博士論文

グラフ畳み込みニューラルネットワークを用いた、疾患標的への結合特異性の高い化合物の網羅的探索手法の開発

宮﨑 優

奈良先端科学技術大学院大学

先端科学技術研究科

情報理工学プログラム

主指導教員：金谷　重彦

計算システムズ生物学（情報科学領域）

令和 3 年 7 月 25 日提出

本論文は奈良先端科学技術大学院大学先端科学技術研究科に
博士（理学）授与の要件として提出した博士論文である。

宮﨑　優

審査委員：

主査　　金谷　重彦　（情報科学領域　教授）

副査　　安本　慶一　（情報科学領域　教授）

MD.ALTAF-UL-AMIN　（情報科学領域　准教授）

小野　直亮　（情報科学領域　准教授）

黄　銘　（情報科学領域　助教）

グラフ畳み込みニューラルネットワークを用いた、疾患標的への結合特異性の高い化合物の網羅的探索手法の開発*

宮﨑　優

内容梗概

医薬品の開発においては、疾患治療の標的となるタンパク質に結合するリガンド化合物を効率的に探索するために、計算化学を用いた手法の開発が長年行われてきた。標的タンパク質への化合物の結合性を仮想的に再現し評価する Structure-Based Drug Design (SBDD) では、膨大な計算量に加え、入手困難な標的タンパク質の三次元構造が必要になる。これに代わる手法として、化合物の構造的特徴の数量表現から結合性を推定する Ligand-Based Drug Design (LBDD) が近年着目されている。特に、化合物の数量表現から結合性を予測するモデルを機械学習により構築する手段は、有用な医薬品候補化合物の迅速な発見を可能にした。しかし機械学習を用いた LBDD にも、化合物の効率的な数量表現に加え、機械学習モデル構築に必要な非リガンド化合物の情報の不足といった課題がある。また、臨床上有用な化合物を探索するためには、治療の標的タンパク質に結合するだけでなく、副作用の原因タンパク質に結合しにくい化合物を特定する必要がある。さらに、特定された化合物の構造を最適化する過程では、標的タンパク質への結合に重要な構造を特定したうえで、その構造を損なわないようにすることが効率化につながる。

本研究では、非リガンド化合物の情報を必要とせずに、副作用の原因タンパク質に比べて標的タンパク質に特異的に結合する化合物を探索する手法を提案する。本手法では、標的タンパク質のリガンドを識別するグラフ畳み込みニューラルネットワーク分類器を構築する。その際、非リガンド化合物の代わりに、副作用の原因タンパク質のリガンド化合物を対照クラスに用いる。さらに、分類器の特徴抽出層を経て得られた化合物の特徴ベクトルを評価することで、標的タンパク質への特異的結合性の構造的要因を考察する。本論文では、アルツ

ハイマー病の治療標的として有力な BACE1 タンパク質のリガンド探索に本手法を適用し、本手法の有用性について検討した。また、天然化合物ライブラリ（KNApSAcK Core Database）を対象として、本手法を用いて BACE1 リガンドの候補化合物の探索を試みた。

キーワード:
ドラッグデザイン、機械学習、リガンドの特異的結合性、グラフ畳み込みニューラルネットワーク、主成分マッピング、天然化合物

---

*奈良先端科学技術大学院大学 先端科学技術研究科 博士論文, 令和 3 年 7 月 25 日.

Comprehensive Exploration of Disease Target-specific Ligands using Graph Convolution Neural Network*

Yu Miyazaki

Abstract

In drug development, methods using computational chemistry has been studied for long time to efficiently explore ligand compounds binding proteins as targets of disease treatment. In structure-based drug design (SBDD), with which binding of compounds toward target proteins can be virtually reproduced, huge calculation costs and unavailability of 3-dimensional structure of target proteins are thought as technical problems. As an alternative, ligand-based drug design (LBDD), in which numerical vectors representing chemical structural information are used to estimate binding affinity, is considered as an efficient way. Especially, approaches to predict binding affinity from numerical vectors of compounds using machine learning methods enabled us to rapidly find promising drug candidates. However, there are also some problems in LBDD with machine learning such as difficulty in effectively representing chemical structure and lack of non-ligand information needed for machine learning model building. In addition, we need to specify compounds not only binding to disease target proteins but also avoiding undesirable binding toward proteins triggering side effects. Moreover, it is beneficial to specify key structures of binding affinity and/or specificity so that we can conserve such structures in lead optimization process.

In this study, I suggested an approach to explore ligand candidates considered to specifically bind to a target protein compared with a protein triggering side effects, without using non-ligand information. In this approach, a classifier distinguishing ligands of a target protein from ligands of a protein causing side effects is built using graph convolution neural network. Moreover, structural factors affecting binding affinity and/or specificity are explored by evaluating feature vectors obtained through feature extraction layers of the neural network.

In this paper, I evaluated the effectiveness of the approach, taking beta-site amyloid precursor protein-cleaving enzyme 1 (BACE1) and cathepsin D as a target protein and a protein triggering side effect. Using the approach, in addition, I explored candidate compounds of BACE1 ligands from KNApSAcK Core Database.

Keywords:
drug design, machine learning, ligand specificity, graph convolution neural network, principal component mapping, natural compound

---

*Doctoral Dissertation, Graduate School of Science and Technology, Nara Institute of Science and Technology, July 25, 2021.

目次

図目次

表目次

第 1 章　背景

1.1　新薬候補化合物の探索を目的とした、計算化学的アプローチの変遷

医薬品の開発においては、現在までに様々な疾患に対する有効な治療薬が創出されてきた。一方で、長期にわたって検討が繰り返されながら依然として決定的な治療薬が見つかっていない疾患も多数あり、医薬品開発の重要性は現在においてなお色褪せることがない。生体機能は体内に存在する受容体や酵素などのタンパク質にシグナル伝達物質が結合することで発現するが、このシグナル伝達物質が過剰に産生されたり枯渇したりすることにより引き起こされる疾患が数多く存在する。このような疾患に対する治療や症状の緩和のために、標的となるタンパク質にシグナル伝達物質の代わりに結合し、生体機能を正常に調節するような化合物を特定することが医薬品開発の第一段階である。しかしながら、特定された化合物が実際に医薬品として開発され製品化される確率は約 30,000 分の 1 と非常に低く、そのため上述の第一段階から数多の試行錯誤を余儀なくされ、このことが医薬品開発の長期化及び膨大なコストの一因となっている。膨大な数の化合物から標的タンパク質への十分な結合力が期待できる候補化合物を選びだす過程は一般にスクリーニングと呼ばれるが、一つの医薬品を製品化するために、スクリーニングにより候補化合物を特定する段階にかかる費用は、約 75 億円程度であるとの報告がされている[1]。以上を鑑みても、より効率的な化合物のスクリーニング手法を開発し医薬品開発の迅速化及び低コスト化に繋げることは極めて有用である。スクリーニングの方法としては、複数の化合物の標的タンパク質に対する *in vitro*（試験管内で体内の状況を人工的に再現した環境）での結合力測定を網羅的に実施する High-Throughput-Screening（HTS）の技術が古くから進展してきた一方、膨大な数の化合物を実際に用意し結合力を評価するのにも限界があった。これを受け、コンピュータを用いて、化合物の標的タンパク質への結合性を予測する手法が注目されるようになった。その有用性が決定的になった出来事として、抗がん作用の期待されるトランスフォーミング増殖因子（Transforming Growth Factor; TGF）-β1 の阻害剤開発が挙げられる。大手医薬品メーカーの Eli Lilly 社が、同阻害剤を従来の HTS により 2003 年に発見した[2]一方、同様の構造を持った化合物を同じく医薬品メーカーの Biogen Idec 社がコンピュータ計算を用いてより低コストで特定し、同年に発表した[3]。このように、標的タンパク質に結合するものとして最適な化合物を、コンピュータを用いて探索・設計する取り組みは、Computer-Aided Drug Design（CADD）と総称される。近年ではコンピュータの計算能力向上にも後押しされ、CADD の技術開発と利用が一層活発になっている。

CADD は、構造に基づいた分子設計（Structure-Based Drug Design; SBDD）とリガンドに基づいた分子設計（Ligand-Based Drug Design; LBDD）の 2 つに大別される。上述の通

り、医薬品分子は標的タンパク質に結合することで効果を発揮することが主である。結合分子と標的タンパク質は鍵と鍵穴の関係にたとえられ、タンパク質内の結合部位（結合ポケット）に構造的に良く適合する化合物が、高い結合力を示すと考えられている。そこで結合ポケットの三次元構造をコンピュータ上で再現し、その構造情報に基づき適当な化合物を設計するというのが、 SBDD の基本的な考え方である。構造的な適合度は、化合物の立体配座に由来する歪みエネルギーや、タンパク質との相互作用エネルギー（結合エネルギー、分子間力、疎水性相互作用）等を基に評価されることが主であり、エネルギーの総和が低ければ低いほど構造が適合していると考えられる。このようなアプローチは一般に Docking Simulation と呼ばれ、30 年以上にわたり検討・活用されてきた。化合物と標的タンパク質の構造的関係を直接利用し医薬品候補化合物を探索・設計する本手法は理にかなっており、神経変性疾患治療薬[4]や抗マラリア薬[5]、リンパ腫治療薬[6]等と多岐にわたる治療薬の開発に応用されてきた。

