

The **Beyond Disciplines** Collection

AI×バイオ DX時代のライフサイエンス・ バイオメディカル研究



国立研究開発法人 科学技術振興機構 研究開発戦略センター
Center for Research and Development Strategy, Japan Science and Technology Agency

目次

はじめに	3
1 画像データ	6
1.1 医用画像（医療・医科学）	6
1.2 イメージングデータ（生命科学）	9
2 分子・化合物データ（in silico創薬）	15
2.1 分子インフォマティクス	15
2.2 タンパク質の3次元構造予測	18
2.3 薬の毒性・副作用予測	19
3 ゲノム・オミックスデータ	21
3.1 ゲノムのSNPs等解析	21
3.2 表現型の層別化	21
3.3 疾患リスク予測	23
3.4 疾患の層別化（予後予測）	23
4 AI活用の課題	27
5 今後の展望（日本が取り組むべき方向性）	30
5.1 今後の展望①戦略的なデータプラットフォーム構築 ～オープンサイエンスの実践～	30
5.2 今後の展望②実験の自動化・デジタル化 ～データ取得のハイスループット化・オンデマンド化～	35
5.3 今後の展望③自然言語処理とEHR（電子カルテ等）の利活用 ～AI医師～	39
5.4 今後の展望④システムとしての生命の理解 ～複雑系・ダイナミクス理解へのAIの活用～	42
5.5 今後の展望⑤脳神経科学ベースの次世代AIアルゴリズムの可能性	45
終わりに 日本への示唆	49
付録	52
付録1 検討の経緯	52

コラム

- ① コンピューショナル・イメージング
- ② 農業におけるAI（画像データ）の応用
- ③ 物質・材料研究（マテリアルズ・インフォマティクス）との比較
- ④ 機械学習プロセスの自動化技術
- ⑤ 健康・医療におけるGoogle社の取組み ～Project Baseline～
- ⑥ ウェアラブル機器からのデータによる健康・医療研究
- ⑦ 第4世代AI

はじめに

JST研究開発戦略センター（CRDS）では、2019年度以降、「研究開発の俯瞰報告書 ライフサイエンス・臨床医学分野（2019年）」、「デジタルトランスフォーメーションに伴う科学技術・イノベーションの変容」の2つの報告書を作成、公開し、デジタル化するライフサイエンス・バイオメディカル研究の動向を注視してきた。その間に、政府では、強化すべき主要技術分野として「AI戦略」、「バイオ戦略」が策定された。いずれにも共通する事項としてデジタル化の進展への対応が挙げられ、その中でAIがキーテクノロジーとなっている。そうした状況を踏まえ、本報告書は、バイオとAIはどの程度クロスしているか、今後の展望はどうかという視点からまとめたものである。

2012年のディープラーニングの登場により、第三次AIブームが到来し、AIの利活用は社会のあらゆる分野に大きな影響を及ぼし始めている。ライフサイエンスや健康・医療分野の研究にはどのように影響しているだろうか。

現在の医療の大きな方向性は、個人ごとに治療の選択や予防を可能とすることであり、これに機械学習やディープラーニングが貢献できることを示す例が次々と出てきている。最も顕著なのは2012年に登場した畳み込みニューラルネットワーク（CNN：Convolutional Neural Network）を多層ニューラルネットワークとして用いた画像解析技術である。初期のCNNは非常に遅く、計算上高コストであったが、いくつかの改良モデルが競われてきた結果、2015年には処理速度や能力において人間の画像認識を超えたとされる。こうした技術の進展の結果、2018年には、FDA（米国食品医薬品局）が世界で初めて自律型AI診断システムを承認した。こういった技術により、早期段階で疾患の可能性を判断することができる。本報告書では、こうした「識別モデル」が医療画像と研究機器からのイメージング画像にいかに関与されているかの例を紹介する。

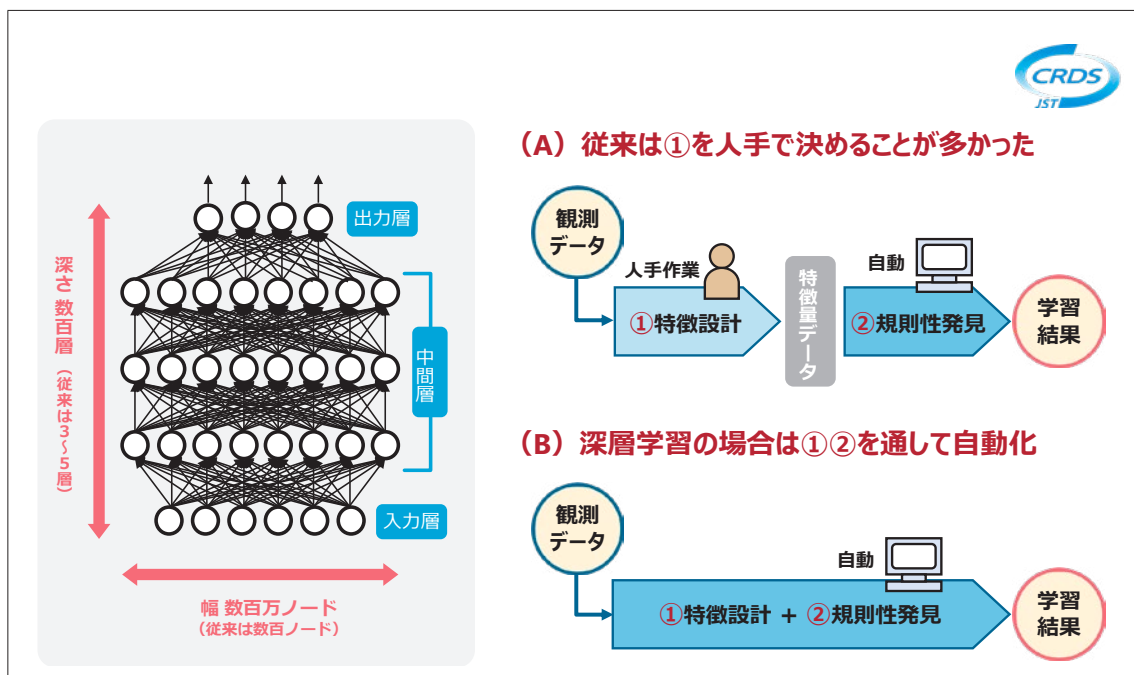
さらに、2014年に敵対的生成ネットワーク（GAN：Generative Adversarial Network）が登場し、本物に見えるが実在しないデータ（フェイク画像等）の自動生成といったことができるようになった。様々な進展を経て、同じ「生成モデル」として2017年には、変分オートエンコーダ（VAE：Variational Auto Encoder）に基づき化合物を生成する技術が登場した。これにより、分子インフォマティクスが再興し、創薬に展開されている。2020年に入り、世界で初めて大手製薬企業がAIを活用して創製された新薬候補化合物の臨床試験の開始を公表した。従来であればリード化合物の同定に4年程度要していた工程を1年未満で実現するものである。

また、2018年頃から、AIとの相性の良さから、物質合成や代謝工学などにおいて、実験の自動化（ロボット利用）、ハイスループット化の研究発表が目立つようになっている。

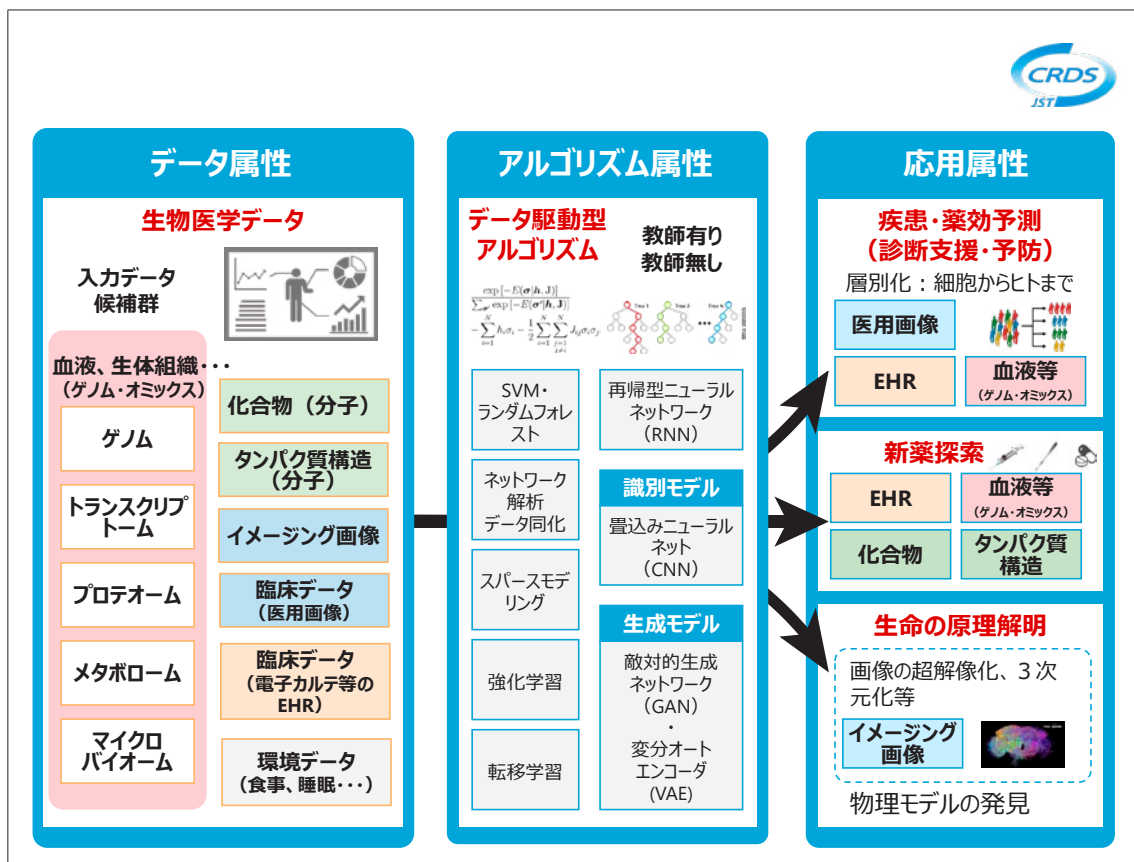
従来のサポートベクターマシン（SVM）、ランダムフォレスト（RF）、スパースモデリングといった手法と機械学習の1種である「ニューラルネットワーク」の階層を深めたディープラーニング（recurrent neural network（RNN）、CNN等）では何が違い、何が変わったのか。

これらを使用する目的は、分類、回帰、クラスタリング等で共通し、実際の作業プロセスとしては、下記のような形となる。

- ①学習用データの準備（アノテーション情報付加）
- ②説明変数（記述子・特徴量・入力変数）の選択
- ③機械学習アルゴリズムの選択
- ④アルゴリズムによる学習用データの学習とモデルの生成
- ⑤得られたモデルをテストデータセットで評価



図a 多層ニューラルネットワーク



図b AI × バイオの全体像

⑥本解析の実施（解析したいデータの適用）

ディープラーニングは、精度の高さと併せて、データを学習させる際に②のプロセスである特徴設計も自動的に行うという点も、従来との違いとして大きい（図a）。

バイオロジー、ライフサイエンス分野は、2010年以降、次世代シーケンサーの急速な技術進展に伴う一細胞トランスクリプトーム技術の浸透、ゲノム編集やオプトジェネティクスのようなゲノム制御技術の出現、さらにAI技術の発展、といった先端技術群の出現により従来の研究からフェーズチェンジが起こった。

上述したようにディープラーニングの登場により、バイオ分野とAI研究の融合は目覚ましく、医用画像を用いた疾患リスク診断支援（病理画像解析）、電子カルテや問診音声データ（EHR）を用いた医療診断支援、血液・尿・唾液中のゲノムやタンパク質など生体分子データからの疾患リスク予測（バイオマーカー探索）、ゲノム・オミックスや化合物データを用いた新薬探索（薬の標的同定、分子設計、ドラッグリポジショニング、薬効（毒性・副作用）予測）、イメージング画像を用いた高度解析など、AI医療やAI創薬等の研究が世界中で活発化している。

今般の新型コロナウイルス感染症、また今後の新興感染症の発生の可能性も考えると、デジタルトランスフォーメーションを加速させていくことは理に適っており、今後もこの方向は重要である。

そこで、本報告書では、主にディープラーニング登場後の研究の進展について述べる。本報告書で述べる研究の全体像は図bのとおりである。

研究動向を概観するにあたって、画像やゲノムなどデータ属性、AIのアルゴリズム属性、応用属性のいずれの視点からも示すことができるが、本報告書ではデータ属性の視点から研究動向を取りまとめた。

1 画像データ

医療画像やバイオイメージングのデータは多次元化が進んでいる。2次元（ピクセル）から3次元（ボクセル）、4次元（ボクセルの時間変遷情報）、さらには5次元（スペクトルのような波長情報）まで含むデータを取得できるようになっており、将来的にデータの次元はますます拡張することが予想される。しかし、人間が目でモノを見ることで判断できる情報は対象物までの距離と色、表面の形に限られる。つまり、2.5次元以上の情報というのは、私たちには認識が困難である。そこで、ディープラーニングによる画像処理が威力を発揮する。

CNN（畳み込みニューラルネットワーク）は、特に画像認識で高い性能を発揮する。あらかじめ与えられた画像データから画像の空間的に相関する特徴量を直接抽出することができる。その結果、分類（その画像に何が写っているのか）、検出（その画像のどこに何があるのか）、セグメンテーション（領域の識別：個々のピクセルが何を意味しているのか）などの解析が可能となる。生命科学や医療の分野では、特に、U-Net（2015年～）に代表されるような画像のSemantic Segmentation手法が用いられることが多い。

GAN（敵対的生成ネットワーク）は、生成器と識別器の2つの学習モデルからなる。生成器では粗い画像からできるだけお手本に近い高精度な画像を生成し、識別器はお手本なのか生成器からの出力なのかを見分ける学習モデルで、両者の精度を高める学習を繰り返すことで、本物と区別できないほどの画像を生成する。データから特徴を学習することで、一見本物のようだが実在しないデータの自動生成や画像の超解像化を行うことが出来る技術である。

ここでは、医用画像（医療・医科学）とイメージングデータ（生命科学）それぞれでの進展を述べる。

1.1 医用画像（医療・医科学）

（1）疾患リスクの検出、予測

米国では2018年4月、眼の網膜を撮影した画像から糖尿病網膜症かどうか診断するAIを用いた医療機器が世界で初めて承認された。それ以降、FDA（米国食品医薬品局）は、AI医療機器を30種類以上承認している。国内でもオリンパス（株）は、大腸の内視鏡画像をAIで解析し、医師の診断を補助する内視鏡画像診断支援ソフトウェア「EndoBRAIN®」を2019年3月から国内で販売している（2018年12月に医薬品医療機器等法の製造販売承認を取得。AI医療機器分野において国内で初めて承認を取得）。このソフトウェアは、昭和大学、名古屋大学、サイバネットシステム（株）により、AMED（日本医療研究開発機構）の支援のもと研究開発が行われた。

このようにすでにAIを搭載した医療機器は社会実装され始めている。以下では、医用画像解析の例を示す。

図1-1は国立がんセンターのグループによるディープラーニングを用いた内視鏡診断サポートシステムである¹⁾。学習用データとして、3つの画像グループを使用している。1,379のポリープ状病変の1,244枚の静止画像、ビデオからの173の連続した病変の891フレームと非癌性組織の134,983フレーム、そして、2,843のわずかに隆起した病変と陥凹した病変564の静止画像（評価用データとして、病変画像705枚、非病変画像4,135枚の結果）、アルゴリズムとしては、R-CNN（Region-based CNN）を用いている。大腸がん検診では特に平坦病変を見落としやすいが、このシステムでは平坦病変も検出できる。現在の精度は、隆起型病変の場合に感度98.0%、特異度99.0%、正診率98.8%

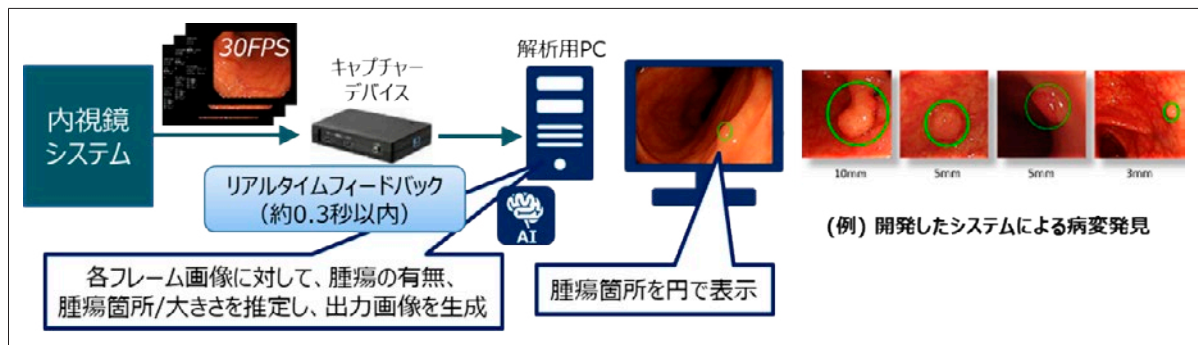


図1-1 ディープラーニングを活用したリアルタイム内視鏡診断サポートシステム

出典：浜本隆二（国立がん研究センター研究所）

である。医師主導治験の開始を予定している。

乳がん検診においてマンモグラフィーは広く採用されているが、マンモグラフィー画像の解釈は難度が高い。そこで、Googleの研究グループは、乳房X線像のスクリーニングによって乳がんを検出できるディープラーニングモデルを開発し、英国（2万5,856例）と米国（3,097例）の2種類の大規模な乳房X線像データセットを用いて、このAIシステムを評価した。偽陽性の絶対的減少（米国5.7%減、英国1.2%減）と偽陰性の絶対的減少（米国9.4%減、英国2.7%減）が起こることを明らかにした²⁾。

今般の新型コロナウイルス感染症においても、胸部レントゲン画像やCT画像から新型コロナウイルス感染症を識別するAIアルゴリズムが多く開発されている。

こういった活動を支援するため、英国における保健福祉の執行機関であるPublic Health Englandは、ケンブリッジ大学の要請に対し、匿名化したCOVID-19の感染患者データ全てを科学者たちに提供することを決めた。これらのデータは、科学者の構築したAIアルゴリズムのトレーニングに活用される。

また、通例このような画像認識には、医師が付与した診断情報をAIが学習する「教師あり学習」と呼ばれる手法が用いられてきたが、これには付与した情報の範囲を超えた分類はできないという限界がある。理化学研究所、日本医科大学の共同研究グループは、医師の診断情報が付いていない1枚あたり100億画素以上の前立腺病理画像から、複数のディープラーニングと非階層型クラスタリングを用いることで、画像上のがんの特徴を人に教わることなく自動で取得し、それを人間が理解できる情報として出力する技術の開発に成功するとともに、がんの再発の診断精度を上げる新たな特徴を見つけることに成功している²⁾。

他にも、AIを用いた際に、理由の説明が難しいという課題（ブラックボックス化）があるが、同じグループにおいて、ジョージア工科大学が開発した画像認識の判断根拠を可視化する技術Grad-CAM（Gradient-weighted Class Activation Mapping）を使用して、ニューラルネットワークがMRI画像の何を見てがんだと判断したのかをマップとして表現する成果を報告している³⁾。放射線科医と病理医が実際に診断を行い、その際に着目した箇所を12のパーツの中から選び、両者の着目した場所を比較した。ニューラルネットワークによる判断はガンの位置の特徴だけでなく周囲の組織も含めた微小特徴などを根拠としていた。これによりMRI画像にはほとんどガンが写っていない場合にも、適切な判断ができていることもあった。

こうした医用画像へのAIの適用は、今後、1) 質的診断、2) 転移予測、3) 予後予測へと発展して

いくことが期待される。

一方、AIの医療・ヘルスケア事業への適用については特有の課題がある。特に学習データを追加すると、その前後で判定結果が変わり得るAI医療機器を臨床意思決定支援に用いることに関して、どのように規制・認可するかについては、従来の規制とは異なる新しいアプローチが不可欠である。それに対し、FDAはAIを含んだ医療機器を把握・規制するための新たな枠組みの検討を開始しており、2019年4月にはArtificial Intelligence/Machine Learning (AI/ML) -Based Software as a Medical Device (SaMD) に関するディスカッションペーパーを公表し⁴⁾、Good Machine Learning Practice (GMLP) という考え方を提起している。

2020年3月現在FDAに承認されているAI医療機器については、パラメータとアルゴリズムを固定（追加の学習は認めない）して承認を受けており、AI機器が確定診断をできるわけではなく、あくまで診断補助（医師が判断の責任をもつ）として用いられている。

