Vol.2019-BIO-57 No.4
2019/3/8

ポケット構造情報を考慮した エンドツーエンド表現学習によるリガンド結合予測

種部俊孝[1,2,a)]　石田貴士[1,b)]

概要：新規薬剤開発の初期工程では, コスト削減のために, 計算機を用いて膨大な数の化合物が収録されたライブラリから薬剤となりうる化合物を選別する化合物バーチャルスクリーニングが行われる. 特に, 既知の薬剤情報が無い場合には, 創薬標的タンパク質の薬剤結合部位 (ポケット) に対して, 薬剤候補化合物 (リガンド) を仮想的に結合させて活性を予測するドッキングシミュレーションが一般的に行われる. しかし, ドッキングシミュレーションでは, 化合物の配座探索や結合エネルギーの評価が行われるため, 計算コストが非常に大きいという問題が存在する. そこで, 本研究では, 既知のタンパク質-化合物間相互作用情報から, 新規標的タンパク質に対する薬剤候補化合物の活性を機械学習を用いて予測する手法を提案する. この際, タンパク質立体構造, 特に, ポケット構造情報を考慮することで, 活性を示すのに重要な結合様式を学習できることが期待される. 本研究では, 化合物構造・タンパク質ポケット構造ともにグラフニューラルネットワークを適用したエンドツーエンド表現学習によって活性予測を行う手法を提案し, AutoDock Vina を用いたドッキングシミュレーションと同等の精度, かつ, 短時間で予測が可能であることを示した. また, アミノ酸配列情報を用いた既存手法よりも高い精度で予測が可能であることも示した.

キーワード：バーチャルスクリーニング, ドッキングシミュレーション, 結合ポケット, エンドツーエンド表現学習, グラフニューラルネットワーク

End-to-end learning based compound activity prediction using binding pocket information

Tanebe Toshitaka[1,2,a)]　Ishida Takashi[1,b)]

Abstract: In the absence of known drug information, activity prediction is generally performed by docking simulation, which calculate the binding energy by virtually combining a pharmaceutical candidate compound with a binding site of a target protein. However, in docking simulation, there is a problem that the calculation cost is high because the comformation search of a compound and evaluation of binding energy are computationally heavy tasks. Therefore, in this research, we propose a machine learning based method to predict the activity of a pharmaceutical candidate compound against a novel target protein using known protein-compound interaction information. In particular, by considering the binding site information, it can be expected to learn the binding mode which is important for the compound to have activity. In this research, we propose a method to predict activity of a compound by end-to-end learning using graph neural network for both compound structure and protein binding site structure. The proposed method showed higher accuracy than a existing method using only amino acid sequence imformation, and achieved equivalent accuracy to docking simulation using AutoDock Vina with much shorter computing time.

Keywords: virtual screening, docking simulation, binding site, end-to-end learning, graph neural network

[1] 東京工業大学情報理工学院情報工学系
Department of Computer Science, School of Computing, Tokyo Institute of Technology, Tokyo, Japan
[2] 産業技術総合研究所・東京工業大学 実社会ビッグデータ活用オープンイノベーションラボラトリ, AIST-Tokyo Tech Real World Big-Data Computation Open Innovation Laboratory
[a)] tanebe@cb.cs.titech.ac.jp
[b)] ishida@cs.titech.ac.jp

1. 序論

創薬にかかるコストは年々増大しており, 現在では, 一つの新規薬剤が開発されるまでに十数年という年月と約三千億円の費用が必要とされている [1]. この問題の解決策の一つとして, 計算機を用いた効率化手法が注目されている. 創薬の初期工程では, 数百万から数千万もの化合物を収録する化合物ライブラリから, 薬剤候補化合物を選別する化合物スクリーニングが行われる. この際, 生化学実験の代わりに, 計算機によって薬剤候補化合物の活性の有無を予測する化合物バーチャルスクリーニングを実施することで, 創薬にかかるコストの削減が期待されている.