Docking Simulation に代表される SBDD のアプローチの欠点は、標的タンパク質の詳細な三次元構造を必要とする点である。タンパク質の三次元構造は核磁気共鳴（Nuclear Magnetic Resonance; NMR）法や X 線結晶構造解析により解明されることが多い。しかしこれらの手法は、血漿や細胞液等に存在する水溶性タンパク質の構造解析には適している一方、多くの医薬品の標的となる膜タンパク質の構造解析を行うにはサンプルの調製が非常に困難である。現状、NMR 法や X 線結晶構造解析に匹敵する精度でタンパク質の構造を解析する手法は存在しないため、特に膜タンパク質を標的とする医薬品を設計する場合には、標的タンパク質の構造情報を必要としない LBDD によるアプローチが SBDD よりも好まれる。LBDD は、化合物の構造的特徴を何らかの方法で数量的に表現することから始まる。この表現を用いて、例えば標的タンパク質のリガンドとして既知の化合物との類似性を定量評価し、類似性の高い化合物を抽出するという手段が活用される。抽出された化合物群は、さらなる *in vitro* 試験を経て標的への結合力や毒性の有無等の観点から既知のリガンドより優れている点を模索されることになる。このようなアプローチは Similarity Search と呼ばれ、過去にはエストロゲン関連疾患の治療薬開発への活用事例[7]等がある。もう一つのアプローチとしては、数量表現された化合物の構造的特徴から、結合性等の特性値を算出する関数を構築するというものが挙げられる。構築された関数を評価することで、化合物のどのような構造的特徴が特性値にどの程度影響を及ぼすのかを定量的に推測することが可能になる。このようなアプローチを、定量的構造特性相関（Quantitative Structure-Property Relationship; QSPR）モデルと呼ぶこともある。さらに、構築された関数を用いれば、未知の化合物について特性値を定量的に予測することが可能になるため、化合物スクリーニングの効率化に極めて有効であると考えられる。予測の対象となる特性値については、結合力のような連続値である必要はなく、結合の有無（無：0、有：1）のような離散値を分類問題として扱うことも可能である。関数は特性値に関する過去のデータを基に構築され、伝統的には単純線形回帰、主成分回帰（Principal Component Regression; PCR）、部分的最

小二乗（Partial Least Square; PLS）回帰、分類問題においてはロジスティック回帰等の線形回帰モデリングが活用されてきた。しかしながら、化合物の構造的特徴と特性値の複雑な関係性を線形回帰モデルで説明するには限界があった。そのような状況下、非線形的な関係の説明にも適した機械学習アルゴリズムの開発と応用が加速し、QSPR モデルの有用性が飛躍的に向上した。特に、生物の神経系ネットワークを模したと言われる数理モデルを応用した人工的ニューラルネットワーク（Artificial Neural Network: ANN）は、結合性を予測する事例において、旧来の主流な非線形モデリング手法の一つであった決定木モデルを有意に上回る精度を示し、大きな注目を浴びた[8]。以上の技術変遷を経て、現在でも機械学習的アプローチを活用した LBDD の取り組みが活発に行われている。

以上に述べた、CADD を構成する SBDD と LBDD の比較について、図 1.1 に要約した。

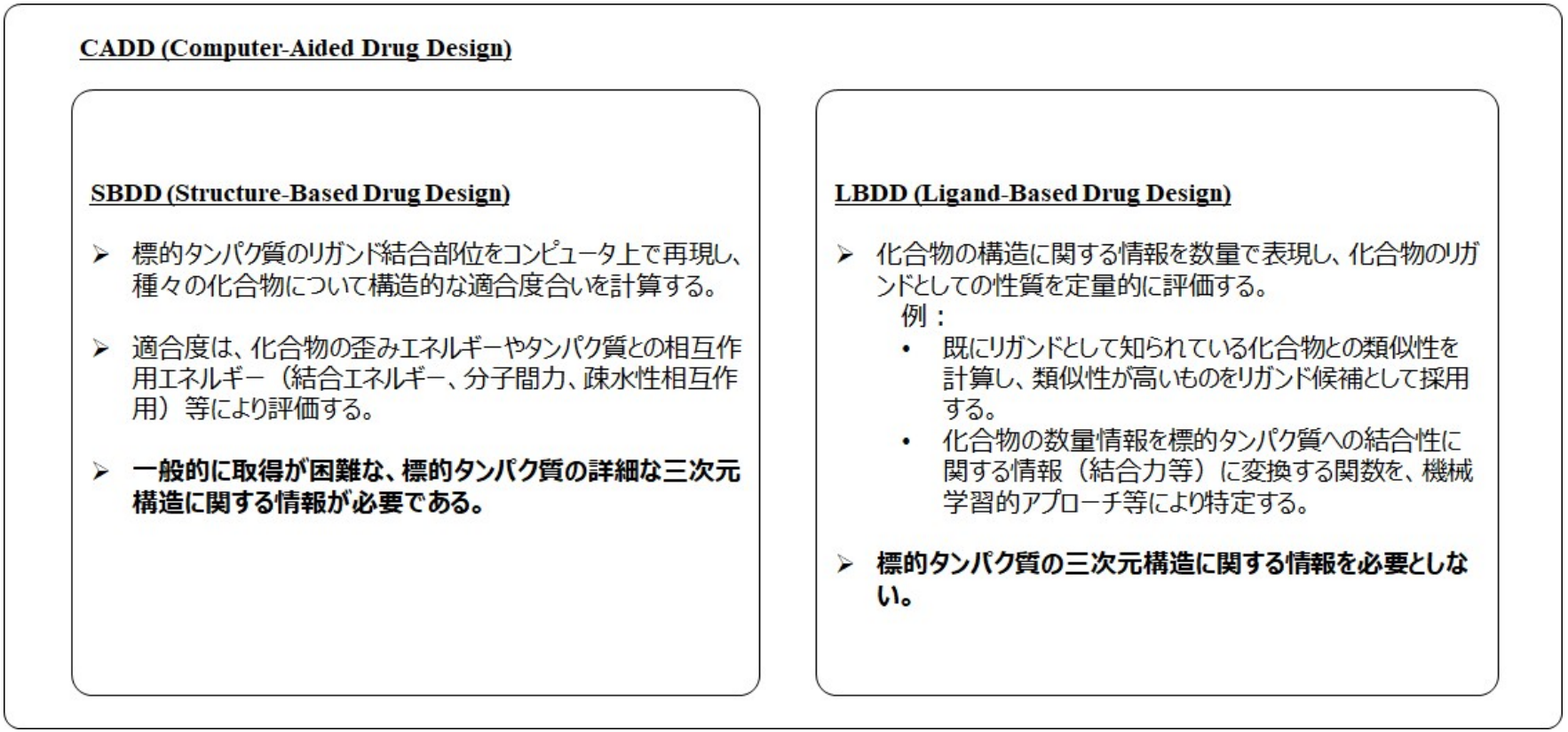

図 1.1　Computer-Aided Drug Design（CADD）を構成する、Structure-Based Drug Design（SBDD）と Ligand-Based Drug Design（LBDD）の概要

1.2　機械学習的アプローチを活用した LBDD の、技術的課題

機械学習的アプローチを用いた LBDD は効率的な新薬候補化合物の探索に有用とされている一方、より実用的な手法として確立するためには克服するべき技術的課題も存在する。以下では、これらの技術的課題について詳述する。

1.2.1　リガンドの特徴の数量表現方法

既に述べた通り、機械学習を用いた LBDD では、化合物の構造的特徴を数量的に表現した入力ベクトルから、標的タンパク質への結合力の値に変換する関数を特定することで、リガンド化合物の標的タンパク質への活性を予測することが可能になる。関数は、化合物の構造と標的タンパク質への結合力に関する過去のデータから学習することで特定される。ここで、化合物の構造的特徴をどのように数量表現するかが問題になる。結合力を正しく予測できる関数を特定するためには、数量表現ベクトルが化合物の重要な構造的特徴に関する情報を漏れなく含んでいる必要がある。一方で、表現が冗長になり過ぎると、関数の特定に必要な学習が非効率的になってしまう。化合物の構造的特徴をどのように表現するかについては、これまでにも様々な研究がなされてきた。古くは、化合物の構造の結果として得られる特性値（分子量、原子数、水／オクタノール分配係数、極性表面積等）が入力に用いられていたが、多くの場合において精度の高い予測モデルを得ることはできなかった。これに代わる手法として、化合物から構造的特徴を直接抽出するアプローチが研究されてきた。代表的なものが分子フィンガープリント法であり、各原子から一定の距離内に存在する原子で構成される部分構造の有無を数量ベクトルで表現する。このようにして得られた記述子を入力に用いたニューラルネットワークは、化合物のタンパク質への結合活性予測において劇的な精度向上を可能にした[8]。しかしながら、フィンガープリント法では、ベクトルの次元は入力に用いられる化合物に含まれる部分構造の数に対応するため、入力データが増えるにつれ非常に次元の高いベクトル表現が得られる。しかも部分構造の大半は、特定の化合物にしか見られないという意味であまり重要でなく、そのような構造の情報も表現される分冗長になってしまう。この問題を解決するために、重要度の低い構造に関する情報をハッシュ関数により圧縮することで、任意長のベクトルに変換する方法が提案されている。しかしハッシュ化されたベクトルでは、各要素がどの構造的特徴を表現しているかを追跡できなくなる。そのため、機械学習による予測の結果に寄与している構造的因子を特定できない等の問題が残る。

1.2.2 True-negative data の不足

リガンド化合物の標的タンパク質への活性を予測する関数を構築するためには、化合物のタンパク質への結合力に関する情報を学習データとして用いる必要がある。この時、関心のタンパク質に結合する化合物だけでなく、結合しない化合物に関する情報も学習に用いることが理想的である。例えば、結合性の有無を判別する分類器を構築する際には、「非結合」のラベルを付与する化合物が明らかに必要であるし、結合力を予測する回帰モデリングを行う上でも、一定の範囲の結合力を示す化合物群から学習された回帰モデルでは、結合力が範囲外の化合物について正確な予測値を得ることが困難になる。医薬品候補として探索の対象となる化合物は、大半が標的タンパク質への活性を有しないことが通常であるため、これらの化合物について正しく結合能を評価することができる予測器を構築することが化合物スクリーニングには必要である。

結合力に関する情報としては、リガンド結合アッセイ（Ligand Binding Assay）等の試験により測定されるデータが用いられるのが一般的である。しかし、このような試験データにおいて、ある化合物が特定のタンパク質に結合しないことを示すデータ（true-negative data）は圧倒的に不足しているのが現状である。分類器の構築を例にとると、リガンド群のデータ数に対し非リガンド群のデータ数が著しく少ない場合、それらのデータを用いて構築された分類器は、多くの非リガンド化合物をリガンドと判定してしまう[9]。これを解決するために、リガンド群から非リガンド群のデータ数と同じだけのデータを無作為に抽出し、分類器の構築に用いることで、データ数を均一にする手段が一般に採用されている[10]。しかしこの手段では、分類器を構築するためのデータ数が十分でなくなり、分類の精度そのものが低くなってしまう。以上のように、true-negative data の不足は、決定的な対処法が確立されておらず、機械学習的アプローチを活用した LBDD における大きな問題として認識されている。