(2) fMRI等脳活動計測

人間の脳の活動情報（必ずしも画像ではない）を計測し、AIを用いて様々な形で応用する研究が行われている。

UCSF（University of California, San Francisco）のグループでは、てんかん患者の脳の活動に加え、唇・声帯・舌などの動きをAIに学習させて、実際の音声を使用せずに音声を合成することに成功している⁵⁾。てんかん治療のために頭蓋内モニタリングを受けた5人の参加者が数百の文章を声に出して話す時の脳活動（ECoG）信号を記録、次に発話音声から声道がいかにか動いているかを特徴づける声道パラメータ（喉頭や舌先などの位置情報）を推定するためのモデルを用意し、RNNの一種である長短期記憶（LSTM）ネットワークを用いて両者を学習させた。脳神経疾患などにより、言葉が話すことができない人が、声（合成音声）でリアルタイムにコミュニケーションを取ることができるようになることが期待されている。

NICT（情報通信研究機構）、東京大学、自治医科大学のグループでは、脳波からてんかん発作を自動検出できるAIの開発に成功している⁶⁾。脳波データ（1124.3時間の脳波データのなかに、97回のてんかん発作が含まれており、それらの発作の持続時間は6,950秒）を脳波計の画面に出力されるよ

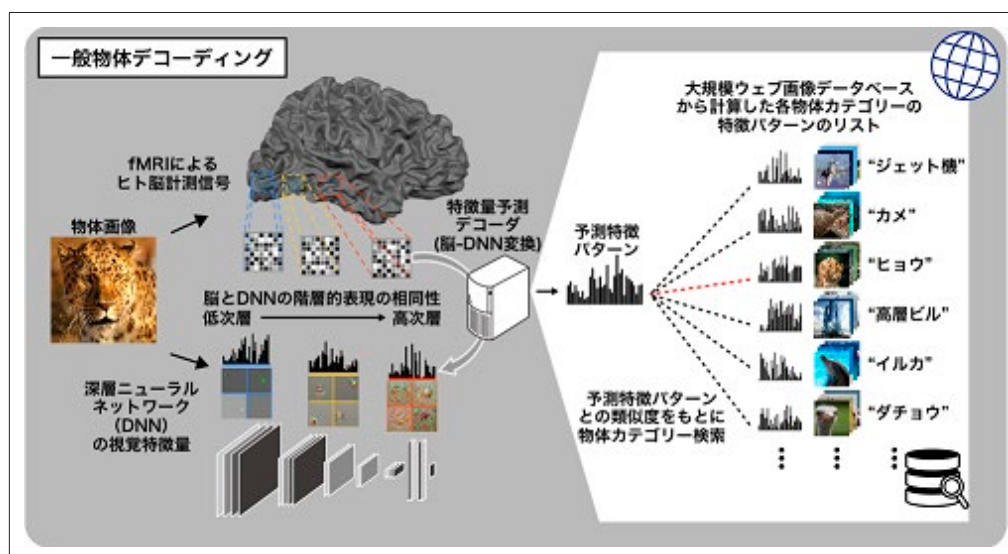


図1-2 人間の脳活動情報からイメージの再構築

出典：神谷之康（京都大学）

うな画像に変換し、これらの膨大な量の画像をCNNに学習させることで、てんかん発作を高い精度で検出できる。

fMRI等により計測されるヒトの脳活動情報を機械学習によるパターン認識で解析し、人間の思考を解読する技術はブレインデコーディングと呼ばれている。京都大学では、fMRIによって視覚化した脳の活動記録（入力）から、人間が実際に頭に思い浮かべたイメージを機械学習を用いて画像に出力することに成功している⁷⁾。様々な物体画像の視覚特徴脳活動パターンを学習させ、特徴パターンと大規模画像群のデータベースから、モデルの学習に用いていない画像カテゴリーを含む任意のカテゴリー情報を脳活動から予測する技術の開発を行った（図1-2）。

1.2 イメージングデータ（生命科学）

同一の対象・サンプルに対して、イメージング解析と次世代シーケンサーや質量分析器を用いたオミックス解析との知見をつなげるには、スループットが低いと言われているイメージング技術のスループットを上げていくことによって要素の多様さに対応できるような全網羅、半網羅的な手法が必要になっている。

電子顕微鏡や光学顕微鏡像に対しても、CNN及びGANを用いた画像変換が威力を発揮する。

（1）電子顕微鏡

電子顕微鏡で近年大きく進展している手法の一つは、走査電子顕微鏡（SEM）の連続断面観察法による電顕画像からの三次元再構成技術であり、ここにディープラーニングが使われている。サンプル（細胞や組織片、臓器等）をイオンビームで削りながら二次元画像（3nm前後のスライス）を連続的に得て、AIを用いて、各々の画像を重ね合わせることによって3次元画像とする技術である（図1-3）。

これにより細胞生物学の分野では、細胞内のオルガネラのコンタクトやネットワークが見えてきており、神経科学の分野では、脳全体の神経の結合（ネットワーク）を見ようというコネクトミクス研究に新たな潮流が生まれている。より広くより正確な情報を取得するということで扱うデータの量が

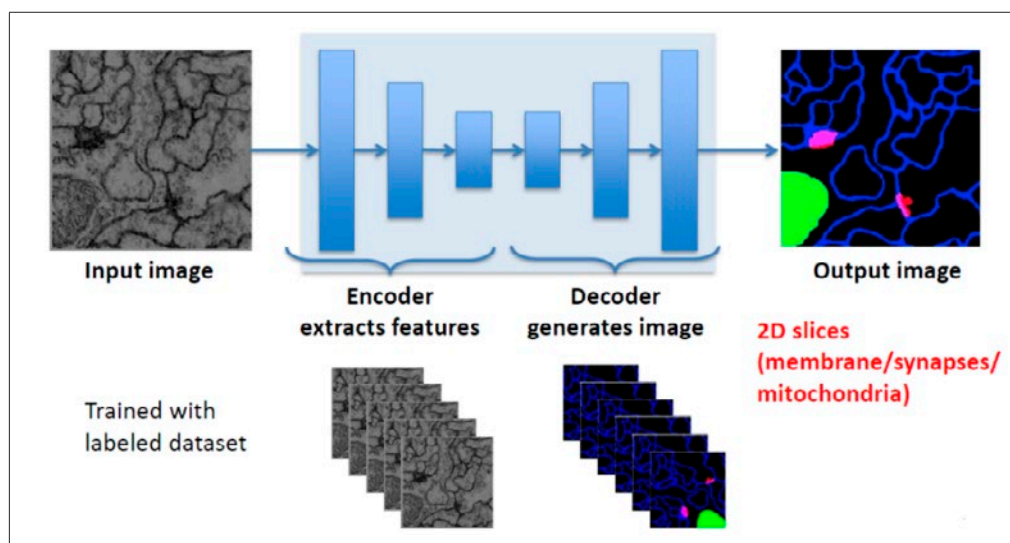


図1-3 電顕画像の畳み込みニューラルネットワーク

出典：石井信（京都大学）

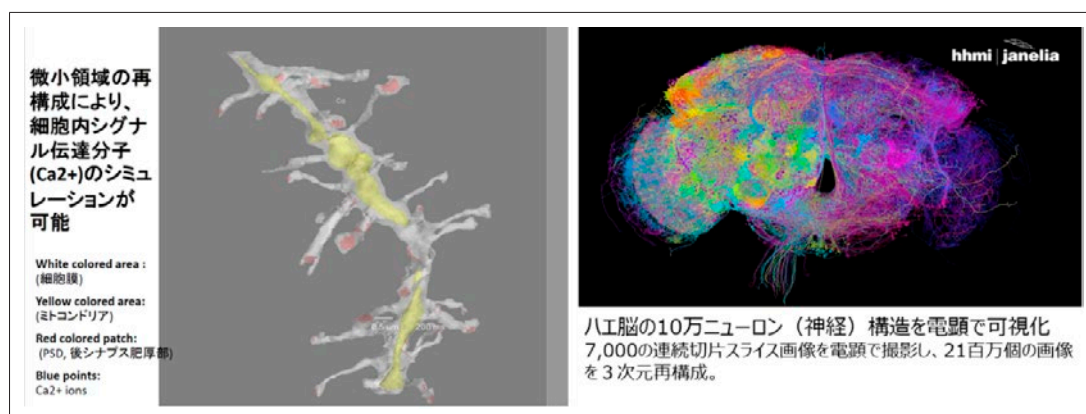


図1-4 神経細胞、ハエ脳全体の電顕3次元再構成モデル⁸⁻⁹⁾
出典：(左) 石井信（京都大学）、(右) 米国ジャネリア研究所

ペタバイトクラスになっており、従来の研究のスタイルから変化が求められている。

こうした手法を用いて、単一の神経細胞の三次元像が作成されている⁸⁾ 他、米国ジャネリア研究所では、ハエの脳全体の電子顕微鏡像を公開している⁹⁾ (図1-4)。

三次元再構成技術により、将来的に細胞や組織における空間解像度の高いシミュレーションを行うことが可能になることが期待されている。一方で、電子顕微鏡像は基本的にはサンプルが死んだ状態での静止画像なので、ダイナミクス（時間）を観るために、光学顕微鏡像との統合解析が模索されており、相関顕微鏡の開発も行われている。

(2) 光学顕微鏡

光学顕微鏡で高い分解能で大きいものを見ようとすると非常に時間を要する。空間分解能と撮像に要する時間とのトレードオフ関係が明確で、大きい対象を撮るには何十時間もかかる。

UCLA等のグループでは、この高速化のために、ディープラーニング（GAN + U-net）を使って、1分子ベースの超解像顕微鏡の撮像時間を従来の20倍速くすることに成功している¹⁰⁾。学習データ用に取得した低解像度画像をGANを使用して高解像度画像に変換している。

これまで1枚の撮像に分単位を要していたが、秒単位を可能にする超解像イメージングが行われ、30ナノメートル以下の分解能でダイナミクスを見る時代が到来しつつある。こうした技術はディープラーニング超解像と呼ばれ、2018年に多くの論文が出ている。このような生成モデルを用いた超解像画像は推論した画像であることから、どの程度正確かという評価手法も今後必要になってくる。

一方で、2次元像の積層（＝3次元）の積み重ねが3次元ダイナミクス（＝4次元）であるため、4次元画像の作成には現実的には非常に時間がかかる。今、4次元画像の取得を高速化する方向として、3次元の一発撮像から、それを積み重ねる高速化が模索されている。この場合、3次元のボクセルの数に対し、カメラ撮像のピクセルの数の桁の違い（少ない問題）をいかにマッチングさせるかという問題がある。

そのための手法のひとつとして、3次元計測が可能なライトフィールド顕微鏡があるが、3次元画像が高速で取得できる反面、空間分解能や画質の面で制約があった。2017年、脳の神経活動の撮像において、計算機的手法による画像処理により脳の散乱の影響を取り除ききれいにシングルニューロンが見えたという論文が、ロックフェラー大学のグループから出ている¹¹⁾ が、今後、機械学習の適用も期待される。

もう一つの例は、ディープラーニングの手法を使って、構造を予測する取組である。例えば、透過光顕微鏡を用いた明視野の画像情報から、その明視野画像に、あたかも蛍光顕微鏡で見たような色付けをするという方法がある。

米国アレン研究所では、3次元透過光（明視野および微分干渉観察DIC）といくつかの主要な細胞内構造（細胞膜、核膜、核小体、DNA、ミトコンドリアなど）に対応する蛍光生細胞画像との関係を学習させることによって、明視野画像情報から色付けをした3次元画像を再構成す

ることに取り組んでいる¹²⁾(図1-5)。電子顕微鏡で撮った画像から構造情報を予測することもできる。同じ頃にGoogleの研究者がCellに同様の論文を出している¹³⁾。

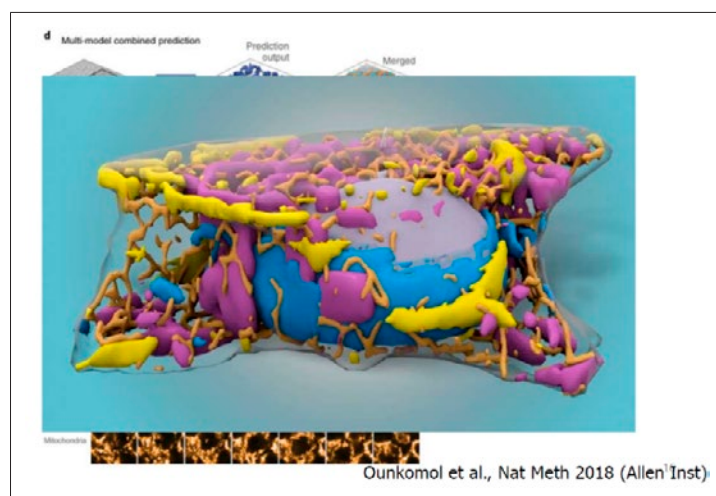


図1-5 光学顕微鏡のラベルフリー3次元再構成¹²⁾

他にも細胞の高速イメージング（共焦点顕微鏡）とディープラーニングを用いた画像解析で細胞を一つ一つ網羅的に高速識別し、その解析結果に応じて所望の細胞を分取する技術¹³⁾が発表されている。同様に、培養細胞（例えばiPS細胞）の評価にAIを用いて、治療に適した細胞の選定や品質管理を行う取組みも多い。

画像を用いた技術の今後の展望としては、画像特徴量の網羅的・定量的抽出（画像ベースのプロファイリングの強化）を行うことが考えられる。電子カルテの自然言語処理は主観が入るため、解析（情報処理）が難しい面があるが、画像は客観的に数値化することができる。例えば、図の画像について臨床医が言葉で表現すると、不整形の腫瘍性病変、分葉状、多房性で前方に充満性進展などとなるが、実際には言葉で遺漏なく表すことは難しい。しかし、コンピュータを用いて特徴を数値化することにより、経時的变化を網羅的かつ客観的に処理することができる。

CT、MRIなどの放射線医用画像は、腫瘍全体を非侵襲的に、反復して解析することが可能な手段であり、日々、膨大なデジタルデータが蓄積されている。例えば、図1-6は脳腫瘍のMRI画像であるが、ヒストグラム、形状、テクスチャー、周波数成分などの画像特徴量を網羅的に定量化することで、臨床医には識別困難な予後や治療効果、遺伝子型を推定できる可能性がある。

一例として、大阪大学では、これまで手術によって腫瘍の一部を採取する必要があったところを、脳腫瘍のMRIの特徴を抽出することで遺伝子変異の有無を予測する技術を開発した¹⁵⁾。悪性脳腫瘍である199例の神経膠腫のMRIと遺伝子情報を11施設から集めデータベースを作成し、CNNを画像の特徴を抽出するフィルターとして用いることで脳腫瘍のMRIの特徴を抽出し遺伝子変異の有無を63.1%の精度で推定することに成功した。

今後電子カルテ等情報の自然言語処理と相まって、個別化・精密化された医療体系の実現に貢献することが可能となる。

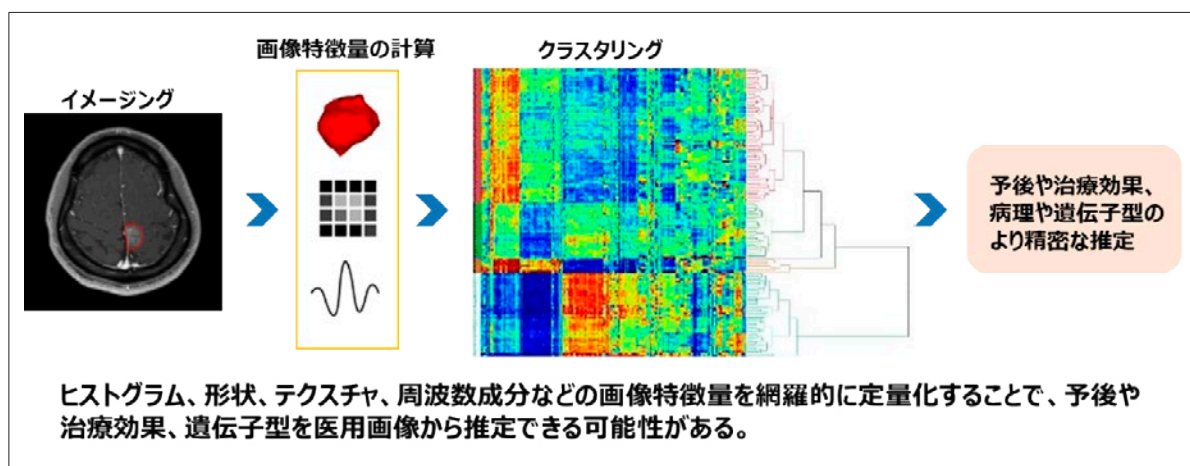


図1-6 医用画像特徴量の網羅的・定量的抽出
出典：浜本隆二（国立がん研究センター研究所）

【コラム①】 コンピューテーショナル・イメージング

レンズで結んだ像をカメラ（2次元検出器）で撮像するのではなく、試料からの散乱光を何らかの形で検出し、その検出結果から物理モデルを逆演算して画像を再構成する方法（機械学習技術ではない）が提案されており、コンピューテーショナル・イメージングと呼ばれている。

例えば、CMOSチップ上に試料を置いて様々な角度から照明することにより試料の散乱情報を取得し、レンズフリーで3次元画像を再構成するというアプローチに、UCLAのグループ等が取り組んできた。結像光学系を使わないので、システムの小型化・広視野化が達成できる。

本手法のような計算機的手法を組み合わせたいメージングは、機械学習と相性が良く、フローサイトメーター、細胞識別分取（セルソーター）や細胞培養などに応用されている。東京大学のグループからは、圧縮センシングを用いた圧縮計測信号を直接機械学習で超高速リアルタイム解析（>万枚/秒）し、「画像を見ずに形を見る」超高速・高精度な光イメージ認識型のリアルタイム細胞分離技術¹⁾が報告されている（図1-7）。

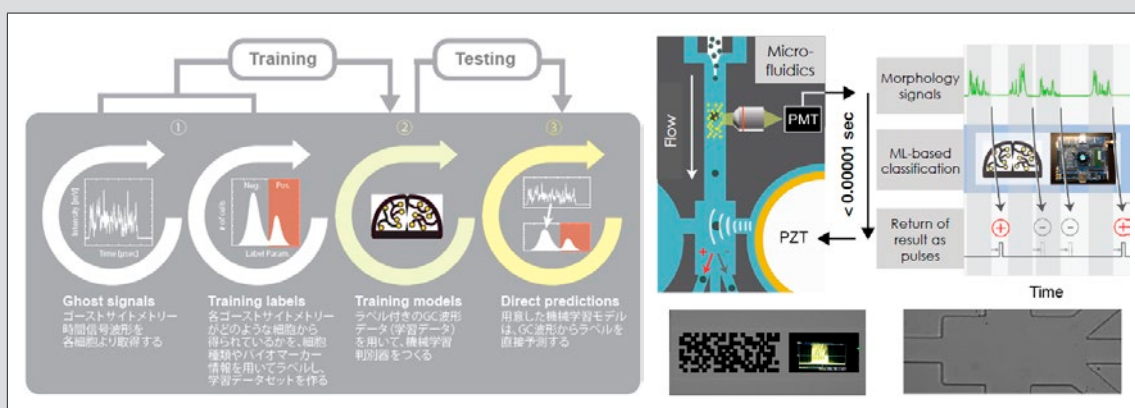


図1-7 機械学習が駆動するイメージ無しイメージセルソーター
出典：太田禎生（東京大学）

参考文献

- 1) Sadao Ota et al., Ghost cytometry Science 15 Jun 2018

【コラム②】農業におけるAI（画像データ）の応用

農業においては以下のようなAI活用事例が見られ、中でも作物と圃場について画像データの解析が用いられている。

- ・作物と土壌の監視：コンピュータビジョンとディープラーニングアルゴリズムを活用して、無人偵察機やセンサでキャプチャされたデータを処理し、作物と土壌の健康を監視。
- ・農業ロボット：自律ロボットのAI制御により、人間の労働者よりも大量かつ速いペースで作物を収穫するなどの農業タスクを処理。
- ・予測分析：気象変化などの作物収量に対するさまざまな環境影響を追跡および予測。

農研機構（農業・食品産業技術総合研究機構）のグループでは、ドローンとAIを利用したスマート育種評価法の開発を報告している¹⁾。これまで育種家が畑を2時間以上歩いて肉眼観察で牧草を一株ずつ評価していた作業を、ドローンで撮影した画像から、あらかじめ学習させておいたAIが5分程度で行えるようになる。

また、農作物の病気診断時に、その根拠となる画像の特徴を表示するAIを開発している²⁾。健全な葉の画像と病気の葉の画像をそれぞれ学習させ、健全、病気、共通部分という3つの領域の特徴を抽出する。そして、オートエンコーダーを用いて病気診断の根拠となる特徴を可視化することができる。

参考文献

- 1) 秋山 征夫 et al., 無人航空機の空撮画像を利用した深層学習によるオーチャードグラス個体選抜法の開発. 育種学研究.
- 2) Harshana Habaragamuwa et al., Plant Disease Identification using Explainable Features with Deep Convolutional Neural Network, 2019 International Joint Conference on JSAM and SASJ, and CIGR VI

