

適用領域を考慮した薬剤標的親和性予測

杉田 駿也¹ 大上 雅史^{1,a)}

概要： 計算機による創薬支援のため、化合物とタンパク質の親和性を網羅的に予測する薬剤標的親和性予測が行われている。化合物-タンパク質間の相互作用の既知情報、化合物の構造情報、タンパク質の配列情報などを用いた教師あり学習に基づく予測手法が、これまでに数多く報告されてきた。しかし、化合物や標的分子の化学空間は広範であるため、常に予測が上手くいくわけではない。本研究では薬剤標的親和性予測において予測モデルの出力の信頼性を評価する適用領域を用いて予測結果を分類する手法を提案する。薬剤標的親和性予測における最新手法である勾配ブースティング木とグラフ畳込みニューラルネットワークに基づく予測手法で検証した結果、構築した適用領域が実際に回帰予測精度を改善することを示した。

1. 序論

創薬研究開発において、薬剤候補化合物の標的親和性を計算機で予測することは、初期段階での化合物スクリーニングや未知の標的に対する結合能の検証に有用である。特に、薬剤と標的の両方の情報を利用した機械学習による予測手法は、薬剤と標的の親和性を迅速かつ包括的に予測することができる [1]。

畳込みニューラルネットワーク (convolutional neural network, CNN) は、タンパク質-リガンド複合体の3次元構造から得られる詳細な情報を用いて、薬剤-標的ペアの親和性値を予測する深層学習研究で採用されている [2, 3]。3次元構造から得られる情報は、薬剤-標的相互作用の構造的メカニズムをよく表すことができるが、これらの研究は複合体構造データの入手に依存している。

薬剤と標的の両方の情報を統合して薬剤標的親和性 (drug-target affinity, DTA) を予測するケモゲノミクス [1] は、必ずしも3次元構造を必要としない。ケモゲノミクス手法は、既存の薬剤標的相互作用データから、検証されていない薬剤標的ペアの相互作用を予測できる点で魅力的である。ケモゲノミクスの手法は、従来、2値分類問題としてアプローチされてきた [1, 4, 5, 6]。近年では、利用可能なデータ数の増加や機械学習の予測性能の向上に伴い、回帰問題としてDTAを予測するアプローチも用いられる [7, 8, 9, 10, 11]。2017年に提案されたSimBoost [7] は、薬剤間および標的間の類似性由来する特徴を持つ勾配ブースティング法を用いてDTAの予測に用いる手法であり、高い予測精度を実現している。

一方で、化合物空間は巨大であり、薬剤と標的のペアの学習による予測は必ずしも信頼できるとは限らない。実際、定量的構造活性相関 (quantitative structure-activity relationship, QSAR) の分野では、合理的に予測できる化合物空間の領域である適用領域 (applicability domain, AD) の研究が盛んに行われている [12]。QSARは、化学構造と生物活性の関係を明らかにするための予測・解析手法であり、より良い化学物質を発見するために開発・発展してきた分野である [13]。ADとは、機械学習モデルがそれなりの性能を発揮できる領域のことである。モデルからサンプルまでの距離を定量化することで、予測モデルが信頼性の高い予測を行う領域を推定できる。これにより、未知のサンプルに対する予測の不確かさを推定し、モデルによる誤りを回避することができる。

DTA予測問題にADを導入した手法としてSimBoostを派生させたSimBoostQuant [7]がある。SimBoostQuantは交差検証における標準誤差と信頼区間からADを決定するアンサンブル学習ベースの方法である。しかしアンサンブル学習に基づくADは、学習データが少ない場合、すなわちデータ密度が低い場合には、信頼区間自体の信頼性が低くなるという問題がある。信頼性の高いADを構築するためには、データ密度を考慮することが重要である [14]。

本研究では、学習データのデータ密度によって決まるADを考慮したDTA予測手法を提案する。データ密度を用いた直感的な手法であるk-nearest neighbor (k-NN)法を用いて、薬剤標的親和性予測のADを構築した。ADを構築することで、予測対象のサンプルがADの内側にある(In-AD)か外側にある(Out-AD)かを事前に推定することが可能となる。勾配ブースティング木およびグラフCNN

