大規模反応データベースを用いた文字列化した化学反応の基盤モデル構築

佐川 達也[1] 小島 諒介[2]
[1]京都大学大学大学院 薬学研究科 [2]京都大学大学大学院 医学研究科
sagawa.tatsuya.82s@st.kyoto-u.ac.jp kojima.ryosuke.8e@kyoto-u.ac.jp

概要

化学反応の高精度な予測は，実験を行う前に実験結果の予測ができることから，創薬をはじめとする分野で実験コストの削減の視点から注目されている．これまでに化学反応予測モデルの開発は活発に行われてきたが，利用可能な学習データが限られていたことから，分布外データへのモデルの微調整（ファインチューニング）を想定した事前学習モデルに関する研究は限定的であった．本研究では，大規模化学反応データベースを用いた事前学習を通じて化学反応基盤モデルを構築した．本モデルを活用することで，従来のモデルと比較して非常に少ないファインチューニングデータで，優れた予測性能を実現できることを示した．

1 はじめに

薬の開発には莫大なコストがかかり，1 つの薬が市場に出るまでに 10 年以上の期間と 10 億ドル以上の資金が必要になるとされている [1]．さらに，薬の開発コストは増加し続けており [2]，創薬コストの削減に向けて AI の活用が期待されている [3]．創薬における課題の一つとして，候補となる化学構造が実際に合成可能かどうかを評価することや，合成のための反応において十分な収率が得られるかどうかを判断する必要がある．通常，これらの問題への対処には多くの実験コストがかかる．これに対し，合成経路の設計を補助するために生成物から反応物を予測する逆反応予測モデルや，反応条件の最適化を目的とした収率予測モデルを活用することで、実験効率の向上が期待されている．

化学構造を扱う問題の多くは，化合物を Simplified Molecular-Input Line-Entry System (SMILES)[4] という文字列で表現することで，自然言語処理分野の技術を活用することができる．例えば，化学反応を反応物から生成物への変換ととらえることで，言語翻訳

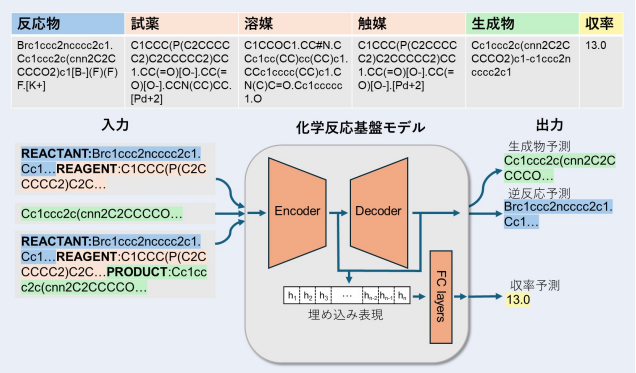

図 1 化学反応基盤モデルの概略図

と類似するタスクとみなすことができ，系列変換モデルである text-to-text transfer transformer (T5) [5] を化学反応の予測に活用した例が報告されている [6]．

ベンチマーク性能において優れた化学反応予測モデルであっても，実際の現場での利用においてはしばしば次のような課題が存在する．第一に，学習データと現場で扱う新規反応の間に乖離があり，これがモデルの適用を困難にする．第二に，モデルの微調整に必要な数千件の反応データを現場で用意するのが難しい．実際に我々の予備実験においても，既存手法で良く用いられる化合物データベースである ZINC[7] を学習データとして使用し，T5 を Masked Language Model (MLM) による自己教師あり学習で事前学習した結果は，下流タスクでのベンチマーク予測性能が向上するものの外部データに対する予測性能には課題があり，そのモデルの微調整には多くの下流タスク用学習データを必要とすることが明らかとなった．

本研究では，従来の事前学習済みモデルが下流タスクで多くの学習データを必要とする原因について，事前学習に用いるデータと下流タスクのデータおよびタスク間に大きな乖離が存在することが影響しているという仮説を立て，これに対処するために，大規模な化学反応データベースである Open Reaction