参考文献

- 1) Masayoshi Yamada et al., Development of a real-time endoscopic image diagnosis support system using deep learning technology in colonoscopy *Scientific Reports* volume 9, Article number : 14465 (2019)
- 2) Scott Mayer McKinney et al., International evaluation of an AI system for breast cancer screening *Nature* volume 577, pages89–94 (2020)
- 3) Yoichiro Yamamoto et al., Automated acquisition of explainable knowledge from unannotated histopathology images *Nature Communications* volume 10, Article number : 5642 (2019)
- 4) Jun Akatsuka et al., Illuminating Clues of Cancer Buried in Prostate MR Image : Deep Learning and Expert Approaches *Biomolecules* 2019, 9 (11) , 673
- 5) Gopala K Anumanchipalli et al., Speech synthesis from neural decoding of spoken sentences *Nature* volume 568, pages493–498 (2019)
- 6) Ali Emami et al., Seizure detection by convolutional neural network-based analysis of scalp electroencephalography plot images *Neuroimage Clin.* 2019 ; 22 : 101684.
- 7) Guohua Shen et al., Deep image reconstruction from human brain activity *PLoS Comput Biol.* 2019 Jan 14 ; 15 (1)
- 8) Hidetoshi Urakubo et al., UNI-EM : An Environment for Deep Neural Network-Based Automated Segmentation of Neuronal Electron Microscopic Images *Scientific Reports* volume 9, Article number : 19413 (2019)

- 9) Zhihao Zheng et al., A Complete Electron Microscopy Volume of the Brain of Adult *Drosophila melanogaster* Zheng et al., 2018, *Cell* 174, 730–743
- 10) Hongda Wang et al., Deep learning enables cross-modality super-resolution in fluorescence microscopy *Nature Methods* volume 16, pages103–110 (2019)
- 11) Tobias Nöbauer et al., Video rate volumetric Ca²⁺ imaging across cortex using seeded iterative demixing (SID) microscopy *Nature Methods* volume 14, pages811–818 (2017)
- 12) Eric M. Christiansen et al., In Silico Labeling : Predicting Fluorescent Labels in Unlabeled Images *Cell* VOLUME 173, ISSUE 3, P792-803.E19, APRIL 19, 2018
- 13) Chawin Ounkomol et al., Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy. *Nat Methods*. 2018 Nov ; 15 (11) : 917-920.
- 14) Nao Nitta et al., Intelligent Image-Activated Cell Sorting *Cell* Volume 175, Issue 1, 20 September 2018, Pages 266-276.e13
- 15) Ryohei Fukuma et al., Prediction of IDH and TERT promoter mutations in low-grade glioma from magnetic resonance images using a convolutional neural network *Scientific Reports* volume 9, Article number : 20311 (2019)

2 分子・化合物データ (in silico 創薬)

ここでは、創薬に関係する分子インフォマティクスとアミノ酸配列からのタンパク質の構造予測を取り上げる。ヒトを構成するタンパク質は数万～十数万種類存在するといわれており、実際に上市されている医薬品の標的分子（タンパク質）数は、300～500程度でしかない。これに対し、薬剤候補となる化合物の数は 10^{23} ～ 10^{60} ほどあると試算されており、現在までに実際に合成されたのは 10^8 程度といわれている。

2.1 分子インフォマティクス

従来から創薬のターゲット（多くはタンパク質）に対する候補分子を探索するため、タンパク質と低分子化合物とのドッキングシミュレーションがバーチャルで行われてきた。ケモインフォマティクスという概念である。従来は計算化学ベースのモデル生成（シミュレーション）が主であったが、近年機械学習ベースのモデル生成のウェイトが大きくなってきている。

化合物に対して機械学習を適用することを考えた場合、コンピュータに対して、化合物をどのように表現して入力するかということが問題となる。従来、特徴量設計において、分子の特徴ベクトルは molecular fingerprints と呼ばれ、化合物の特徴的な一部分（フラグメント）がその分子にあるかどうかを0/1で表したbitを並べて作られた、多くの種類の fingerprints が提案されてきた。

2017年、Variational Auto Encoder (VAE) に基づく化合物自動生成技術が現れたことで^{1, 2)}、分子設計にディープラーニングが持ち込まれ、「分子インフォマティクス」という形で再興した。SMILES 表記（分子の化学構造を英数字で文字列化する表記法）から連続潜在変数を取り出すエンコーダと、潜在変数から SMILES 表記を復元するデコーダを用意し学習を行い、予測器は分子の潜在空間の連続的なベクトル表現から目的の物性をもつ分子を生成する（図2-1）。（SMILES空間を表す文字列の確率分布が出るので、その確率分布を学習し、ベイズ最適化で最適化することによって生成ができるようになっている）。

これまでの化合物自動生成法は、あらかじめ定義されたフラグメントを組合せるものが多かったが、こうした手法により de novo 生成が出来るようになったことで、探索可能な空間が広がった。

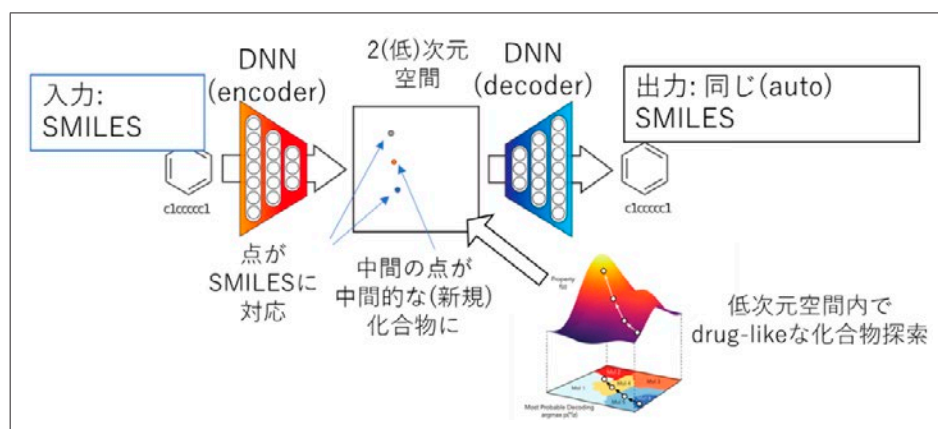


図 2-1 Variational Auto Encoder (VAE) に基づく化合物自動生成

出典：津田宏治（東京大学／理研）

関連して、IBMチューリッヒ研究所では、SMILESと機械翻訳にも使われているリカレントニューラルネットワーク（RNN）を用いた有機化学反応予測（反応物から生成物を予測）の試みを行った³⁾。さらに彼らはMolecular Transformerという、データセット内に存在する反応物、試薬、および生成物における化学モチーフの存在と非存在の間の相関を推測することによって予測を行う別のモデルを開発した⁴⁾。微妙で選択的な化学変換を正確に予測し、正しい化学選択性、位置選択性、およびある程度の立体選択性を得ることができる他、不確実性も推定できる。こうした取組はIBM RXN for chemistryという形で公開され無償で使うことができるようになっている。

他にも、ケンブリッジ大学のグループでは、化学反応を、反応物候補から実際に使う反応物を選択するタスクと前述のような反応物から生成物を予測するタスクに分解し、前者をMolecule Chefと名付け、新しい分子生成モデルを提案している⁴⁾。

また、上述したVAE等によって提示された百単位の候補分子群のどれをどのように合成するかのプロセスが研究開発の律速になる。従来の逆合成経路探索はルールベースの手法（エキスパートシステム）であり、新分子・反応機構が予測できず汎化性能が低かったが、近年、記号とパターンを組み合わせた手法で進展があった。ドイツの大学では、AlphaGoと同じようにモンテカルロ木探索（MCTS）とシンボリックAIを用いた生成化合物の合成手順を予測する試みも行われている⁵⁾。これまでに知られている一段階有機化学反応のほとんど全て約1,240万をAIに学習させた結果、2015年に報告された薬物候補合成の中間体の6段階のルートの例は、アルゴリズムによって5.4秒で発見された。

こうした技術の誕生により、数多くの創薬AIベンチャーが誕生し、多くの世界的な大企業が投資を行っている。

2019年に入り、AIを用いた創薬（候補化合物発見から前臨床までの実証）に関する論文が2報相

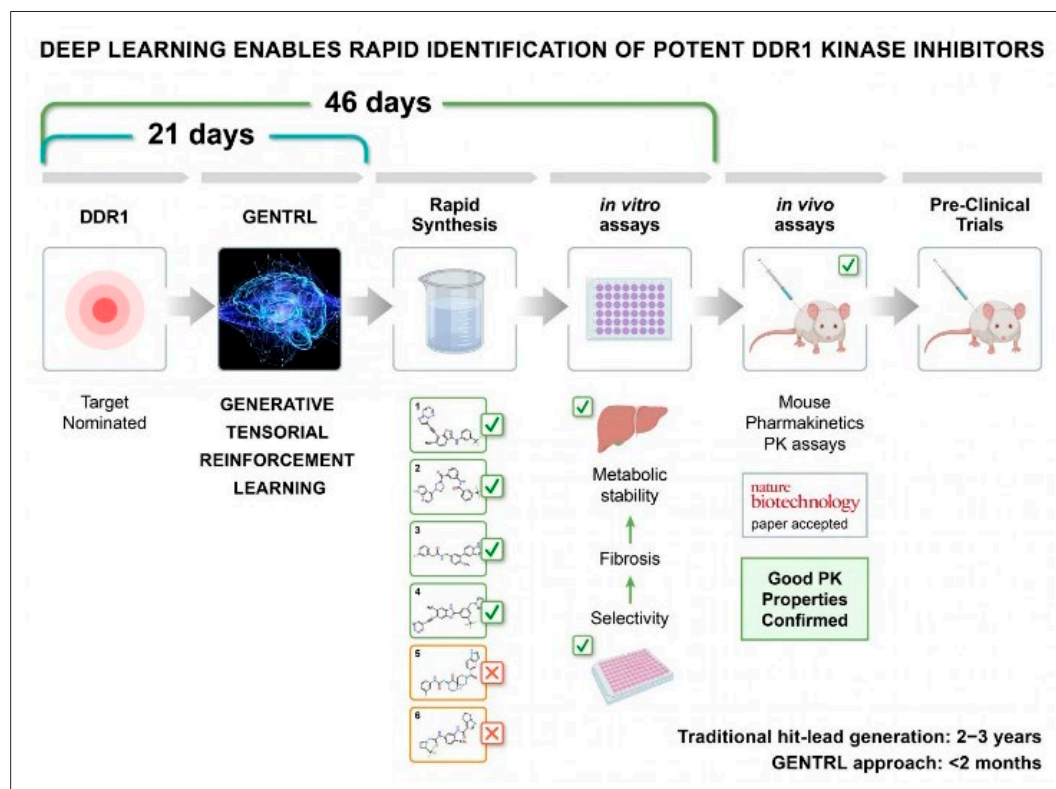


図2-2 AI創薬概念図⁶⁾

次いで報告された。

AI創薬ベンチャーのInsilico Medicine社（米国）では、新規の低分子設計のために、生成テンソル強化学習（GENTRL）を開発した（図2-2）。GENTRLは強化学習、変分近似、および、テンソル分解を融合し、自己組織化マップ（SOM）に基づく評価を組み込んだ機械学習アルゴリズムであり、これにより、合成の実現可能性、新規性、および生理活性の最適化をすることができる。2019年、これを使用して、線維化やその他の疾患に関係するキナーゼターゲットであるディスコイジンドメイン受容体1（DDR1）の強力な阻害剤を20日間程で発見したとしている⁶⁾。そのうち4つの化合物が生化学アッセイで有効で、2つの化合物が細胞ベースのアッセイで検証された。1つのリード候補がテストされ、マウスで好ましい薬物動態が示された。実験開始から2ヶ月でここまでのプロセスを終えたとしている。

MIT（マサチューセッツ工科大学）の研究者は、複数の化学ライブラリーで予測を行い、新規な抗生物質となる化合物（ハリシン）を発見した⁷⁾。この分子は、従来の抗生物質とは構造的に異なり、マウス実験において、結核菌やカルバペネム耐性腸内細菌を含む幅広い病原体に対する殺菌活性を示した。従来の抗生物質は細菌の細胞壁の構造を標的とするか、感染のための酵素またはタンパク質合成活性を妨げるものがあるが、この新しい分子は細胞膜のプロトンポンプをブロックすることを目的としており、機序が異なることが示された。プロセスとしては、化合物のグラフ表現と抗菌活性を紐付けるニューラルネットワーク（Message Passing Neural Network）を用いて、大腸菌を殺すのに効果的であることが知られている約2,500の化学構造で最初に訓練され、これらのデータのパターンと共通性を特定した後、Broad InstituteのDrug Repurposing Hubにある約6,000の化合物ライブラリーを用いてスクリーニングし、他の既知の抗生物質とは異なる構造を持つ潜在的なリード化合物を特定した。より大きなライブラリーとしてUCSFが公開しているZINC15データベースからは、1億個を超える分子群の中から既知の抗生物質から構造的に遠い8つの化合物を発見したとしている。

時を同じくして企業側での発表も始まった。

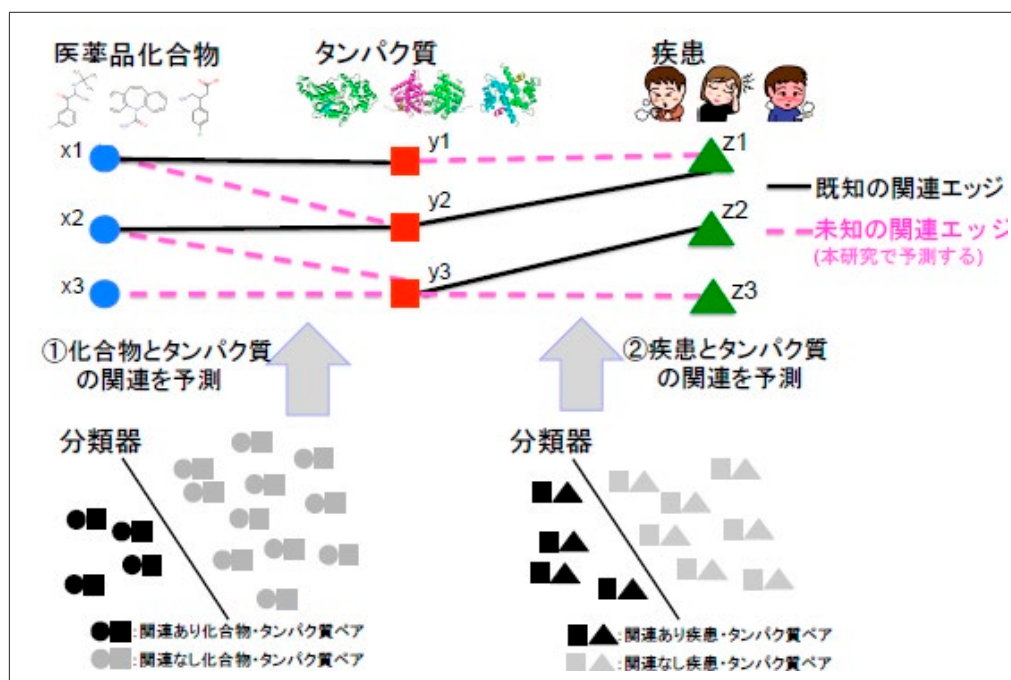


図2-3 ドラッグリポジショニングの概念図

出典：山西芳裕（九州工業大学）

2020年1月には、大日本住友製薬とAI創薬ベンチャー Exscientia社（英国）でも、AIを活用して創製された新薬候補化合物のフェーズ1試験を開始すると発表している⁸⁾。強迫性障害に対するGタンパク質共役受容体（GPCR）をターゲットとした化合物の発見で、平均で4年半を要するとされる探索研究を12ヶ月未満で完了したとしている。

同時期に、がん治療向けウイルスベース免疫治療の設計開発を手掛けるバイオテクノロジー企業であるTransgene社とNECは、AIによって開発されたがんワクチンの治験を開始したと発表した⁹⁾。ウイルスベクターを基にした、固形がんに対する個別化免疫療法のプラットフォームにおいて、NECのネオアンチゲン予測システムで特定・選択された患者固有の変異をコードし、ウイルスベース免疫治療を非常に短期間で始めることを可能ととしている。

既に承認された化合物の新しい効能を発見し、別の疾患の治療薬として開発するドラッグリポジショニングと相性が良い（図2-3）。

今般の新型コロナウイルス感染症（COVID-19）でも早い段階で複数の研究成果が発表されている。一例として、韓国のグループでは、2019-nCoVのウイルスタンパク質に作用する市販の薬物として、抗レトロウイルス薬であるアタザナビルが最も有望な化合物であることを示した¹⁰⁾。事前トレーニング済みのディープラーニングベースの薬物標的相互作用モデルを使用して、ウイルスタンパク質に作用する可能性のある薬物を特定している。BenevolentAI社（英国）は、機械学習によって科学文献から抽出された多数の接続（ナレッジグラフ）を構築している。ここから、関節リウマチ治療薬のバリシチニブがウイルス感染プロセスをブロックする可能性のある治療薬となる可能性を予測する研究が報告されている¹¹⁾。

一方で、AIを用いた研究の評価・検証の問題が課題となる。第三者が各論文などで使用されたアルゴリズム間の相互比較を可能にすることやトレーニングデータセットの公開など、結果をより適切に公開するためのガイドラインが求められている。

2.2 タンパク質の3次元構造予測

アミノ酸配列からタンパク質の3D構造を予測するという取組は古くから行われてきた（図2-4）。その一環として、1994年から「タンパク質構造予測コンペティション（CASP: (Critical Assessment of Structure Prediction)）」が行われている。これが実現するとX線結晶学などの既存の技術よりも安価で高速であり、極論的には放射光施設やクライオ電子顕微鏡で計測しなくともタンパク質の結晶構造が予測できることになる。

タンパク質の折りたたみ（タンパク質が特定の3次元構造に折りたたまれる現象）は、構成するアミノ酸間の相互作用の数が膨大になるため、対応する遺伝子配列だけから決定することは難しいとされてきたが、2018年にGoogleのAI企業であるDeepMind社がAlphaFoldと呼ばれるアルゴリズムを発表した¹²⁾。他、2019年以降も複数のチームから新しいアルゴリズムが発表されている。アミノ酸同士の距離や構成元素の化学結合の間の角度、エネルギー的に安定な配置などが入力変数として用いられている。

AlphaFoldの予測は正しい構造から中央値で6.6オングストローム誤差があり、クライオ電子顕微鏡の解像度である2オングストロームにはまだ及ばない。薬物設計（化合物がタンパク質に結合する機構の解明）や、酵素の触媒メカニズムの解明などには、2～3オングストロームの分解能を必要とするが、今後の精度の向上により、薬への応用の他、変異タンパク質が疾患にどのように寄与するかを理解したりすることができるようになると期待されている。

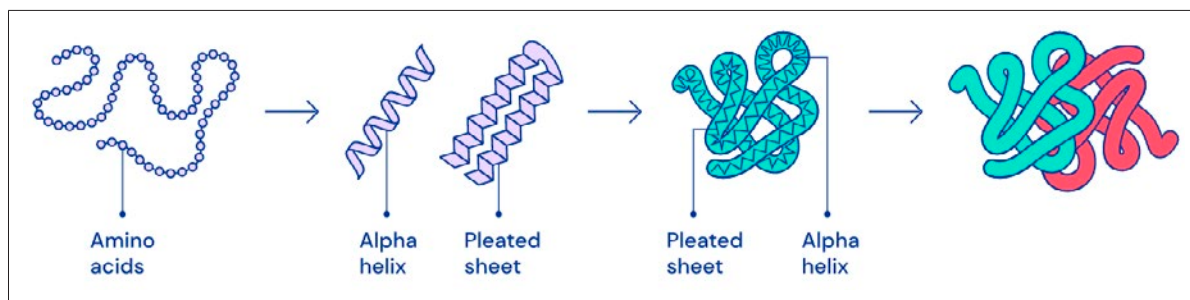


図 2-4 アミノ酸配列からタンパク質の複雑な3次元構造の予測
出典：DeepMind 社

上記の分子インフォマティクス（化合物自動生成）と組み合わせると、将来的に完全なバーチャルベースで、タンパク質と未知な化合物の膨大な組合せをコンピュータ上でドッキングシミュレーションを行うことも可能となる。

2.3 薬の毒性・副作用予測

創薬シーズの体内での吸収、分布、代謝、排泄、および毒性（ADMET）といった薬物動態について、初期の段階で評価を行うことは、そのシーズの実用化の可能性を評価する意味で重要であり、従来より薬物候補分子の分子構造から、ADMETを推測する研究が行われてきた。

一例として、AIベンチャーのAtomwise社（米国）のAtomNetは、標的タンパク質と薬物候補となるリガンド（化合物）が様々な化学構造から構成されていることをChEMBL（低分子化合物データベース）およびPDB（タンパク質構造のデータベース）からの数百万のデータをディープラーニングを用いて学習し、選択した標的タンパク質に作用する、毒性のない分子を探索するシステムである¹³⁾。人間の体内でどのように振る舞うか、薬剤としての効果、毒性と副作用などが予測される。

細胞、臓器、またはモデル動物を使用した前臨床で検証された薬物の80%以上は、予想しない臨床結果のために臨床試験で失敗しているといわれ、創薬の莫大な開発費の元凶ともなっている。そこで、ヒトでの臨床結果を予測するための機械学習アルゴリズムの応用もトレンドとなっている。

