Vol.2022-BIO-69 No.18
2022/3/11

機械学習による化合物の逆合成解析可能性予測手法の開発

小澤 真実[1]　安尾 信明[2]　関嶋 政和[1,2]

概要：創薬の分野における研究・開発コスト削減手段の一つとして，計算機科学を利用する手法が注目されている．化合物探索におけるその一例として，分子生成モデルによる膨大な量の化合物の提案，そしてその膨大な量の化合物をウエットな実験に進む前に合成可能かどうかでフィルタリングする，というものが考えられる．フィルタリングには逆合成解析によるものとスコアリングによるものがあり，既存のスコアリング手法の一つとして RAscore が挙げられる．これは，逆合成解析可能性を機械学習によって予測するスコアである．本研究では，RAscore モデルの実用性について議論し，パラメータチューニングによってより正確な予測が行えるモデルの開発を提案する．元々のモデルの学習に用いられた ChEMBL のデータのほかに，分子生成モデルを用いて作った合成困難なデータを用いて重みパラメータを学習したモデルを 2 種類作成し，既存の RAscore モデルを用いたスコアをベースラインとして比較することでこの手法を評価した．その結果，新たに作成したモデルの一つはベースラインモデルベースラインと比較して負例についての判別性能の向上が見られ，AUC，Binary Accuracy, loss について同等以上の精度をもつモデルとなった．

キーワード：逆合成解析，機械学習

Development of a method for predicting the accessibility of retrosynthesis of compounds by machine learning

Abstract: In the field of drug discovery, the use of computer science is gaining increasing attention as a means of reducing R&D costs. One example of this in compound discovery is the method where a molecular generation model proposes a large number of compounds and filters the large number of compounds based on their synthetic potential before proceeding to wet experiments. Filtering can be done either by inverse compositional analysis or by scoring. One of the existing scoring methods is RAscore, a score that predicts the likelihood of inverse compositional analysis by machine learning. In this study, we discuss the practicality of the RAscore model and propose to develop a model that can make more accurate predictions by tuning its parameters. In addition to the ChEMBL data used to train the original model, two models with weight parameters trained on difficult-to-synthesise data generated using molecular generation models were created, and the scores were compared using the existing RAscore model as a baseline to evaluate the method. The results showed that one of the new models had better discrimination performance for negative cases than the baseline model, and had similar or better accuracy for AUC, Binary Accuracy and loss.

1. はじめに

ヒット化合物探索，Hit-to-Lead，Lead 最適化などの創薬の様々なプロセスにおいて，新規化合物の探索は欠かすことのできないプロセスである．新たな医薬品の開発は我々の生活に欠かせないものになっており，例えば現在では Covid-19 の治療薬の開発が急がれている．[1] しかし，その一方で創薬の研究にかかる時間的コスト，金額的コストは莫大である．一つの薬を上市するまでには平均的に約 10〜15 年もの開発期間，約 26 億ドルもの開発資金が必要 [2] とも言われており，そのコストは年々増加している．このため，計算機科学の技術を用いることにより効率的に開発研究を行う手法が注目されている．

分子生成モデルを利用した新規化合物の探索のユースケースとして，新規の化合物を分子生成モデルによって多

[1] 東京工業大学情報理工学院
School of Computing, Tokyo Institute of Technology
[2] 東京工業大学物質・情報卓越教育院
Tokyo Tech Academy for Convergence of Materials and Informatics, Tokyo Institute of Technology

数提案し，そしてウェットでの化合物合成やアッセイ試験を行う前に，コンピュータによってその化合物の合成可能性を予測してフィルタリングする，ということが考えられる．近年の化合物生成手法の進歩によって，提案ステップでは膨大な量の提案化合物が得られるようになったが，同時にフィルタリングのステップでこの膨大な量の提案化合物を高速かつ正確に合成不能な化合物を排除する必要性が増大している．

このフィルタリングで用いられる手法に，逆合成解析と呼ばれるものがある．これはコンピュータ上で化合物の合成経路を探索することで合成可能性を予測する手法であり，合成経路に基づいているため予測の信頼性を裏付けられるというメリットがある．一方でその計算コストの削減には限界があり，化合物一つあたりの逆合成解析には数分程度の時間がかかるため膨大な量の提案化合物について全て予測するのは困難であるというデメリットが残る．そこで，さらに事前に大まかな合成可能性予測によるフィルタリングを行うことでこのデメリットを解決している．