1.2.3 副作用の原因タンパク質への結合

ここまでの議論では、LBDD に機械学習的アプローチを適用することで、化合物の標的タンパク質への結合力を予測・評価することについて述べてきた。しかし創薬においては、標的タンパク質への結合力が高い化合物を設計するだけでは不十分である。実際に、十分な結合力を有すると評価された医薬品候補化合物の多くは、その後の動物を対象とした試験や臨床試験において、効果不十分又は安全性に問題があるとして脱落を余儀なくされる。こういった問題の原因の一つとして、化合物が生体内に存在する標的以外のタンパク質にも結合してしまうことが挙げられる。このような非特異的な結合は、化合物の標的への結合力を相対的に弱めるだけでなく、予期せぬ副作用の発現を引き起こす。このような現象は、近

年活発に行われているアルツハイマー病の治療薬開発においても見られる。アルツハイマー病は認知症の最も典型的な病型であり、脳神経の変性により記憶喪失や思考能力の低下を引き起こす。このアルツハイマー病においては、β-セクレターゼ(β-site amyloid precursor protein [APP]-cleaving enzyme 1; BACE1）が治療のための標的候補と考えられている。BACE1 はアルツハイマー病の原因物質であるアミロイドβタンパク質（Aβ）の生成過程を開始する酵素であるため、BACE1 を阻害する化合物はアルツハイマー病の治療薬になり得ると考えられている[11]。これを受け BACE1 阻害薬の開発が行われてきたが、いずれも失敗に終わっている。その主な原因は安全性上の問題であると考えられており、動物を対象とした前臨床試験では BACE1 阻害化合物が眼毒性を発現したことが報告されている。さらに近年の研究では、眼毒性の原因が cathepsin D という化合物に対する結合であることが示唆された[12]。以上の例からも、医薬品の候補化合物を設計するにあたっては、標的タンパク質への結合力を高めるだけでなく、副作用の原因タンパク質への結合を回避することが極めて重要であることが理解できる。

1.2.4　結合性・特異性の化学構造的要因の特定

スクリーニングにより医薬品候補化合物として特定された化合物が、そのまま次の開発段階に移行することは少ない。例えば、医薬品として投与された化合物が効率よく治療標的部位に行きわたるよう、化学構造の微修正が必要になることがある。アルツハイマー病治療薬の例では、治療標的は脳内に存在するため、標的部位に化合物が分布するためには血液脳関門を透過するような物性を付与する必要がある。化学構造を修正する際、標的への結合性や結合特異性が損なわれるような編集は避けるべきである。無論、化合物の編集も網羅的に行い、LBDD に基づくスクリーニングによって編集後の化合物の結合性等を評価することも可能ではある。しかし、結合性や結合特異性の要因となる部分構造を特定し、その構造は保存したうえで、全体の構造を最適化する方が明らかに効率的である。以上の理由から、リガンド化合物の結合性や特異性の要因となる構造を特定することは重要であると考えられる。

1.3　課題を解決するための、手法の提案

本研究では、BACE1 と cathepsin D それぞれのリガンドを例にとり、上述の技術的課題を解決するための手法を検討した。まず、BACE1 に結合すると考えられる化合物を識別する分類器を、グラフ畳み込みニューラルネットワーク（Graph Convolution Neural Network; GCNN）によって構築した。GCNN は、化合物の構造をグラフ化し畳み込みニューラルネ

ットワークを適用する手法として近年注目を集めている。GCNN では、原子をノード、原子間の結合をエッジとして表現し、ノード内に原子の種類や価数、異性体や混成軌道の種類等といった情報のベクトルを割り当てる。またエッジを介して各原子の隣接原子を定義し、原子の情報ベクトルを隣接原子間で畳み込むことにより、分子内の局所的な構造的特徴を表現する。GCNN では特徴抽出層で畳み込みを経て化合物の領域毎の特徴を効率的に抽出し、任意のサイズのベクトルとして全結合層に渡すことができるため、フィンガープリント法に比べ冗長でない特徴抽出が可能と考えられる。近年ではリガンドとタンパク質の相互作用解析に GCNN を適用することにより、従来の機械学習的アプローチに比べ結合力の予測精度が向上したという報告もされており[13]、このことからもリガンド活性予測への GCNN の有用性が示唆されている。

次に、GCNN 分類器における BACE1 リガンド化合物群の対照クラスとして、非リガンド化合物群の代わりに、cathepsin D リガンド化合物群を用いた。これにより、入手困難な negative data を用いる必要がなくなると同時に、副作用の原因タンパク質である cathepsin D への望ましくない結合を回避しやすい化合物(BACE1 への特異的リガンドの候補化合物)をスクリーニングできるようになると考えられる。構築された GCNN 分類器の特徴抽出層を経て得られた化合物の特徴ベクトルは、全結合層を経て分類クラスの数と同じ次元（本例では二次元）を持ったベクトルに変換できる。この二次元出力ベクトルの要素の値を評価することで、BACE1 リガンドと考えられる化合物のみをスクリーニングすることが可能と考えられる。

GCNN 分類器により識別されるそれぞれのリガンド化合物群について、分類の根拠となる部分構造を特定することの重要性は既に述べた通りである。画像認識等で用いられる畳み込みニューラルネットワーク（Convolution Neural Network; CNN）においては、画像のどのピクセル領域が最終的な分類の決め手になったかを可視化する手法として、Grad-CAM（Gradient-weighted Class Activation Mapping）が知られている[14]。本研究では、構築した GCNN 分類器に Grad-CAM を適用するだけでは、リガンド間の明確な構造的差異を特定するのが困難であったことを後述する。そこで、GCNN 分類器の特徴抽出層を経て得られた化合物の特徴ベクトルを直接評価するというアプローチを検討した。具体的には、各化合物の特徴ベクトルの主成分を二次元平面上にマッピングし、各リガンドの Chemical Space を特定した。この Chemical Space を活用することで、リガンド間の化学構造の差異に関する有益な知見が得られることを、本例において確認した。

次章以降では、アルツハイマー病治療薬として意義のある BACE1 リガンド候補化合物の探索を想定し、検討手法の有用性を評価した。探索対象の化合物ライブラリには、KNApSAcK Core Database [15] を用いた。

第 2 章　手法

2.1　データセット

BACE1 及び／又は cathepsin D に対する結合活性を有する化合物を Binding DB (https://www.bindingdb.org/)から検索し、10,084 種類の BACE1 リガンド、及び 3,042 種類の cathepsin D リガンドからなるデータセットを得た。データセットには結合力に関する様々な測定値（結合定数、50%阻害濃度（IC50）等）が含まれており、これらの値は直接比較することができない。そこで本研究では、データセットに含まれる K_d (解離定数)、K_i (阻害定数)、IC50（50%阻害濃度)、EC50（50%有効濃度）の最小値が 1μM 以下である化合物を「1: 高度又は中等度に結合活性を有する化合物」、1μM より大きい化合物を「0: 低い結合活性を有する、又は結合活性を有しない」としてクラス分けを行った。以上の処理により、それぞれの化合物について BACE1 及び cathepsin D に対する結合活性に関するラベル情報を有するデータセットを得た。化合物の構造は、データセット内では化合物の構造表記法の一つである SMILES（simplified molecular-input line-entry system）で記録されているが、Binding DB には、ID が異なるが SMILES が同一である行がいくつか含まれている。これらの重複を除去するために、以下の処理を行った。1) SMILES が等しく、結合力に関する測定値も等しい場合、完全な重複と考えられるため、1 つを残して全て除去した。2) SMILES のみ重複し結合力に関する測定値に違いがみられる場合、これは同一の化合物に対し複数の結合試験結果が報告されたことに起因すると考えられるが、K_d、K_i、IC50、EC50 の全ての値の中で最も小さいものを含む行のみを残し全て除去した。以上の処理の結果、BACE1 のみに対し高度又は中等度の活性を有する 4,603 の化合物 (BACE1 リガンド群)、cathepsin D のみに対し活性を有する 471 の化合物（cathepsin D リガンド群)、両方のタンパク質に対し活性を有する 268 の化合物（非特異的リガンド群）からなるデータセットを得た。分類器の学習には、BACE1 リガンド群と cathepsin D リガンド群を用いた。さらにクラス間のデータ数の不均衡を解消するために BACE1 リガンド群からランダムサンプリングを行い、最終的に 533 の BACE1 リガンド群と 471 の cathepsin D リガンド群からなるデータセットを得た。

疾患標的化合物と副作用の原因タンパク質それぞれのリガンド化合物を識別する分類器を構築することの有用性について、一般性を検討するために、非選択的リガンドの存在が知られているその他のタンパク質の組み合わせについても、同様の基準でリガンド化合物の情報を収集した。具体的には、アドレナリン β 1 受容体（ARB1）とアドレナリン β 2 受容体（ARB2）それぞれのリガンドについて 519 の化合物、両方に活性を有する 909 の化合物を得た。加えて、ヒスタミン H 1 受容体（H1）とヒスタミン H 2 受容体（H2）それぞれのリガンドについて 264 の化合物、両方に活性を有する 75 の化合物を得た。

2.2　グラフ畳み込みニューラルネットワーク（GCNN）

GCNN は画像認識等に用いられる CNN と同様に特徴抽出層と全結合層からなり、特徴抽出層は Convolution 層と Pooling 層の繰り返しにより構成されることが一般である。今、ある化合物に含まれる i 番目の原子に割り当てられたベクトルを $\boldsymbol{v_i}$ とすると、Convolution 層及び Pooling 層におけるベクトルの操作は以下の式(1)により表される。

$$\begin{aligned} &\text{Convolution: } \boldsymbol{v_i^C} = f_{ReLU}\left(\sum_{j\in Adj(i)} \boldsymbol{W_c}(d)\boldsymbol{v_j}\right), \\ &\text{Pooling: } \boldsymbol{v_i^P} = max_{j\in Adj(i)}\left(\boldsymbol{v_j^{Convolved}}\right) \end{aligned} \tag{1}$$