一例として、日本のATR（国際電気通信基礎技術研究所）のグループでは、被験薬のヒトでの副作用・有害事象および効能を、薬の種類を問わずに予測できるAIシステムの開発を行った¹⁴⁾。被験薬を投与したマウスの全身網羅的なトランスクリプトームの変動のみを指標とし、被験薬の構造、標的などの情報を必要としないシステムとなっている。

ここで挙げてきたような従来型の標的分子（受容体・酵素・イオンチャネル・トランスポーター）と鍵と鍵穴の関係で結合しやすい分子量500以下の低分子化合物の創薬モダリティには限界が来ているとも言われている。新しい標的分子探索には、疾患との関係でゲノミクスやケミカルゲノミクスが用いられている。新しい標的として、タンパク質間相互作用、タンパク質-核酸相互作用等が注目されている。タンパク質間相互作用に関与するタンパク質数は、数万のオーダーで存在すると報告されている。今後、こうした中分子医薬や抗体医薬などにもAIを用いた設計、探索が展開されていくことが期待されている。

参考文献

- 1) Gómez-Bombarelli et al. “Automatic chemical design using a data-driven continuous representation of molecules,” ACS Cent. Sci. 2018, 4, 2, 268–276
- 2) Kusner et al., “Grammar Variational Autoencoder,” ICML, 2017.
- 3) Philippe Schwaller et al., “Found in Translation” : predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models Chem. Sci., 2018, 9, 6091–6098
- 4) Molecular Transformer : A Model for Uncertainty-Calibrated Chemical Reaction Prediction ACS Cent. Sci. 2019, 5, 9, 1572–1583
- 5) John Bradshaw et al., A Model to Search for Synthesizable Molecules arXiv : 1906.05221
- 6) MHS Segler et al., Planning chemical syntheses with deep neural networks and symbolic AI. Nature 555, 604–610 (2018)
- 7) Alex Zhavoronkov et al., Deep learning enables rapid identification of potent DDR1 kinase inhibitors Nature Biotechnology volume 37, pages1038–1040 (2019)
- 8) Jonathan M. Stokes et al., A Deep Learning Approach to Antibiotic Discovery Cell Volume 180, Issue 4, 20 February 2020, Pages 688-702.e13
- 9) 大日本住友製薬とExscientia Ltd.の共同研究 人工知能（AI）を活用して創製された新薬候補化合物のフェーズ1試験を開始（2020年1月30日）
- 10) Transgene社とNEC、AIによって開発されたがんワクチンTG4050の2つの治験を卵巣がんと頭頸部がんで開始（2020年1月7日）
- 11) Peter Richardson et al., Baricitinib as potential treatment for 2019-nCoV acute respiratory disease Lancet. 2020 Feb 15 ; 395 (10223) : e30-e31
- 12) Bo Ram Beck et al., Predicting Commercially Available Antiviral Drugs That May Act on the Novel Coronavirus (SARS-CoV-2) Through a Drug-Target Interaction Deep Learning Model Comput Struct Biotechnol J. 2020 Mar 30 ; 18 : 784-790.
- 13) Andrew W. Senior et al., Improved protein structure prediction using potentials from deep learningNature volume 577, pages706–710 (2020)
- 14) Izhar Wallach et al., AtomNet : A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery arXiv : 1510.02855
- 15) Satoshi Kozawa et al.,Predicting Human Clinical Outcomes Using Mouse Multi – Organ Transcriptome iscience Volume 23, Issue 2, 21 February 2020, 100791

3 ゲノム・オミックスデータ

画像データ解析へのディープラーニングの適用が進んだ一方で、ゲノム、オミックスなど生体内の網羅的な分子・細胞データへの適用はどうであろうか。一細胞遺伝子発現プラットフォームの開発により、数十万個、さらには数百万個の個々の細胞からトランスクリプトーム情報全体を取得することはほぼ日常的になりつつある。次世代シーケンサーなどの計測機器から産出されるデータが膨大になるという意味では、こちらの方がデータ科学手法適用の期待が大きい。

この分野の機械学習の適用については、血液などからのゲノムや分子情報データを用いたヒト（患者）の層別化や疾患の予測といったことが行われている。これはバイオマーカーの発見と密接な関係にある。バイオマーカーとは、疾患特異的な分子などの指標のことで、疾患を特徴付けたり、疾患の発見や進行・経過をフォローしたりするのに用いられる血液や尿中の物質（群）を指す。

3.1 ゲノムのSNPs等解析

従来より、次世代シーケンサーを用いた大規模ゲノム解析のためのインフォマティクスは、バイオインフォマティクスの中心的存在であった。現在の技術をもってしても、シーケンサーで解読したヒト全ゲノムの約35億塩基もの配列から、遺伝子発現に関する情報を正確に解析することは比較的時間とコストを要する。

そこで、遺伝子間の関係を探求するため大量の遺伝子発現データを、画像のようなものに変換し、畳み込みニューラルネットワーク（CNN）に持ち込むことが行われ始めている。非画像データから画像データへの変換を行うことで、非画像データの解析に対してもディープラーニングの適用が可能になる。

Google社は、ディープラーニングを使用して、ゲノムシーケンスを従来の方法よりも大幅に高い精度で再構築し、ゲノムの変異（DNA配列の小さな変化）を検出するために、DNA鎖をコンピューターが認識できる画像に変換し、次に、参照ゲノムとの比較から、変異が知られているDNA SNPs情報を学習した¹⁾。2017年末に、「DeepVariant」として公開している。

理化学研究所等の国際共同研究グループは、遺伝子発現などのゲノミクスデータを使って診断や予測のためのクラス同定や分類を高精度に行うために、適切な変数（画素）の配置、特徴抽出、そして適切な分類モデルの構築、という3つのステップを行う「DeepInsight法」を開発した²⁾。この手法をがん遺伝子発現などの実データに適用した結果、ランダムフォレスト法などの機械学習よりもはるかに高精度で分類できることがわかったとしている。

3.2 表現型の層別化

ゲノムのような高次元データに対して機械学習を適用することによって、潜在変数の数百次元に落としながら生体機能の面から分類を行うことができる。

一例として、ヒトには多数の種類のHLA（ヒト白血球抗原）遺伝子が存在し、各遺伝子が数十種類以上の配列パターンを有するため、白血球の血液型の組み合わせが膨大な数になってしまう。大阪大学のグループでは、これらのゲノム配列の膨大な組み合わせを効率的に分類する目的で、得られたHLA遺伝子ゲノム配列情報に対して、高次元データの可視化に適している次元削減アルゴリズムt-SNE（T-distributed Stochastic Neighbor Embedding）という（ディープラーニングではないが）機

機械学習手法の一つを適用している（図3-1）。これにより、日本人集団の白血球の血液型が11パターンの組み合わせで構成されることを解明し、このような白血球の血液型の個人差が、病気や量的形質を含む50以上の表現型に関わることを示した³⁾。

基礎的なところでは、国際的なヒューマンセルアトラス（HCA）プログラムにおける一細胞トラン

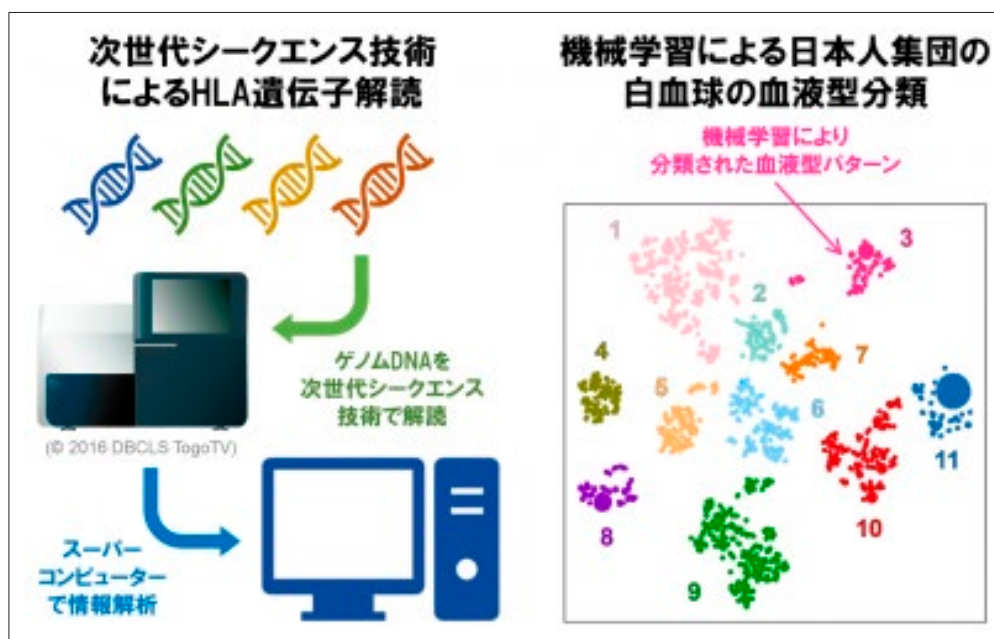


図3-1 日本人集団の白血球の血液型の個人差³⁾

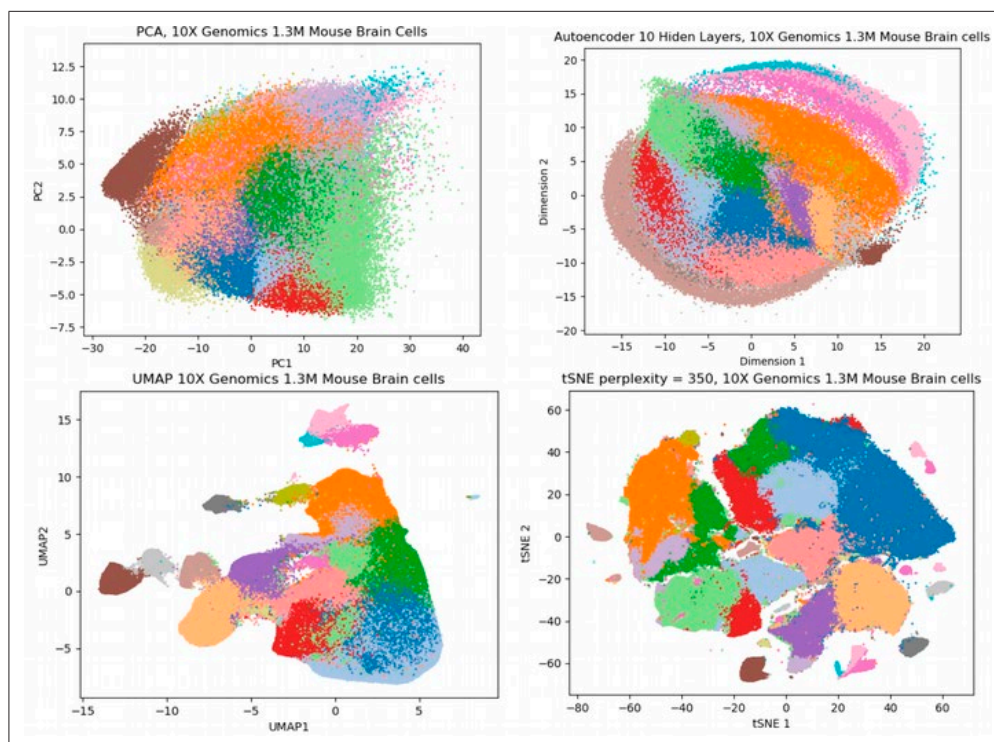


図3-2 scRNA-seqデータの次元削減手法の比較

スク립トームデータによる細胞の分類も同様である（図3-2）。例えば百万個の細胞の種類を2次元マップに落とし込む際に、t-SNEとディープラーニングの手法を用いた次元圧縮アルゴリズムであるUMAP（Uniform Manifold Approximation and Projection）で比較するといった試行が行われている⁴⁾。

3.3 疾患リスク予測

簡単にアクセスできる体液（血液、尿など）から疾患に関連するバイオマーカーを測定し、疾患蓋然性を予測するリキッドバイオプシーは、個別化医療に革命をもたらしつつある。従来のバイオマーカー研究は特定の疾患と関係の深い分子を一对一で探索することが主流であった。機械学習は測定されたバイオマーカーの多重化された測定値を組み合わせること（複数の分子の濃度などの特徴パターンを自動的に検出すること）が得意であり、関連した多くの論文が発表されている。特に血液や尿に含まれる細胞外微粒子であるmiRNA（マイクロRNA）が様々な疾患の新たなバイオマーカーとして有望視されている。

名古屋大学等のグループでは、ナノメートル単位のワイヤを使って尿に含まれるmiRNAを集めるデバイスを作成し、尿1ミリリットルから1,000種類以上検出できることを確認した。肺、脾臓、肝臓、膀胱、前立腺のがん患者と健康な人の尿を調べてそれぞれのがん患者で増減しているmiRNAを発見した⁵⁾。マイクロアレイでmiRNAの発現プロファイルを蛍光強度として測定し、あらかじめ機械学習によって開発しておいた判定アルゴリズムに基づき、蛍光強度のプロファイルから、癌に罹患しているかどうか判定することが想定されている。

国立長寿医療研究センターらの研究チームは、血中miRNAを用いた認知症発症リスク予測モデルの構築を行い、三大認知症（アルツハイマー病、血管性認知症、レビー小体型認知症）を高感度かつ特異的に判別可能にした⁶⁾。三大認知症を含む認知症群と認知機能正常高齢者群1,569例の血清を高感度DNAチップを用い、網羅的にmiRNA発現解析を行い、得られたデータを、教師あり機械学習手法のひとつである、主成分分析ロジスティック回帰法（supervised PCA logistic regression method）に応用し、予測モデルの構築を行った。予測モデルで用いられたmiRNAのターゲット遺伝子の関係ハブ遺伝子としてEXOC5、DDX3X、YTHDF3を抽出した（EXOC5はアルツハイマーと、DDX3XとYTHDF3は脳腫瘍との関連が知られている）。

今後、miRNA、メタボローム、マイクロバイーム等の研究などと相まって、血液や尿、唾液などから多くの疾患のリスク（蓋然性）が簡便に検出できる時代が到来するであろう。

今般の新型コロナウイルス感染症においても、ニューヨーク大学、中国Wenzhou Central Hospital等のグループが、重度肺疾患の発症を予測できるAIツールを開発し、肝酵素アラニンアミノトランスフェラーゼALT値の上昇・筋肉痛の存在・赤血球ヘモグロビン値の上昇という3つの因子が重症疾患の最も正確な予測因子となることを発見した⁷⁾。米カリフォルニア州に本拠を置くBitscopic社の研究チームでは、一般的な血液検査項目の組み合わせから、新型コロナウイルス感染を識別するAIモデルを構築した。検査を受けた75,991人からなる退役軍人コホートを利用し、感染の有無をラベル付けした上で、20の血液検査項目の結果から機械学習モデルをトレーニングしたところ、感度82.4%・特異度86.8%・正確度86.4%で予測することができたとしている⁸⁾。

3.4 疾患の層別化（予後予測）

理化学研究所のチームでは、血液検査データに基づく、機械学習（ランダムフォレスト）による卵

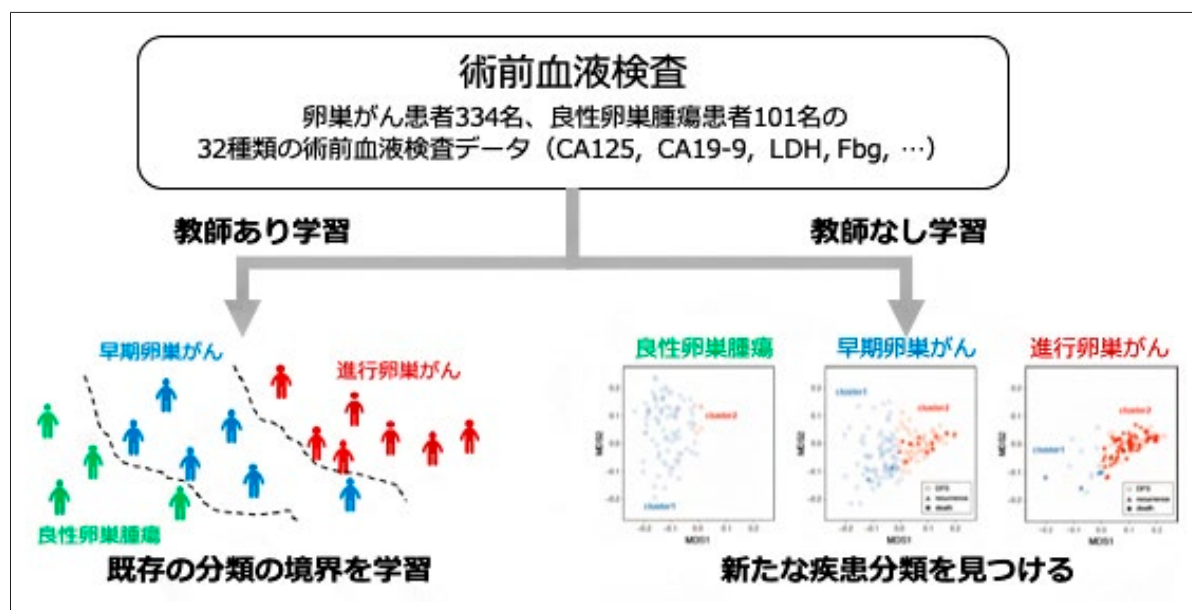


図 3-3 機械学習でがんの特性を術前予測⁹⁾

巣がんの良性／悪性を予測する「卵巣がんの術前予測アルゴリズム」を開発した⁹⁾。これにより診断時の年齢および術前血液検査データ 32 項目のデータに基づいて、悪性腫瘍と良性腫瘍を非常に精度良く予測できることを証明した。またこの研究では、進行期予測において、教師なしランダムフォレスト法を用いて、今まで临床上気づかれなかった血液検査のパターンも発見している (図 3-3)。あるクラスは再発がほとんどなかったのに対して、別のクラスは再発率と死亡率が高いという予後との強い関連を見出した。

また、こういった研究によって層別に重要な特徴量選択ができる。毎回全部の項目を網羅的に測定するというのは大変だが、機械学習である因子が重要だということがわかれば (それが一つに絞れなくても、10 個重要だというのがわかれば)、その 10 個を計測していけばいいということも考えられる。例えば、この研究では、既に知られている腫瘍マーカーに加えて CRP と LDH の値が重要であることが示され、進行期と炎症との関連が示されている。

がん患者の予後予測は、疾患管理にとって重要である。プレプリントの段階であるが、UCSD (カリフォルニア大学サンディエゴ校) 等の研究グループでは、ディープラーニングモデルと機械学習モデルのアンサンブルを使用して、複数のタイプのオミックスデータに基づいて患者の生存サブタイプを予測する新しいフレームワークである DeepProg を開発している¹⁰⁾。TCGA (Cancer Genome Atlas) の 32 の癌データセットに DeepProg を適用し、複数の癌生存サブタイプを特定した。この方法が他の方法よりも有意に優れていることを証明した。

以上のように、ゲノムやオミックスデータにおいては、多因子の変動をパターンとして捉えることができるという機械学習の特徴が疾患研究や予測・予防研究と非常に相性が良く、多くの研究成果が見られる。

一方で、ゲノム等の解析の成果をどのように活用すれば疾患病態や新規創薬に活用できるのかはほとんどわかっていない。また、ディープラーニングの適用法という意味では、画像データに対してのような方法論が確立しておらず、途上である。データ数に対して、パラメータ (変数) が多すぎるという問題があることに加え、実験の複雑さから、画像に比して、データ・情報科学者が参入しにくい

という構造上の問題もあると考えられる。何でもディープラーニングというより、その要所要所に数理、機械学習の最適な技術を適用（選択）することで規則や秩序を抽出するような汎用的な手法を開発していく必要があると考えられる。

また今後、個別医療による高精度な予測という方向に向かっていくためには、時系列のデータが必要になる。データ取得の時系列が散発的で、一人一人の患者だと時系列と呼べないようなものしか集まらない場合は状態空間モデル（潜在変数が遷移していくモデル）を用いた少しずつ追加される観測データに基づいたモデル修正も考えられる。多くの患者のデータを次々にそこに入れていくと、患者一人一人のデータが少なくとも、例えばそれを数万人規模で集めると、一つの遷移モデルを構築することができるとも考えられる。また、ウェアラブルデバイスのように、計測機器の工夫によって時系列データを取得しやすくすることも必要であろう。

【コラム③】物質・材料研究（マテリアルズ・インフォマティクス）との比較

マテリアルズ・インフォマティクスという概念が浸透してきた。冒頭に述べたように、そのプロセスは基本的にはバイオ研究と変わらないが、物質・材料研究においては、研究目的当たり存在するデータ量が限定的であることから、現時点ではデータを学習させる際に、特徴設計をヒトが与えることが通常である。すなわち記述子（ディスクリプター）として、関係する各種の物理・化学パラメータを用意するというプロセスが一つ加わることになる。

- ①学習用データ準備（入力、出力などのアノテーション情報）
- ②記述子（ディスクリプター）の選択
- ③アルゴリズム選択
- ④得られたモデルをテストデータセットで評価