Database (ORD)[8] を用いた二段階の事前学習を提案する．ORD には特許文書，学術論文，製薬企業の実験ノートから収集された化学反応が集約されており，広範な化学反応空間を網羅するデータベースである．我々の提案モデルは，一段階目で，SMILES 表記された単一化合物を学習するために，化合物データベースである ZINC[7] を学習データとして使用し，二段階目では反応内の複数化合物間の関係を学習するために，ORD を用いた事前学習を行う．

本稿では，この二段階学習によって得られたモデルを化学反応基盤モデルと呼び，基盤モデルとして，生成物予測，逆反応予測，収率予測の 3 つ下流タスクにおいて微調整を行い，ベンチマークで評価する．さらに，微調整用の学習データの件数を変化させた評価実験を行い，利用可能なデータが少数しかない場合におけるモデルの挙動を検証する．

2 モデル構築

化学反応基盤モデル（以下，「提案モデル」という）は，化合物事前学習と化学反応事前学習の二段階の事前学習を通じて作成される．本章では，それぞれの事前学習方法について説明する．

2.1 化合物事前学習

化合物事前学習では，学習データとして市販の低分子化合物を大規模に収録したデータベースである ZINC[7] を利用する．ZINC 中には SMILES 形式で表現された約 2300 万の化合物が含まれており，この SMILES 文字列をトークナイザによってトークン化し，t5-v1_1-base[i]のパラメータ初期化後，span-masked language modeling (span-MLM) [5] を用いて学習を行う．span-MLM では入力テキスト内の連続するトークン（span）をマスクトークンに置換し，モデルに元のトークンを予測する．

後の実験では，学習時のハイパーパラメータは T5 の既報論文 [5] に倣って設定した．具体的には，入力文中の 15%のトークンをランダムにマスクし，span の平均長を 3 に設定した．

トークナイザは，SentencePiece unigram tokenizer[9] をもとに，ZINC データセットを用いて学習を行う．このトークナイザを用いることで，化合物データベースでの学習を通じて，化合物中で頻出する構造を少数のトークンで効率的に表現できる [10]．

2.2 化学反応事前学習

化学反応事前学習では，学習データとして大規模公共反応データベースである ORD[8] を使用する．ORD は，卓上実験から自動化されたハイスループット実験まで含む，約 150 万件の多様な化学反応が登録されている．各反応中の化合物には，「反応物」，「試薬」，「溶媒」，「触媒」，「生成物」の化学構造情報とタグ，加えてその反応の収率，反応温度などの情報が記載されている．

このデータベースを利用する際，データ提供元によって「試薬」，「溶媒」，「触媒」のタグ付けが曖昧であるという課題がある．さらに，実際の反応においてもこれら 3 つの化合物の役割を明確にタグ付けすることが困難な場合が多い．そのため，本研究ではこれらの化合物をすべて「試薬」としてタグ付けして処理を行う．また，一部の反応には，反応物と生成物間の原子対応情報が含まれているが，データセット内で統一するためすべて削除して利用した．また，後述する評価実験では，下流タスクで使用するベンチマークデータセットと重複する反応を除去したうえで，データを 8:1:1 の比率で学習・検証・テストデータに分割して評価を行った．

データセット中の文字出現頻度をみると，ORD には ZINC よりも多様な元素が含まれている（図 4）．ZINC で学習したトークナイザはこれらの未知の元素を扱うことができず，モデル性能の低下を招く可能性がある．そのため，特に金属元素を中心にトークンの追加を行い，トークナイザが ORD 内の全ての化合物を認識できるようにした．

化学反応事前学習は，下流タスクに合わせて生成物予測，逆反応予測，収率予測の三つのタスクで行う．

生成物予測　生成物予測は系列変換の枠組みを採用し，反応物と試薬の文字列から生成物の文字列を予測するタスクである．同じ役割を持つ複数の化合物は「.」で結合して表現する．また，モデルが入力中の化合物の役割を認識できるよう，反応物と試