この大まかなフィルタリングを行う既存手法には，SAscore[3]，SYBA[4], SCscore[5]，RAscore[6] などによるスコアリングが存在する．SAscore は分子のフラグメントごとの報酬と複雑さに対するペナルティで算出されるフラグメントベースのスコア，SCscore は化合物の合成ステップ数に相関する反応ベースのスコア，SYBA はベイズ推定と負例の追加によってフラグメントスコアの算出方法に改良を加えているフラグメントベースのスコア，そして RAscore は上記の手法のみではそれぞれのスコアの閾値を決めるのが困難であり正確な予測ができないデータが多いという問題が残るため，それらを解決するために開発された機械学習による手法で，逆合成解析モデルで解決できるかどうかを正解ラベルとした分類の予測スコアである．

RAscore モデルは分子の記述子として FCFP6 を用いたニューラルネットワークであり，そのハイパーパラメータである層数，層の大きさ，活性化関数，ドロップアウト率，学習率は Optuna[7] を用いて設計されている．これらのハイパーパラメータやモデルのネットワーク構造，モデルの重みパラメータなどは，逆合成解析モデル AiZynthFinder[8] を用いてラベリングされた ChEMBL[9] のデータセット約 200000 個を利用する学習，検証によって作成されている．作成のワークフローが図 1 に示されており，この結果として入出力層のほかに 10 層の全結合層，9 層のドロップアウト層を持つモデルが構築されている．モデルの概形は図 2 に示す．

実際に RAscore が利用されるシチュエーションとして，他の既存のスコアリング手法とも組み合わせて分子生成モデルによる提案化合物のフィルタリングを行うことが考えられる．そこで，分子生成モデル MERMAID[10] によって生成された化合物のうち，SAscore が 3 以下になるよう