ここで、$\boldsymbol{W_c}(d)$ は第 c 層の Convolution 層に固有かつ i 番目の原子からの距離 d に依存する重みベクトルであり、$\mathrm{A}dj(i)$ は i 番目の原子に隣接する原子の集合である（ただし、i 番目の原子自身も含む）。f_{ReLU} は rectified linear unit function と呼ばれる活性化関数であり、以下の式(2)により表される。

$$f_{ReLU}(x) = \begin{cases} x & (x > 0) \\ 0 & (x \leq 0) \end{cases} \tag{2}$$

また Pooling の過程では、第 c 層における i 番目の原子ベクトルについて、自身を含み隣接する原子ベクトルの要素の値から最大のものが各成分の値として選択され、置換される（max pooling）。化合物のグラフ構造に対する Convolution と Pooling の過程を模式的に表したものを、図 2.1 に示す。

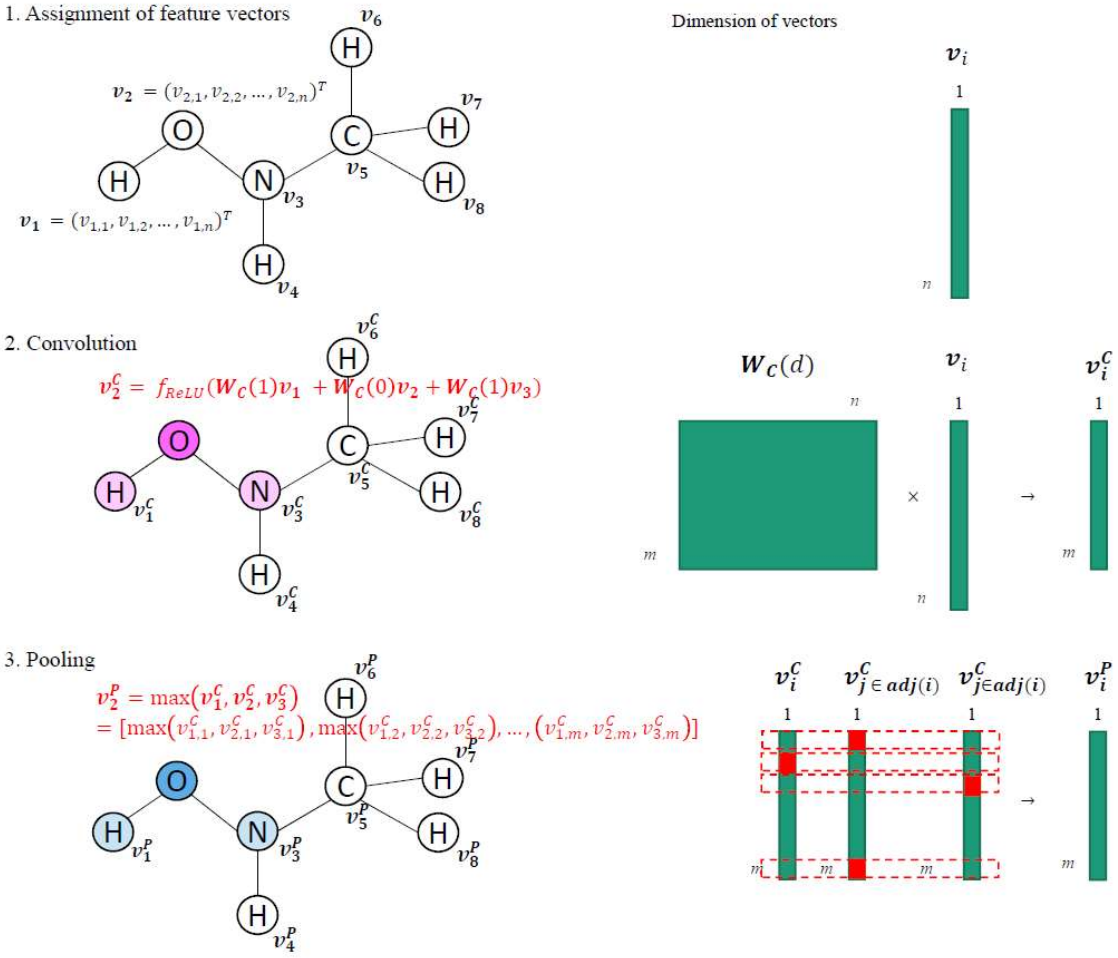

図 2.1　化合物のグラフ構造に対する Convolution と Pooling

Convolution や Pooling を経て構築されたベクトルは、依然として原子毎に割り当てられている。しかしながら、ベクトルとして特徴を表現したい対象は化合物の分子構造であるため、各原子のベクトルの情報を分子毎に集約する過程が必要である。これは、図 2.2 に示す Graph Gathering という処理により実現できる。Graph Gathering では、分子に含まれる全ての原子ベクトルの、各成分の合計をとったものを分子の特徴表現ベクトルとして得るのが一般である。これに加えて、本研究では、同じく分子に含まれる全ての原子ベクトルの各成分の最大値をとり、二つのベクトルを結合した。この操作により、原子の特徴表現ベクトルから分子の特徴表現ベクトルに変換される際の、情報損失がより少なくなると考えられる。

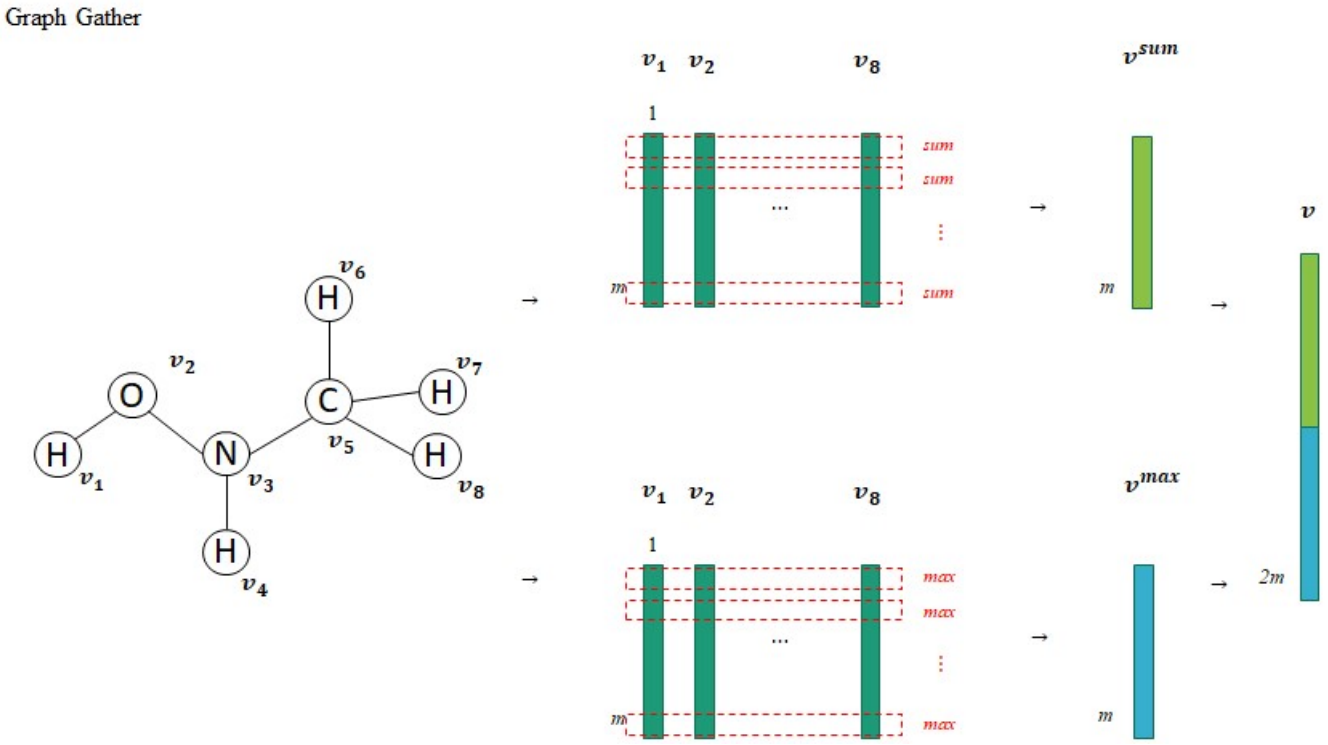

図 2.2　各原子のベクトル情報を、分子毎に集約する Graph Gathering

本研究では、各原子に以下の特徴を反映した 75 次元のベクトルを割り当てた。

- 原子の種類
- 隣接する原子の数
- 結合する水素原子の数（GCNN では、水素原子には特徴ベクトルを割り当てない）
- 電荷
- ラジカルの数
- 混成軌道の種類（sp3、sp2、sp、sp3d、sp3d2）
- 芳香族であるか否か

特徴抽出層には Convolution と Pooling を 3 回繰り返す構造を採用し、出力ベクトルの次元は 64 とした。Convolution と Pooling を経たベクトルは 128 次元の全結合層に渡したのち Graph Gathering の処理を行い、結果、化合物毎に 256 次元の特徴ベクトルを得た。さらに、これらのベクトルを 2 次元のベクトルに濃縮した後に Softmax 化を行い、最終的な出力ベクトル $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2)$を分類の予測値として得た。この予測値をラベル情報$\mathbf{y} = (y_1, y_2)$と比較し、分類の精度を評価した。ここで、$\mathbf{y} = (1, 0)$は「BACE1 のリガンドである」ことを表し、$\mathbf{y} = (0, 1)$は「cathepsin D のリガンドである」ことを表す。精度の評価には、式(3)に示す損失関数を用いた。

$$L(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{k=1}^{K}\{y_{k1}\log(\hat{y}_{k1}) + y_{k2}\log(\hat{y}_{k2})\} \qquad (3)$$

Convolution 層と全結合層に用いた重みベクトルについては、勾配降下法を用いて上記の損失関数を最適化することで得た。より具体的には、以下に示す繰り返し手順により損失関数 L を最小化することを試みた。

1. t 回目の繰り返しにおける重みベクトル群を$\boldsymbol{W}^t = [W_1, W_2, \dots, W_C]$とする。
2. t 回目の繰り返しで予測値 $\hat{\mathbf{y}}$ が得られた際に、以下の式(4)に示す計算により$\boldsymbol{W}^t$を更新する。

$$\boldsymbol{W}^{t+1} = \boldsymbol{W}^t - \alpha^t \frac{\partial L}{\partial \boldsymbol{W}^t}$$