¹ 東京工業大学 情報理工学系

^{a)} ohue@c.titech.ac.jp

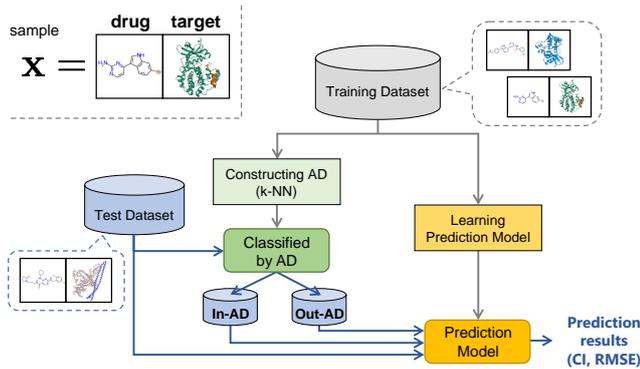


図 1: 適用領域を用いた予測の枠組みと評価方法

による予測モデルを用いた検証の結果、サンプルが AD の内側にあるか外側にあるかによって予測結果の信頼性が異なり、AD の内側に分類されたサンプルの予測精度がより良好であることがわかった^{*1}。

2. 手法

本研究で行ったデータセットの分割、学習、適用領域 (Applicability Domain, AD) の構築、推論の概要を図 1 に示す。評価は、テストデータ全体、AD により適用領域内 (In-AD) として分類されたサンプル、AD により適用領域外 (Out-AD) として分類されたサンプルそれぞれを対象として算出した。

2.1 データセット

提案手法の精度の評価のために、表 1 に示す 2 つのデータセット、Davis [16], kinase inhibitors bioactivity (KIBA) [17] を用いた。 K_i , K_d , IC_{50} の目標活性値 (実数値) のラベルまたは、これらを利用したスコアでラベル付けされている。なお、Davis データセットは全ての薬物-標的ペアに活性値のラベルがあるが、KIBA データセットは全ての薬剤-標的ペアに活性値のラベルがあるデータセットではない (表 1 を参照)。

Davis データセットは、68 の薬剤と 442 のターゲットの全ペアの結合親和性が K_d 値 (キナーゼ解離定数) で測定されている。データは、先行研究に従い、 $pK_d = -\log(K_d \cdot 10^{-9})$ で pK_d 値に変換した [7, 8]。KIBA データセットは、 K_i , K_d , IC_{50} といった異なるソースの活性を組み合わせて、KIBA スコアという形で統一されている [17]。

各データセットを 4 : 1 の割合でトレーニングデータセットとテストデータセットに分割し、トレーニングデータセットで予測モデルの学習と AD の構築を行った。

表 1: 薬剤標的親和性 (drug-target affinity, DTA) データセット。括弧内の数値は全組合せ ($\#drug \times \#target$) に対する割合である。

Dataset	#drug	#target	#sample
Davis [16]	68	442	30,056 (100%)
KIBA [17]	2,116	229	118,234 (24.4%)

2.2 予測モデルの学習

DTA の予測モデルには、SimBoost [7] および DGraphDTA [10] を使用した。実装は公開されているコード [18, 19] を使用した。

SimBoost [7] は、各薬剤標的ペアに対して薬剤類似性、標的類似性、薬剤標的親和性から特徴量を生成し、98 次元の特徴ベクトル \mathbf{x} を利用する決定木による機械学習手法である。ハイパーパラメータは文献 [7] で最適化された値を使用した。

DGraphDTA [10] は、グラフ畳み込みニューラルネットワーク (graph convolutional neural networks, GCNN) を利用する深層学習手法である。薬剤は原子同士の、標的はアミノ酸同士の結合をグラフ表現として入力特徴量に変換する。ハイパーパラメータは文献 [10] で最適化された値を使用した。

2.3 適用領域の構築

本研究では、予測モデルがテストデータセット中のサンプルをカバーしているかどうかを事前に推定するために、AD を利用した。本研究では、データ密度に基づいて AD を求めるために、k-NN の手法を利用する。データセット中の学習サンプルに含まれる Out-AD の割合をあらかじめ ν で与える。

k-NN は、あるテストサンプルとそのテストサンプルの近傍にある k 個の学習サンプルとの距離を用いて、データ密度から AD を決定するために用いられる手法である。サンプル間の距離は、ユークリッド距離で $\mathbf{w}^T \mathbf{x}$ を用いて計算する。これは、scikit-learn の RobustScaler メソッドに従って定義されたスケール付き特徴ベクトル \mathbf{x} に重み \mathbf{w} を乗算したものである。 \mathbf{w} は学習モデルで計算される特徴量の重要度である。

