令和 6 年度　厚生労働行政推進調査事業費補助金（化学物質リスク研究事業）
分担研究報告書

分担研究課題名：　深層学習による QSAR モデルの深化に関する研究

研究分担者　水野忠快　　東京大学　助教

**研究要旨**

本研究では、深層学習を用いてQSARモデルを深化することを目的としている。深層学習モデルのQSAR応用としては、分子構造をグラフや文字列表現として入力して化合物構造を扱うニューラルネットワークの研究が急速に拡大している。一方、毒性予測モデル分野において、モデル構造の整理や評価指標の統一といった体系的な整備が進んでおらず、まずは分野横断的な知見の集約が求められている。

そこで2024年度の本研究では、深層学習による毒性予測に関する文献サーベイを行い、研究領域の全体像と課題の明確化を試みた。PubMedおよびプレプリントサーバーarXivから2015年以降に発表された論文を収集し、タイトルと要旨に基づいて、「研究論文であり、分子構造のグラフまたは文字列表現に基づく深層学習モデルを用いて、ヒトに関係する毒性あるいは毒性関連アッセイの予測を行っている」論文に絞り込んだ。

最終的に155件の論文が抽出され、そのうち63件がグラフやSMILES表記を入力とする深層学習モデルを採用していた。入力形式とモデル構造は図1に示す通りであり、グラフを用いるモデルは、例外2件を除き、すべてMessage-Passing型のアルゴリズムに基づいていた。SMILESを入力とするモデルは、事前学習を行ったEncoderからの転移学習形式、またはEnd-to-Endで構成されるものに大別された。

予測対象は、臨床毒性ではなく、主に前臨床段階の毒性関連アッセイ結果であり、これは利用可能な大規模データセットの整備状況を反映していると考えられる。また、論文数は年々増加しており、特に2023年以降は、ChatGPTに代表される大規模言語モデルの登場以後の技術波及により、急増している傾向が確認された。一方で、不均衡データに対してAUROCのみで性能評価を行う、あるいは評価指標の誤記・誤用といった統計的な問題も複数の論文で見受けられた。これらは、モデルの信頼性や再現性の観点から課題である。

2024年度での本研究は、深層学習による毒性予測という研究領域の現状と傾向を明らかにするとともに、今後の研究展開に向けた課題の洗い出しを目的とした。現時点では、モデル開発および性能評価の妥当性確保に向けたリテラシー向上が急務であり、分野横断的な基盤整備が求められる。

## A．研究目的

創薬のさらなる効率化と低コスト化が求められる中、機械学習技術を応用した毒性予測への期待が急速に高まっている。特に近年では、分子構造を表現するグラフや SMILES といった文字列表現を入力とするニューラルネットワークに基づいた深層学習手法が急速に進展しており、情報科学を中心とした分野において、化合物を対象とする機械学習モデルの研究開発が活発に行われている。

こうした状況を踏まえ、本分担課題では、深層学習を活用した QSA モデルの深化を通じて、毒性予測の信頼性と有用性を高めることを目的とする。研究開始当初は、医学薬学領域における深層学習を応用した毒性予測モデルに関して、モデル構造の整理や入力形式の体系化といった基本的な情報の整備が十分に進んでいない状況であった。そこで 2024 年度の研究では、まず深層学習を活用した毒性予測に関する学術的知見の収集と現状把握を目的とし、包括的な文献調査を実施することとした。

## B．研究方法

まず医学論文データベースPubMedとプレプリントサーバーarXivから下記の検索クエリで2015年以降の論文を取得した。

Query: ( drug / chemical ) & ( toxicity ) & ( neural network / deep learning ) & (graph / SMILES / strings)

次にタイトルと要旨から、「研究論文であり、グラフ、または文字列表現由来の記述子を利用した深層学習モデルを用いて、人間に関係する毒性、並びに毒性関連アッセイ結果の予測を行っている」ものに絞り込んだ。候補論文を対象に内容、特にモデルへの入力形式やモデルの種類、予測対象や評価指標等を確認し、収集した。