化合物バーチャルスクリーニングは大別して, 既知の薬剤情報を用いるリガンドベース手法と創薬標的タンパク質の立体構造情報を用いる構造ベース手法の 2 種類が存在する. このうち, 構造ベース手法では, 創薬標的タンパク質表面の薬剤結合部位 (ポケット) に薬剤候補化合物 (リガンド) を仮想的に結合させて, 薬剤としての所望の性質があるかないかを推定するドッキングシミュレーションが一般的に行われている. ドッキングシミュレーションは, 既知薬剤情報を必要とするリガンドベース手法とは異なり, 既知の薬剤情報が存在しない, あるいは, 十分に無い場合であっても薬剤候補化合物の選別が行える有用な方法である. また, 既知の薬剤情報を使用しないため, 既知薬剤と性質が大きく異なる新規薬剤候補化合物を選別できる可能性があることも利点の一つとして挙げられる.

ドッキングシミュレーションを実行できるソフトウェアは, AutoDock Vina[2], Glide[3], eHiTS[4] など数多く存在しており, 広く利用されている. しかし, ドッキングシミュレーションでは, タンパク質ポケット内で化合物を回転・平行移動させながら配座探索を行う必要や, スコア関数を用いてタンパク質—化合物間の相互作用の強さを評価する必要があるため, 計算コストが非常に大きい (CPU1 コアを用いて 1 つの化合物を評価するのに Glide では 0.2〜2.4 分程度 [3], eHiTS では最速で数秒 [4]) という問題が存在する. 1 つの化合物のドッキングは比較的高速であるが, 化合物バーチャルスクリーニングでは大量の化合物ライブラリから僅かなヒット化合物を探索する必要がある. そのため, もし, 1000 万化合物からなる化合物ライブラリ全体に対して, 10 秒で 1 つの化合物を評価できるドッキングシミュレーションを行なった場合, 1200 CPU days もの時間が必要となり, 金銭的・時間的負担が大きい.

上述の問題を解決するために, 本研究ではドッキングシミュレーションの代わりに機械学習を用いて, リガンドが標的タンパク質に結合するかどうかを予測する手法を提案する. その際, 構造ベース手法と同じくタンパク質の立体構造, 特に, タンパク質ポケット構造を考慮することにより, ドッキングシミュレーションと同様に, 活性を示すのに重要な結合様式を学習できることを期待する. また, 本研究では, 機械学習に入力する特徴量の設計方法にも着目した. 従来は, 入力特徴量を人手によって設計するのが一般的であったが, 近年では, 自然言語処理などの分野で, これまで独立していた特徴量設計とタスクの学習を一つの大きなニューラルネットワークに置き換えて学習するエンドツーエンド表現学習が盛んである. 入出力の関係を直接学習することにより, 人手により設計するよりも優れた入力データの表現を獲得できる可能性があるとされ, 実際に, 機械翻訳などでは従来手法を上回る精度を出している [5].

エンドツーエンド表現学習を活性予測問題に適用した既存研究としては, Tsubaki ら [6] の研究が存在する. Tsubaki らは, 化合物にグラフニューラルネットワーク, タンパク質アミノ酸配列に一次元畳み込みニューラルネットワークを適用したエンドツーエンド表現学習によって, 新規標的タンパク質に対するリガンドの結合有無を予測している. しかし, アミノ酸配列情報よりもポケット構造情報の方が, リガンド結合に関してより多くの有用な情報を含んでいると考えられることから, Tsubaki らの手法には予測精度改善の余地があると思われる. また, アミノ酸配列情報を用いる場合, 標的タンパク質と訓練データセット内のタンパク質との間に配列相同性が存在することを仮定しているが, ポケット構造情報を用いることで, ドッキングシミュレーションと同様に, 配列相同性の無い新規標的タンパク質に対しても予測が可能になると考えられる.

以上を踏まえて, 本研究では, 化合物・タンパク質ポケット構造ともにグラフニューラルネットワークを適用したエンドツーエンド表現学習によって活性予測を行う手法を提案する. 提案手法を用いて, バーチャルスクリーニング性能評価用データセットである MUV データセットに対して予測を行なった結果, AutoDock Vina によるドッキングシミュレーションと同等の精度, かつ, 短時間での予測に成功した. また, Tsubaki らの既存研究と比較しても高い精度で予測が可能なことを示した.