[i] https://huggingface.co/google/t5-v1_1-base

表 1　生成物予測と逆反応予測の性能比較

	生成物予測				逆反応予測			
	Top 1	Top 3	Top 5	invalidity	Top 1	Top 3	Top 5	invalidity
Molecular Transformer[11]	88.8	-	94.4	-	43.5	60.5	-	-
T5Chem[6]	90.4	-	96.4	-	46.5	64.4	70.5	-
化合物事前学習モデル	86.6	90.4	91.2	19.2	44.2	61.4	67.3	4.75
提案モデル	**97.5**	**98.8**	**99.0**	**7.1**	**71.0**	**85.2**	**86.9**	**0.45**
提案モデル (zero-shot)	92.8	96.4	97.1	12.5	13.8	21.4	26.2	3.08

薬の前にそれぞれ「REACTATNT:」および「REAGENT:」という特殊トークンを付与する．このようにして結合された入力文字列を学習データとして用い，損失関数にはクロスエントロピー損失を採用して学習を行う．生成時には，ビーム幅を 5 に設定したビームサーチを用いて，確率の高い上位 5 つの予測を出力する．

逆反応予測　逆反応予測は，生成物の文字列から反応物の文字列を予測するタスクである．このタスクは生成物予測の逆方向であり，実験の設定もほぼ同じである．ただし，先行研究 [6] に基づき試薬の予測は行わず，入力は生成物の単一の役割を持つ文字列のみで構成される．そのため，特殊トークンを用いた入力化合物の役割の区別は行わない．

収率予測　収率予測は，反応物，試薬，生成物の文字列から収率を回帰予測するタスクである．このタスクでも特殊トークン「REACTATNT:」，「REAGENT:」，「PRODUCT:」を付与することで，入力中の化合物の役割を明確にする．収率予測は 0 から 1 の連続値を予測する回帰タスクであるため，事前学習済みモデルのエンコーダーおよびデコーダーから得られる埋め込み表現を基に，全結合層を追加して予測を行う．損失関数には平均二乗誤差を採用する．

3 結果

3.1 生成物予測

生成物予測では，USPTO_MIT データセット [12] を用いてベンチマーク評価を実施した（表 1 左）．学習データと評価データの分割方法は先行研究 [6] に従い，Top k (k = 1, 3, 5) の予測が正解と完全一致するかどうかを評価する「top k accuracy」と，予測が化合物として破綻していない割合を示す「SMILES invalidity rate」で評価した．

提案モデルは，Top 1 accuracy を既存モデルから

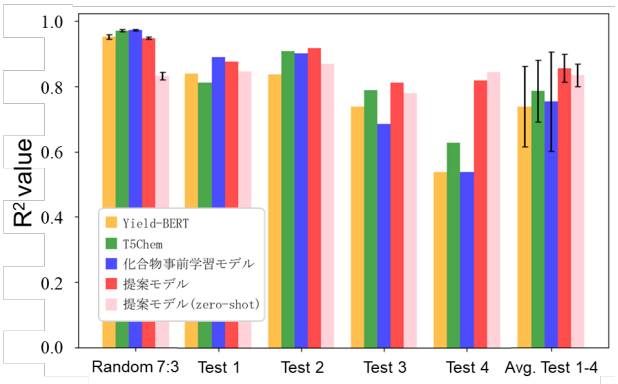

図 2　収率予測の性能比較

7.1 ポイント改善し，97.5%の精度を達成した．さらに，Top 5 accuracy は 99.0%と非常に高く，化学反応事前学習によって SMILES invalidity rate は 19.2%から 7.1%に大きく改善した．加えて，USPTO_MIT の学習データで微調整を行わないゼロショット予測において，提案モデルは 40 万件の反応データで微調整を行った既存モデルよりも優れた予測性能を示した．

3.2 逆反応予測

逆反応予測では，USPTO_50k データセット [13] を用いてベンチマーク評価を実施した（表 1 右）．生成物予測と同様に，データ分割方法は先行研究 [6] に基づき，評価指標として「top k accuracy」と「SMILES invalidity rate」を用いた．

提案モデルは，top 1 accuracy 71%を達成し，これは既存モデルと比較して 24.5 ポイントもの大幅な改善に相当する．また，SMILES invalidity rate は化学反応事前学習の適用により約 10 分の 1 程度にまで低下した．

