令和 7 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、「深層学習モデルによる QSAR の深化」の一環として、SMILES 表記揺れが化学言語モデルを用いた QSAR 解析の信頼性に与える影響を検討した。SMILES を入力とする深層学習モデルは、毒性予測や化合物物性予測に広く利用されている。一方、Canonical SMILES であっても、使用するデータベースやソフトウェアにより同一化合物が異なる文字列で表されることがある。この表記揺れは、文字列を直接入力とする化学言語モデルにおいて、予測結果や潜在表現に影響する可能性がある。

本年度は、文献調査、公開データセット解析、分子翻訳モデルを用いた検証を行った。PubMed で抽出した化学言語モデル関連論文 264 件を調査した結果、対象となった 131 件のうち約半数で SMILES の canonicalization 手順が明記されていなかった。MoleculeNet、および TDC の解析では、同一化合物に対する複数の SMILES 表記、Kekulé 表記と aromatic 表記の混在、立体化学情報の欠落が確認された。

分子翻訳モデルを用いた実験では、Raw SMILES を入力した場合、標準化した SMILES と比べて構造再構成性能が低下し、潜在表現にも明確な差が生じた。一方、物性・毒性関連の分類タスクでは、表記揺れの影響が下流モデルの特徴選択により見かけ上吸収される場合があった。さらに ClinTox および BBBP では、特定の表記形式がラベルと偶然相関し、予測性能を人工的に高く見せる「表記レベル交絡」が認められた。

以上より、SMILES表記揺れはQSARモデルの性能評価、再現性、解釈性に影響する重要な要因であることが示された。深層学習QSARモデルの信頼性を高めるためには、使用ツール、バージョン、canonicalization設定、および立体化学情報の扱いを明示することが不可欠である。

A．研究目的

創薬、および化学物質リスク評価では、機械学習を用いた QSAR モデルによる毒性予測の高度化が期待されている。近年は、分子構造をグラフや SMILES 文字列として入力する深層学習モデルが急速に発展しており、毒性関連アッセイや化合物物性の予測に応用されている。

本分担研究では、深層学習モデルによる QSAR の深化を通じて、毒性予測モデルの信頼性と有用性を高めることを目的としている。特に本年度は、SMILES を入力とする化学言語モデルに着目し、同一化合物に対する表記揺れがモデルの潜在表現、および予測性能に与える影響を解析した。SMILES 表記揺れは、単なる前処理上の差異ではなく、文字列を直接学習するモデルにとって入力情報そのものの違いとなる。そのため、表記揺れが QSAR モデルの性能評価を歪める可能性を明らかにすることは、化学物質リスク評価に用いる AI モデルの妥当性を確保する上で重要である。

B．研究方法

化学言語モデル研究における SMILES 標準化手順の報告状況を把握するため、PubMed から「SMILES AND strings」により 264 件の論文を抽出し、PRISMA の考え方に基づいてスクリーニングした。機械学習と無関係な論文、SMILES を主要入力としない論文、グラフ構造のみを扱う論文、記述子ベースの QSAR 研究を除外し、化学言語モデルを用いた 131 件について、canonicalization 手順、使用ツール、バージョン、立体化学情報の扱いを確認した。次に MoleculeNet、および Therapeutics Data Commons（TDC）を対象に、公開データセット中の SMILES 表記揺れを解析した。データセットに登録された SMILES を Raw SMILES とし、RDKit v2024_03_6 により canonical=True、isomeric=True で変換したものを Standardized SMILES とした。両者を比較し、同一化合物に対する複数表記、芳香族環の Kekulé/aromatic 表記の混在、立体化学情報の欠落を評価した。

Transformer encoder と VAE を組み合わせた分子翻訳モデルを用い、上記の影響を評価した。具体的には Randomized SMILES を入力、Canonical SMILES を出力として学習したモデルに対し、Raw SMILES および Standardized SMILES を入力して構造再構成性能を比較した。評価指標には完全一致率および文

字レベルの部分一致率を用いた。潜在表現については、Raw SMILES と Standardized SMILES 間の Levenshtein 距離と潜在ベクトル間距離を算出し、t-SNE により可視化した。

下流予測性能の評価では、MoleculeNetおよびTDCの二値分類タスクを対象とし、分子翻訳モデルのencoderから得た潜在表現をXGBoost分類器に入力した。Raw SMILES由来の潜在表現とStandardized SMILES由来の潜在表現でAUROCを比較した。さらに、XGBoostの特徴重要度を用いて、表記揺れの影響が下流モデルの特徴選択で吸収されるかを検討した。ClinToxおよびBBBPについては、Kekulé/aromatic表記不一致とラベルの関連を解析し、表記上の特徴が性能評価を押し上げる可能性を検証した。