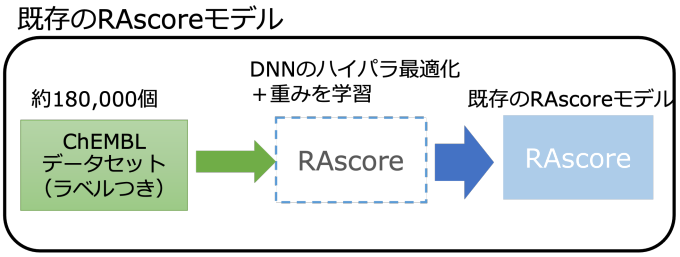

図 1 RAscore モデルの学習ワークフロー

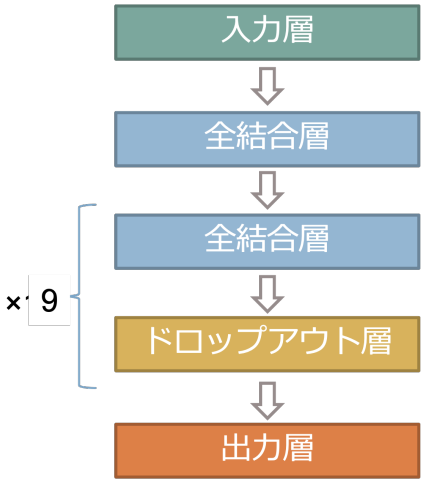

図 2 RAscore モデルの概形

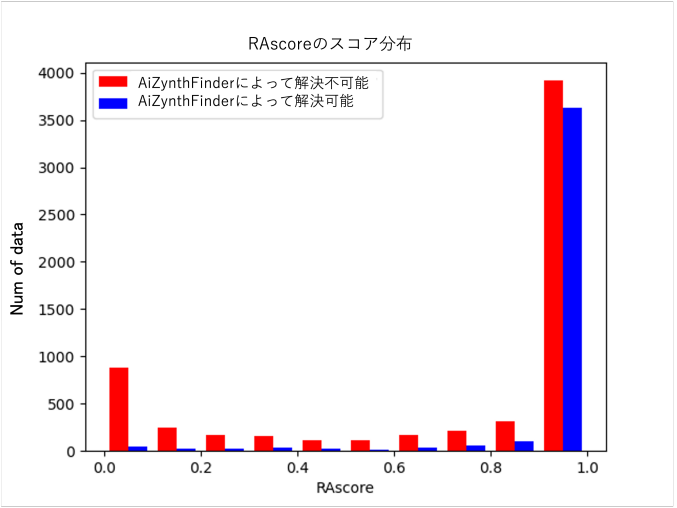

図 3 MERMAID によって作成したデータセットを SAscore によってフィルタリングしたものに対する RAscore の分布．青：AiZynth Finder によって逆合成解析可能，赤：逆合成不可能

フィルタリングされた化合物データセットに対して，さらに RAscore で分類を行なった結果が図 3 である．

図 3 では，AiZynthFinder によって逆合成解析が可能な分子（正例）数を青，不可能な分子（負例）数を赤で表現しているが，理想的には青のグラフが 1，赤のグラフが 0 に偏るようになるべきところを，実際にはどちらも 0 に偏るように分布してしまっている．これは，RAscore は正例

は正しく分類できている一方で，負例を正しく弾けていないことを示している．

以上の分類の失敗に対する考察として，RAscore が使用される文脈は分子生成モデルによる膨大な量の化合物の探索であり，ほとんどが逆合成解析不可能なもの（負例）となるデータセットがスコアリング対象になると考えられる一方で，Thakkar ら [6] によれば，AiZynthFinder による ChEMBL 中の分子の逆合成解析可能なもの（正例）の割合は 75%程度と高くなっており，学習データである ChEMBL と実際に利用対象となるデータセットとの性質が大きく異なっている可能性が考えられる．

本研究では，分子生成モデルによる提案分子も併用して RAscore モデルの学習を行い直すことで，重みパラメータの更新によるモデルの性能を向上を目的とする．

2. 提案手法

2.1 提案手法に用いるデータセット

提案手法の学習と検証のために，RAscore の学習に用いられた既存のデータセットと，分子生成モデル MERMAID による新たに作成したデータセットを利用した．

既存のデータセットは, ChEMBL から取得した約 200,000 の化合物と GDBMedChem[11] から取得した約 100,000 の化合物について，AiZynthFinder によってラベリングを行ったものである．さらに，これらのデータセットはそれぞれ 9：1 の割合で学習データセットとテストデータセットに分割されている．データセットは RAscore の実装 [12] から得ることができ，ラベル作成時に AiZynthFinder を利用する際の探索ポリシーに関わる詳細な設定等は論文 [6] から得ることができる．

新たに作成したデータセットは，分子生成モデル MERMAID を用いて 1100 の化合物をベースに生成した化合物のうち，SMILES の文法が正しいものを 200,000 個ランダムに選択し，AiZynthFinder でラベリングを行うことで作成した．さらにこのデータセットを 9：1 の割合で学習データセットとテストデータセットに分割した．AiZynthFinder を利用する際の探索ポリシーなどの設定は，上記の既存データセットを作成する際に使われたものを利用した．

本手法では学習データとして ChEMBL データセットと MERMAID データセット，テストデータとして ChEMBL データセット, GDBMedChem データセットと MERMAID データセットを利用した．

既存の ChEMBL と GDBMedChem の学習データ・テストデータ間に重複がないことは，元の論文で InCHI-Key に変換後の重複確認によって確認されており，新たに作成した MERMAID のテストデータと MERMAID・ChEMBL・GDBMedChem の学習データ間についても同様の方法で重複がないことを確認した．