提案モデルにおける性能向上幅は，逆反応予測のほうが生成物予測よりも大きかった．この結果は，逆反応予測で用いた USPTO_50k データセットが生成物予測で用いた USPTO_MIT データセットに比べてデータ数が約 10 分の 1 程度と少なく，モデルが逆

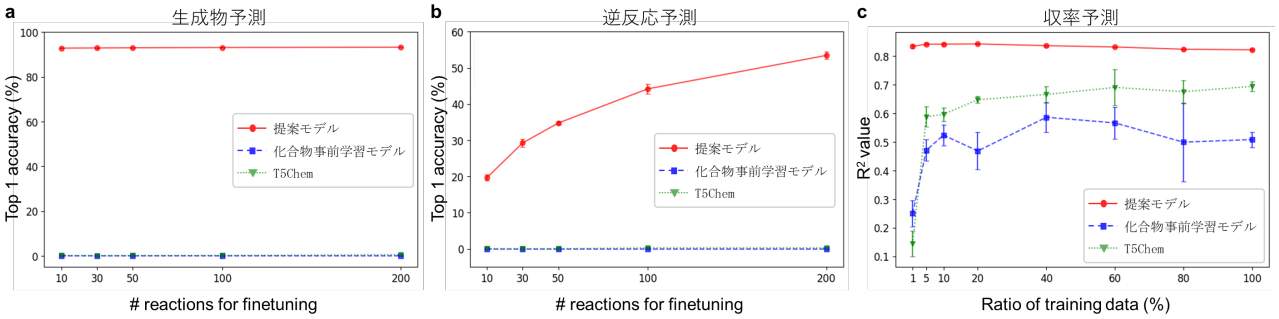

図 3　少データ学習時の性能比較

反応のパターンを学習するには十分でなかったため既存モデルでは学習が難しかったのに対し，提案法では，大規模な化学反応事前学習を行ったことで，データ不足の影響を効果的に緩和することができたためと推測される．

3.3 収率予測

収率予測では，Buchwald-Hardwig クロスカップリングデータ[14]でベンチマーク評価を実施した（図 2）．評価方法としては，先行研究[6]に基づき，10 回のランダムなデータ分割による評価（Random 7:3）と，学習データとテストデータにドメインシフトが生じるよう分割をしたデータを用いた評価（Test 1-4）を行った．また，評価指標として予測と正解値の間の決定係数を算出した．

提案モデルは，Random 7:3 の評価条件下では有意な性能向上を示すことはなかったが，より困難な Test 2-4 の条件下では既存モデルを上回る予測性能を達成した．特に，最も困難とされる Test 4 において，ゼロショット予測の条件下でも既存モデルと比較して 0.214 ポイントの性能向上を示した．

これらの結果は，あらかじめ広い化学反応空間のデータセットで化学反応事前学習を行うことで，データセットのドメインシフトに対して堅牢なモデルが構築できることを示している．この効果は，特に学習データとテストデータ間のドメインシフトが大きい場合に顕著に表れることが確認された．

3.4 少データでの検証

提案モデルおよび既存モデルを，少数件数の反応データで微調整し，学習データ件数と予測性能の関係を分析した（図 3）．生成物予測および逆反応予測では，ベンチマークに含まれる学習データから 10~200 件サンプリングし，これら少数データのみで微調整を行い，性能評価を実施した．収率予測では，学習データから 1~100%をサンプリングし，同様に微調整後の性能を評価した．図 3 にはサンプリングを 3 回繰り返して得られた結果の平均と標準偏差を示している．

生成物予測　提案モデルは，微調整データの件数に依存せず，高い予測性能を維持した．対照的に，既存モデルは 200 件の微調整データを用いた場合でも学習が成功せず，性能を向上させることはできなかった．

逆反応予測　提案モデルでは，微調整データ件数に応じて性能が向上し，およそ 100 件での微調整を行うだけで，全データを使用して微調整された既存モデルの性能を上回ることが確認された．一方で，既存モデルは少数件数の微調整では学習に失敗し，性能改善は見られなかった．