2. 手法

本研究の提案手法は, タンパク質側の特徴量として, 既存手法で用いられたアミノ酸配列情報ではなくポケット構造情報を使用し, 化合物だけではなくタンパク質に対してもグラフニューラルネットワークを適用したものである. ここで, 化合物特徴量の生成方法は Tsubaki らの手法と同じである. 提案手法は以下の 3 つの部分からなる.

(1) **化合物・タンパク質ポケットからグラフ生成**
化合物を頂点 (原子種) と辺 (化学結合) からなるグラフ構造, タンパク質ポケットを頂点 (アミノ酸残基種) と辺 (アミノ酸残基の $C\alpha$ 原子間距離タイプ) からなるグラフ構造であると考え, それぞれをグラフへと変

換する.

(2) グラフニューラルネットワークによる表現学習
1 の操作で変換された化合物グラフ・タンパク質ポケットグラフに対し, グラフニューラルネットワークによる表現学習を行う. この操作により, 化合物グラフとタンパク質ポケットグラフはそれぞれ固定長の実数値ベクトルへと変換される.

(3) 分類器によるリガンド結合予測
2 の操作で変換された化合物ベクトルとタンパク質ポケットベクトルを用いて, 分類器によるリガンド結合予測を行う.

2.1 化合物・タンパク質ポケットのグラフ生成

2.1.1 化合物グラフ生成

化合物は, SMILES 形式の文字列を RDKit[7] によってグラフ構造に変換した. この際, SMILES にドット (非連結を表す) を含む場合はグラフ生成ができないため, データセットから除外した.

2.1.2 タンパク質ポケットグラフ生成

タンパク質は, タンパク質-リガンド複合体構造から結合部位を判別する LPC software[8] によって同定されたリガンド接触残基の情報 (残基種, 座標) を元に, 以下の頂点と辺を持つグラフ構造に変換した (図 1).

- 頂点: アミノ酸残基種 (20 種類)
- 辺: $C\alpha$ 原子間距離タイプ (5 種類: I〜V)
  I : 1.0 Å〜4.8 Å, II : 4.8 Å〜7.0 Å
  III : 7.0 Å〜9.2 Å, IV : 9.2 Å〜11.4 Å
  V : 11.4 Å〜13.6 Å

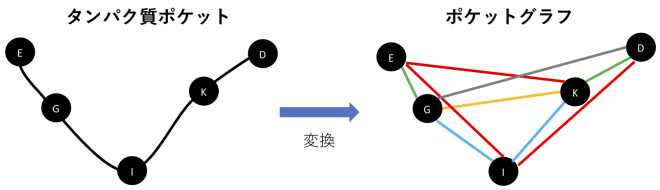

図 1 タンパク質ポケット構造のグラフへの変換

辺の種類は Ito ら [9] の研究を基に設定した. ここで, $C\alpha$ 原子間距離にある程度の幅を持たせて種類分けを行うことで, タンパク質ポケットの構造変化にも対応できることを期待している.

2.2 グラフニューラルネットワークによる表現学習

先述した手順により生成された化合物グラフとタンパク質ポケットグラフは各々グラフニューラルネットワークにより表現学習が行われる. グラフニューラルネットワークでは, 化合物側もタンパク質側も手順は同じく以下の 3 つの部分からなる (図 2).

(1) **Embedding**
頂点と辺の概念をそれぞれ拡張した r-radius vertex と r-radius edge を定義する. そして, 各 r-radius vertex・各 r-radius edge にランダムにベクトルを割り当てる.

(2) **Transition**
各 r-radius vertex・各 r-radius edge ごとに以下の操作を任意回繰り返す.

- 隣接する r-radius vertex・r-radius edge のベクトルを加える
- 非線形関数に入力してベクトルを更新する

(3) **Averaging**
全 r-radius vertex ベクトルを平均して, 一つの実数値ベクトルを出力する.