記述子とは、「化合物や分子」をどのようにコンピュータに入力するか的方式であり、組成・分子式、構造、物性等に関する情報など、ある物質・材料の特徴を様々な形で機械可読にする方法である。ケモインフォマティクスの分野では、従来から有機化合物や高分子の記述子（方法）に関する研究が一分野として進展してきた。

マテリアルズ・インフォマティクスの代表的な研究として、ランダムフォレスト（RF）等を用いたバーチャルスクリーニングやスパースモデリング等を用いた計測インフォマティクス（計測のハイスループット化・高解像度化）が行われている。大きく分類すると、「安定した材料の発見とその結晶構造の予測」、「多数の定量的構造とプロパティの関係」、および「第一原理計算を機械学習で置き換えるための研究」といったことがある。

その際に特に無機材料の分野においては、同質の実験データが決して多くないところ、第一原理計算によって、構造と物性のデータセットを作ることができるというところが、バイオ研究との違いである。一方で、計算によっても計測によっても、学習データを整備するのに多大な時間とコストを要する点が課題である。

物質・材料研究においても、ディープラーニングは用いられ始めている。例えば結晶構造予測や結晶構造を考慮した特性予測、2-D/3-D微細構造画像等である。もう一つは創薬のところで述べたVariational Auto Encoder（VAE）に基づく化合物自動生成手法（分子インフォマティクス）が有機太陽電池や有機EL材料など、有機エレクトロニクス材料や高分子物性の研究に用いられている点で共通する。

ブラックボックス化の解消が課題となることに対し、NEC等は、実験データ生成の省力化技術と、材料特性に対する高精度の予測とその解釈性を与える技術に加え、材料の特性向上に関わる無数の要因から主要因を効率良く抽出する手法を開発した¹⁾。さらに、この手法を用いてスピン熱電材料の熱電性能向上の実証に成功している。物理・化学などの専門的な知見を持った開発者

がAIの予測結果の背後にある支配的な物理現象や因果関係をひもとくヒントが得られる、としている。

その他、化学触媒開発において、JAIST等のグループでは、ハイスループット触媒評価装置を設計し、これを用いて過去30年間に蓄積されたデータ数を一桁上回る12,000点ものデータをわずか3日で取得した。得られた触媒ビッグデータを機械学習などによって分析し、その結果に基づいて固体触媒や反応プロセスを通して、メタンの酸化カップリング反応の反応収率を大きく改善することに成功している²⁾。AIとロボティクスを用いた研究の自動化が盛り上がりを見せつつある点などは、ライフ・バイオ分野と共通する。

参考文献

- 1) Yuma Iwasaki et al., Identification of advanced spin-driven thermoelectric materials via interpretable machine learning, Nature Partner Journal Computational Materials. (2019)
- 2) Thanh Nhat Nguyen et al., High-Throughput Experimentation and Catalyst Informatics for Oxidative Coupling of Methane ACS Catal. 2020, 10, 2, 921-932

参考文献

- 1) Ryan Poplin et al., A universal SNP and small-indel variant caller using deep neural networks Nature Biotechnology volume 36, pages983–987 (2018)
- 2) Alok Sharma et al., DeepInsight : A methodology to transform a non-image data to an image for convolution neural network architecture Scientific Reports volume 9, Article number : 11399 (2019)
- 3) Jun Hirata et al., Genetic and phenotypic landscape of the major histocompatibility complex region in the Japanese population Nature Genetics volume 51, pages470–480 (2019)
- 4) Nikolay Oskolkov ., Deep Learning for Single Cell Biology Resolving cellular architectures by Deep Learning
- 5) Takao Yasui et al., Unveiling massive numbers of cancer-related urinary-microRNA candidates via nanowires Science Advances 15 Dec 2017 : Vol. 3, no. 12, e1701133
- 6) Daichi Shigemizu et al., Risk prediction models for dementia constructed by supervised principal component analysis using miRNA expression data Communications Biology volume 2, Article number : 77 (2019)
- 7) Xiangao Jiang et al., Towards an Artificial Intelligence Framework for Data-Driven Prediction of Coronavirus Clinical Severity CMC-Computers, Materials & Continua, 63 (1) , 537–551.
- 8) Vafa Bayat et al., A SARS-CoV-2 Prediction Model from Standard Laboratory Tests Clinical Infectious Diseases, ciaa1175
- 9) Eiryo Kawakami et al., Application of artificial intelligence for preoperative diagnostic and prognostic prediction in epithelial ovarian cancer based on blood biomarkers Clinical Cancer Research Published first April 11, 2019.
- 10) Olivier B Poirion et al., Multi-omics-based pan-cancer prognosis prediction using an ensemble of deep-learning and machine-learning models

4 AI活用の課題

前章までに、ディープラーニングがライフサイエンス研究にブレークスルーをもたらしている状況を見てきた。しかも2016年以降に急激に研究成果が出始めている新しい潮流であることがわかる。

生命科学や医学研究のあり方を変えている駆動力となっているのは、ディープラーニング等のデータ解析技術もさることながら、計測技術や計算機技術の進展によって生命の時空間に亘ってデジタルデータがハイスループットにとれるようになったことが大きい。

また、ヒト等の生命システムやヘルスケア研究などは、研究対象が絶えず変動するオープンな複雑系システムであり、そういったシステムは確率論で論じることが妥当であることからビッグデータや統計といったデータ駆動型と相性が良い。さらに人間は一般的に直接には2.5次元の世界までしか認識することが難しいが、計算機はある意味無限次元からパターンを抽出できることが挙げられる。

今後もこの方向は進展していくことが予期され、データを制するものが、研究開発、産業を制するということがいよいよ現実的になってきた。

一方で、下記の通りディープラーニング（現在のアルゴリズム）の限界も見えてきた。

- 1) 学習に大量の高品質データや計算資源が必要
- 2) 理由の説明が難しい（ブラックボックス化）
- 3) 学習範囲外のケースへの対応が難しい

特に、1) については根幹の問題であり、以下に研究開発一般における課題とバイオ分野固有の課題について述べる。

(1) データの質・量に関わる一般的な課題

問題となるのがデータの質と量の確保である。これまでのライフサイエンス研究における成功事例のほとんどは教師あり学習であることから、アウトプットの品質は入力データの品質（特にラベリング・アノテーション）に依存する。高品質のラベルありデータは不足しており、機械学習を用いる際の重要なステップとして、トレーニングデータセットにおける人によるラベリング・アノテーションのバイアスを回避することがある。量だけではだめで、質、すなわち構造化、標準化されたラベルありデータが強く求められている。

研究機関等には既にたくさんのデータが分散して存在するが、そうした既存のデータ群は、予め機械学習に使うことを想定して産出されていないため、信頼性や様式不統一、ラベル・アノテーション欠損の問題が大きく、実際はほとんど使用できない。また、医学データなどは、大量のラベルなしデータと少量のラベルありデータが存在する。対象疾患でとられたデータは少量（数十～数千）なのに対し、健常者もしくは対象ではない別の疾患向けにとられたデータはある程度存在する（数万～数百万）。

こうした課題に対応するためにも、デジタル計測機器の進展により、様々なデータがハイスループットにとれるようになってきていることから、今後データプラットフォームの設計と構築が不可欠になってきている。

一方で、機械学習は、教師ありで学習するという使い方をされることが多かったが、今回GANや教師なし学習によって、人間が今まで気づかなかった複雑なパターンを発見するという可能性も見え取れた。

(2) バイオ分野におけるデータの課題

生命現象は複雑であるため、一つのモダリティでは分からない情報が多いが、複数の情報の相関解析によって、今まで見えなかった現象の解析が可能になると考えられる。一つのモダリティから得られるデータによる精緻な観察と、複数のモダリティからのデータを組み合わせることによって統合的な情報が得られる。ゲノム、エピゲノム、トランスクリプトーム、プロテオームやイメージング画像など、さまざまな客観的データから、疾患の機構解明、解析につなげようとしている。これらのデータと実臨床データやセンサなどからのリアルワールドデータをリンクさせることによって、最終的には予防や予後、治療効果の推定を可能にすることが期待されている。

一方で、生物という対象の特殊性と現在の計測技術の限界から、生物・医学データについて下記のような本質的な課題が認識されている。

- ・P（特徴・パラメータ）がN（計測データ数）より圧倒的に大きい。測定対象数に比して測定項目の次元数が非常に多い。例えば、遺伝子発現の場合、計測データは数千次元であるのに対し、一塩基多型（SNPs）は数百万次元となる。
- ・因子間の相互作用が非線形。一つ一つの相互作用を見てもある程度のところで限界があるため、一般的な相関解析だとうまく因果関係が抽出できない。
- ・計測時間が散発的で不規則。物理の時系列データだとゼロが決まっていて、サンプリングポイントも一定で、時系列データが大量にある状態でモデルをつくることができる。医学の場合、まず患者が病院に来るという時点でもう病気が始まっている。さらに、その次のデータが患者が再診に来る3カ月後でその間のデータがないといったように、時系列モデルというのが難しい。
- ・内部因子だけでモデルが完結しない。例えば、トランスオミックスで、あるレイヤーでモデルをつくったとしても、それだけではモデルが完結せず、他のレイヤーも考えないと最終的な説明や理解には至らない。
- ・計測データに時空間ギャップが生じる。分子から個人までの空間スケールと短時間のメタボローム変化、長時間のゲノム変化までの時間スケールがある。
- ・系のヘテロ性。同一の分子・細胞であらゆる時空間データを同時に観測することは困難、全細胞の同時観測も困難である。

加えて、疾患研究においては、ヒトの遺伝子・ゲノム、表現型（健康診断、電子カルテ等）、リアルワールド（ウェアラブル等）データなどの知識データベース（学習用データセット、テストデータセット）の構築が必要であり、下記のような課題が挙げられる。

- 1) 電子カルテ等の知識記述における定式化と標準化
- 2) データの大規模収集（既存のデータの連携、コホート）
- 3) 個人情報保護や匿名化

これらの課題を踏まえた上で、以下の章では今後の展望、日本が推進すべき方策を述べる。表4-1に、課題と対応策をまとめた。特に対応策の①～⑤について5章で述べる。

表 4-1 課題と対応策



課題	対応策
データ品質と数量 ● 新NP問題（サンプル数<<パラメータ数） ● 大量のラベルなしデータと少量のラベルありデータ ● 事後アノテーションのコスト高	①データプラットフォーム＋分散統合学習 ②研究環境の自動化（ロボットなどの活用） ③自然言語処理（NLP） ■ 生成モデル（GAN、VAE）[これまでの章で記述]
ブラックボックス	■ 次元削減技術や重み付け可視化技術 [これまでの章で記述]
学習範囲（内挿外挿）	⑤次世代AI ■ 生成モデル（GAN、VAE）[これまでの章で記述] ■ データ同化（実験とのサイクル）
複雑系生命理解（時空間問題）	④計測技術＋AIアルゴリズム

【コラム④】機械学習プロセスの自動化技術

研究者が機械学習を用いて研究を実施しようとした時に、困るのが適切なアルゴリズムの選択ではなかろうか。そこで、注目されているのは機械学習プロセスの自動化技術 AutoML (Automated Machine Learning) である。機械学習を用いたデータ分析作業において、従来、分析のための前処理や構造設計（特徴エンジニアリング、モデル選択、ハイパーパラメーターチューニング等）のステップは、人手で試行錯誤しなくてはならなかった。しかし、AutoMLは、それら各ステップにおいて考えられるさまざまな組み合わせの中から最適なものを効率よく探索することを自動化する。代表的なシステムとしては、DataRobot、dotData、Google Cloud AutoML等が知られている。AutoML技術によって、機械学習・データ分析のプロセスに要する期間が大幅に短縮されることを加えて、必ずしも高い専門知識を持たない人でも簡単に実施可能になり、機械学習を用いたデータ分析の適用場面が広がる。

5 今後の展望（日本が取り組むべき方向性）

5.1 今後の展望① 戦略的なデータプラットフォーム構築 ～オープンサイエンスの実践～

デジタル化の波と技術進展のサイクルの短期化により、研究開発の方法論も大きく変化してきており、変化に対して研究体制を柔軟に変化させていくことが求められる。また、こうした先端研究に要するコスト（特に計測や解析のための費用）も高騰し、大学等においても個別ラボでできることは相当限定的になっていくことは想像に難くない。

データの数量不足、様式・形式不統一、収集の高コストを克服するため、ゲノムや脳科学のような分野（下記参考を参照）では、プロジェクト毎に、皆でデータを共有して（Shared）、オープンサイエンスで進めようということが世界的な潮流になっている。このようなプロジェクトでは、①集約化された先端計測装置でデータを取り、②標準化されたデータセットを効率よく利用し、③理論、数理やデータ科学研究者も参画しながらモデルに仕上げていくために、④データ・コーディネーション・プラットフォームをクラウド上に構築している。これは、理論研究者やデータ・情報研究者を巻き込むには信頼性の高い価値のある実験データが必要であることから、理にかなっている。

ここでいうデータプラットフォームとは、単なるデータの置き場所だけではなく、データ解析のためのソフトウェアライブラリやコーディングフレームワークなども含む概念である。

【参考】海外動向

- Cancer Genome Atlasは、2006年に米国 NCI（米国国立がん研究所）とNHGRI（国立ヒトゲノム研究所）の共同で開始されたがんゲノム解析プロジェクトである。合計33種類のがん腫、11,000例を超える患者検体のゲノム、エピゲノムやトランスクリプトームなど統合的遺伝子解析が実施され、各がん腫において包括的に遺伝子異常の全体像が解明されただけでなく、それらの分子分類が提唱された。2018年には、その成果をまとめて、細胞起源、共通のがん化プロセス、がん腫ごとの異常パスウェイの違いなどのテーマごとにPan-Cancer Atlasとして発表している。さらに、治療に難渋することがある転移がんの全ゲノム解析も2019年に報告している。
- NIHのLINCS L1000プロジェクトでは数千の医薬品候補化合物を様々な培養細胞に暴露した時の遺伝子発現データ（トランスクリプトームの変動）等を公開している。2014年度に開始されたフェーズ2では、6つのLINCSデータおよび署名生成センターをサポートしている。
- Allen Brain Atlasは、2003年に開始され、遺伝子と脳機能の関係をよりよく理解するために、脳スキャン、医療データ、遺伝情報、および心理データを収集して統合するプロジェクトである。最初のアトラスは2006年に公開された。哺乳類の脳の生物学的複雑さを探求するためのデータ、ツール、知識を生成するためにマウス脳アトラス、ヒト脳アトラス等が公開されている。
- Human Cell Atlas（HCA）は、2018年に開始された国際的な科学者の共同コミュニティで、人間の健康を理解し、病気を診断、監視、治療するための基礎として、人体のすべての細胞の包括的なリファレンスマップを作成する取組である。データ調整プラットフォーム（HCA DCP）は、Human Cell Atlasを構成するデータを整理、標準化、およびアクセス可能にするために開発されたオープンソースのクラウドベースのプラットフォームとなっている。

見方を変えると、科学研究においても産業と同様、プラットフォーマーが現れていると見ることも

できる。脳神経科学研究では、ジェネリア研究所やアレン研究所のデータプラットフォームの活用なしに進まない、がんや神経疾患研究ではブロード研究所のゲノムデータプラットフォームなしには進まないという時代が来つつある。

スイスのロッシュグループは、癌創薬分野のプラットフォーマーになるべく、癌遺伝子パネル・治療成績とリアルワールドのカルテ情報を統合したデータベースを構築することによって、癌新薬開発に不可欠なインフラ整備を推し進めている。ロッシュ傘下のFoundationOneは癌遺伝子パネル検査で、世界中の癌患者の変異情報を収集している。エーザイ、米国Pfizer社、スイスNovartis社などほとんど全てのビッグファーマやベンチャー企業がFoundationOneを活用して治験を始めている。ロッシュは2018年米国の医療機関からカルテ情報と医療費情報を集めた米flatiron社も買収した。

関連して、デジタルデータ研究時代に重要なことのひとつに、計測技術の世界標準（デファクトスタンダード）をどう握るかという問題がある。欧米では新しい技術開発が非常に盛んに行われており、標準化（サンプルも含め）をどうするかに取り組んでいる。

例を挙げると、工業用のX線CTは日本のメーカーが最も性能の良いものを作っていたが、ドイツが国策としてFraunhoferとMax-Planck、さらに産業界と一緒に規格を作った。それを今、ISOに登録しようとしている。自動車メーカーに直接部品を供給するTier 1企業であるボッシュが、ドイツの規格に則って撮影されたCTデータで保証されていない部品は受け入れないとした。つまり、国際的な規格ができる前に、ドイツにデファクトを作られてしまうと、この規格から外れた日本のメーカーのX線CTは販売できなくなる。

同様のことが、バイオイメーキングのデータの分野で起こる可能性がある。現にEMBL（欧州分子生物学的研究所）がハブの中心になって運営されているユーロバイオイメーキング（欧州の11カ国、29のバイオイメーキング拠点が参加）ではヨーロッパの生命科学者に高度なイメージング取得・解析技術を提供している。トレーニングコースを行ったり、画像データの共有や管理、活用をするネットワークであり、これはオープンアクセスである。インダストリーボードがあり、日本からもオリンパス、ニコン、浜松ホトニクスといった企業が名を連ねている。

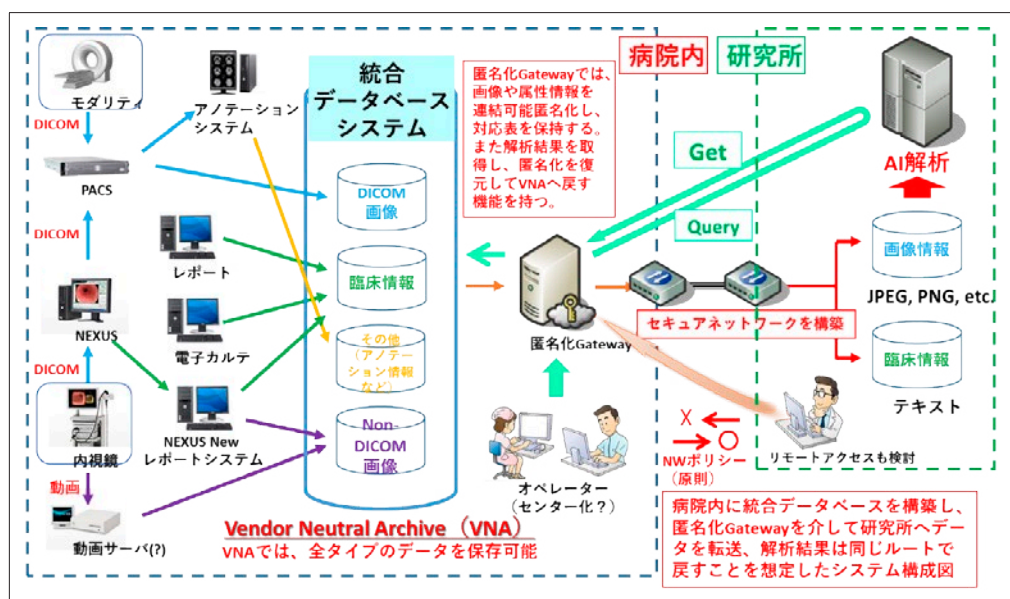


図5-1 病院と研究所のネットワーク構築
出典：浜本隆二（国立がん研究センター）

画像処理の進展には、良いデータが自動的に集まる仕組みが必要である。さらに、高解像度化、圧縮、特徴抽出などをパイプライン的に行うプラットフォームが必要であり、クラウド上でアプリケーションやデータを共有して解析ができる環境を作ることが求められる。

一例として、国立がん研究センターでは、国内で先駆けてVendor Neutral Archive (VNA) システムを構築した(図5-1)。VNAにはさまざまなモダリティ、例えばPACS (Picture Archiving and Communication System) のデータ、内視鏡の動画、電子カルテのデータ、さらにゲノムやエピゲノムのデータといったものを合わせて保存可能である。これらのデータは匿名化ゲートウェイにおいて画像や属性情報を連結可能匿名化し、対応表を保持している。また、解析結果を取得して匿名化を復元し、VNAへ返す機能もある。さらに、アノテーションツールを開発し、臨床の現場で出てきた診断結果にアノテーション付けして、構造化した状態で保存するというシステムを構築している。

AMED (日本医療研究開発機構) とNII (国立情報学研究所) は「医療ビッグデータ研究センター」を新設した。AMEDの「医療のデジタル革命実現プロジェクト」事業の中で、医療健康分野での人工知能(AI)などを開発していくICT基盤をつくる課題があり、これまでに日本消化器内視鏡学会、日本病理学会、日本医学放射線学会、日本眼科学会の4医学会が採択され、NIIがパートナーになった。各医学系学会が病院等から医療画像を集め、それを匿名化し、「医療画像ビッグデータクラウド基盤」に投入する。このクラウドにはITの研究者、医療の研究者がアクセスでき、協力してその解析をすることが可能になっている。

理化学研究所では、クラウド上で画像を管理するサーバ、画像処理をするソフトウェア、何をどのように処理したかという履歴、これらを一括して管理するシステムがプロトタイプとして動いている(図5-2)。このシステムの上にデータを載せると、入力データは何か、どのような撮影条件かという情報が記録される。したがって、異なるモダリティで撮影された画像データを、情報処理によって同じ空間分解能に落とすといった操作が実行しやすい。また、どのような画像処理をすれば望みの領域を抜き出せるか、実際におこなった手順の記録が残るので、方法論をシェアすることができる。例えば脳の研究で、脳幹を抜き出すアルゴリズムを開発したとすると、それを別の人が使えるようになり、さらに、別の人が使うことで改良された内容をお互いにシェアし、画像処理の方法論自体を革新していくことができる。