ここで、

$$\frac{\partial L}{\partial \boldsymbol{W}^t} = [\frac{\partial L}{\partial W_1^t}, \frac{\partial L}{\partial W_2^t}, \dots, \frac{\partial L}{\partial W_C^t}]^T \qquad (4)$$

3. 1. に戻る。

上記において、α は学習率であり、損失関数の重みベクトルに対する勾配に従い重みをどの程度更新するかを決定するものである。本研究では、α の決定に Adaptive Moment Estimation 法（Adam）を採用した[16]。本手法は多くの研究においてニューラルネットワークの学習を最も効率的にするものとして知られている。重みの更新は、損失関数の推移を表す曲線が十分に飽和したとみられるところまで繰り返した。

以上に示した GCNN について、2.1 で構築したデータセットの 80%をトレーニング用として用い、残りの 20%のデータセットをテスト用として分類の性能評価に用いた。分類の性能については、古くから用いられてきたフィンガープリント法（フラグメント半径 = 2）により構築した 1024 次元のハッシュ化された特徴ベクトルにランダムフォレスト法、サポートベクターマシン、ニューラルネットワークを適用した分類結果との比較も行い評価した。ニューラルネットワークについては、GCNN とパラメータ数及び隠れ層の数が大きく異ならないように設計した。

2.3 GCNN を通して得られた特徴量ベクトルのマッピング

2.2 で示した GCNN を用いることで、各化合物について 256 次元の特徴ベクトルを Graph Gathering 層から得ることができる。これらのベクトルに主成分分析（Principal Component Analysis; PCA）を適用し、第一主成分と第二主成分を二次元平面上にマッピングした。また、第二主成分までの累積寄与率を算出した。BACE1 リガンド、及び cathepsin D リガンドが主成分マップ上で高密度に分布する領域を、カーネル密度推定法を用いて推定し、二次元平面状にカラーマップとして可視化した。カーネル密度推定におけるバンド幅の選択には、Scott 則を採用した。

以上に述べた GCNN 分類器、及び特徴ベクトルの可視化のプロセスを含んだ全体の構造図を、図 2.3 に示した。併せて、GCNN 分類器の比較対照として用いた、フィンガープリント法により得た特徴ベクトルを入力に用いたニューラルネットワークの構造を図 2.4 に示した。

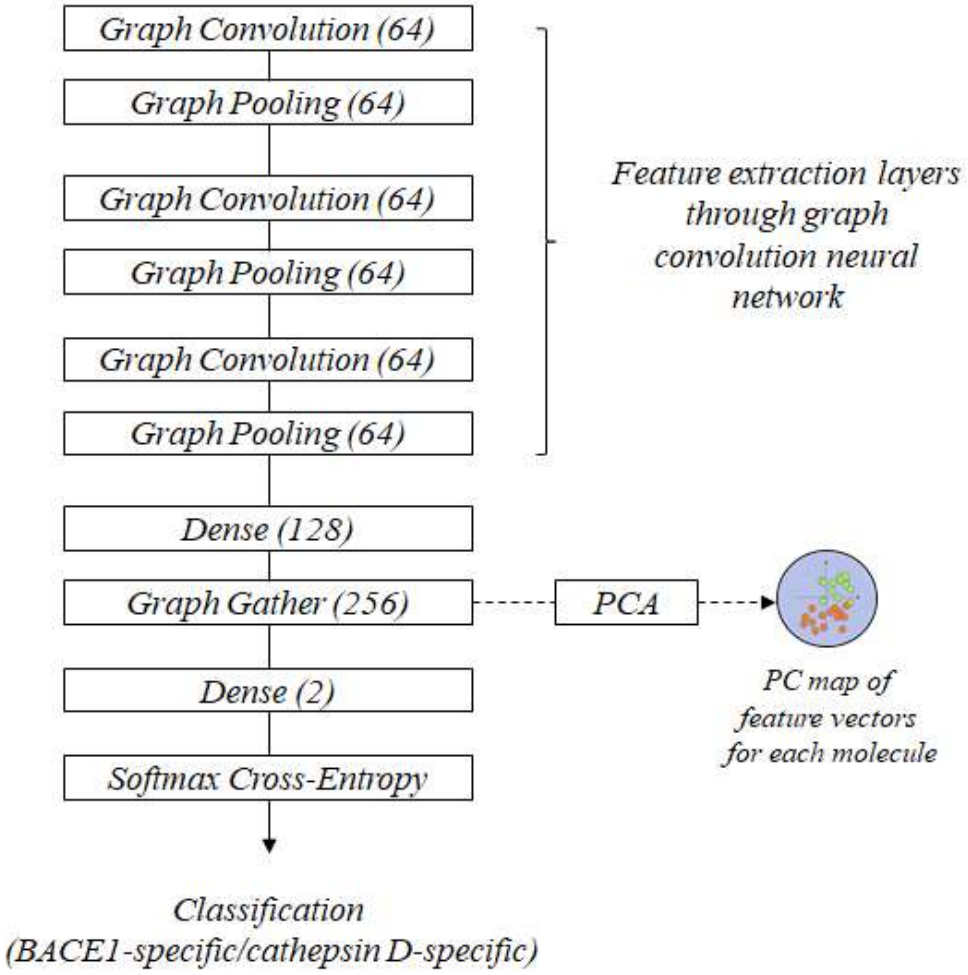

図 2.3　主成分マッピングを伴う GCNN 分類器の構造図

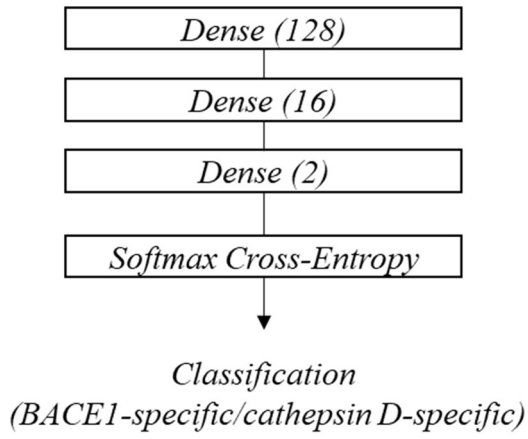

図 2.4　比較に用いたニューラルネットワークの構造図

2.4　Grad-CAM を用いた各リガンドの化学構造的差異の推定

Grad-CAM では、畳み込みニューラルネットワークの特徴抽出層における最後の畳み込み層から得られるベクトルに関する、出力ベクトルの勾配を評価することにより、各分類クラスへの分類に対するピクセル領域の重要性を定量化し、ヒートマップ等で可視化する。ここでは、同様の手法を、GCNN による化合物の分類器にも適用する。I 個の原子からなる化合物 k がクラス c に分類された場合に、その判断に対する化合物 k 内の各原子の寄与の大きさを計算することを考える。図 2.3 の GCNN 構造図における、Dense 層から得られる二次元出力ベクトルの、クラス c の要素を y_k^c とする。また、特徴抽出層の最後の畳み込み層から得られるベクトル群 $\boldsymbol{A}_k (\in \mathbb{R}^{I \times J})$ のうち、$i (\in I)$ 番目の原子に帰属するベクトルを $\boldsymbol{a}_k^i$ とする。ただし、 $J (\ni j)$ は畳み込み層から得られるベクトルの次元数である。このとき、出力ベクトルの勾配は以下の通り算出される。

$$\alpha_k^c = \frac{1}{I} \sum_{i \in I} \frac{\partial y_k^c}{\partial \boldsymbol{a}_k^i} \tag{5}$$

以下の式のように、得られた勾配を重みとして用いることで、クラス c への分類に対する各原子の寄与率を値として得ることができる。

$$L_k^c = f_{ReLU}\left(\sum_j \alpha_k^c \boldsymbol{A}_k\right) \tag{6}$$

寄与率の大きさを円の直径に反映し、原子に円を重ねて表示することで、各クラスへの分類に重要と考えられる部分構造を可視化することが可能である。

第 3 章　結果と考察

3.1　BACE1 と cathepsin D それぞれのリガンドを判別する GCNN 分類器の作成

はじめに、BACE1 と cathepsin D それぞれのリガンドを判別する GCNN 分類器を作成した。学習曲線を図 3.1 に示した。

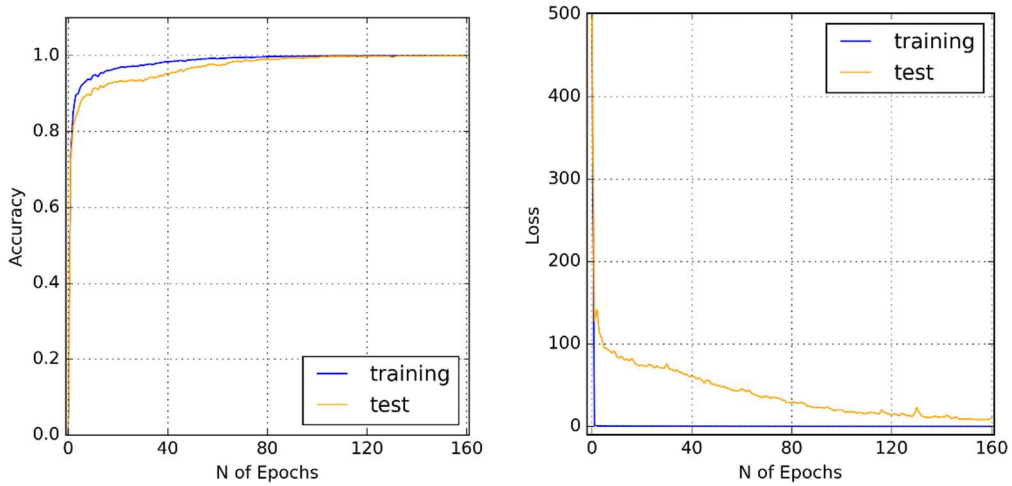

図 3.1　　BACE1/cathepsin D リガンドの GCNN 分類器の学習曲線

損失関数の値はエポック数 160 で十分に収束した。作成した GCNN 分類器の精度はトレーニング用・テスト用それぞれのデータセットに対しほぼ 100%（0.999）であり、比較のために行ったランダムフォレスト（トレーニングセット：0.877、テストセット：0.881）並びにサポートベクターマシン（トレーニングセット：1、テストセット：0.965）よりも高かった。フィンガープリント法で表現した特徴ベクトルを入力に用いたニューラルネットワーク（Fingerprint-Neural Network; FP-NN）の精度（トレーニングセット：1.0、テストセット：0.995）と比べると、ほぼ同等であった。