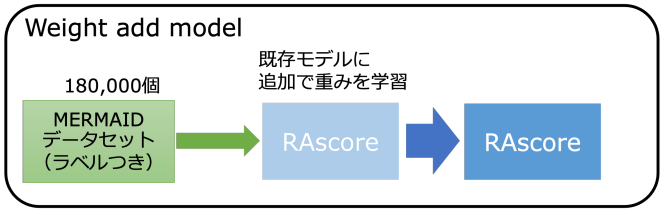

図 4 weight add model の学習ワークフロー

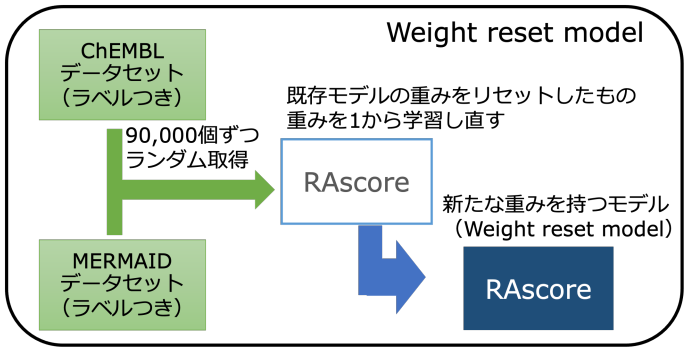

図 5 weight add model の学習ワークフロー

2.2 学習方法

既存の RAscore モデルは 1 章で述べた通り，図 1 のように作成されている．本手法ではこれを改良することで weight add model, weight reset model の 2 種類のモデルを作成した．モデルの再学習には Keras with Tensorflow[13] を用いた．

2.2.1 weight add model

weight add model と呼ばれるモデルは，図 4 のように作成した．具体的には既存モデルのネットワーク構造とハイパーパラメータ，重みパラメータを初期値として，MERMAID 学習データ 180,000 個を用いて後からモデルの学習を行い直すことで重みパラメータを更新した．

2.2.2 weight reset model

weight reset model と呼ばれるモデルは，図 5 のように作成した．既存モデルのネットワーク構造とハイパーパラメータのみを利用し，重みパラメータは一度初期化した後に ChEMBL の学習データセットと MERMAID の学習データセットからそれぞれランダムに 90,000 個ずつ取得し，連結した 180,000 個のデータで学習を一から行うことで，新たな重みパラメータを得た．

3. 実験

新たに作成したモデルの性能を評価するため，以下の 3 つの実験を行った．

3.1 実験 1

RAscore モデル，weight add model，weight reset model について様々なメトリクスで ChEMBL データセット，GDBMedChem データセットと MERMAID データセット

について評価を行い，比較した．モデルコンパイル時の評価メトリクスは RAscore の実装コードで用いられていたものと同じ，Loss(損失)，AUC（ROC 曲線下の面積），Accuracy（ラベルの完全な正解率），Binary accuracy（2 値分類における正解率），Precision（適合率），Recall（再現率），True negatives（真陰性），True positives（真陽性），False negatives（偽陰性），False positives（偽陽性）の 10 種類を利用した．

3.2 実験 2

実験 1 においてより精度が改善されたモデルを用いて，第 1 章で提示したグラフに用いたものと同じデータセット（分子生成モデル MERMAID によって生成された化合物のうち，SAscore が 3 以下になるようフィルタリングされた化合物データセット）について，同様の分布図を作成し，定性的な比較を行う．

3.3 実験 3

実験 1 においてより精度が改善されたモデルに対し，最適な閾値を設計する．閾値の設計に用いるデータセットは，第 2.1 節で述べた ChEMBL と MERMAID の学習データセットの中から，最適化対象のモデルの学習に用いたものと同じ構成になるようにランダムに選んだ．詳細は第 4.3 節でも述べる．最適化は Optuna を用いて行われ，最適化対象は f1-score，試行回数は 100 回とし，この最適化を 10 セット行ってその平均値を取ることで最適化閾値を取得した．これと，デフォルトの閾値である 0.50 を用いて ChEMBL データセット，GDBMedChem データセットと MERMAID データセットの 3 つのテストデータセットに対して f1-score を計算し，評価を比較した．f1-score の算出には scikit-learn の f1_score メソッドを利用した．

4. 結果

実験 1, 2, 3 を行った結果を以下の章で示す．

4.1 結果 1

作成した 2 種類のモデルと既存モデルの各テストデータセットに対する評価結果はそれぞれ表 1, 2, 3 のようになった．

weight add model では，base line（既存モデル）に比べて負例に対する正答率が向上し，正例に対する正答率・AUC が低下している．また，loss が MERMAID 由来のテストデータ以外で大きく悪化している．

weight reset model では，base line に比べて負例に対する正答率・ AUC・ loss がほぼ同等かそれ以上になり，正例に対する正答率が低下している．