（倫理面の配慮）

本研究では動物を用いた実験は実施しない。

## C．研究結果

最終的に155件の論文が抽出され、そのうち63件がグラフおよびSMILESを入力とした深層学習モデルを用いた毒性予測を行っていた(図1)。これらの論文における入力形式とモデルの種類に関する分析を行った結果、グラフを用いるモデルのほとんどが、例外的な2件を除いてすべてMessage-Passingアルゴリズムを基本構造として採用していた。グラフニューラルネットワークにおいては、同一層で繰り返し情報集約を行うRecGNNではGRUを活用したGGNNが用いられていた。一方、異なる層で集約を進めるConvGNNでは、2020年以前はGCN（Kipf et al., 2017）の利用が主流であったが、それ以降はGIN（Xu et al., 2019）やChemprop（Yang et al., 2019）に代表されるD-MPNNなど、モデルの多様化が進んでいた。中でも集約時にattention重みを導入するGAT（Velickovic et al., 2017）の使用頻度が高く、近年の注目を集めていることが明らかとなった。

SMILESを入力とするモデルにおいては、GRUやLSTMといったRNNに加えて、並列計算が可能なTransformer（BERTを含む）や1D-CNNなど、多様なアーキテクチャが使用されていた。これらのモデルは、学習形式に基づいて、対象データのみでモデル構築を完結するEnd-to-End型と、大規模な化合物データベースで事前学習したEncoderに予測器を接続し、転移学習（Encoderの重みを固定したfreeze学習またはfine-tuning）を行う型に大別された。後者では予測器にニューラルネットワークだけでなく、ランダムフォレストなどの古典的機械学習手法を併用する事例も確認された。

また、グラフおよびSMILESの両方を活用するマルチモーダル型モデルも一定数存在しており、それぞれの表現によって抽出された特徴ベクトルを結合し、場合によっては次元削減を行った上で予測器に入力する構成が一般的であった。

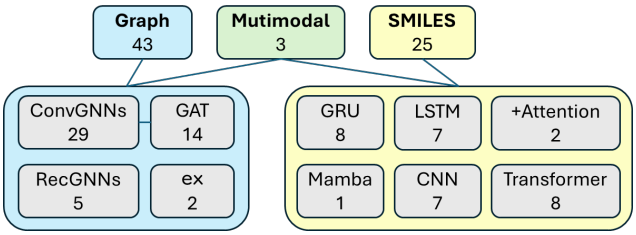


図1. モデル種別

性能評価指標についてまとめる(図2)。予測手法としては2値分類が主流であり、63件中59件が該当した。分類モデルの評価にはAUROCおよびaccuracyが多く用いられていたが、これらの指標はクラス不均衡に対して過大評価されやすく、実際にAUROCのみによって性能を報告している論文が17件確認された。特に少数クラスが10%以下である場合に、これらの指標はモデルの性能を適切に評価できないことが知られており、MCC（マシューズ相関係数）など、陽性・陰性双方の予測性能を反映する評価指標の利用が望まれる。毒性予測は創薬初期のスクリーニング段階で用いられることが多く、毒性のある化合物を取りこぼさずに検出することが重要である。その意味で、sensitivityの使用頻度が高かったことは妥当であり、続いて多く用いられていたF1-scoreも、precisionとsensitivityのバランスを取った実用的な指標であるといえる。