逆反応予測　提案モデルはデータ件数に関係なく高い性能を維持したのに対し，既存モデルは性能のプラトーに達するために 20%（約 600 件）程度のデータが必要とした．また，既存モデルは，予測性能の分散が大きいことから，微調整データに影響されやすいということが明らかになった．

4 おわりに

本研究では，大規模化学反応データベースを活用した二段階事前学習による化学反応基盤モデルを開発した．本モデルは生成物予測、逆反応予測、収率予測といった化学反応予測タスクにおいて既存モデルを上回る性能予測を実現し，特に，少数データ環境やドメインシフトが大きい状況下での高い汎化性能が示された．

謝辞

本研究は JSPS 科研費 JSPS KAKENHI Grant 21H05207, 21H05221 (Digi-TOS) および CREST JPMJCR22D3 の助成を受けたものです.

参考文献

1. Hinkson, I. V., Madej, B. & Stahlberg, E. A. Accelerating therapeutics for opportunities in medicine: A paradigm shift in drug discovery. *Front. Pharmacol.* **11**, 770 (2020).
2. 厚生労働省. (n.d.). 医薬品産業の現状. Accessed 3 Jan. 2025. https://www.mhlw.go.jp/content/10807000/001036959.pdf.
3. Rehman, A. U. *et al.* Role of artificial intelligence in revolutionizing drug discovery. *Fundamental Research* (2024) doi:10.1016/j.fmre.2024.04.021.
4. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).
5. Raffel, C. *et al.* Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv [cs.LG]* (2019).
6. Lu, J. & Zhang, Y. Unified Deep Learning Model for Multitask Reaction Predictions with Explanation. *J. Chem. Inf. Model.* **62**, 1376–1387 (2022).
7. Irwin, J. J. *et al.* ZINC20-A free ultralarge-scale chemical database for ligand discovery. *J. Chem. Inf. Model.* **60**, 6065–6073 (2020).
8. Kearnes, S. M. *et al.* The Open Reaction Database. *J. Am. Chem. Soc.* **143**, 18820–18826 (2021).
9. Kudo, T. & Richardson, J. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. *arXiv [cs.CL]* (2018).
10. Mielke, S. J. *et al.* Between words and characters: A Brief History of Open-Vocabulary Modeling and Tokenization in NLP. *arXiv [cs.CL]* (2021).
11. Schwaller, P. *et al.* Molecular Transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS Cent. Sci.* **5**, 1572–1583 (2019).
12. Jin, W., Coley, C. W., Barzilay, R. & Jaakkola, T. Predicting organic reaction outcomes with Weisfeiler-Lehman network. *Adv. Neural Inf. Process. Syst.* 2607–2616 (2017).
13. Liu, B. *et al.* Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS Cent. Sci.* **3**, 1103–1113 (2017).
14. Ahneman, D. T., Estrada, J. G., Lin, S., Dreher, S. D. & Doyle, A. G. Predicting reaction performance in C-N cross-coupling using machine learning. *Science* **360**, 186–190 (2018).