3.2　GCNN 分類器の、Softmax 化前の二次元出力ベクトルの評価

次に、テスト用データ（BACE1 又は cathepsin D のリガンド化合物）、天然化合物群（Natural Products; 大半は非リガンド化合物と考えられる）、非特異的リガンド群（Non-specific Ligands）を、学習済みの GCNN 分類器の入力として用いた場合の、Softmax 化前の二次元出力ベクトルの値の分布を評価した結果を図 3.2 に示した。

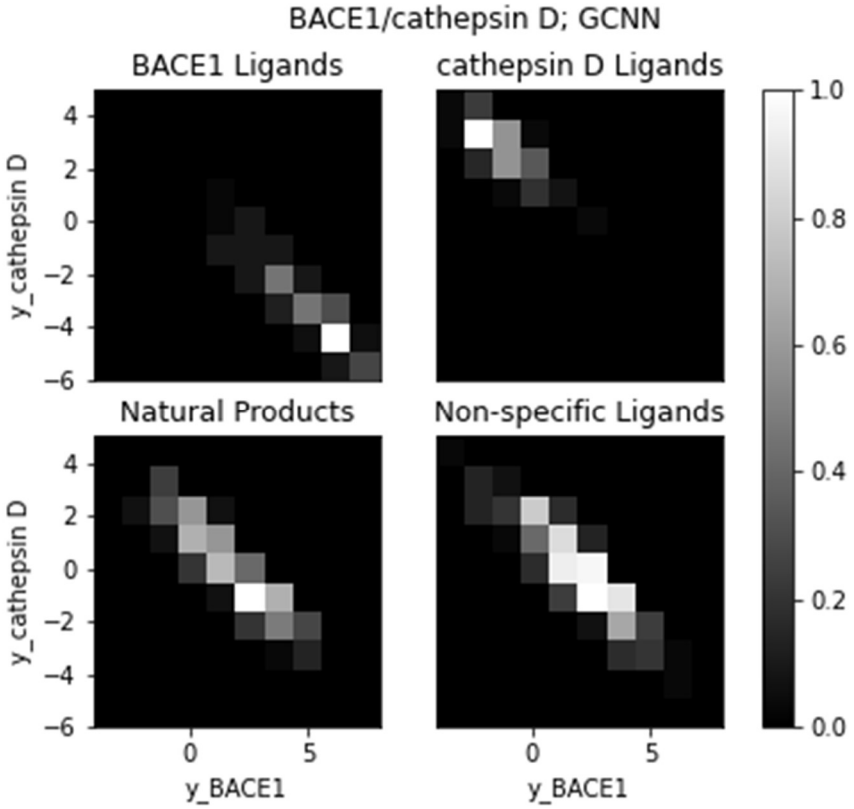

図 3.2　BACE1/cathepsin D リガンドの GCNN 分類器における、
Softmax 化前の二次元出力ベクトル値の分布

図 3.2 を見ると、それぞれのリガンド化合物の二次元出力ベクトルについて、自身への分類に寄与する要素の値が大きく、対照クラスの要素の値が小さい領域に分布していることが確認できる。一方、どちらのクラスにも属さないと考えられる天然化合物群と非特異的リガンドについては、それぞれのリガンド化合物の中間領域に多く分布ことが分かる。以上から、Softmax 化前の二次元出力ベクトルの値を評価することで、リガンド化合物をスクリーニングすることは可能であると考えられる。

比較として、FP-NN で構築した分類器の、Softmax 化前の二次元出力ベクトルの値を評価した結果を図 3.3 に示した。FP-NN では、天然化合物群と非特異的リガンドの分布が各リガンドの分布と顕著に重なっている様子が確認できる。このことから、リガンドのスクリーニングに GCNN 分類器を用いることの優位性が示唆された。

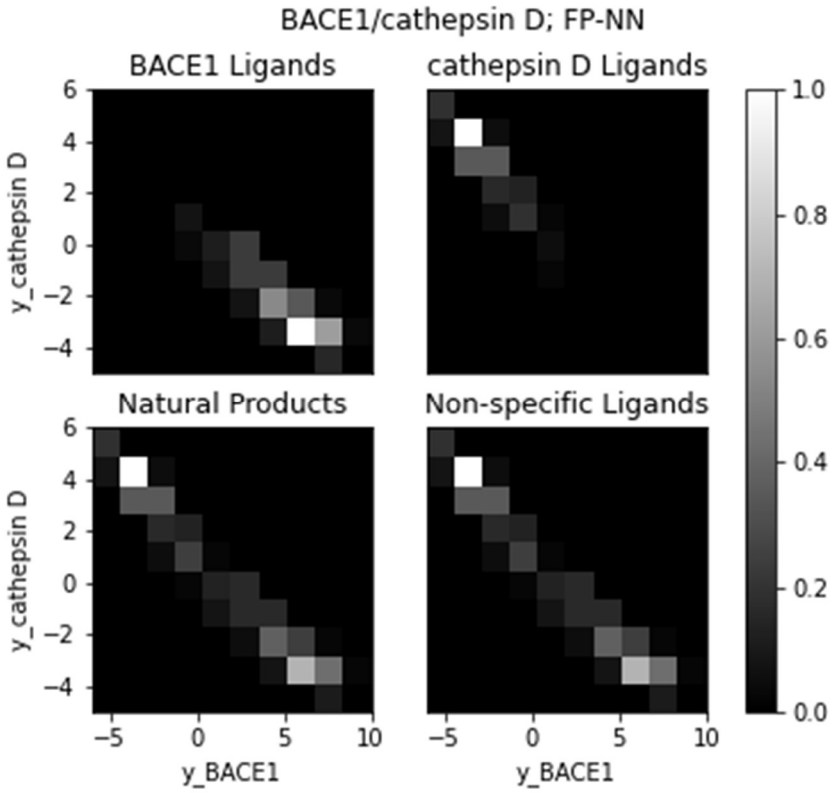

図 3.3　BACE1/cathepsin D リガンドの FP-NN 分類器における、
Softmax 化前の二次元出力ベクトル値の分布

上述の結果より、BACE1/cathepsin D それぞれのリガンドの GCNN 分類器を構築し、Softmax 化前の二次元出力ベクトルの値を評価することで、各リガンドをスクリーニングできることが示された。この結果の一般性について、すなわち非選択的リガンドが知られている他のタンパク質の組み合わせについても同様の結果が得られるのかを検討した。アドレナリンβ 1 受容体（ARB1）／アドレナリンβ 2 受容体（ARB2)、ヒスタミン H1 受容体（H1）／ヒスタミン H2 受容体（H2）の組み合わせにおける、GCNN と FP-NN によるリガンド分類器の識別精度を表 3.1 に示した。また、Softmax 化前の二次元出力ベクトルの値を比較した結果を図 3.4-図 3.7 に示した。これらの結果から、それぞれのタンパク質に特異的に結合すると考えられるリガンドの識別に限れば、FP-NN 分類器においても高い精度が得られる一方、出力チャネルの値から非リガンド化合物や非特異的リガンド化合物を区別することは困難であることが分かる。一方で GCNN 分類器においては、各リガンドの出力ベクトルの値の分布が天然化合物群、並びに非特異的リガンド群と異なるという結果が、これらのタンパク質についても再現されている。以上のことから、本研究で提案するリガンドスクリーニング手法が、一般に GCNN 分類器においてのみ有効に機能することが示唆された。