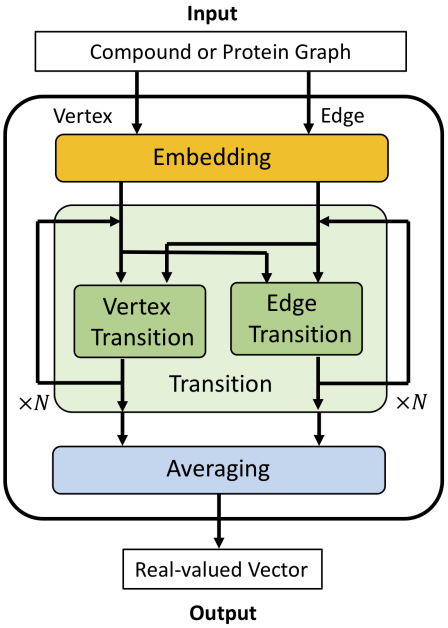

図 2 グラフニューラルネットワークによる表現学習

ここで, それぞれの操作は Tsubaki らの手法と同等である. 以下では Embedding についてのみ説明を行う.

r-radius subgraphs による Embedding

頂点集合 V と辺集合 E からなるグラフ $G = (V, E)$ を考える (化合物の場合, $v_i \in V$ は i 番目の原子, $e_{ij} \in E$ は i 番目と j 番目の原子間の化学結合である. タンパク質ポケットの場合, $v_i \in V$ は i 番目のアミノ酸残基, $e_{ij} \in E$ は i 番目と j 番目の $C\alpha$ 原子間距離タイプである).

ここで, 表現学習の効率を高めるために, 頂点 v_i と辺 e_{ij} にそれぞれ r-radius subgraphs[10] の概念を導入する.

i 番目の頂点から半径 r 以内 (r はホップ数) の全近傍頂点の集合を $\mathcal{N}(i, r)$ とすると, 頂点 v_i に対する r-radius subgraphs は以下の式 (1) で書ける.

$$v_i^{(r)} = \left(V_i^{(r)}, E_i^{(r)}\right) \tag{1}$$

ここで, $V_i^{(r)} = \{v_j | j \in \mathcal{N}(i,r)\}$,
$E_i^{(r)} = \{e_{mn} \in E | (m,n) \in \mathcal{N}(i,r) \times \mathcal{N}(i,r-1)\}$ である.

また, 辺 e_{ij} に対する r-radius subgraphs は以下の式 (2) で書ける.

$$e_{ij}^{(r)} = \left(V_i^{(r-1)} \cup V_j^{(r-1)}, E_i^{(r)} \cap E_j^{(r)}\right) \tag{2}$$

ここで, $v_i^{(r)}$ を r-radius vertex, $e_{ij}^{(r)}$ を r-radius edge とそれぞれ呼ぶ.

Embedding では, 各 r-radius vertex・各 r-radius edge にランダムに初期化したベクトルを割り当てる. また, r=0 の場合は, 各頂点・各辺ごとにランダムに初期化したベクトルを割り当てるのと同値である.

2.3 分類器によるリガンド結合予測

ここでは, グラフニューラルネットワークを用いた表現学習によって得られた d 次元の化合物ベクトル $\mathbf{y}_{molecule}$ とタンパク質ベクトル $\mathbf{y}_{protein}$ を用いて活性予測をする.

まず, $\mathbf{y}_{molecule}$ と $\mathbf{y}_{protein}$ を単純連結する. ここでは, $[\mathbf{y}_{\text{ molecule}}; \mathbf{y}_{\text{ protein}}]$ のように表記する. 次に, 以下の式 (3) によって, softmax 層への入力 $\mathbf{z} \in \mathbb{R}^2$ を得る.

$$\mathbf{z} = \mathbf{W}_{\text{output}} [\mathbf{y}_{\text{molecule}}; \mathbf{y}_{\text{protein}}] + \mathbf{b}_{\text{ output}} \tag{3}$$

ここで, $\mathbf{W}_{\text{output}} \in \mathbb{R}^{2 \times 2d}$, $\mathbf{b}_{\text{output}} \in \mathbb{R}^2$ である.

最後に, $\mathbf{z} = [y_0, y_1]$ を softmax 層に入力して, リガンドが結合するかどうかの 2 値分類を行う (式 (4)).

$$p_t = \frac{\exp(y_t)}{\sum_i \exp(y_i)} \tag{4}$$

ここで, $t \in \{0, 1\}$ は結合するかしないかの 2 値ラベルであり, p_t は t の確率である.