決定木では特徴量の重要度は、ノードに到達したサンプル数をサンプル総数で割った値から算出される。深層学習による特徴量の重要度は、ある特徴量 x_i をサンプル間でランダムにシャッフルした場合の精度の変化分から算出される [20]。この値が大きいほど、その特徴はより重要である。

2.4 評価指標

提案手法の性能評価には、DTA 予測値の順序関係の正しさを評価する C 統計量 (concordance index, CI) と、DTA 予測値の総合的な精度を評価する平均二乗偏差 (root mean

*1 本稿の一部 (SimBoost による実験結果および考察) は、国際会議 18th IEEE International Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2021) の Conference Full Paper として出版されている [15]。

squared error, RMSE) を使用した。

CI は以下の式で計算する。

$$CI = \frac{\sum_{y_i > y_j} h(\tilde{y}_i - \tilde{y}_j)}{\sum_{y_i > y_j} 1} \quad (1)$$

$$h(x) = \begin{cases} 1 & (x > 0) \\ 0.5 & (x = 0) \\ 0 & (x < 0) \end{cases} \quad (2)$$

ここで、 y_i, y_j は親和性の真の値であり、 \tilde{y}_i, \tilde{y}_j は、親和性のモデルによる予測値である。

RMSE は一般的に予測との誤差を測るために利用される指標の一つであり、以下の式で計算する。

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \tilde{y}_i)^2} \quad (3)$$

ここで、 n は評価に用いたサンプルの数である。

3. 実験と考察

3.1 In-AD と Out-AD における予測精度

各データセットにおける適用領域内 (In-AD) および適用領域 (Out-AD) の精度を表 2 および表 3 に示した。k-NN のパラメータは $k = 5$ に設定した。これは、Out-AD のサンプルの RMSE が k の値が小さいほど悪くなるためである。なお、表 2 および表 3 について、訓練データとテストデータの分割が異なるため、結果の単純な比較はできないことに注意する。

提案手法は、AD が狭くなるほど RMSE の精度を良くする。Out-AD における RMSE はベースラインの精度より悪く、予測の難しいサンプルが Out-AD に分類できていることが分かる。In-AD と Out-AD の分類により、より確実な予測ができることを示した。一方、CI への寄与は限定的である。予測精度が低いものを取り除けば順序関係の予測はより高精度になる。原因は、取り除いたサンプルの割合が比較的低いことが考えられる。

図 2 および図 3 に、In-AD と Out-AD に分類されたサンプルの予測と真値の散布図を示す。本来なら Out-AD データは有効な予測ができず、予測誤差も大きくなるはずだが、散布図からは大幅な相関の悪化は起こっていないように見える。SimBoost や DGraphDTA の予測精度が高く、また各データセットのデータ密度が高いことに起因していると考えられる。ただし、RMSE や CI の値からもわかるように、実際には Out-AD における相関は In-AD よりも悪化しており、AD の分類結果によって予測性能の信頼性が向上できているといえる。

3.2 各 ν がテストサンプルを Out-AD に分類する割合

ν は訓練データが Out-AD に分類される比率である。パ

ラメータ ν がどのようにテストサンプルを分類するかを調べた。SimBoost における結果を表 4 に示す。表 4 は、各 ν (訓練データ中で Out-AD に分類される比率) を与えたときの、テストサンプル中の Out-AD と判別された標本の比率を示す。その結果、 ν と Out-AD と判別されたテストサンプルの比率はほぼ同じであることがわかった。これは、訓練データとテストデータの標本が類似した分布を示していることを意味する。しかし、実際には、訓練用データと親和性を知りたいサンプルの分布が似ていることは稀である。したがって、 ν を決定するためのトレーニングデータは、利用するターゲットによって慎重に選択する必要がある。