（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

文献調査の結果、化学言語モデルを用いた131件の論文のうち、約半数でSMILESのcanonicalization手順が記載されていなかった。記載がある論文ではRDKitやOpenBabelが多く用いられていたが、バージョン情報や立体化学情報の扱いはほとんど明記されていなかった。これは、深層学習QSARモデルの入力前処理に関する報告が十分に標準化されていないことを示す。

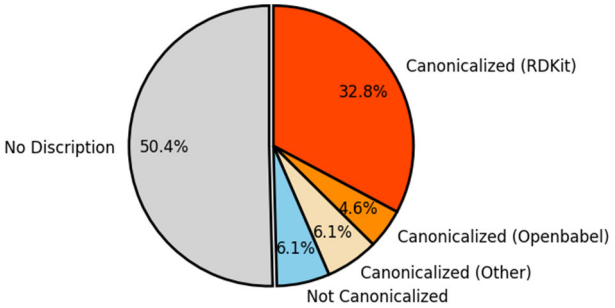

図1. 前処理に関するサーベイ結果

公開データセットの解析では、MoleculeNet、およびTDCにおいて、同一化合物に対する複数のSMILES表記が確認された。特に芳香族環についてKekulé表記とaromatic表記が混在していた。立体化学情報についても欠落が多く、MoleculeNet全体ではエナンチオマーに関係する構造の約半数、cis-trans異性体に関係する構造の約3割で情報が十分に指定されていなかった。RDKitによる標準化は表記の一部を統一できるが、元データに存在しない立体化学情報を復元することはできなかった。

分子翻訳タスクでは、Raw SMILESを入力した場合、Standardized SMILESと比べて構造再構成性能が低下した。潜在空間解析でも、Raw SMILES由来の表現とStandardized SMILES由来の表現は異なる分布を示し、SMILES文字列の差が大きいほど潜在表現の差も大きくなる傾向が認められた。以上より、SMILES表記揺れは化学言語モデルの内部表現に直接影響することが示された。

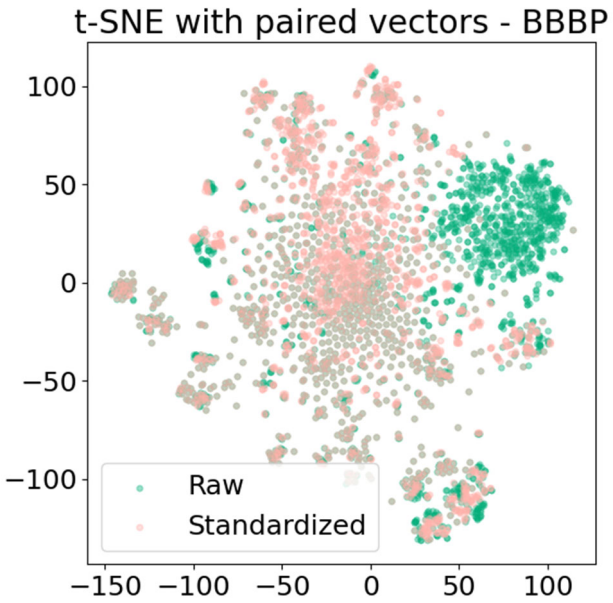

図2. 同一空間に埋め込まれた正規化前・後のSMILESの潜在表現

一方、物性・毒性関連の二値分類タスクでは、Raw SMILESとStandardized SMILESの間でAUROCに大きな差が見られない場合があった。特徴重要度解析の結果、下流のXGBoost分類器は、表記揺れにより変動しやすい潜在特徴ではなく、予測に有用で比較的安定した特徴を選択している可能性が示された。このため、予測性能が保たれている場合でも、化学言語モデル自体が表記揺れに頑健であるとは限らない。

さらにClinToxおよびBBBPでは、Raw SMILESを用いた方がStandardized SMILESより高いAUROCを示した。ClinToxを詳細に解析したところ、陽性ラベルの化合物にKekulé/aromatic表記不一致が強く偏っていた。ROC曲線上でも、これらの化合物が低いfalse positive rateの領域でtrue positiveとして現れ、AUROCの上昇に寄与していた。これは、モデルが化学構造そのものではなく、表記上の偏りを分類の手がかりとして利用した可能性を示す。

clintox

Types of mismatch	Positive	Negative	Positive rate
K/A mismatch	93	0	1.000
Other mismatch	18	1318	0.013
No mismatch	1	58	0.017
Total	112	1376	0.075

図3. Kekulé/aromatic表記不一致とデータセットのラベル(陽性)との交絡

D．考察

本研究により、SMILES 表記揺れは深層学習 QSAR モデルの信頼性を左右する重要な要因であることが示された。分子翻訳タスクでは、同一化合物であっても表記が異なるだけで再構成性能が低下し、潜在表現にも差が生じた。この結果は、化学言語モデルが SMILES を単なる化学構造の代理表現としてではなく、文字列として学習していることを示している。