3. 評価実験

3.1 データセット

3.1.1 訓練データセット

本研究では, 訓練データセットとして DUD-E (A Database of Useful Decoys: Enhanced) データセット [11] を使用した. DUD-E は Mysinger らが作成した構造ベース手法の性能評価用データセットである. 標的タンパク質は多様性を考慮した 102 種類が選ばれており, それぞれに対して活性化合物とデコイ化合物が用意されている. 全体で, 活性化合物は 22,886 個, デコイ化合物は 100 万個以上存在する. 本研究では, Tsubaki らと同様に, 活性化合物とデコイ化合物の比率が 1:1 となるようにダウンサンプリングしたものを訓練データセットとして使用した.

3.1.2 テストデータセット

テストデータセットとして Roher ら [12] が作成した Maximum Unbiased Validation (MUV) データセットを使用した. Tsubaki らは Liu ら [13] が作成したデータセットを使用していたが, 本研究ではタンパク質の立体構造が必要なため, このデータセットを使用した. Rohrer らは PubChem[14] に収録されている生理活性データから, 17 個の標的タンパク質に対するアッセイデータを取得し, それぞれの標的タンパク質に対して, 活性化合物を 30 個, デコイ化合物を 15,000 個割り当てた. この際, バーチャルスクリーニングの結果が過大評価されないように, 独自の手順に従いデコイ化合物を選んでいる.

本研究では, MUV データセットを構成する 17 個の標的タンパク質のなかで, タンパク質—リガンド複合体構造が解かれていた 9 個の標的タンパク質からなる以下のデータセットを使用した (表 1). これは Ragoza ら [15] の研究で用いられていたデータセットと同等である.

表 1 MUV データセットの詳細

MUV ID	PDB ID	Ligand	Assay Type
600	1yow	P0E	cell
692	1yow	P0E	cell
859	5cxv	0HK	cell
852	4xe4	NAG	biochemical
548	3poo	S69	biochemical
832	1au8	0H8	biochemical
689	2y6o	1N1	biochemical
846	5exm	5ST	biochemical
466	3v2y	ML5	cell

3.2 評価指標

本研究では評価指標として AUROC (Area Under Receiver Operating Characteristic) を使用する. AUROC は, ROC 曲線の曲線下面積を用いた指標であり, 主に 2 値分類問題に用いられる評価指標である. ROC 曲線は, 正例/負例の予測の閾値を変化させながら, 縦軸に TPR, 横軸に FPR をとった曲線である. TPR は, データセット中の正例の中で正しく正例と判別できたものの割合であり, FPR は, データセット中の負例の中で誤って正例と判別されたものの割合である. TPR と FPR はそれぞれ以下の式 (5) と式 (6) で求められる.

$$TPR = \frac{\#TP}{\#TP + \#FN} \tag{5}$$

$$FPR = \frac{\#FP}{\#FP + \#TN} \tag{6}$$

3.3 ハイパーパラメータ探索

提案手法のハイパーパラメータはタンパク質特徴量に関連するパラメータを除いて, Tsubaki らの研究で最も予測精度が高いと報告されていた値と同じものを用いた. 提案手法のハイパーパラメータの詳細は表 2 に示した.

表 2 から分かるように, 提案手法ではタンパク質 r-radius subgraphs の r の値のみを探索し, タンパク質グラフニューラルネットワークの層数は 3 と固定した. これは, Tsubaki らの研究で, グラフニューラルネットワークの層数は予測精度にほぼ影響しなかったと結論づけられていたためである.

表 2 提案手法のハイパーパラメータ

ハイパーパラメータ	値
特徴量の次元数	10
化合物 r-radius subgraphs	2
化合物 GNN の層数	3
タンパク質 r-radius subgraphs	0 or 1 or 2
タンパク質 GNN の層数	3
学習率	0.001
学習率の減衰	0.5
減衰の間隔	10
最適化関数	Adam
エポック数	100
バッチサイズ	1

本研究では訓練データをさらに訓練データと検証用データにタンパク質単位で分割 (e.g. 72 個訓練 : 30 個検証用) し, 検証用データに対する AUROC が最も大きくなるハイパーパラメータの値をテストデータに対する予測の際に使用した.