付録

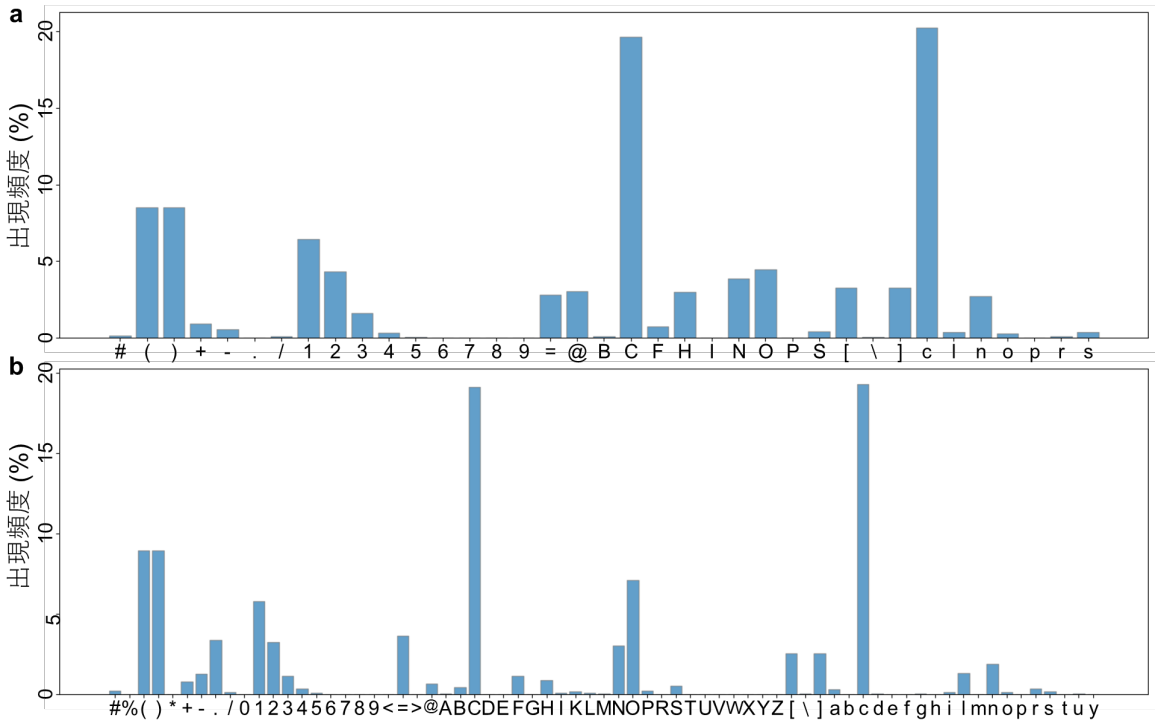

図 4　データセット中の文字の出現頻度分布．ZINC 中の文字の種類(a)は ORD 中の文字の種類(b)よりも少なく，ORD のほうが多様な元素を含んだデータセットであることが分かる．

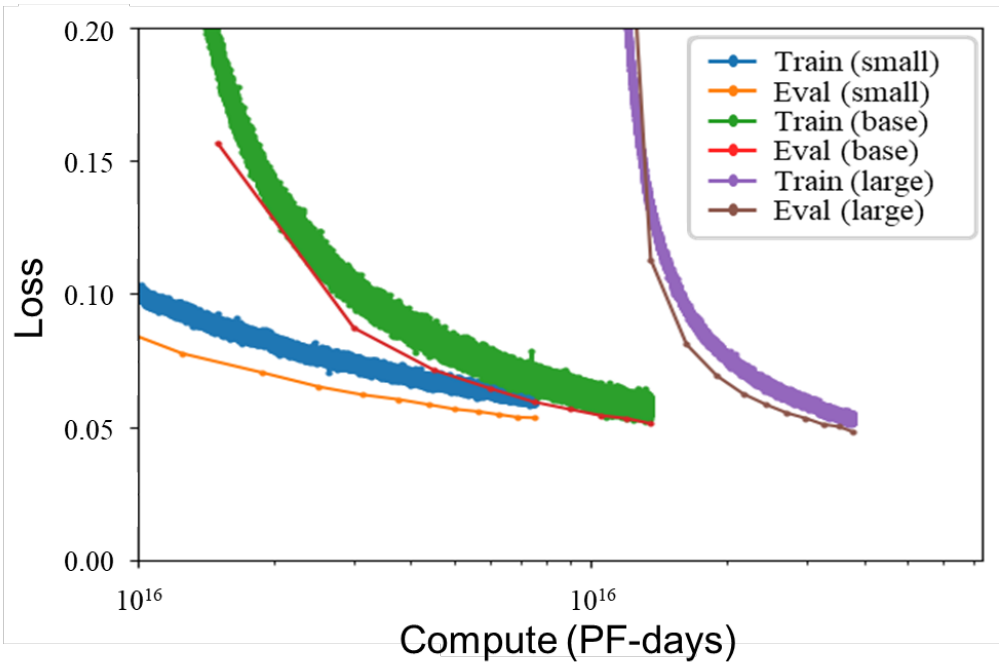

図 5　化合物事前学習の学習曲線（縦軸がクロスエントロピー損失，横軸に計算量）．small，base，large と T5 モデルのサイズが大きくなるほどわずかだが事前学習時の損失は低い傾向がみられる．