3.3 適用領域の範囲パラメータ ν の効果

表 2 に示す通り、In-AD に分類された RMSE は ν が増加する (AD が狭くなる) に従い、より精度が良くなる傾向がある。検討のため、Davis データセットに対して、k-NN 法 ($k = 5$) により、様々な ν の場合について検討した。SimBoost における結果を表 5 に示す。

k-NN 法では、全ての ν に対して RMSE を改善する AD が決定できている。RMSE だけでなく CI も若干改善される傾向にある。 ν が大きいとき、AD が狭いことを示すため、データ密度に基づく AD が狭いほど、AD 内のサンプルの予測は信頼性が高いことが示唆される。また、 $\nu = 0.10$ の比較的広い AD では、Out-AD と分類されるサンプルの数は少ない。このような少数の Out-AD サンプルの予測精度はかなり低い。逆に、 $\nu = 0.90$ の比較的小さな AD では、In-AD のサンプルの予測精度は良好である。したがって、予測を適用する状況に応じて、緩い AD を求めるか、厳しい AD を求めるかを考慮し、所望の AD が得られるようにパラメータ ν を設定することが必要である。

3.4 k-NN パラメータ k の効果

表 6 は SimBoost モデルにおいて Davis データセットの In-AD と Out-AD それぞれの k の値を変化させたときの RMSE の値を示す。表 6(a) から、 k の値は $k = 1$ から 5 の間、AD におけるサンプルの信頼性にほとんど影響を与えない。逆に、表 6(b) では、 k の値が小さいほど Out-AD のサンプルの RMSE が悪くなっており、 k の値が大きくなるほど適用領域が狭くなることが示唆される。しかし、 k を 10 や 100 にしても In-AD の RMSE は改善されないことから、 $k = 1$ から 5 が適当であることが示唆される。本研究で構築した k-NN ベース AD では、パラメータ k を 5 とし、 ν は実際のデータやアプリケーションの状況に応じてユーザが調整することを推奨する。

3.5 SimBoost と DgraphDTA における結果の差異

SimBoost[7] 及び DGraphDTA[10] に対する AD の構築

表 2: SimBoost における各テストデータセットの予測性能 ($k = 5$). ベースラインは, SimBoost で学習した予測モデルを AD なしで直接全テストデータセットに適用した場合の予測精度である. 太字は各データセットにおける In-AD の RMSE が最良であり, 斜体は各データセットにおける Out-AD の RMSE が最悪である値を表す.

Method	ν	Davis dataset				KIBA dataset			
		In-AD		Out-AD		In-AD		Out-AD	
		CI	RMSE	CI	RMSE	CI	RMSE	CI	RMSE
Baseline	-	0.901	0.488	-	-	0.835	0.452	-	-
k-NN	0.02	0.901	0.482	0.873	<i>0.742</i>	0.835	0.437	0.791	<i>0.944</i>
	0.04	0.899	0.476	0.873	0.705	0.836	0.429	0.801	0.841
	0.06	0.898	0.476	0.879	0.664	0.836	0.422	0.810	0.796
	0.08	0.899	0.470	0.878	0.653	0.836	0.417	0.814	0.749
	0.10	0.900	0.468	0.879	0.633	0.834	0.411	0.821	0.731

表 3: DGraphDTA における各テストデータセットの予測性能 ($k = 5$). ベースラインは, DGraphDTA で学習した予測モデルを AD なしで直接全テストデータセットに適用した場合の予測精度である. 太字は各データセットにおける In-AD の RMSE が最良であり, 斜体は各データセットにおける Out-AD の RMSE が最悪である値を表す. 表 2とは学習とテストサンプルの分割が異なる.

Method	ν	Davis dataset				KIBA dataset			
		In-AD		Out-AD		In-AD		Out-AD	
		CI	RMSE	CI	RMSE	CI	RMSE	CI	RMSE
Baseline	-	0.894	0.485	-	-	0.892	0.384	-	-
k-NN	0.02	0.894	0.481	0.871	0.871	0.892	0.380	0.886	<i>0.564</i>
	0.04	0.892	0.476	0.873	0.873	0.892	0.377	0.892	0.530
	0.06	0.893	0.472	0.874	0.874	0.892	0.374	0.891	0.519
	0.08	0.893	0.468	0.878	<i>0.878</i>	0.892	0.370	0.888	0.520
	0.10	0.894	0.466	0.873	0.873	0.892	0.365	0.886	0.530

表 4: SimBoost における各 ν の Out-AD と分類されたテストサンプルの比率

ν	Davis	KIBA
0.02	0.018	0.019
0.04	0.043	0.039
0.06	0.055	0.058
0.08	0.082	0.079
0.10	0.104	0.098

において, 同様の傾向の結果が得られた. これはモデルの精度が良く, モデルによるデータセットの予測がデータセット毎に似通っていることが一つの原因と考えられる.