以下では，本結果において既存モデルに比べてより負例に対する正答率と AUC の向上，loss の低下がみられた weight reset model を「本モデル」と呼ぶこととし，RAscore と本モデルとで評価が分かれた化合物の例は図 6 に示す．これらの化合物は GDBMedChem のデータセットから選択され，正例・負例についてそれぞれの評価が 0.8 以上のものと 0.2 以下のものとを選択した．なお，図中の True/False については本モデルの分類を基準とした記載である．

表 1 ChEMBL 由来のテストセットへの評価

評価メトリクス	base line	add model	reset model
loss	**0.38**	7.09	0.5
AUC	**0.93**	0.69	0.91
Accuracy	0.07	0.01	**0.14**
Binary accuracy	**0.9**	0.43	0.88
Precision	0.92	**0.98**	0.91
Recall	**0.95**	0.25	0.94
True negatives	3647	**4883**	3522
True positives	**14254**	3784	14062
False negatives	**740**	11210	932
False positives	1294	**58**	1419

表 2 MERMAID 由来のテストセットへの評価

評価メトリクス	base line	add model	reset model
loss	0.87	0.18	**0.14**
AUC	0.85	0.75	**0.86**
Accuracy	0.0	**0.04**	0.0
Binary accuracy	0.79	**0.98**	0.97
Precision	0.09	**0.63**	0.34
Recall	**0.82**	0.29	0.39
True negatives	15399	**19408**	19116
True positives	**411**	147	199
False negatives	**93**	357	305
False positives	4097	**88**	380

表 3 GDBMedChem 由来のテストセットへの評価

評価メトリクス	base line	add model	reset model
loss	2.14	3.28	**1.16**
AUC	0.74	0.62	**0.83**
Accuracy	**0.01**	0.0	**0.01**
Binary accuracy	0.55	**0.82**	0.72
Precision	0.3	**0.78**	0.41
Recall	**0.85**	0.78	0.82
True negatives	3647	**7807**	5497
True positives	**1769**	343	1698
False negatives	**305**	1731	376
False positives	4198	**97**	2407

4.2 結果 2

本モデルによるスコアリングの分布は図 7 のようになった．既存モデル（図 3）と比較し，負例（赤）の分布が正しく 0 に偏るようになった一方で，正例（青）の分布は 1 への偏りが小さくなった．