3.3.1 AutoDock Vina によるドッキングシミュレーション

本研究では, タンパク質-リガンド複合体構造が得られている標的タンパク質を対象に実験を行なった. そのため, リガンドの中心座標がポケットの中心座標であると仮定し, その座標を中心とした 24 Å× 24 Å× 24 Å の立方体をドッキングシミュレーションの探索範囲と定めた. また, *num_modes* は 100, *energy_range* は 3, *exhaustiveness* は 8 とした.

4. 結果と考察

4.1 予測精度の結果

MUV データセットに含まれる個々の標的タンパク質に対して予測した際の結果を以下の表 3 と図 3 に記載する. ここでは, 提案手法と Tsubaki らの手法, AutoDock Vina によるドッキングシミュレーションと比較を行なっている. 表 3 の結果から, 9 個中 6 個の標的タンパク質で提案手法が最も良い精度を示すことがわかる. また, 提案手法と Tsubaki らの手法, AutoDock Vina について, それぞれ対応のある t 検定を行ったが, 提案手法はどちらの手法とも $P < 0.05$ で有意とはならなかった.

表 3 MUV データセットに対する AUROC

	提案手法	Tsubaki	Vina
id600	**0.574**	0.539	0.555
id692	**0.542**	0.531	0.470
id859	0.508	0.498	**0.509**
id852	**0.647**	0.643	0.482
id548	**0.721**	0.707	0.482
id832	**0.612**	0.599	0.535
id689	0.467	0.481	**0.547**
id846	**0.631**	0.630	0.461
id466	0.409	0.404	**0.613**
Average	**0.568**	0.559	0.517

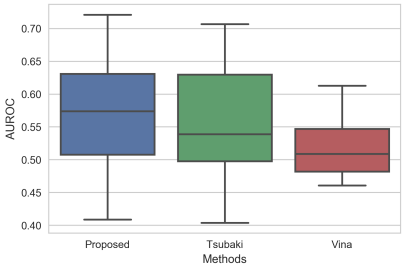

図 3 MUV データセットに対する AUROC の箱ひげ図

4.2 予測時間の結果

以下の表 4 に詳細を示す TSUBAME3.0 の f ノードを使用して, 提案手法と Tsubaki らの手法, AutoDock Vina の予測時間を計測した. 提案手法と Tsubaki らの手法は, 特徴量生成にかかる時間とテストデータセットの予測にかかる時間の合計を予測時間とした. 1 化合物あたりにかかる予測時間を以下の表 5 に示した. この表から, 提案手法は Tsubaki らの手法に比べると予測に時間がかかるが, AutoDock Vina と比べると短時間で予測可能であることが分かる.

表 4 TSUBAME3.0 の f ノードの詳細

CPU	Intel Xeon E5-2680 v4 2.4GHz × 2
コア数	28 コア
メモリ	240GB
GPU	NVIDIA TESLA P100 for NVlink-Optimized Servers × 4

表 5　1 化合物あたりの各手法の予測時間

提案手法	Tsubaki らの手法	AutoDock Vina
0.011 [sec]	0.0034 [sec]	14.37 [sec]

4.3　考察

実験の結果, アミノ酸配列情報を用いた Tsubaki らの手法と比べて, ポケット構造情報を考慮した提案手法では AUROC の値がわずかに向上した. また, AutoDock Vina によるドッキングシミュレーションと比較しても, 提案手法は高い AUROC を示した. しかし, 表 3 から分かるように, 提案手法と Tsubaki らの手法には相関がある (ピアソンの積率相関係数を使用した場合, 提案手法と Tsubaki らの手法は 0.88, 提案手法と AutoDock Vina は-0.39). すなわち, ポケット構造情報を考慮することで, アミノ酸配列情報を使用した場合よりもわずかに高い予測精度を出すことは出来るが, ドッキングシミュレーションと同様な新規化合物を発見することは難しい可能性があると考えられる.

5.　まとめ

5.1　結論

本研究では化合物・タンパク質ポケット構造ともにグラフニューラルネットワークを適用したエンドツーエンド表現学習によって, 新規標的タンパク質に対してリガンドが結合するかどうかを予測する手法を提案し, MUV データセットを用いた評価実験を行なった. その結果, AutoDock Vina によるドッキングシミュレーションと比べて同等の精度, かつ, 短時間での予測が可能であることを示した. また, アミノ酸配列情報を用いた Tsubaki らの手法と比べてわずかに予測精度が向上することも示した.