	FP-NN	GCNN
ARB1/ARB2	0.989	0.936
H1/H2	0.999	0.976

表 3.1　非選択的リガンドの存在が知られているタンパク質における、
GCNN と FP-NN によるリガンド分類器の識別精度

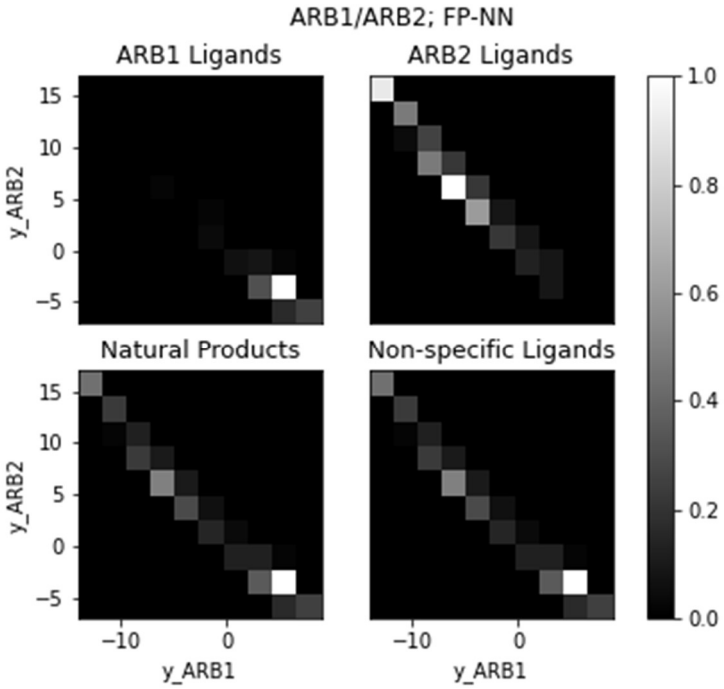

図 3.4　ARB1/ARB2 の FP-NN 分類器における、Softmax 化前の
二次元出力ベクトルの値の分布

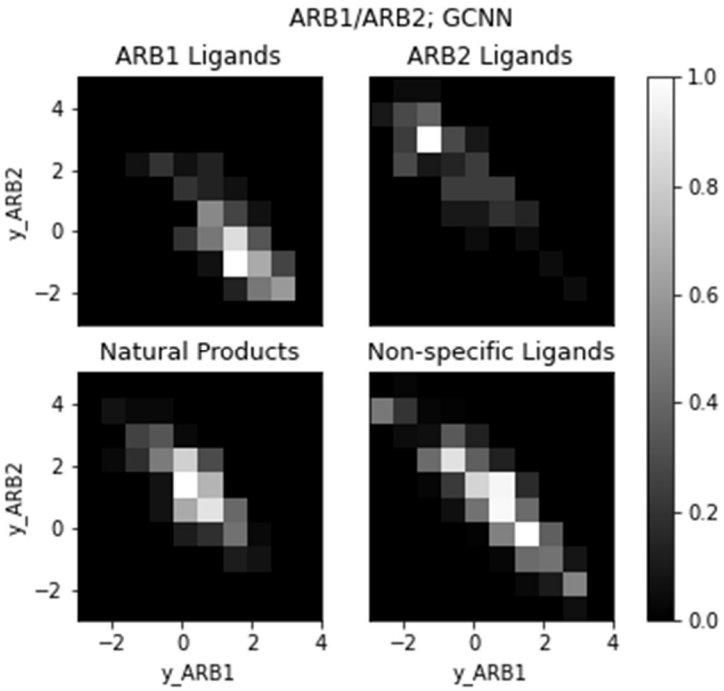

図 3.5　ARB1/ARB2 の GCNN 分類器における、Softmax 化前の
二次元出力ベクトルの値の分布

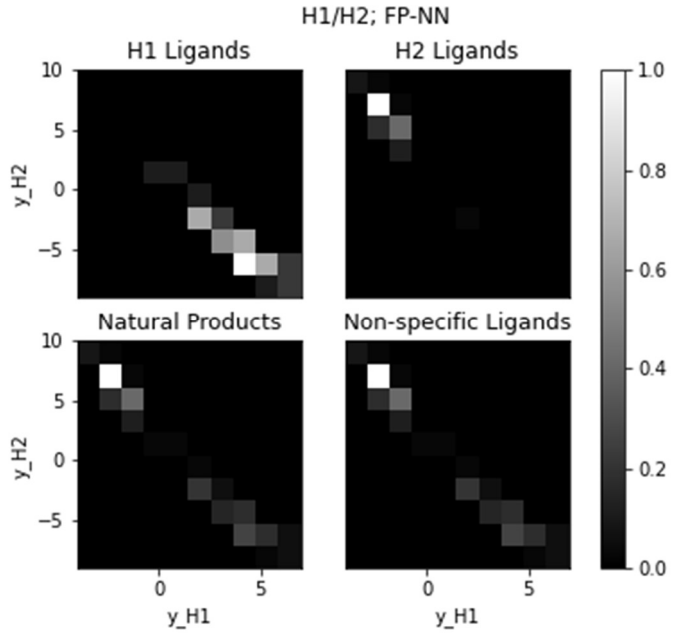

図 3.6　H1/H2 の FP-NN 分類器における、Softmax 化前の
二次元出力ベクトルの値の分布

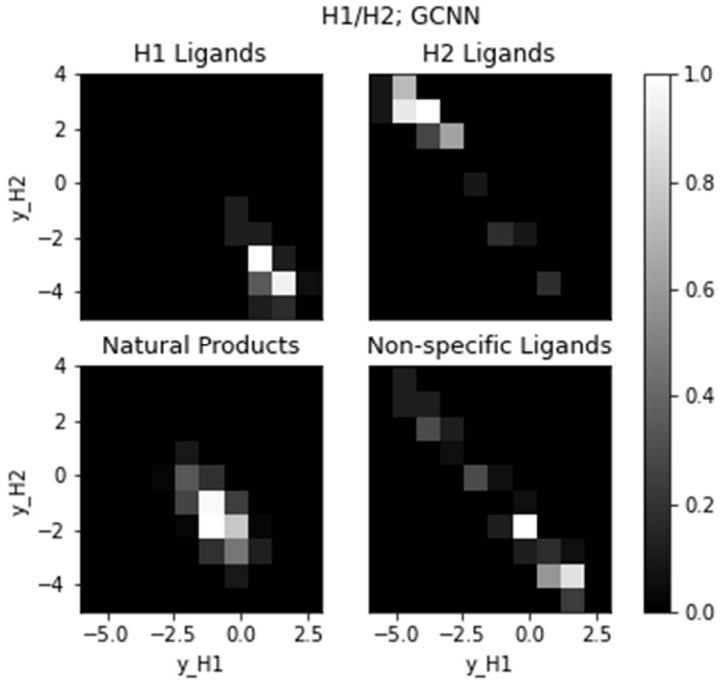

図 3.7　H1/H2 の GCNN 分類器における、Softmax 化前の
二次元出力ベクトルの値の分布

3.3 Grad-CAM を用いた、各リガンドへの分類に寄与する構造の可視化

BACE1 と cathepsin D それぞれのリガンドを識別する GCNN 分類器に Grad-CAM を適用することで、識別に寄与すると考えられる部分構造の可視化を試みた結果の例を、図 3.8 及び図 3.9 に示した。広範にわたりあらゆる原子がハイライトされていることから、この結果を用いて分類に寄与する部分構造を明確に特定することは困難であることが分かる。