一方、回帰モデルについては4件に留まっていたが、$R^2$（決定係数）が主要な評価指標として用いられていた。学習時に損失関数として用いられるRMSEやMAEをそのままスコアとして報告する事例も多く見られた。なお、線形回帰においてはピアソンの相関係数Rの二乗が$R^2$と一致するが、非線形性を持つ深層学習モデルでは一致しない点に注意が必要である。16件の$R^2$を報告する論文のうち、$R^2$を「決定係数」と明示したのは1件のみであり、その他は「ピアソンの相関係数Rの二乗」として記載していた。また、Rの二乗であるにも関わらず、負のスコアが示されている論文も存在し、評価指標の取り扱いに関するリテラシーの向上が研究の信頼性を確保する上で喫緊の課題である。

| Classification | 59 |
|---|---|
| AUROC | 46 |
| AUPRC | 8 |
| Accuracy | 29 |
| Balanced Accuracy (BA) | 10 |
| Sensitivity (Recall・TPR) | 25 |
| FNR (1-TPR) | 1 |
| Precision (PPV) | 17 |
| Specificity (TNR) | 11 |
| FPR (1-TNR) | 2 |
| NPV | 1 |
| F1 | 21 |
| MCC | 18 |
| **Regression** | **21** |
| $R^2$ | 16 |
| R (Spearman・Pearson) | 3 |
| RMSE | 9 |
| MAE | 9 |

図2. 評価指標の分布

予測対象としては、「毒性関連バイオアッセイ結果（31件）」と「毒性試験・臨床毒性（35件）」に大別された。前者ではTox21やToxCast、hERGなど、多数のアッセイ結果を含む大規模なデータセットが利用されており、対象とするアッセイ数も多かった。後者では、肝毒性や心毒性、変異原性、生殖毒性、LD50など、対象が一種であるケースが多く、データセットの入手が容易でないことから、複数のデータベースを組み合わせた独自データセットの構築が多く見られた。同一名称のデータセットであっても出典が異なることで化合物数に相違があることも確認され、たとえばClinToxデータセットではMoleculeNetでは1484件、TDCでは1478件の化合物が登録されていた。出典を明記していない論文も一部存在し、再現性の確保のためにはデータセットの明示と統一的な取り扱いが不可欠である。

## Ｄ．考察

発表年別の論文数を見ると、近年の深層学習の普及とともに関連研究が急増している。特に 2023 年以降に顕著な増加が認められており、ChatGPT など大規模言語モデルの一般公開など深層学習の認知度との関連も考察される()。一方で、不均衡データに対して AUROC のみで性能評価を行うなど、統計学的に不適切な指標の使用や記述の誤りも散見された。

また対象論文の約半数が開発したモデルを GitHub 等で公開していたが、残る半数以上はモデル非公開であった。中にはユーザー利便性を考慮してウェブアプリケーションとして提供している例もあったが、公開から１年未満でアクセス不能となっているケースもあり、持続可能な情報公開の在り方が課題として浮かび上がった。

## Ｅ．結論

本研究では、深層学習を用いた毒性予測に関する学術文献の包括的なサーベイを通じて、研究領域の現状と課題を明らかにした。活発な研究開発が進む一方で、性能評価指標の誤用やデータセットの不統一、再現性の欠如といった問題点も明確となった。今後、深層学習モデルの信頼性と妥当性を担保するためには、予測対象に応じた適切な評価指標の選定、データ出典の明示、ならびに研究者間のリテラシー共有が不可欠である。

## F. 健康危険情報

該当なし。

## G. 研究発表

### １．論文発表

（発表誌名・頁・発行年等も記入）

なし

### 2.　　学会発表

1. 第7回医薬品毒性機序研究部会, 李澤昇、水野忠快、根本駿平、楠原洋之,“医学薬学領域における深層学習を用いた化合物毒性予測研究の現状と課題”, 2025/1/8, グランシップ静岡, 静岡
2. 日本薬学会第145年会, 李澤昇、水野忠快、根本駿平、楠原洋之,“深層学習を用いた化合物毒性予測モデルのサーベイと体系化”, 2025/3/27, マリンメッセ福岡, 福岡

## H.知的所有権の取得状況

### １．特許取得

なし

### ２．実用新案登録

なし

### ３．その他

なし