一方, Davis データセットにおいて, SimBoost[7] では, ν の増加 (AD を狭くする) に伴い RMSE が良化した, DGraphDTA では k が小さい ($k \leq 5$) において極端に ν が狭い場合 ($\nu \sim 0.8$) に, RMSE が悪化し, In-AD における精度が Out-AD における精度を下回る結果となった. 実用上 AD を極端に狭くすることは考えにくい, もし AD を厳しく設定したい場合には, k をある程度大きく取る必要がある可能性がある.

また, DGraphDTA に AD を利用する際に特徴量ベクトルを特徴量の重要度で重みづけは, SimBoost に AD に利

表 5: Davis データセットによる様々な ν 値に対する予測性能. ベースラインは, SimBoost で学習した予測モデルを AD なしで直接テストデータセットに適用した場合の予測精度である. 太字は In-AD の RMSE が最小のとき, 斜体は Out-AD の RMSE が最大を表す.

method	ν	In-AD		Out-AD	
		CI	RMSE	CI	RMSE
Baseline	-	0.901	0.488	-	-
k-NN	0.1	0.900	0.468	0.879	<i>0.633</i>
	0.2	0.903	0.459	0.879	0.591
	0.3	0.903	0.450	0.885	0.566
	0.4	0.905	0.438	0.888	0.555
	0.5	0.906	0.433	0.891	0.536
	0.6	0.906	0.404	0.894	0.533
	0.7	0.907	0.386	0.896	0.524
	0.8	0.910	0.376	0.897	0.512
	0.9	0.916	0.324	0.897	0.502

用する際に特徴量ベクトルを特徴量の重要度で重みづけよりも上手くはたらかなかった. 考えられる原因は, 深層学習における特徴量の重要度の算出では, 特徴ベクトルに相関とき上手くはたらかない場合があることである.

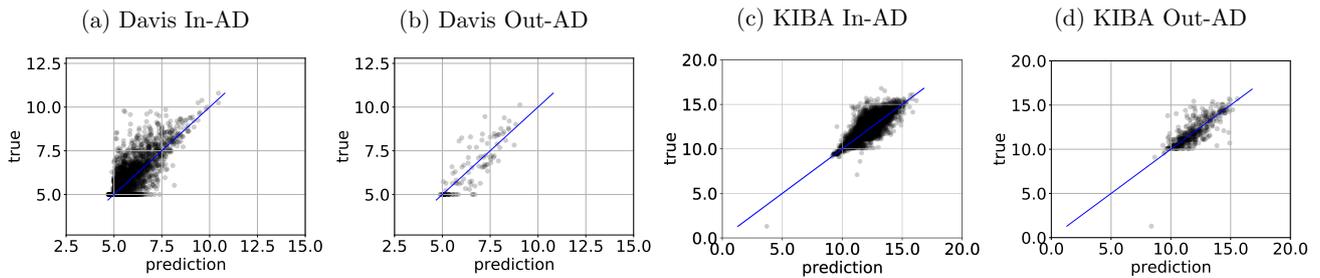


図 2: SimBoost における各データセットの散布図 ($\nu = 0.02, k = 5$)

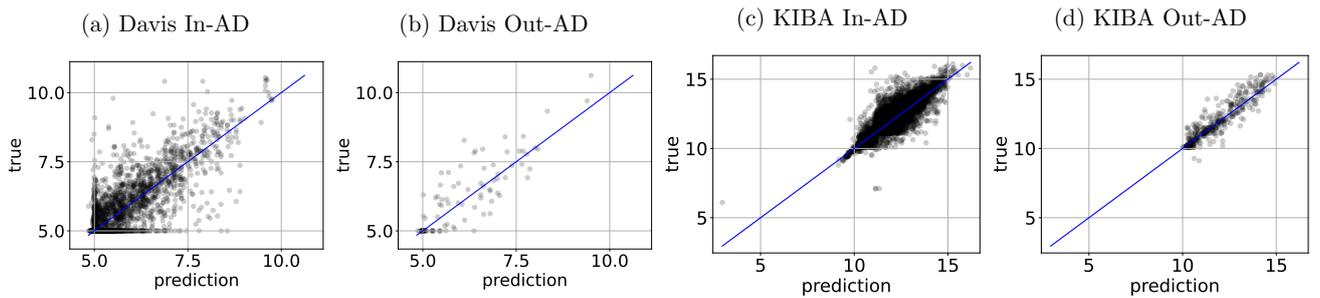


図 3: DGraphDTA における各データセットの散布図 ($\nu = 0.02, k = 5$)

表 6: SimBoost モデルにおいて Davis データセットの k の値を変化させたときの RMSE. 太字は In-AD の RMSE の最小値, 斜体は Out-AD の RMSE の最大値を示している.