5.2　今後の課題

5.2.1　複合体構造が存在しない標的タンパク質への対応

本研究ではポケット構造をタンパク質-リガンド複合体を用いて決定した. そのため, 複合体構造が解かれていない標的タンパク質に対しては本研究の手法をそのまま適用することは出来ない. そこで, fpocket などのポケット検知ソフトウェアによってポケットを定義するなどの対応策が考えられる. しかし, ポケットが明確ではなく検知しづらい場合も数多く存在しており, その場合は依然として本手法を適用することは困難であると考えられる.

5.2.2　柔軟なポケット構造への対応

タンパク質ポケットの形状は結合するリガンドに応じて変化する. 全く異なる形状のリガンドが結合するような柔軟なポケットの場合, 適切にポケット構造を定義することは難しいと考えられる. 本研究では, アミノ酸残基の $C\alpha$ 原子間距離にある程度の幅を持たせてグループ化することで, 柔軟なポケット構造に対応しようとしているが, この手法により適切にポケットの柔軟性に対応できているのかは検証の必要がある.

謝辞　本研究の一部は JSPS 科研費 (18K11524), 産総研・東工大 実社会ビッグデータ活用オープンイノベーションラボラトリ (RWBC-OIL) の支援を受けて行われた.

参考文献

[1] Mullard, A. New drugs cost US$2.6 billion to develop. *Nature Reviews Drug Discovery*, 13(12):877, (2014).
[2] Trott, O., Olson, A. J. (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31(2), 455–461. (2010).
[3] Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., ⋯ Shenkin, P. S. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *Journal of Medicinal Chemistry*, 47(7), 1739–1749. (2004).
[4] Zsoldos, Z., Reid, D., Simon, A., Sadjad, S. B., Johnson, A. P. eHiTS: A new fast, exhaustive flexible ligand docking system. *Journal of Molecular Graphics and Modelling*, 26(1), 198–212. (2007).
[5] 敏明中澤. 機械翻訳の新しいパラダイム：ニューラル機械翻訳の原理. 情報管理, 60(5), 299–306. https://doi.org/10.1241/johokanri.60.299. (2017).
[6] Tsubaki, M., Tomii, K., Sese, J. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics*, 35(2), 309–318. (2019).
[7] G. Landrum. RDKit: Open-source cheminformatics.
[8] Sobolev, V., Sorokine, A., Prilusky, J., Abola, E. E., Edelman, M. Automated analysis of interatomic contacts in proteins. *Bioinformatics* (Oxford, England), 15(4), 327–332. (1999).
[9] Ito, J.-I., Tabei, Y., Shimizu, K., et al., PDB-scale analysis of known and putative ligand-binding sites with structural sketches. *Proteins: Structure, Function, and Bioinformatics*, 80(3), 747–763. (2012).
[10] Costa, F., De Grave, K. (n.d.). Fast Neighborhood Subgraph Pairwise Distance Kernel. (2010)
[11] Mysinger, M. M., Carchia, M., Irwin, J. J., Shoichet, B. K. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *Journal of Medicinal Chemistry*, 55(14), 6582–6594. (2012).
[12] Rohrer, S. G., Baumann, K. Maximum Unbiased Validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioactivity Data. *Journal of Chemical Information and Modeling*, 49(2), 169–184. (2009).
[13] Liu, H., Sun, J., Guan, J., Zheng, J., Zhou, S. Improving compound–protein interaction prediction by building up highly credible negative samples. *Bioinformatics*, 31(12), i221–i229. (2015).
[14] Wang, Y., Xiao, J., Suzek, T. O., Zhang, J., Wang, J., Zhou, Z., ⋯ Bryant, S. H. PubChem’s BioAssay Database. *Nucleic Acids Research*, 40(Database issue), D400-12. (2012).
[15] Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J., Koes, D. R. Protein–Ligand Scoring with Convolutional Neural Networks. *Journal of Chemical Information and Modeling*, 57(4), 942–957. (2017).