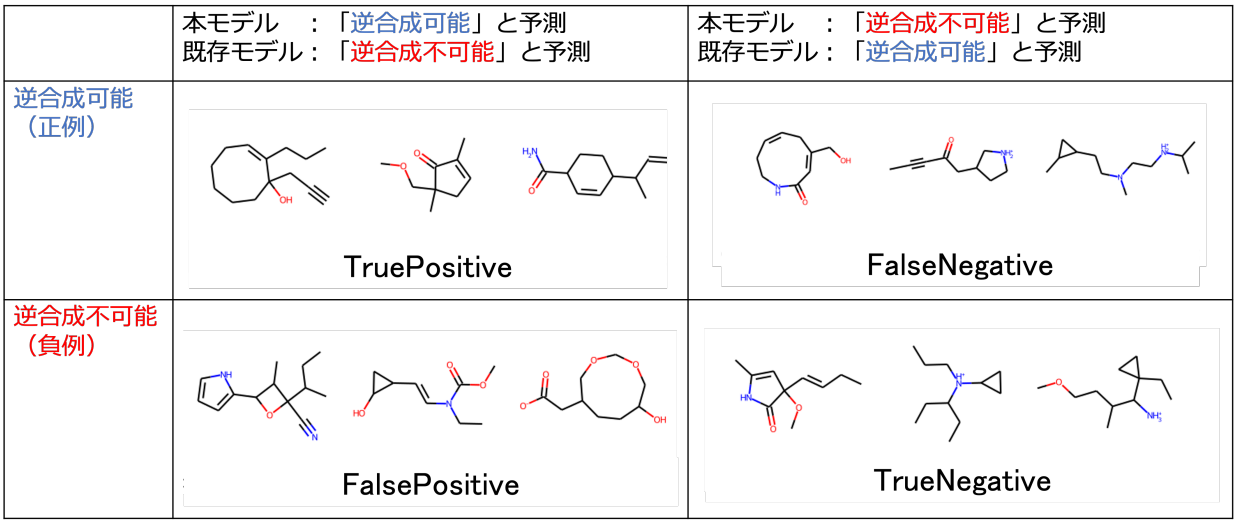

図 6　RAscore と本モデルとで評価が分かれた化合物の例

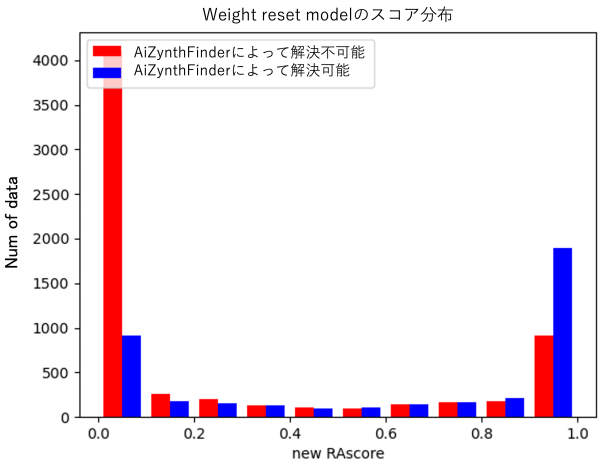

図 7　MERMAID によって作成したデータセットを SAscore によってフィルタリングしたものに対する本モデルの分布．
青：実際には逆合成解析可能（正例），赤：逆合成不可能（負例）

表 4　各データセットに対する本モデルの f1-score

閾値	ChEMBL	MERMAID	GDBMedChem
0.50(default)	**0.9228**	**0.3674**	0.5496
0.59(optimized)	0.9219	0.3624	**0.5525**

4.3　結果 3

本モデルに対し，閾値の最適化を行った結果，0.59 を得た．これをデフォルトの閾値である 0.50 と f1-score を比較した結果が表 4 で，ほぼ同等の結果となった．

5.　考察

以上の結果より，Weight add model では負例の正答率が最も良いにも関わらず，loss や AUC が悪化している．この理由としては，Weight add model では RAscore モデルによる ChEMBL データセットの学習結果である重みを，パラメータの初期値としてしか利用していないことでそのため学習した情報をうまく引き継ぐことができなかったことなどが考えられる．さらに後から負例を多く含む MERMAID のデータを多数学習したために，重みパラメータの上書きのような状態になり，負例に過剰適合してしまった可能性が挙げられる．

Weight reset model の正例の正答率の低下は正例の学習数の減少，負例の正答率の向上は負例の学習データ数の増加が単純な理由として考えられる．そのほか，AUC が向上していることやほぼ全てのデータセットに対して loss がの結果が良くなっていること，既存モデルや Weight add model に学習データへの過剰適合があった可能性，などを考慮すると，他の 2 モデルのような過剰適合を避けることができた結果として，正例単体や負例単体での精度はそれぞれのモデルより劣るものの，全体としては性能が向上したとも考えられる．

しかし，全体の AUC や Binary Accuracy などを見れば Weight reset model は RAscore モデルより優れているが，一方で RAscore モデルでは偽陽性が多いというデメリットの一方で偽陰性が少ないというメリットはあり，逆に本モデルでは AUC や Binary Accuracy が向上しているものの，偽陽性を減らした結果偽陰性が増えてしまっている．したがってユースケース次第では既存モデルの方が利用価値が高いということも考えられる．

また，閾値の最適化によってあまり性能の向上が見られなかったことについては，そもそもデフォルトの閾値と最適化閾値の差はあまり大きくなかったこと，加えて図 3 や図 7 に見られるように，0.5 付近にはあまり分布がない事から閾値を 0.5 付近で変化させてもあまり変化が見込めないことなどが考えられる．