(a) In-AD (baseline RMSE = 0.488)						
$k \setminus \nu$	0.01	0.02	0.03	0.04	0.05	0.10
1	0.482	0.478	0.477	0.477	0.475	0.469
2	0.484	0.478	0.477	0.476	0.476	0.468
3	0.484	0.479	0.477	0.476	0.476	0.469
4	0.485	0.480	0.477	0.475	0.475	0.468
5	0.485	0.481	0.476	0.476	0.476	0.468
10	0.485	0.482	0.479	0.477	0.477	0.470
100	0.487	0.487	0.485	0.483	0.483	0.472

(b) Out-AD (baseline RMSE = 0.488)						
$k \setminus \nu$	0.01	0.02	0.03	0.04	0.05	0.10
1	<i>0.901</i>	0.831	0.752	0.708	0.687	0.631
2	0.806	0.831	0.760	0.712	0.672	0.641
3	0.758	0.807	0.756	0.695	0.670	0.634
4	0.733	0.790	0.760	0.709	0.687	0.640
5	0.741	0.746	0.773	0.718	0.675	0.640
10	0.695	0.726	0.708	0.706	0.664	0.624
100	0.607	0.544	0.560	0.546	0.565	0.614

3.6 関連研究との比較 - SimBoostQuant

SimBoostQuant モデルは、分位点回帰に基づいて AD を設定する方法である。このモデルは SimBoost と共に文献 [7] で提案された。Davis や KIBA はデータセットが高密度であるため、SimBoostQuant は効果的である。データセットが疎な場合、データ密度の低い領域での予測区間が信頼できないため、分位点回帰は有効でない。逆に、

SimBoostQuant モデルは文献で指摘されているように、予測区間が出力であり、予測精度は信頼区間の中央値で計算されるため、オリジナルの SimBoost と比較して正確な予測は得られないと考えられる。提案手法は、SimBoost モデルを使用しながら、学習データとモデルを使ってテストデータの結果を予測する能力に信頼性を持たせているため、優れていると言える。しかし、表 2 に示すように、テストデータ全体と In-AD に分類されたデータの予測精度に大きな差はない。ただし、ベンチマークデータセットの状況は、現実とはやや離れていることに注意が必要である。創薬分野では実際のデータで高いデータ密度を維持することは難しく、低いデータ密度の状況でも有効に機能する手法が求められている。その意味で、提案手法にはメリットがある。

4. 結論

本研究では、薬物標的親和性予測の精度と信頼性を向上させるために、適用領域 (AD) を用いた予測手法を提案した。k-NN を用いてデータ密度に基づく AD を構築し、2 つのデータセットで予測精度の検証を行った。その結果、k-NN によって構築された AD は、予測が難しいテストサンプルを Out-AD として分類できることが示された。異なる 2 つの予測モデルでほとんど同様の結果が得られており、AD の利用によって機械学習による予測結果の信頼性を高めることができることを示した。

今回用いたデータセットはいずれも高密度であった。学習データがカバーする薬物空間や標的空間がより広範で疎になった場合、AD にどのような影響を与えるかは未

知数である。複数のデータセットが混在する場合や BindingDB [21] のようなデータベースを利用する場合、学習データが構築する化合物空間やターゲット空間は広範かつ疎になる。この場合、本研究で提案したデータ密度ベースの AD はより効果的に働く可能性がある。

また、他の AD 構築のアプローチ、例えばアンサンブル学習とデータ密度を組み合わせた手法 [22] を用いることもある。アンサンブル学習は学習データ密度の高い領域で有効に働くため、アンサンブル学習とデータ密度の組合せ手法も有効である [14]。

謝辞 本研究は、JST ACT-X JPMJAX20A3, 科研費基盤研究 (B) 20H04280, AMED BINDS JP20am0101112 の支援を受けて行われた。