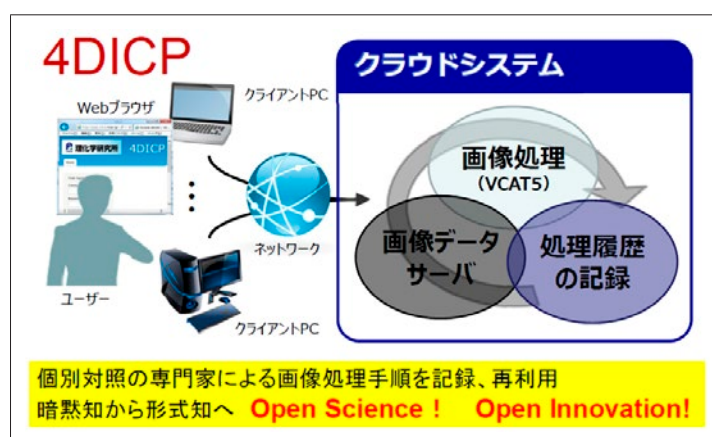


図5-2 画像を収集・解析・再利用を実現する仕組み
出典：横田秀夫（理化学研究所）

ゲノム、臨床データとして、AMEDでは、未診断疾患患者の情報共有と診断確定、そして治療を見据えた病態解明やシーズ創出を目的として、「難病克服プロジェクト」のもとで未診断疾患イニシアチブ Initiative on Rare and Undiagnosed Diseases (IRUD) を2015年から推進し、国際連携可能なデータベース構築等による積極的なデータシェアリングを行う体制を構築している。国立精神・神経医療研究センターをとりまとめ機関（コーディネーティングセンター）として、全国各地の37ヶ所

(2018年4月現在)のIRUD拠点病院を中心とした地域連携・専門分野連携を行っている。

国立がん研究センターのがんゲノム情報管理センター (Center for Cancer Genomics and Advanced Therapeutics : C-CAT) では、がんゲノム医療中核拠点病院やがんゲノム医療連携病院等で行われる、患者のゲノム解析の結果得られる配列情報および臨床情報を集約し保管する「がんゲノム情報レポジトリ」を構築するとともに、ゲノム解析結果の解釈・臨床的意義付けを行うためのがんゲノム知識データベース (Cancer Knowledge DataBase : CKDB) を構築している (2018年～)。

このようにデータ駆動型で研究を進めるに際しては、目的に応じて、プログラム、あるいはプロジェクト単位でデータコーディネートプラットフォームを構築することはもはや必要不可欠となっている。大規模なサンプルから標準 (リファレンスデータ) を作るための取組と、疾患とゲノム等の関係を理解するための取組に大別される。

なお、ディープラーニングなどを用いたデータ解析には、大規模な計算資源が必要になる。例えば、産総研のABCI (AI Bridging Cloud Infrastructure) は、4,352個のNVIDIA Teslaを搭載し、550ペタフロップスというスパコン速度性能ランキング世界5位のスペックを持つAI専用のスパコンである。これらを活用しながらデータ解析を進めている。

こうしたハードウェア資源は数年に一度更新する必要がある、数年毎に一定の投資を要する。財政上の持続可能性や外部のクラウドソースの利活用など検討していく必要がある。

【コラム⑤】医療におけるGoogleの取組み ～Project Baseline～

Verily Life Sciences (Googleからスピンオフ、Alphabet傘下) と、米デューク大学、スタンフォード大学の共同研究プロジェクト (2017年始動し約5年間)。

どんな経緯で、なぜ、いつ、特定の人が病気になるのか、病気にならない人がいるのはなぜかを把握すること、さらに、生活習慣と遺伝子の情報に基づき、すべての人々に該当する知見を得ることを目指している。これにより、もっと個人に応じた医療を実現し、ひとりひとりに適した治療を判断すること。また、予防を強化すること、つまり、事後対応型の医療システムから先取的な医療体系に移行するにはどうすればいいかを知ることができるとしている。

成人してから老齢に至るまでに人の健康状態がどんな過程をたどるのか記録することを目指している。ボランティア1万人の登録を目標。ボランティアには、自分の身体の状態に関するあらゆる要素について、毎日データを共有してもらう。例えば、睡眠時間、運動量、食事の内容、ワクチンの接種、血液検査や聴力測定の結果といった情報が該当する。ゲノム配列決定、マイクロバイオーム (身体の微生物群)、イミュノーム (免疫系を構成する一連の遺伝子とタンパク質) などを含む。データの多くはプロジェクト参加者が装着するVerily独自の医療研究用機器スマートウォッチからも送られてくる。膨大な量に及ぶデータ (変数) の間の複雑な関係を解き明かすには、AI・機械学習が活用される。

2019年5月には、Novartis、大塚製薬、Pfizer、Sanofiと提携し、「Project Baseline」のプラットフォームやツールを利用した臨床試験プログラムを開発すると発表した。

一方で、これまで紹介してきた事例において、学習用データとなる画像の準備に時間とコストを要するという課題がある。これを回避するために、GANを用いて少ない学習データから人工的にそれらしいデータを自動生成し、人工的に作成されたデータを学習に用いるということが行われ始めている。

例えば、NVIDIA、Mayo Clinic、およびMGH等のグループは、既存のアルツハイマー病のニューロ

イメージングデータや脳腫瘍画像データを用いて、データ上で腫瘍のサイズと場所を変更したり、健康な脳のスキャンに腫瘍を移植したりすることで、トレーニングに使用できる脳画像を生成できることを示した¹⁾。実際の脳スキャンと合成脳スキャンの組み合わせを使用して機械学習モデルをトレーニングした場合、実際のデータのみでトレーニングしたモデルよりも優れていた。今後、希少疾患のより多くの画像例を生成することなどが期待される。

多くの場合、研究者はさまざまなソースからデータを収集するが、医療の場合、患者データのプライバシーポリシーにより、データの中央管理は現実的ではない。そこで、Federated Learning（分散協働学習）の考え方が生まれた。この分散協働学習によって、機密の臨床データを互いに共有することなく、さまざまな機関がAIモデルの訓練をローカルに行い、モデルの開発に協力できるようになった。

2018年にIntelとペンシルバニア大学から脳腫瘍を検出する画像診断アルゴリズム開発が発表され、データ共有型のモデルと同等の性能を示している²⁾。他にもNVIDIA、米国放射線医学会、Diagnosticos da America、Partners HealthCare、オハイオ州立大学、およびスタンフォード大学は、乳房組織密度を評価するためのより優れた予測モデルを開発した³⁾。各機関は分散協働学習を使用して2Dマンモグラフィ分類モデルを改善し、ローカルデータのみでトレーニングされた元のニューラルネットワークよりもローカルデータセットの予測能力を向上させた。

データのアノテーション（ラベリング）の重要性として、学習用データに人種差や性差などの偏りが入り込むことで、感度が上がらなくなる可能性が指摘されてきた。アルゼンチンの研究グループにおいて、米国NIH（国立衛生研究所）およびスタンフォード大学で管理されている数万人規模の胸部X線画像データセットにおいて、「男性比率が高いデータセット」でトレーニングしたAIアルゴリズムを女性に適用させた場合、診断能力が明確に低下することが確認されている⁴⁾。

参考文献

- 1) Hoo-Chang Shin et al., Medical Image Synthesis for Data Augmentation and Anonymization Using Generative Adversarial Networks arXiv : 1807.10225
- 2) NVIDIA Clara Federated Learningが病院にAIをもたらし、患者のデータを保護 <https://blogs.nvidia.co.jp/2019/12/02/clara-federated-learning/>
- 3) Intel Works with University of Pennsylvania in Using Privacy-Preserving AI to Identify Brain Tumors <https://newsroom.intel.com/news/intel-works-university-pennsylvania-using-privacy-preserving-ai-identify-brain-tumors/#gs.6i416a>
- 4) Agostina J. Larrazabal et al., Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis PNAS first published May 26, 2020

5.2 今後の展望② 実験の自動化・デジタル化 ～データ取得のハイスループット化・オンデマンド化～

前節①のデータプラットフォーム構築に関連するもう一つの流れとして、AIを用いる前段階で必要となるデータ取得の自動化が挙げられる。これは、データ取得作業の自動化（Automation）に加え、繰り返し実験における条件設定等含めた取得プロセスの自律化（Autonomous）を志向するものである。

データインフラの構築には、データ取得の自動化（ロボティクス）とAIの連携が有効で、そのためには、これまで測定技術・機器セントリックな環境を構築していたものを、一つのサンプルに対して、様々な計測機器で物理、化学的データを取得するようなサンプルセントリックな研究環境に移行していくことも視野に入れるべきである。ハイスループット化のキーテクノロジーとしてロボット工学とマイクロフルイディクスが挙げられる。2018年頃から化学合成とバイオ生産の分野でScience誌やNature誌等で論文が相次いで発表されている¹⁻⁵⁾。

化学合成では有機合成のデジタル化、ロボットケミストなどの概念が出てきており、データベース等過去のデータから学習するだけでなく、ロボットを用いて、実際に合成実験を行いながらデータを解析して予測したり、半自動で合成したりする研究論文が海外から出始めている。

ファイザー製薬のグループでは、複数の離散変数（触媒、リガンド、塩基）および連続変数（温度、滞留時間、圧力）（24時間あたり1,500反応を超える速さで合計5,760反応ポイントでのLC-MSデータを生成）を使用した鈴木-宮浦カップリングのハイスループット反応スクリーニングを実証した。これにより、フロー装置が薬物発見プロセスの早期段階での反応プロセスの最適化を加速できること、および他の研究者が機械学習を利用するために信頼できるデータを提供できることを示した¹⁾。

グラスゴー大学のグループでは、約1,000個の化学反応組合せのわずか10%を実現することで、残りの実験を実施することなく、90%の結果を80%以上の精度で予測するシステムを構築した²⁾。

MITのグループでは、最初にAIが分子を合成するためのルートを提案し、化学者がこのルートを

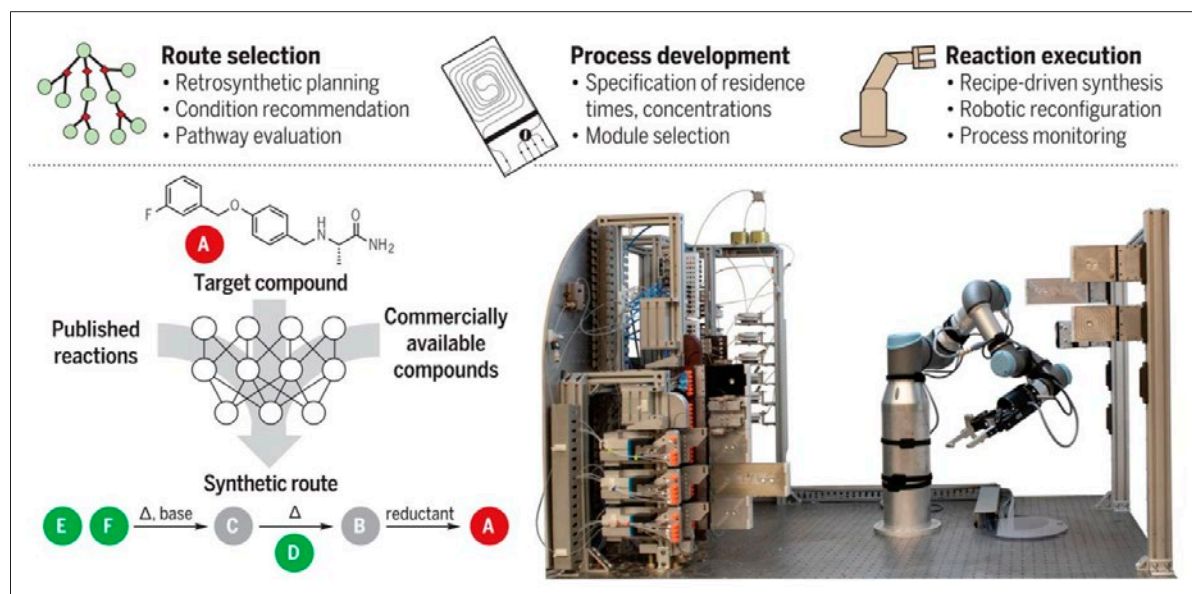


図5-3 化学合成の半自動化（MIT）³⁾

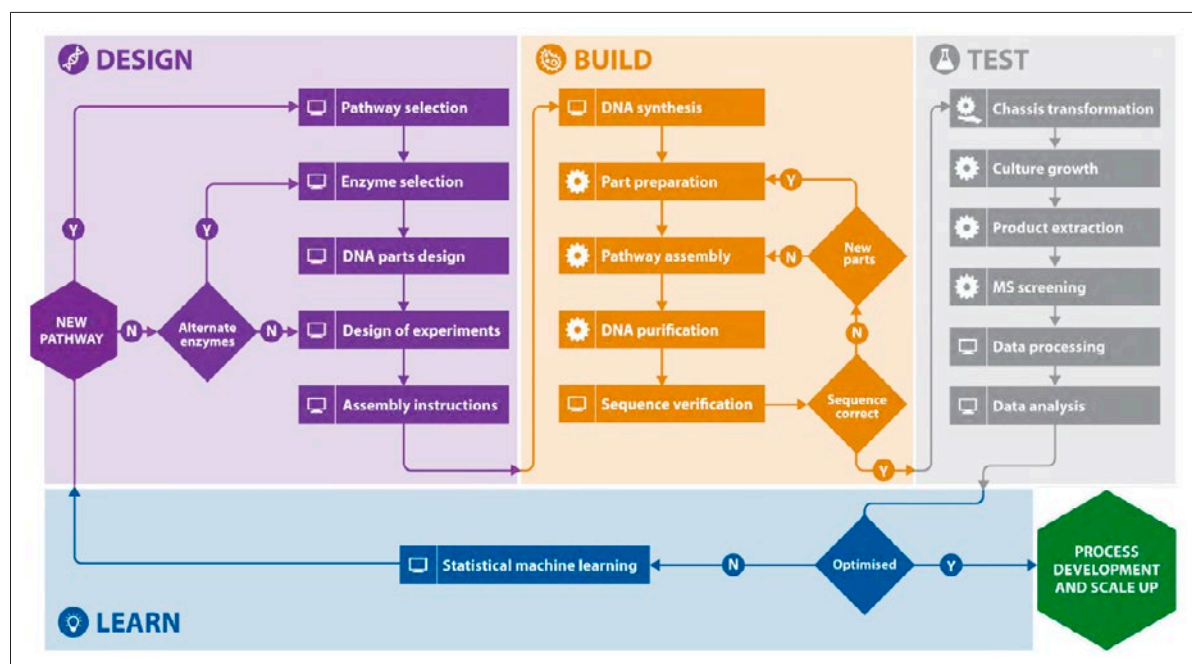


図5-4 DBTLサイクルのパイプライン化
出典：高野恵理子（マンチェスター大学）

確認して化学的なレシピに仕上げ、最後にハードウェアを自動的に組み立てて反応を実行するロボットプラットフォームにレシピを送信する、というプラットフォームを構築した³⁾ (図5-3)。

微生物を用いたバイオ生産にも活用されている。微生物等の生化学的経路を利用して有用な化合物を生成するために、設計、構築、テスト、学習のDBTLのサイクルを何度も回して、最適なデザイン（遺伝子の改変）につなげていくというプロセスが必要とされている。一方で、大規模なデータと分析が必要になるため、実験コスト、生物の変動性、人間のバイアス、ハイスループットな評価システム等が課題となっている。

マンチェスター大学等のグループでは、DBTLサイクルを回す各要素のパイプライン化を行い（図5-4）、大腸菌によるフラボノイド（2S）-ピノセンブリンの生産に適用し、最初の仮説プロセスに基づき、設計、実験を行い、2回のDBTLサイクルの適用により、最大88mg/L-1に到達している⁶⁾。最初の設計から効率（単位当たりの収量）が500倍改善された生産経路を確立したことになる。

イリノイ大学のチームでは、ロボットと機械学習で実行できるBioAutomataプラットフォームを作成し、概念実証として、リコピン生合成経路を最適化する（ランダムスクリーニングを77%上回る）ことにより、その能力を実証した⁷⁾。イリノイ州の支援の下、2014年に開設されたiBioFabでは、20以上の計測等機器を統合し、バイオエンジニアリングの効率化を狙っている。

日本でこのような包括的な動きはほとんど見られない。一例として、東京大学、産業技術総合研究所のグループでは、生命科学研究へのLabDroidの活用（図5-5）および将来的な研究の方向性として、「ロボット群集」で生命科学を自動化するRobotic Crowd Biologyを提案するコンセプト論文を発表している⁸⁾。

大阪大学のグループは細胞内の分子の動態を観る、1分子ライブイメージングの自動化を行っている⁹⁾。96穴のウェルの中の細胞に、どの時間にどのタイミングでどの溶液を与えるかを全てプログラムでき、自動で焦点を合わせて細胞を探して、得られた画像から、1分子を自動で認識して、位置と

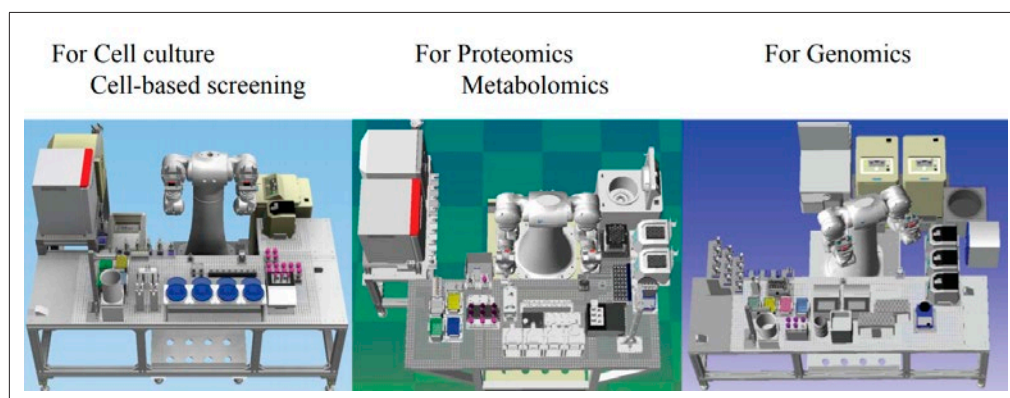


図 5-5 LabDroidの汎用性⁸⁾

蛍光強度の時間変化を追跡することができる。生きた細胞内での蛍光1分子のビデオ画像を撮る。複数の分かれたウェルの中に異なる条件での細胞の観察が可能であり、1日で8,000細胞まで見るができるようになってきている。今後、広視野化の開発による高速化や同時に複数の分子を見ることなどが考えられるとしている。均一な細胞集団であれば、複数の分子種から得られたデータを全てコンピュータに入れて、コンピュータの中で統合することは可能であろう。

米国DOE (Department of Energy) では、2018年に、新規物質・材料発見の自動化に向けて、AI、ロボティクス、逆設計、データインフラ、時空間スケールの統合が重要な要素となる、という主旨のMaterials Acceleration Platformレポートを発行している¹⁰⁾ (図5-6)。

自動車業界におけるCASEは、Connected (コネクテッド)、Autonomous (自律)、Shared (シェアリング)、Electric (電気自動車) の頭文字をとった造語であるが、物質材料研究やライフサイエンス、バイオロジー研究においても、Connected (コネクテッド)、Autonomous (自律化)、Shared (共有) といった考え方は共通する。

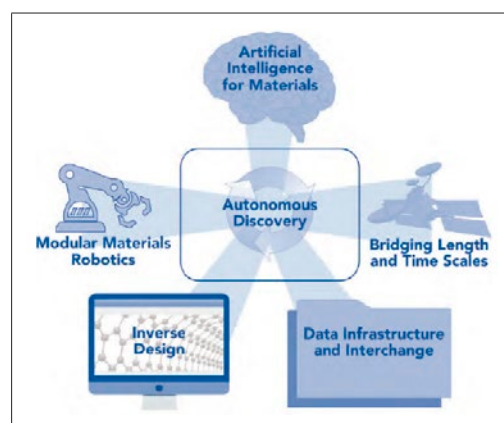


図5-6 DOE Materials Acceleration Platform¹⁰⁾

以上のように、データは研究プロジェクト毎に合目的的にとっていくというのが時代の潮流に合っているように思われ、今後、機器の高度化やデータ産出の低コスト化により、バイオフィアウンドリ (必要な装置・ソフトウェア群を集積し自動化・パイプライン化した施設) を構築して、オンデマンド (On-Demand) で大量のデータをとって研究を進めていくという環境の構築が進んでいくことが考えられる。

参考文献

- 1) Damith Perera et al., A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow Science 26 Jan 2018 : Vol. 359, Issue 6374, pp. 429-434
- 2) Jarosław M. Granda et al., Controlling an organic synthesis robot with machine learning to search for new reactivity Nature volume 559, pages377–381 (2018)

- 3) Connor W. Coley et al., A robotic platform for flow synthesis of organic compounds informed by AI planning *Science* 09 Aug 2019 : Vol. 365, Issue 6453, eaax1566
- 4) Sebastian Steiner et al., Organic synthesis in a modular robotic system driven by a chemical programming language *Science* 11 Jan 2019 : Vol. 363, Issue 6423, eaav2211
- 5) Elizabeth Bilsland et al., Plasmodium dihydrofolate reductase is a second enzyme target for the antimalarial action of triclosan *Scientific Reports* volume 8, Article number : 1038 (2018)
- 6) Pablo Carbonell et al., An automated Design-Build-Test-Learn pipeline for enhanced microbial production of fine chemicals *Communications Biology* volume 1, Article number : 66 (2018)
- 7) Mohammad Hamedirad et al., Towards a fully automated algorithm driven platform for biosystems design *Nature Communications* volume 10, Article number : 5150 (2019)
- 8) Robotic crowd biology with Maholo LabDroids *Nature Biotechnology* volume 35, pages310–312 (2017)
- 9) Masato Yasui et al., Automated single-molecule imaging in living cells *Nature Communications* volume 9, Article number : 3061 (2018)
- 10) Aspuru-Guzik et al., Materials Acceleration Platform : Accelerating Advanced Energy Materials Discovery by Integrating High-Throughput Methods and Artificial Intelligence.

