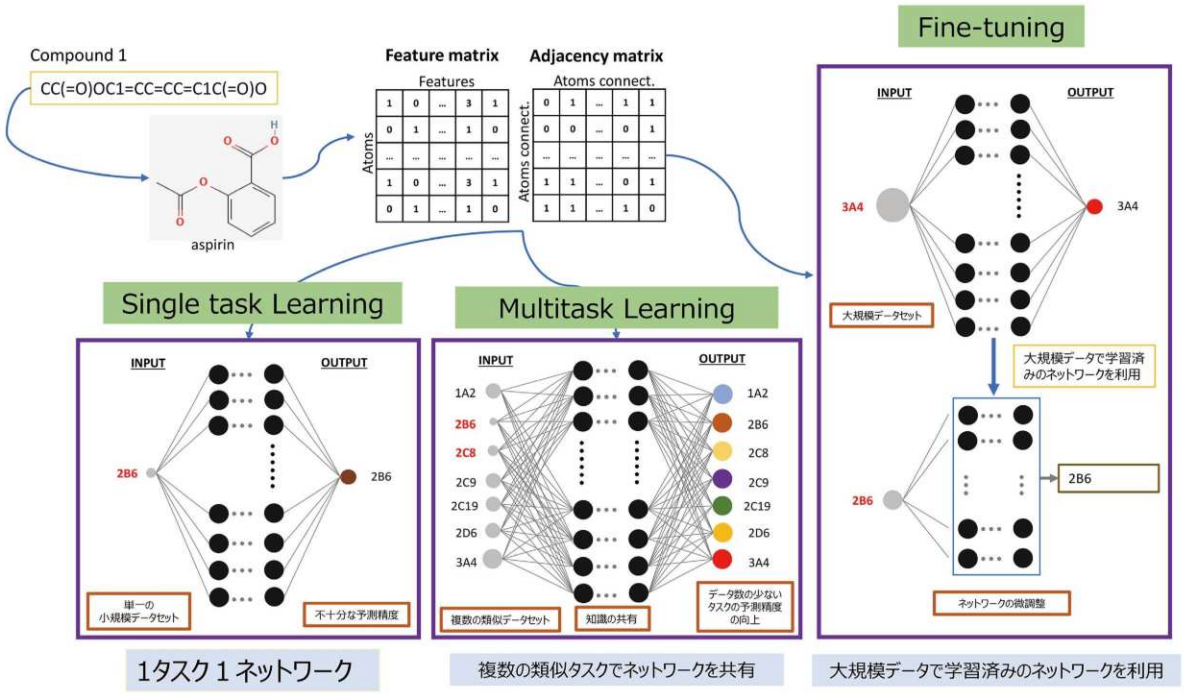


図3　データが不十分なCYP種に対する阻害予測の精度向上の試み

きりしている分野に比べて、生命科学におけるハイブリッドモデリングは、まだ一般的ではない印象がある。しかし、関連する考え方自身は古くからあり、例えば我々が行った、DNA二重鎖の分子動力学によってデータを生成し、それを学習データとして予測モデルを構築する研究[10]もその一例と言えるだろう。さらに、化学平衡の原理に基づく数理モデルの係数の一部を機械学習モデルで決めることにより、投与した化合物が脳に移行する度合いを予測するモデル[11]についても、ハイブリッド的なアプローチと言える。

分子動力学などの物理化学原理に基づくシミュレーションと機械学習との融合は、液—液相分離への応用など生物物理学的な問題での有用性を発揮しつつある。一方で、細胞レベルの数理モデルに基づくシミュレーションと、原子・分子レベルの機械学習との組み合わせについては、まだこれから挑戦すべき課題と考えられる。

事前学習は、ChatGPTなどの生成AIの進展に大きな寄与をしている。タンパク質の言語モデルはその直接的な応用だが、これを他の種類の生物学的データに拡張して、「基盤モデル」を構築しようとする試みが広がりつつある（文献12、https://www.riken.jp/research/labs/trip/agis/ など）。

おわりに

本稿では、データに焦点をあて、生命科学・創薬研究におけるAI活用の基盤としての重要性を議論した。これらの分野では、実験研究によって多くのデータが産生されているにも関わらず、それらの集積・活用がなされていない場面が散見される。伝統的な生物学は仮説駆動型であり、仮説に合わないデータは「ネガティブデータ」として、捨ててしまうマインドが見られる。一方で、データ駆動型の研究においては、ネガティブデータは、ポジティブデータと同様あるいはそれ以上の価値を持つ場合がある。データを掘り起こして整理することで、新たな融合研究に繋がることを期待したい。

謝辞

図の作成に協力頂いた渡邉怜子博士に感謝する。本研究の一部は国立研究開発法人日本医療研究開発機構（AMED）および、OUマスタープラン実現加速事業の支援を受けて行われた。

参考文献

1) Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493－500 (2024).
2) Burley, S. K. *et al.* Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res* **47**, D520－D528 (2019).
3) Simon, E., Swanson, K., Zou, J., Biohub, C.-Z. & Francisco, S. Language models for biological research: a primer. *Nature Methods* **21**, 1422－1429 (2024).
4) Esaki, T. *et al.* Data Curation can Improve the Prediction Accuracy of Metabolic Intrinsic Clearance. *Mol Inform* **38**, 1800086 (2019).
5) Watanabe, R. *et al.* Predicting Fraction Unbound in Human Plasma from Chemical Structure: Improved Accuracy in the Low Value Ranges. *Mol Pharm* **15**, 5302－5311 (2018).
6) Mulpuru, V. & Mishra, N. In Silico Prediction of Fraction Unbound in Human Plasma from Chemical Fingerprint Using Automated Machine Learning. *ACS Omega* **6**, 6791－6797 (2021).
7) Kawashima, H. *et al.* DruMAP: A Novel Drug Metabolism and Pharmacokinetics Analysis Platform. *J Med Chem* **66**, 9697－9709 (2023).
8) 陳怡安, 李秀栄 & 水口賢司. TargetMine による生物学的知識の発見. 医学のあゆみ **278**, 641－645 (2021).
9) Permadi, E. E., Watanabe, R. & Mizuguchi, K. Improving the accuracy of prediction models for small datasets of Cytochrome P450 inhibition with deep learning. *J Cheminform* (in press).
10) Andrabi, M. *et al.* Predicting conformational ensembles and genome-wide transcription factor binding sites from DNA sequences. *Sci Rep* **7**, 4071 (2017).
11) Watanabe, R. *et al.* Development of an In Silico Prediction Model for P-glycoprotein Efflux Potential in Brain Capillary Endothelial Cells toward the Prediction of Brain Penetration. *J Med Chem* **64**, 2725－2738 (2021).
12) Eisenstein, M. Foundation models build on ChatGPT tech to learn the fundamental language of biology. *Nat Biotechnol* **42**, 1323－1325 (2024).