参考文献

- [1] X. Chen *et al.*: Drug-target interaction prediction: Databases, web servers and computational models, *Brief Bioinform*, Vol. 17, No. 4, pp. 696–712 (2016).
- [2] M. M. Stepniewska-Dziubinska *et al.*: Development and evaluation of a deep learning model for protein–ligand binding affinity prediction, *Bioinformatics*, Vol. 34, No. 21, pp. 3666–3674 (2018).
- [3] J. Jiménez *et al.*: K_{DEEP} : Protein-Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks, *J Chem Inf Model*, Vol. 58, No. 2, pp. 287–296 (2018).
- [4] Y. Yamanishi *et al.*: Prediction of drug-target interaction networks from the integration of chemical and genomic spaces, *Bioinformatics*, Vol. 24, No. 13, pp. i232–i240 (2008).
- [5] T. Ban *et al.*: NRLMF β : Beta-distribution-rescored neighborhood regularized logistic matrix factorization for improving the performance of drug–target interaction prediction, *Biochem Biophys Rep*, Vol. 18, p. 100615 (2019).
- [6] M. Ohue *et al.*: Link Mining for Kernel-Based Compound-Protein Interaction Predictions Using a Chemogenomics Approach, In *Proc ICIC 2017, Lecture Notes Comput Sci*, Vol. 10362, pp. 549–558 (2017).
- [7] T. He *et al.*: SimBoost: a read-across approach for predicting drug-target binding affinities using gradient boosting machines, *J Cheminform*, Vol. 9, No. 1, p. 24 (2017).
- [8] H. Öztürk *et al.*: DeepDTA: Deep drug-target binding affinity prediction, *Bioinformatics*, Vol. 34, No. 17, pp. i821–i829 (2018).
- [9] T. Nguyen *et al.*: GraphDTA: Predicting drug target binding affinity with graph neural networks, *Bioinformatics*, Vol. 37, No. 8, pp. 1140–1147 (2021).
- [10] M. Jiang *et al.*: Drug-target affinity prediction using graph neural network and contact maps, *RSC Adv*, Vol. 10, No. 35, pp. 20701–20712 (2020).
- [11] K. Wang *et al.*: DeepDTAF: A deep learning method to predict protein-ligand binding affinity, *Brief Bioinform*, Vol. 22, No. 5 (2021).
- [12] S. Dimitrov *et al.*: A stepwise approach for defining the applicability domain of SAR and QSAR models, *J Chem Inf Model*, Vol. 45, No. 4, pp. 839–849 (2005).
- [13] T. Scior *et al.*: How to Recognize and Workaround Pitfalls in QSAR Studies: A Critical Review, *Curr Med Chem*, Vol. 16, No. 32, pp. 4297–4313 (2009).
- [14] H. Kaneko, K. Funatsu: Applicability domain based on ensemble learning in classification and regression analyses, *J Chem Inf Model*, Vol. 54, No. 9, pp. 2469–2482 (2014).
- [15] S. Sugita, M. Ohue: Drug-target affinity prediction using applicability domain based on data density, In *Proc IEEE CIBCB 2021*, pp. 1–6 (2021).
- [16] M. I. Davis *et al.*: Comprehensive analysis of kinase inhibitor selectivity, *Nat Biotechnol*, Vol. 29, No. 11, pp. 1046–1051 (2011).
- [17] J. Tang *et al.*: Making sense of large-scale kinase inhibitor bioactivity data sets: A comparative and integrative analysis, *J Chem Inf Model*, Vol. 54, No. 3, pp. 735–743 (2014).
- [18] SimBoost, <https://github.com/hetong007/SimBoost/>.
- [19] DGraphDTA, <https://github.com/595693085/DGraphDTA>.
- [20] A. Altmann *et al.*: Permutation importance: a corrected feature importance measure, *Bioinformatics*, Vol. 26, No. 10, pp. 1340–1347 (2010).
- [21] M. K. Gilson *et al.*: BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology, *Nucleic Acids Res*, Vol. 44, No. D1, pp. D1045–D1053 (2016).
- [22] R. P. Sheridan: Using Random Forest To Model the Domain Applicability of Another Random Forest Model, *J Chem Inf Model*, Vol. 53, No. 11, pp. 2837–2850 (2013).