5.3 今後の展望③ 自然言語処理とEHR（電子カルテ等）の利活用 ～AI医師～

医療研究における代表的な課題の一つに電子カルテ等の知識の構造化と自然言語処理（NLP：Natural Language Processing）が挙げられる。これまでに挙げた機械学習・ディープラーニングの活用事例は主にパターン処理に関するものである。自然言語には固有の曖昧さや特性など、コンピュータにとって扱いづらい点が多くある。

従来、言語研究のためにテキスト情報を大量に集積した「コーパス」や、単語の意味、類義関係、上位・下位の関係を構造化した「シソーラス」といった言語のデータベースを用いて、文章の理解を行うアプローチが行われてきた。このような手法は人間が手動で構築することが大半であり、また言語の体系は時代とともに変遷もするものであることから、常にアップデートが必要であり、高コストになるもの事実であった。

医療分野では、従来より電子カルテ等（EHR）の様式の構造化、標準化が進んでおらず、また個人情報（プライバシー）の問題もあり、デジタルでの共有が進んでいないことから、活用が遅れている。

中国では、AI医療、AI医師の取組が多く見られる。広州医科大学広州女性小児医療センターでは、ディープラーニングを使用した自動化された自然言語処理システム（hierarchical logistic regression models）を適用して、EHRから臨床的に関連する情報を抽出する研究が報告されている¹⁾（図5-7）。136万人の小児患者からの計1,016万のデータポイントを分析して、トレーニングし、検証を行っている。その結果、一般的疾患（インフルエンザ、手足口病など）と、危険もしくは生命を脅かすような疾患（急性の喘息発作、髄膜炎など）の診断など、複数の臓器システムにわたって、55種の小児疾患の診断において経験豊富な小児科医に匹敵する高い診断精度を実証したとしている。

このシステムで3万276回の診断を行い、診断と臨床の一致度は87.4%にのぼる。他にも上海交通大学医学院附属上海児童医学センターの外来で、小児科の「AI医師」が勤務を開始したと発表している。

すでに民間ベースでも商用されており、世界最大の保険会社「平安保険グループ」では、コアとなる医療サービスとして、オンライン問診や遠隔診療により診察の効率化を実践している。平安医好生（Ping An Good Doctor）と名付けられたサービスは、登録ユーザーは2億8,900万人、月間6,270万

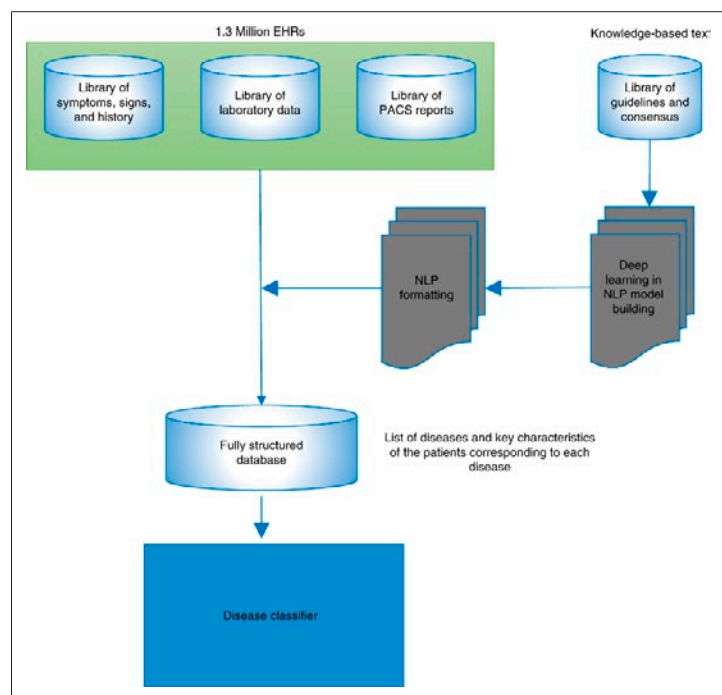


図5-7 小児疾患のAIシステム診断のフローチャート¹⁾

人の利用がある（2019年6月）。診察データは累計で3億件以上が蓄積されており、これをもとにAI診断を行い、人間の医師をサポートすることに役立てている。これにより、従来の5倍近い効率で診察が行われている。契約するドクターは4万人に達する。さらに、3,100の病院、1,100の診療所、500のデンタルクリニック、1万の薬局と提携をしており、プラットフォームの体をなしている。

テンセント（騰訊控股）傘下の医療企業「WeDoctor（微医）」でも、自治体との契約によって、農村部に定期的に移動診療による健康診断を行い、ここで集められたデータを、AI訓練用として同社のデータベースに蓄積。データベースには2,000種の疾患と5,000以上の症状が記録されており、AIによる診断の精度は90%を超える。自治体の医療従事者はこのシステムを利用することができる。

中国は、政府の体制や社会のAIや個人の医療情報（プライバシー）に対する許容度など様々な側面からこのような医療応用がいち早く可能になっていると考えられる。

英国のBabylon Healthが開発中の医学的な因果関係を抽出するAI手法によって、筆記試験のスコアで、英国の家庭医（GP：general practitioner）の平均診断精度71.4%を上回る結果を示した²⁾。筆記試験には1,671例の現実即した症例をテストセットとして用い、相関に基づく従来のアルゴリズムと、新しい因果推論のアルゴリズムの両方を使用し、後者の方が高い精度を示した。

他の自然言語処理の利用の取組の一例として、日本において、AIを活用した認知症診断支援システムが開発されている。診察の際に医師と患者が交わした5～10分程度の会話をテキスト化し、独自の自然言語解析AI「Concept Encoder」で分析する。少量の会話から認知症患者に特有の会話パターンを見だし、認知機能障害の有無や重症度の判定を行う。AIベンチャーのFRONTEO社と慶応義塾大学が研究開発を行い、FRONTEOと共和薬品工業が実用化に向けて協業している。

このように医師とAIが互いに補完し合あう時代は着実に到来しつつあり、将来的には個人が判断する医療の時代も視野に入れておく必要がある。

【コラム⑥】ウェアラブル機器からのデータによる健康・医療研究

ウェアラブルデバイスはGPSや加速度センサーを備え、心拍や運動などの物理的な生体活動、睡眠状態等を記録できることから健康・医療研究、とりわけ集団研究からの個別化医療への活用が期待されている。

韓国の大学の研究では、米国のFitbit（Google傘下Fitbit社のウェアラブルデバイス）で集めた活動量や睡眠に関連するデータから、Neural Networkを用いて認知症の前段階である認識機能障害を予測、81%の確度で予測可能との結果が得られている¹⁾。

NIHでは、感染症に罹った際に休息時の心拍が上昇する傾向があるという観察結果をベースに、Fitbitで47,249人の休息時心拍数をモニターしたデータから、2016-2018年の米国でのインフルエンザ流行発生を予測した²⁾。この結果を受けて、米国Scripps Research Instituteが、各社のウェアラブルデバイスから集めたデータより新型コロナも含めた感染症流行の追跡を行うことを狙ったDETECT（Digital Engagement & Tracking for Early Control & Treatment）という取組みを行っている。

現在はスマートウォッチ型が主流で、取得できる生体情報に限界があるが、他の様々な生体信号を取得可能なウェアラブルデバイスの開発、普及が進めば、利用の幅も広がることが期待される。

参考文献

- 1) Sang-ho Lee et al., Lifelog-based classification of mild cognitive impairment using artificial neural networks 2018 International Conference on Electronics, Information, and

Communication (ICEIC)

- 2) Cecile Viboud et al., Fitbit-informed influenza forecasts Lancet Digital Health 2020 ; 2 : e54–55

参考文献

- 1) Liang H et al., Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence Nature Medicine volume 25, pages433–438 (2019)
- 2) Jonathan G. Richens, Ciarán M. Lee & Saurabh Johri Improving the accuracy of medical diagnosis with causal machine learning Nature Communications volume 11, Article number : 3923 (2020)

5.4 今後の展望④

システムとしての生命の理解 ～複雑系・ダイナミクス理解へのAIの活用～

シュレディンガーは、生命について、「複雑な秩序ある構造体」ということで一般物体の物理が当てはまらないことをエントロピーの観点から示した（1944年）。その後クリックはセントラルドグマを提唱した（1953年）。その後の遺伝子工学（クローニング技術）の出現（1970年）により、生命を構成し、機能させる様々なビルディングブロックが発見されてきた。これまでにほとんどの重要な因子は発見されてきたといえるが、いまだにシステム（アルゴリズム）としての生命の理解は一細胞レベルでさえ解明できていない。細胞から個体のマルチスケール性については言うまでもない。システムとしての理解には、これまで生物学の中心であった生化学・分子生物学だけでは不十分で、数理、統計、情報、そして物理による解析が必要となる。

AIの潮流とともに世界的な流れとして、アンバイアスな網羅的解析・ビッグデータから新しいものを抽出してくるという方法が注目を浴びている。とくに期待が大きいのは、ゲノムを始めとした膨大に産出されるオミックスデータからの創薬ターゲット候補の発見、またこれらと疾患との関係の解明である。例えば臨床ゲノムでは、網羅的分子情報と臨床表現型との相関関係は膨大なマトリクスになり、予防医学の場合は、遺伝情報と臨床・環境情報の相関マトリクスが大きくなるので、基軸的な関係を探ることが求められている。

生命科学の今後の方向性としては、生命の普遍的機構を理解するため、マルチスケールで分子から個体までを記述する（モデル化する）流れがある。そのためには、オミックスデータ（主に次世代シーケンサー、質量分析器から産出されるゲノム、エピゲノム、トランスクリプトーム、マイクロバイオーム、プロテオーム、メタボロームなどデータ）とイメージング等画像データ（電子顕微鏡、光学顕微鏡、PET、MRI等から産出される画像やスペクトルデータ）、臨床健康記録などの他の情報ソースを統合し、定量的に時空間情報を解析することが期待されている。

今後も多くのオミックスデータが一細胞単位でとれるようになっていくことで、一細胞単位での生命や疾患の再定義が行われる。CRISPRのようなゲノム編集やiPSのような細胞初期化技術を用いて、細胞群に人為的にハイスループットに変化を与えて、タンパク質・細胞毎に変容（表現型）を見るようなスクリーニング技術も出てきている。

一方で、現状は生体（細胞あるいは細胞群）のある時点におけるスナップショットの研究になっているのが実情である。先に紹介した神経コネクトームの研究は電子顕微鏡を用いて、脳の構造を細胞レベルで可視化するものであり、Cellアトラス（細胞地図）では、臓器等を細胞種毎に一細胞単位で地図を描くことが、世界の研究トレンドになっている。生体組織（臓器）は億単位の細胞群から構成され、個別の細胞の発生段階や周期は様々である。こういったスナップショットのオミックスあるいはイメージングビッグデータから生命を理解しようという研究が盛んである。

このような組織・臓器レベルでは、単一細胞の発生の時系列推論の研究が盛んである。ある臓器の数千の単一細胞からのオミックスデータを分析し、発生時系列に沿ったこれらの細胞の順序を計算により推測する研究が行われている。70以上の推論ツールがすでに開発されているが、それらが必要とする入力とそれらが生成する出力モデルが大幅に異なっている¹⁾。

細胞レベル、つまり、細胞システムのモデル化に向かっていく研究として、以下のようなものが見られる。

遺伝子発現機構に関する研究成果を2例紹介する。米国コールド・スプリング・ハーバー研究所では、遺伝子調節（発現機構）を理解するために、ゲノミクスの問題（超並列レポーターアッセイ

(MPRA) データ) にディープラーニング技術を適用し、熱力学モデルの推論を行った²⁾。大腸菌 lac プロモーターでの転写活性化の熱力学的モデルでは、シス調節のすべての熱力学モデルは3層ニューラルネットワークとして定式化できる。第1層ノードは相互作用エネルギーを表し、第2層ノードは状態確率を表し、第3層ノードは活動を表す。ここでは、ゲノム全体のシス調節コードではなく、個々のシス調節配列のモデリングに使用するに留まっている。

UCSDのグループでは、トランスクリプトームとプロテオミクスのトレーニングデータに機械学習を使用して、ゲノム全体のエピジェネティックマーク、つまりすべての推定タンパク質コーディング遺伝子をサイレントまたは発現可能に分類した (99.4%の精度)³⁾。また遺伝子本体内のDNAメチル化パターンは、タンパク質を発現する遺伝子とRNAのみを発現する遺伝子を区別できることがわかった。本調査は特定の詳細な分子機能を含むトウモロコシ遺伝子の30,000近くのサンプルセットに基づく。

細胞内のタンパク質の網羅的解析の例を2つ示す。免疫蛍光顕微鏡法によって細胞内のタンパク質の局在を視覚化できる。これを利用して、チューリッヒ大学では、数千の単一細胞で40種類以上の異なるタンパク質の検出 (発現レベルを同時に定量化) を実現するハイスループット法を開発した⁴⁾。各細胞内のさまざまな器官、オルガネラ、および細胞構造への詳細な細胞内分布を把握することができ、そこから新しい細胞状態の識別、異なる細胞周期状態、および薬物治療における単一細胞間の細胞内組織の定量的比較を可能にする。

また、こうした顕微鏡画像から細胞内タンパク質の局在を自動で特定する取組もある。Human Protein Atlas (スウェーデン) の画像コレクション (12,390 タンパク質の画像データ) に基づいて、ディープラーニングを用いて、顕微鏡画像から細胞内タンパク質の局在を特定するコンテストが開催された。世界の2,172のチームが、タンパク質の染色された細胞といくつかの参照マーカーを調べ、タンパク質の空間分布を計算できるディープラーニングモデルの開発を競った⁵⁾。

他にも細胞の回路レベルでの統合として2例を示す。UCSDのグループでは、これまでの細胞生物学の広範な知識により、モデルの内部の仕組みと実際のシステムの仕組みを結びつける可視ニューラルネットワーク (VNN) として、真核細胞を構成する2,526個のサブシステムの階層構造をモデル化したDCellを開発している⁶⁾。文献情報をもとにつくられた細胞・遺伝子の機能を説明するための語彙 Gene Ontology を用いている。これにより遺伝子への摂動が階層を介して伝播し、それらを含む親サブシステムに影響を与え、タンパク質複合体、生物学的プロセス、オルガネラ、そして最終的には予測される応答に機能変化を引き起こし、細胞増殖表現型のレベルを理解できるとしている (図5-8)。

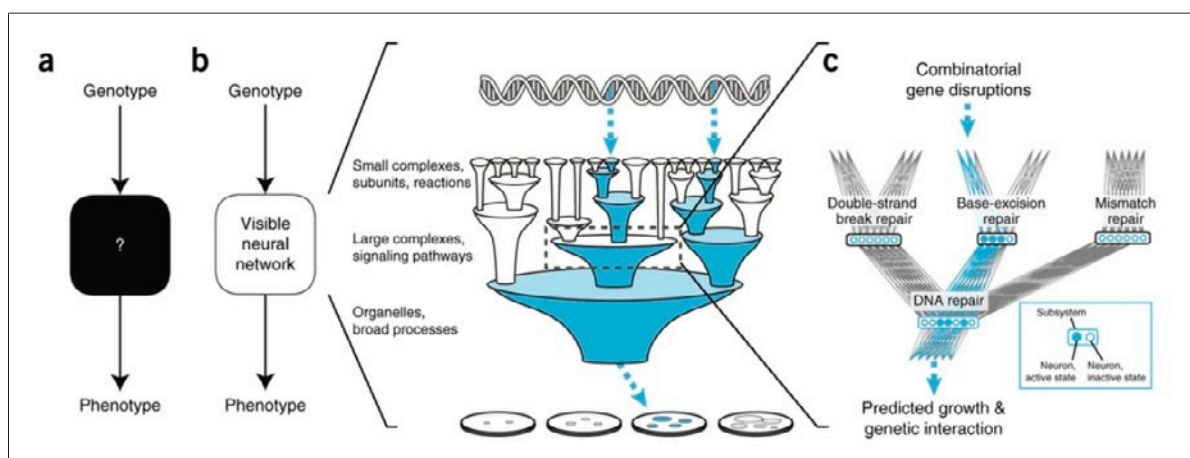


図5-8 細胞システムの構造と機能のモデリング⁶⁾

デューク大学では、人工細菌の複雑な変数間で発生する相互作用をモデル化するために機械学習、Long-Short-Term Memory (LSTM) ネットワークを使用した。細胞の増殖率、拡散、タンパク質分解、細胞運動、および遺伝子発現に関連する速度論パラメーターなど、13の異なる細菌の生命活動に関する変数を使用。1台のコンピューターでパラメーターごとに6つの値を計算するには少なくとも600年は必要だが、新しい機械学習システムでは数時間で完了することができた⁷⁾。

分子スケールと細胞のスケールをつなぐというのは、時間的にも空間的にも 10^5 の隔たりが、分子と個体では 10^{10} の隔たりがあり、これを技術的につなげていくという課題がある。生命の理解のためには、ダイナミクス（時間変遷）のデータが必要であり、今後の計測技術の発展が待たれる。

参考文献

- 1) Wouter Saelens et al., A comparison of single-cell trajectory inference methods Nature Biotechnology volume 37, pages547–554 (2019)
- 2) Ammar Tareen et al., Biophysical models of cis-regulation as interpretable neural networks bioRxiv doi : <https://doi.org/10.1101/835942>
- 3) Ryan C. Sartor et al., Identification of the expressome by machine learning on omics data PNAS September 3, 2019 116 (36) 18119-18125
- 4) Gabriele Gut et al., Multiplexed protein maps link subcellular organization to cellular states Science 03 Aug 2018 : Vol. 361, Issue 6401, eaar7042
- 5) Wei Ouyang et al., Analysis of the Human Protein Atlas Image Classification competition Nature Methods volume 16, pages1254–1261 (2019)
- 6) Jianzhu Ma et al., Using deep learning to model the hierarchical structure and function of a cell Nature Methods volume 15, pages290–298 (2018)
- 7) Shangying Wang et al., Massive computational acceleration by using neural networks to emulate mechanism-based biological models Nature Communications volume 10, Article number : 4354 (2019)

5.5 今後の展望⑤ 脳神経科学ベースの次世代AIアルゴリズムの可能性

ディープラーニングだけで産業・社会的な課題の全てが解決できるわけではなく、より広汎に用いることができるAI技術のために、ディープラーニングを超えたAIのアルゴリズムが求められている。新しいAIのモデルは必ずしも脳のモデルと同じである必要はないが、脳のアルゴリズムと実装の理解から新たなAIの実装が生まれてくることも期待されている。

視覚の分野では、計算論的神経科学（脳を情報処理機械に見立ててその機能を調べる）の源流を創ったDavid Marrが、神経系がどのように情報を表現しているかを考える上での枠組みとして、計算理論（何を計算しているのか）、アルゴリズム（計算の手続き）、神経回路（ハードウェア）実装（どうやって実行するのか）の3つの階層からなる脳の階層モデルを1976年に提唱した。HubelとWieselが大脳皮質に視覚中枢を発見し皮質領野による視覚情報の処理を明らかにしたことや、脳の経験依存学習の知見と層構造を取り入れたパーセプトロンの概念を合わせ発展させて福島邦彦が提唱したネオコグニトロンは現在のディープラーニングの基盤となった。

強化学習の分野では古典的条件づけやオペラント条件づけなど心理学・脳科学の分野からのコンセプトを抽象化する形で、BertやSuttonらがTD学習と呼ばれる近代的な強化学習の理論を提唱したことに始まり、ドーパミン細胞を中心とした働きの解明や大脳基底核の強化学習モデルの提唱が行われ、その検証のための実験が行われた、という流れがある。