6. 結論

6.1 本研究のまとめ

既存の RAscore モデルの学習における問題点を指摘し，より実用的なモデル構築のために，新たに分子生成モデルを利用したデータセットを構築し，学習しなおした．その結果，新たなモデルでは負例の正答率・AUC・loss が向上した．これにより実際に逆合成解析を行うよりも早く既存の RAscore モデルよりも AUC 等が優れた予測が可能になったが，実際は RAscore モデルに比べ偽陽性の数は減らせたものの，偽陰性の数は増加しており，現状ではユースケース次第で使い分けることで化合物の効率的なフィルタリングに貢献すると考えられる．

6.2 今後の課題

今後の課題として，負例への正答率を下げずに正例の正答率をさらに改善する必要が挙げられる．この展望としては，単純にもっと多くのデータを利用して学習することや，今回はモデルの改善の寄与とデータセットの改善の寄与を分けて考える都合上行わなかったが，モデルのハイパーパラメータやネットワーク構造の再最適化をすることなどが考えられる．また，改良によってスコア 0.5 付近にも分布するようになった場合，閾値の最適化がより重要な要素になりうるため，より優れた閾値の最適化方法についても議論の余地があると考えられる．そのほか，既存モデルと本モデルとの間で評価が分かれた化合物についても何らかの関連性を見出すことができれば，よりモデルの理解に貢献し，モデルの改良につながる可能性がある．

参考文献

[1] Yamamoto, K. Z., Yasuo, N. and Sekijima, M.: Screening for Inhibitors of Main Protease in SARS-CoV-2: In Silico and In Vitro Approach Avoiding Peptidyl Secondary Amides, *Journal of Chemical Information and Modeling*, Vol. 62, No. 2, pp. 350–358 (2022).

[2] DiMasi JA, Grabowski HG, H. R.: Innovation in the pharmaceutical industry: New estimates of R&D costs., *J Health Econ*, Vol. 47, No. 4, pp. 20–33 (2016).

[3] Ertl, P. and Schuffenhauer, A.: Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions, *Journal of Cheminformatics*, Vol. 1, No. 1, p. 8 (2009).

[4] Voršilák, M., Kolář, M., Čmelo, I. and Svozil, D.: SYBA: Bayesian estimation of synthetic accessibility of organic compounds, *Journal of Cheminformatics*, Vol. 12, No. 1, p. 35 (2020).

[5] Coley, C. W., Rogers, L., Green, W. H. and Jensen, K. F.: SCScore: Synthetic Complexity Learned from a Reaction Corpus, *Journal of Chemical Information and Modeling*, Vol. 58, No. 2, pp. 252–261 (2018).

[6] Thakkar, A., Chadimová, V., Bjerrum, E. J., Engkvist, O. and Reymond, J.-L.: Retrosynthetic accessibility score (RAscore) – rapid machine learned synthesizability classification from AI driven retrosynthetic planning, *Chem. Sci.*, Vol. 12, pp. 3339–3349 (2021).

[7] Akiba, T., Sano, S., Yanase, T., Ohta, T. and Koyama, M.: Optuna: A Next-Generation Hyperparameter Optimization Framework, pp. 2623–2631 (2019).

[8] Genheden, S., Thakkar, A., Chadimová, V., Reymond, J.-L., Engkvist, O. and Bjerrum, E.: AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning, *Journal of Cheminformatics*, Vol. 12, No. 1, p. 70 (2020).

[9] Gaulton A, Hersey A, N. M. e. a.: The ChEMBL database in 2017, *Nucleic Acids Res*, Vol. 45, No. D1, pp. D945–D954 (2017).

[10] Erikawa, D., Yasuo, N. and Sekijima, M.: MERMAID: an open source automated hit-to-lead method based on deep reinforcement learning, *Journal of Cheminformatics*, Vol. 13, No. 1, p. 94 (2021).

[11] Awale, M., Sirockin, F., Stiefl, N. and Reymond, J.-L.: Medicinal Chemistry Aware Database GDBMedChem, *Molecular Informatics*, Vol. 38 (2019).

[12] A-Thakkar: RAscore, *https://github.com/reymond-group/RAscore* (2020).

[13] F.Chollet: Keras, *https://Github.Com/Fchollet/Keras* (2015).