一方、下流の分類タスクでは、表記揺れの影響が見かけ上小さい場合があった。この点は、化学言語モデルが表記揺れに本質的に頑健であることを意味しない。下流分類器が、予測に有用な一部の特徴を選択することで、潜在表現全体の不安定性を部分的に吸収していると考えられる。したがって、AUROC が維持されていても、潜在表現や構造理解が安定しているとは限らない。

特に重要なのは、表記揺れが性能低下ではなく性能上昇として現れる場合である。ClinTox、および BBBP では、Kekulé/aromatic 表記不一致がラベルと相関し、Raw SMILES を用いたモデルの性能を人工的に高く見せていた。このような表記レベル交絡が存在すると、モデルは汎化可能な構造活性相関ではなく、データセット固有の表記上の偏りを学習する。毒性予測モデルでは、こうした性能の過大評価がリスク評価やスクリーニング判断に影響する可能性がある。

本研究の結果は、深層学習モデルによる QSAR の深化において、モデル構造の高度化だけでなく、入力表記の標準化と前処理手順の透明化が不可欠であることを示す。今後、SMILES を用いる QSAR 研究では、使用した化学情報処理ツール、バージョン、canonicalization 設定、および立体化学情報の扱いを明示する必要がある。公開データセットを利用する場合でも、Raw SMILES をそのまま用いるのか、単一のツールで再標準化するのかを明確にし、表記揺れとラベルの偶然相関を確認することが望まれる。

E．結論

本研究では、SMILES 表記揺れが化学言語モデルを用いた QSAR 解析の信頼性に与える影響を検討した。文献調査では、化学言語モデル研究の約半数で canonicalization 手順が明記されていないことを見出した。公開データセット解析では、同一化合物に対する複数の SMILES 表記や立体化学情報の欠落が確認された。分子翻訳モデルを用いた実験では、表記揺れが構造再構成性能と潜在表現を変化させることを示した。

物性・毒性予測タスクでは、特徴選択により表記揺れの影響が見かけ上小さくなる場合があった。一方、ClinTox、および BBBP では、表記揺れがラベルと相関し、予測性能を人工的に高く見せる表記レベル交絡が認められた。以上より、SMILES 表記揺れは深層学習 QSAR モデルの再現性、解釈性、性能評価の妥当性に関わる基盤的課題である。化学物質リスク評価に資する深層学習 QSAR モデルを構築するためには、前処理手順を明示し、標準化された入力表記に基づく評価を行うことが重要である。

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. 水野忠快, NAMs に向けた化合物構造の潜在表現抽出の現状. 日本薬学会第 146 年会 (2026.3)
2. 水野忠快, 観測から潜在へ：毒性学に向けた生命科学データに内在するパターンの抽出. 第 8 回医薬品毒性機序研究会 (2025.12)
3. 菊池陽佑, 吉開泰裕, 古濱彩子, 根本駿平, 楠原洋之, 山田隆志, 水野忠快: SMILES 表記揺れに起因する化学言語モデルの予測バイアスの検証; 第 42 回メディシナルケミストリーシンポジウム (2025.11)
4. 菊池陽佑, 吉開泰裕, 古濱彩子, 根本駿平, 楠原洋之, 山田隆志, 水野忠快: The impact of SMILES notation inconsistencies on chemical language model property prediction; CBI 学会 2025

(2025.10)

5. 水野忠快, Study on Pattern Recognition of Chemical Structures by Language AI; 第 40 回日本薬物動態学会 (2025.10)
6. Tadahaya Mizuno, Representation Learning for Drug-oriented Life Science Data; The 33rd Annual Meeting of the Korean Society of Applied Pharmacology (2025.10)
7. Tadahaya Mizuno, Current Status of Understanding and Application of AI for Handling Chemical Structures; GSRS2025 (2025.09)
8. Tadahaya Mizuno, Shumpei Nemoto, Yasuhiro Yoshikai, Yosuke Kikuchi, Takuho Ri, Hiroyuki Kusuhara; Investigation of Chirality Recognition by Encoder-Decoder Chemical Language Models; EuroTox 2025 (2025.09)
9. 水野忠快, 潜在的言語構造に基づく生命科学データのパターン認識; 2025 年日本バイオインフォマティクス学会年会・第 12 回生命医薬情報学連合大会 (2025.09)
10. 水野忠快, 毒性学発展に向けた生命科学データのパターン認識研究, 第 52 回日本毒性学会学術年会 (2025.07)
11. 水野忠快, 言語モデルを基盤とした毒性研究の深化, 第 52 回日本毒性学会学術年会 (2025. 07)
12. Tadahaya Mizuno, Shumpei Nemoto, Yasuhiro Yoshikai, Yosuke Kikuchi, Takuho Ri, Hiroyuki Kusuhara, How Chemical Language Models Learn: Architectural Insights into Molecular Structure and Chirality Recognition; QSAR2025 (2025.06)

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