このようにAIの発展には脳科学との相互的な知見やアイデアのやり取りが大きく貢献している²⁾。

現在使われているディープラーニングは莫大な量の教師データが必要であり、画像・パターン認識を基本とする医療診断技術やゲームなどでは強力な力を発揮するが、教師データなしで、状況を認識し判断を下す能力はない。ヒトの幼児は一度犬に吠えられただけで危険な生き物であるという認識を形成できるが、そのようなわずかの学習による認知機能などは欠落している。脳はどのようにして少数データから学習しているのか、という解答が心待ちにされている。文脈に依存する情報処理もまた現行AIでは解けないが、脳はどの様に問題を抽出してやるべきことを決定しているか、という問題も脳科学からの知見が待たれている。

未知のヒトの脳のアルゴリズムの解明により、新たなAIが産み出されることが期待される。今後の課題として脳のエネルギー効率、データ効率、脳内シミュレーション、モジュール性（モジュール自己組織化）や能動性、intuitive physics、transfer learning、imagination/planningなどが挙げられる。

国際的にも次世代AIを目指す組織が活動を開始している。MITのBrains、Minds and MachinesとNYU（ニューヨーク大学）の研究者を中心にしたCanonical Compilation In Brains and Machinesが、日本では東京大学のIRCNがそれにあたる。また、MITは2018年2月にMITの研究者を糾合して発達心理学・認知科学分野から次世代AIを創出しようとする「MIT IQ」プロジェクトを発表した。

米国では2018年にDARPA（Defense Advanced Research Projects Agency）からAI Next Campaignが発表された。次世代AI研究に20億ドル以上の大型投資を実施する。すでに人工知能分野で大きな存在感を示している中国は、2017年7月に国務院が「次世代人工知能発展計画」を発表し、2030年までに中国のAIの理論・技術・応用すべての分野を世界トップ水準に引き上げる目標を掲げた。欧州の研究者も米国や中国の次世代AI研究に対抗するために2018年12月に欧州の関連研究者のコンソーシアムを立ち上げた。

神経科学とAIの接点について最新の研究を紹介する。

東京大学では、目で見た画像に対して視覚野の神経細胞がどのように反応するかを忠実に再現するように、画像と神経細胞の反応（2光子カルシウムイメージング法）の対応関係をコンピュータ上で深層ニューラルネットワークに写し取る手法を開発した³⁾。本手法をマウス一次視覚野の神経細胞の解析に適用し、各神経細胞がどのような画像に対して最も強く反応するかを仮説フリーかつ網羅的に可視化することができた。人間や動物の振る舞いに近いAIの開発が期待される。

脳は多数のニューロンにより構成され、さまざまな情報はニューロンの組み合わせとその活動パターンでコードされると考えられている。こうした活動は発火シーケンスとして知られ、記憶や学習を始めとする高次機能の発揮に不可欠であると考えられている。東京大学の研究グループでは、大規模スパインイメージング法に基づき、海馬ニューロンが受けるシナプス入力を大規模に可視化することで、近傍のシナプスが特定の順番で反復入力（シーケンス入力）を受けることを発見した⁴⁾。新たな脳内アルゴリズムの発見と言え、将来のAI開発への実装も期待される。

脳神経科学の研究者には、次のブレークスルーは自律性を持ったAIだという考え方もある。このためには人が定義した範囲だけではなく、AIが自律して探索空間を拡張していく必要がある。

関連して、将来的にAIが人間と同様の「意識」を持つことが起こりうるのかという問題がある。人の意識や意志決定のメカニズム研究は有効な方法論に限界もありなかなか進んでいない。我々の意識は、結局物理法則に従っているだけなのかという論点もあるが、いまだ科学的に証明されているわけではない（しかしこれに反論する有望な仮説が見当たらないのも事実である）。

スタンフォードやユタ大学等のグループでは、パルス超音波がアカゲザルのスクリーンの左と右に表示されたターゲットのうちどちらを見るかの意思決定に影響を与えることを明らかにした³⁾。具体的には、サルの前頭眼野を刺激した際、サルたちは右のターゲットに目をやることが多く、逆の場合も同様であった。サルたちが超音波に影響されてターゲットに目をやった確率は倍以上だった。この超音波療法を運動皮質（運動を司る脳の領域）に照射しても何の影響も生まなかったという点で、つまり身体的な反射を引き起こしていたのではなく、実際に知覚による選択に影響を与えていることを示唆している。

ウィスコンシン大学では、サルの脳にマルチ電極と記録を行う実験により麻酔状態を維持しているサルのどこをどう刺激すればサルが覚醒するか、260箇所を刺激を繰り返した結果、視床の中央部を50Hzの高い周期の電流で刺激したときだけ、脳各部の高い周波数の活動が上がり、また行動記録からも覚醒していることがわかったとしている⁴⁾。視床の中心部の刺激から、皮質表層、皮質深層へ投射する神経を介して興奮が伝わり、これが次の引き金になって皮質深部と表層、そして逆に視床間の θ 、 α 波を介した相互作用が起こり、意識が維持されるという仮説を提唱している。

人間は合理的で妥当な選択を行わず、理屈に合わない行動をとってしまう場合がある。心理学的には、人が複雑な問題解決などのために、意思決定を行うときに、暗黙のうちに用いている簡便な解法（ヒューリスティックス）を指し、判断結果に一定の偏り（バイアス）を含んでいることが多い。

量子認知学の分野では、こうした人間の思考のゆらぎは、量子力学の確率論による計算と一致すると言われている。そのような研究の一例として、中国科学技術大学の研究チームでは、脳内スキャンによって、量子的作用に関連する可能性の高い脳領域を特定した⁵⁾。今回の研究では、脳の活動領域を比較するために、非喫煙者と喫煙者が使われたが、これが本当に量子的な脳神経の作用を反映したものとして妥当であるかは、さらなる研究が必要であろう。

こうした意識や思考に関する研究は実験を検証する方法論が確立しておらず、どれだけ本質的に対象に迫っているかは曖昧模糊としているところが課題でもある。

次のフロンティアは、人の認知と機械を融合させることだと考えている研究者も多い。脳を制御する技術はどの程度進んでいるだろうか。

ハーバード大学では、脳にメッシュ電極インターフェースを移植し、脳からデータを集めて神経細胞がコミュニケーションする様子を調べ、それを元にして、必要に応じて患部に刺激を与えて、新しく形成された神経細胞を脳の損傷を受けた部位に移動するよう促し、治療する試みを発表した⁶⁾。刺激を与える電極は超小型で、細胞かそれより小さなものとなっている。

以上の動向に鑑みると、上述したような脳神経科学にインスパイアされた自律型AIの誕生にはもう少し時間がかかりそうである。

【コラム⑦】 第4世代AI

第4世代AIの方向性は「深層学習（即応的 AI）と知識・記号推論（熟考的 AI）の融合による、社会に適合し、人に寄り添って成長する次世代AI」と考えられている。

深層学習は脳科学ベースであり、知識・記号推論はエキスパートシステムやデータマイニングの流れを引く記号処理ベースの方法論である。

これについて、CRDSでは、下記図5-9の歴史認識に基づき、戦略プロポーザル「第4世代AIの研究開発—深層学習と知識・記号推論の融合—」という提言を行った¹⁾。

そこでは、以下に示す4つの重要な研究開発課題への取り組みを推進することを提案している。

1. 教師なしでも知識獲得・成長できる融合AI研究
2. 実世界で安全性・頑健性を確保できる融合AI研究
3. コンテキストに応じ適切な推論・アクション可能な融合AI研究
4. 知能の要件・原理の探究に基づき、即応的知能と熟考的知能を統一的に扱う融合AI研究

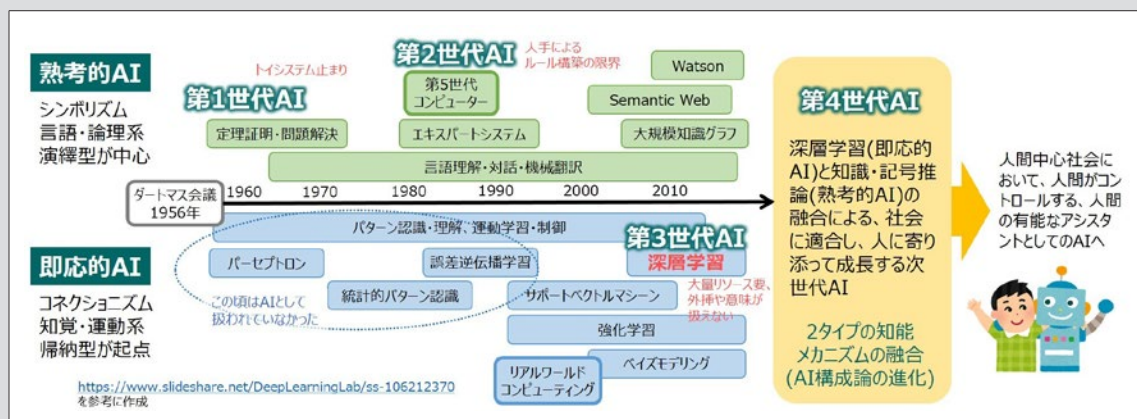


図5-9 AI研究の2つの流れと次なる進化

参考文献

- 1) JST研究開発戦略センター 戦略プロポーザル「第4世代AIの研究開発—深層学習と知識・記号推論の融合—」／CRDS-FY2019-SP-08

参考文献

- 1) JST研究開発戦略センター 調査報告書 ドライ・ウェット脳科学／CRDS-FY2019-RR-06
- 2) Demis Hassabis et al., Neuroscience-Inspired Artificial Intelligence Neuron, VOLUME 95, ISSUE

- 2, P245-258, JULY 19, 2017
- 3) Jan Kubanek et al., Remote, brain region-specific control of choice behavior with ultrasonic waves Science Advances 20 May 2020 : Vol. 6, no. 21, eaaz4193
 - 4) Michelle J. Redinbaugh et al., Thalamus Modulates Consciousness via Layer-Specific Control of Cortex Volume 106, Issue 1, 8 April 2020, Pages 66-75.e12
 - 5) Ji-An Li et al., Quantum reinforcement learning during human decision-making Nature Human Behaviour volume 4, pages294–307 (2020)
 - 6) Shaun R. Patel et al., Precision electronic medicine in the brain Nature Biotechnology volume 37, pages1007–1012 (2019)
 - 7) JST研究開発戦略センター 戦略プロポーザル. 第4世代AIの研究開発—深層学習と知識・記号推論の融合—/CRDS-FY2019-SP-08

終わりに 日本への示唆

ここまで示してきた様々な近年の研究動向から、AIを用いた生物医学研究の進展を掴むことができた。バイオロジー、ライフサイエンス分野は、2010年頃を境に、次世代シーケンサーの急速な技術進展に伴う一細胞トランスクリプトーム技術の浸透、ゲノム編集やオプトジェネティクスのような簡便なゲノム制御技術の出現、さらにAI技術の発展、といった先端技術群の出現により、従来の研究からフェーズチェンジが起こった。研究により生み出されるデータセットはあまりに膨大で、ヒトの脳ではもはや全てを解析できない。

ハイスループットに産出されるデジタルデータによって実験の設計を行い、これらを機械学習・ディープラーニングでハイスループットに解析するといったデジタルバイオロジー、コンピュータショナルバイオロジーの側面が大きくなってきており、今後5年、10年スパンで考えると、あらゆる研究分野で、こうしたデータ、デジタル、コンピュータが研究の中心になっていく変容の方向性は十分に予期される。この理由として、以下が挙げられる。

- ①ヒト等の生命システムやヘルスケア研究などは、研究対象が絶えず変動するオープンな複雑系システムであり、そういったシステムは確率論で論じることが妥当であることから、ビッグデータや統計といったデータ駆動型と相性が良いこと、
- ②人間は直接的には2.5次元を超える世界を完全には認識できないが、計算機はある意味無限次元からパターンを抽出できること、
- ③計測技術や計算機技術の進展によって生命の時空間に亘ってデジタルデータがハイスループットにとれるようになったことにより生命科学データの爆発的に増加したこと、

新型コロナウイルス感染症の問題があろうとなかろうと時代の要請に応じて、科学技術・研究開発の現場において、デジタルトランスフォーメーションを進めていくことは自然の流れであろう。

こうした変化により、現代の生物医学研究の多くにとって、伝統的な研究組織、つまり比較的小さなラボ単位で閉じて研究を進める方法は限界になってきている。今後バイオ研究は、目的に応じて、皆でデータを持ち寄る、あるいは高度に役割分担された異分野連携による大規模チーム型・コラボレーション型の研究が主流になっていくだろう。

戦後の分子生物学の勃興により、生命を構成し、機能させる様々なビルディングブロックが発見されてきており、これまでにほとんどの重要な因子は発見されてきたといえる。一方で、生命のシステム（アルゴリズム）はほとんど理解されていないに等しい。要素還元的な研究からの脱却として、生命システムのアルゴリズムと物理化学法則の関係を解き明かしていくという課題が残されている。細胞内での液液相分離、分子モーター（いずれも熱力学的性質）、あるいは生物に現れるチューリング波（反応拡散系）等が生命が物理法則に従っている例として捉えられている。

今後、膨大な計測・実験結果から帰納的に普遍的と思われる法則を導くという研究活動がより重要になってくる。このために、AIだけでなく、生命科学、化学、数理統計、物理などの様々な研究者が連携して研究を進めていかないと新しい発見が難しい時代になっていくだろう。このような時代には高度研究支援人材やPIコースにはのらないスタッフサイエンティストのような人材が研究者と同等の立場で研究を進めて行く形が考えられる。

このような時代に国や公的な研究機関は何をすべきかを考える必要がある。

日本ではこのような研究のフェーズチェンジ（変容）を先導するどころか、追従することさえ難しくなっている。AI×バイオ分野の研究でも日本は他国に大きく遅れをとった。日本はノーベル

賞受賞にも見られるように、生化学・分子生物学的実験手法は伝統的に強く、いまビッグデータの主戦場となっているゲノム、タンパク質も、神経科学等の分野においても、過去にはゲノムプロジェクト、タンパク 3000、FANTOM（網羅的トランスクリプトーム解析）等のような大規模プロジェクトの結果、2010年以前までは国際的に先導的な役割を果たしてきた。一方で、研究がデジタル化、ハイスループット化し、データ解析が重要（中心）になってくるに従って国際的にほとんど存在感が見られなくなっている。このような分野はデータ科学者の参画のみならず、先端計測機器への集中投資（初期投資として数十億円／拠点）も必要であり、日本でこのような資金を擁する仕組みが持っていないことも大きい。必要とされる技術の変遷も目まぐるしく、このような状況では、政府が単独でAI研究・開発を主導するのは現実的でなく、企業あるいはベンチャーとアカデミアの相互互惠的な連携が必須ではないかと考えられる。今後ますますデータの質と量が核となるが、ものによっては国際コンソーシアムで実施すべきことを検討すべきであろう。

AIやロボットの研究現場への実装は今後もますます進展していくだろう。特に分子スケール、細胞ベースでのアッセイなどは、実装コストの面的問題があるものの、これまでの大きな課題である規格化された、時間分解能の高いデータを大量に得ることに適している。10年以内にはオルガノイドなどを含む組織、臓器の研究にまで普及していくだろう。AIや高度計測技術を用いたシミュレーションでは、生命のシステムとしての研究が進展していくことが期待され、一細胞内、一臓器内のレベルから、分子から個体までのマルチスケールモデル化研究が精緻になっていくことが予想される。

一方で、生物研究は非常に広範にわたることから、要素還元主義の弊害として研究分野の細分化（サイロ化）が起こっており、生物を多角的な視点でシステムとして広く見ることでできる研究者が減少していることも懸念されている。データ科学は、個別の事象をつないでマクロにもものを見ることに適している。今後の我が国の研究者人口の減少も踏まえると、データ科学的な視点は不可欠である。ライフサイエンスもビッグサイエンス化が進み、20を超えるような多分野の多国籍の研究室がデータを持ち寄って、シェアして論文を出すことや、多くの人が1つの成果を生み出すために広くかわる形になっていく分野も増加していくだろう。こうした時代に論文のオーサiershipや個々の業績の評価の取り扱いに関する問題もある。生命科学研究のカルチャーを変え、様々な形で貢献した人材が評価される仕組みをつくらなければならないだろう。

海外も含め人材の流動性を上げることや新しい科学技術をよく知る若手をPI等に登用することが理にかなっているが、日本のアカデミアにおけるシステムや雇用制度がボトルネックになっている。

社会に近いところでは、日本においても医師がAIの助けを借りてより良い医療行為を行うことが通常になるという近未来が予想できる。そのような環境に適応的な医師を育てていくための教育も必須である。医療経済という観点から言えば、人間とAI・ロボットの高度な分業化が必須になり、そもそもの病院のあり方も変わっていくと思われる。バイオテクノロジーとICT等が発展実装されていくに従って、個人の医療（個人が判断する医療）というものも浸透していかざるを得ないようになるだろう。それに即した法制度を検討しないといけないであろう。

AIが医療等で一般向けに社会実装されていく分野においては、ELSIについて深く考慮しなければならない。中でも当座の問題は個人のプライバシー保護と説明責任の問題であろう。前者について、多くのデータ・情報をサービス側（病院や情報サービス機関等）に提供することが求められるが、そうした情報が第三者の手にわたったり、目的外に使用されると個人の権利が侵害される。後者は、AIを用いたサービスにおいて判断の材料として用いられる入力（データ）、判断の論理（アルゴリズム）等は、ステークホルダーの要求に応じて開示され、検証の対象とならなければならないということである。AIの有効性と安全性確保のためには、当面は人間が最終判断を下す必要があると考えられる。

今後の技術の進展と人々の価値観の変化を注視しておく必要がある。

もちろん初等中等高等教育におけるデータ科学やELSIの教養が重要になることは言うまでもない。AI開発や活用にはあらゆる層の人材への投資が必須である。

国はこうした時代における教育・人材育成と科学研究環境の変容を促すことに政策的力点を置くべきではないだろうか。

CRDSで先に発行した、調査報告書「研究力強化のための大学・国研における研究システムの国際ベンチマーク～米国、英国、ドイツおよび日本の生命科学・生物医学分野を例に海外で活躍する日本人研究者に聞く～」(2019年8月発行)でもこうした研究システムや土壌の改革に言及しているので参照されたい。

付録

付録1 検討の経緯

本報告書は、これまでのCRDSの調査活動において、下記の識者へのインタビューやワークショップでの発表から得られた情報も参考にしている。

情報・資料提供

画像データ

石井信（京都大学 大学院情報学研究科 教授）
岡田康志（東京大学 大学院理学系研究科 教授／理化学研究所 生命機能科学研究センター）
浜本隆二（国立がん研究センター 研究所 分野長／理化学研究所 革新知能統合研究センター）
横田秀夫（理化学研究所 光量子工学研究センター チームリーダー）

化合物データ

山西芳裕（九州工業大学 情報工学研究院 教授）

オミックスデータ

川上英良（千葉大学 大学院医学研究院 教授／理化学研究所 科技ハブ産連本部）

医療データ

永井良三（自治医科大学 学長）

AI・機械学習（ディープラーニング）

津田宏治（東京大学 大学院新領域創成科学研究科 教授／理化学研究所 革新知能統合研究センター）
福島俊一（JST研究開発戦略センター（CRDS））

JST研究開発戦略センター（CRDS）

- 1.（研究開発の俯瞰報告書）ライフサイエンス・臨床医学分野（2019年）／CRDS-FY2018-FR-04
- 2.（The Beyond Disciplines Collection）デジタルトランスフォーメーションに伴う科学技術・イノベーションの変容／CRDS-FY2020-RR-01
- 3.（調査報告書）研究力強化のための大学・国研における研究システムの国際ベンチマーク～米国、英国、ドイツおよび日本の生命科学・生物医学分野を例に海外で活躍する日本人研究者に聞く～／CRDS-FY2019-RR-03

■ 作成メンバー ■

○ 島津 博基	フェロー	(ライフサイエンス・臨床医学ユニット)
高島 洋典	フェロー	(システム・情報科学技術ユニット)
永野 智己	総括ユニットリーダー	(企画運営室)
福島 俊一	フェロー	(システム・情報科学技術ユニット)
宮藺 侑也	フェロー	(ライフサイエンス・臨床医学ユニット)
(監修)		
谷口 維紹	上席フェロー	(ライフサイエンス・臨床医学ユニット)
永井 良三	上席フェロー	(ライフサイエンス・臨床医学ユニット)

○主担当

※お問い合わせは下記までお願いいたします。

CRDS-FY2020-RR-03

The **Beyond Disciplines** Collection

AI× バイオ DX時代のライフサイエンス・ バイオメディカル研究

令和2年9月 September, 2020

ISBN 978-4-88890-682-1

国立研究開発法人科学技術振興機構 研究開発戦略センター

Center for Research and Development Strategy, Japan Science and Technology Agency

〒102-0076 東京都千代田区五番町7

電話 03-5214-7481

E mail crds@jst.go.jp

<https://www.jst.go.jp/crds/>

©2020 JST/CRDS

許可無く複写／複製することを禁じます。

引用を行う際は、必ず出典を記述願います。

No part of this publication may be reproduced, copied, transmitted or translated without written permission.

Application should be sent to crds@jst.go.jp. Any quotations must be appropriately acknowledged.

ISBN 978-4-88890-682-1