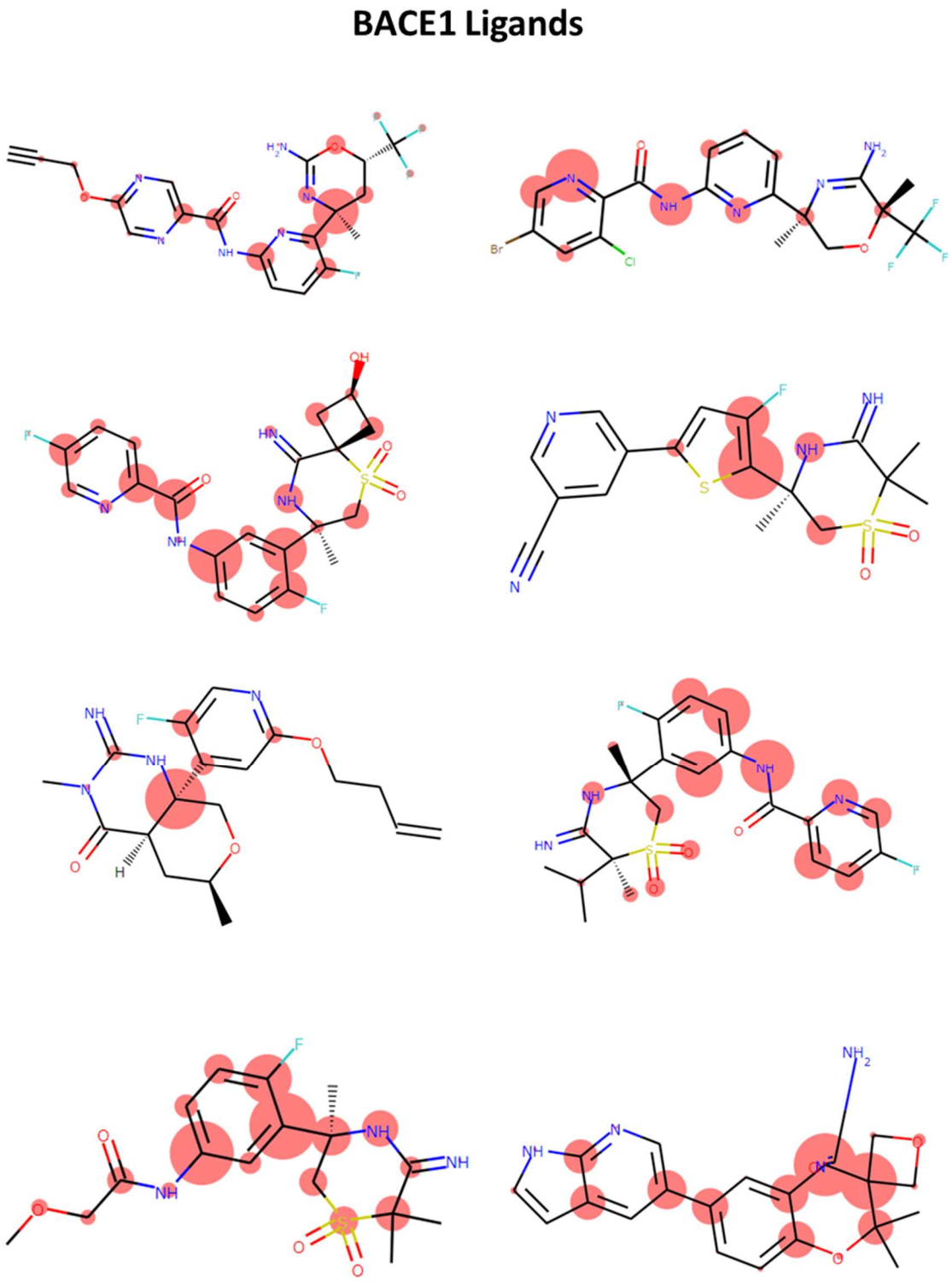

図 3.8　リガンドの分類に寄与する構造の可視化（BACE1 リガンド）

*赤色の円は BACE1 リガンドへの分類に寄与する原子を表し、円の大きさは寄与の度合いの大きさを表す。

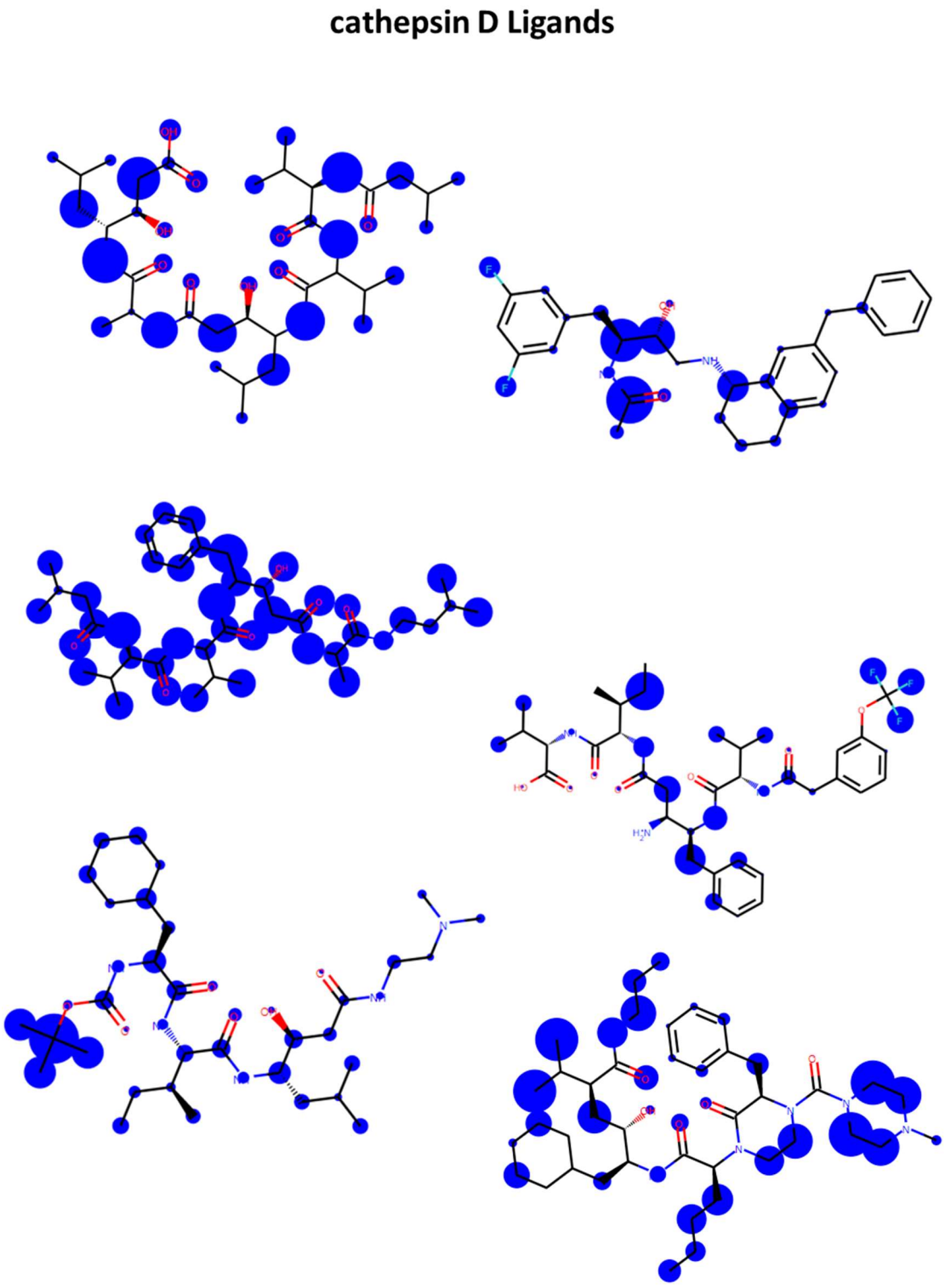

図 3.9　リガンドの分類に寄与する構造の可視化（cathepsin D リガンド）
*青色の円は cathepsin D リガンドへの分類に寄与する原子を表し、円の大きさは寄与の度合いの大きさを表す。

3.4　GCNN の特徴抽出層を経て得られた、化合物の特徴ベクトルの評価

各リガンドの分類器に Grad-CAM を適用するアプローチでは、両リガンドの構造的差異に関する明確な知見を得ることができなかった。そこで、GCNN 分類器の特徴抽出層を経て得られた化合物毎の特徴ベクトルを二次元平面上にマッピングし、各リガンドの高密度領域に含まれる化合物の構造を直接比較することで、構造的差異に関する知見が得られないか検討した。

3.4.1　特徴ベクトルの主成分マッピング

学習データの化合物の特徴ベクトルに主成分分析を適用することで得られた第一主成分と第二主成分を、二次元平面上にマッピングした結果を図 3.10 に示した。

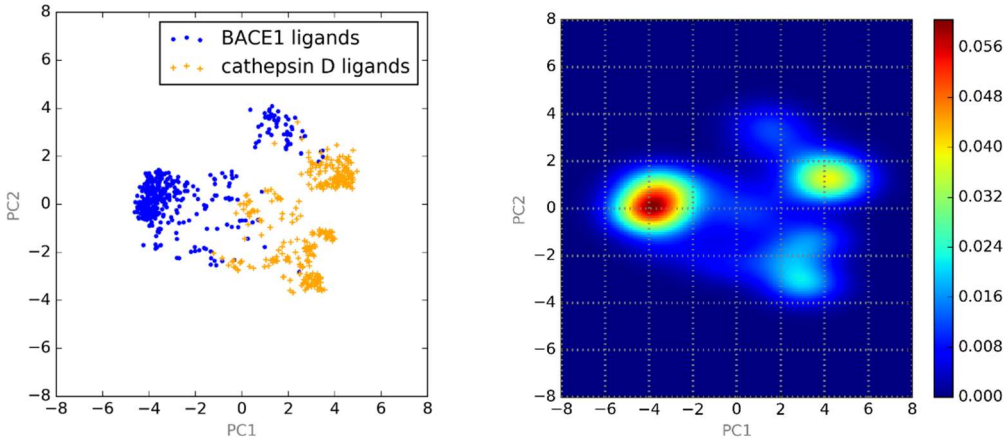

図 3.10　学習用データにおける特徴ベクトルの主成分マップ（左）と、
特徴ベクトルの確率密度マップ（右）

3.4.2　カーネル密度推定を用いた、主成分マップ上の Chemical Space の特定

次に、主成分マップ上における 2 つのリガンド群の Chemical Space を特定するために、カーネル密度推定を用いて確率密度を計算し可視化した結果（確率密度マップ）を図 3.10 に示した。主成分マップと確率密度マップを比較することで、（PC1, PC2) = (-4, 0）を中心に存在する高密度領域が BACE1 の、(PC1, PC2) = (4, 1）並びに（PC1, PC2) = (3, -3) 付近に存在する高密度領域が cathepsin D のリガンド群の Chemical Space であると推察できる。今回構築した GCNN 分類器を用いて作成した主成分マップ上においては、BACE1 と比べて cathepsin D のリガンドの Chemical Space の方がより広い領域に亘ることが示唆さ

れた。

3.4.3　未知の化合物の、主成分マップ上の分布の評価

前節で見てきたように、GCNN 分類器と主成分分析を用いて、各リガンド群の Chemical Space を特定した。次に、未知の化合物が主成分マップ上でどのように分布するのかを確かめた。初めに、テスト用のデータセットに対し学習済み GCNN 分類器を適用し、特徴抽出ベクトルを得た。これらの主成分の、図 3.10 で示した確率密度マップ上の分布を図 3.11 に示した。

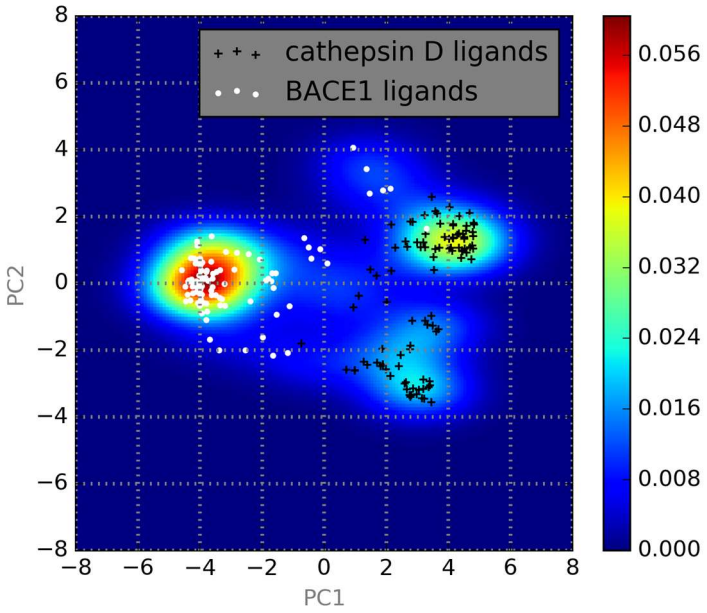

図 3.11　テスト用データの主成分の、確率密度マップ上の分布
*確率密度マップは、学習用データセットから生成したもの（図 3.10）。

図 3.11 のように、GCNN 分類器の学習に用いていない各リガンド化合物についても、学習用データセットを用いて特定したそれぞれの Chemical Space 上に局在することが確認できた。

次に、天然化合物群の化合物に同分類器／次元圧縮器を適用し得られた主成分の、図 3.10 で示した確率密度マップ上の分布を図 3.12 に示した。

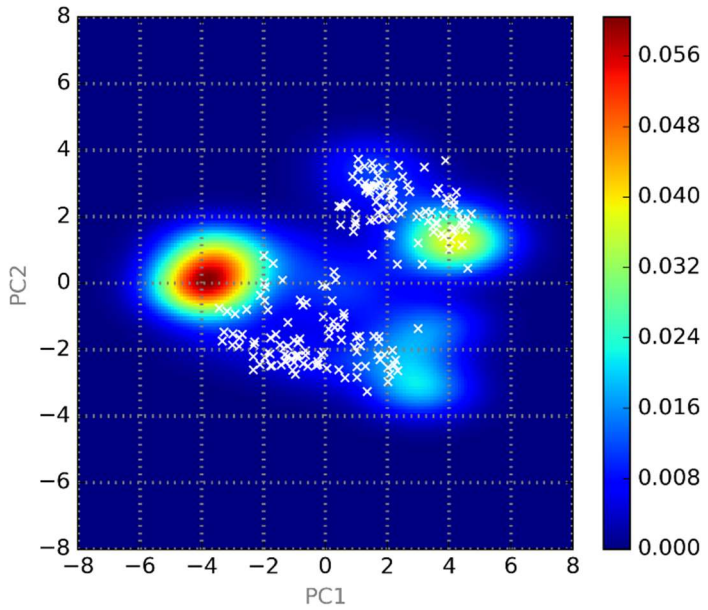

図 3.12　天然化合物群の主成分の、確率密度マップ上の分布
*確率密度マップは、学習用データセットから生成したもの（図 3.10）。

これらの化合物については、BACE1 と cathepsin D のいずれに対しても結合活性に関する情報が得られていないが、一般に存在する殆どの化合物と同様いずれのタンパク質にも結合活性を有しない可能性が高いと考えられる。図 3.12 の結果は、このような化合物についてはいずれの Chemical Space にも局在しないことを示している。

さらに図 3.13 では、非特異的リガンド化合物群に対し、同じ分類器／次元圧縮器を適用し得られた主成分の、図 3.10 で示した確率密度マップ上の分布を示した。

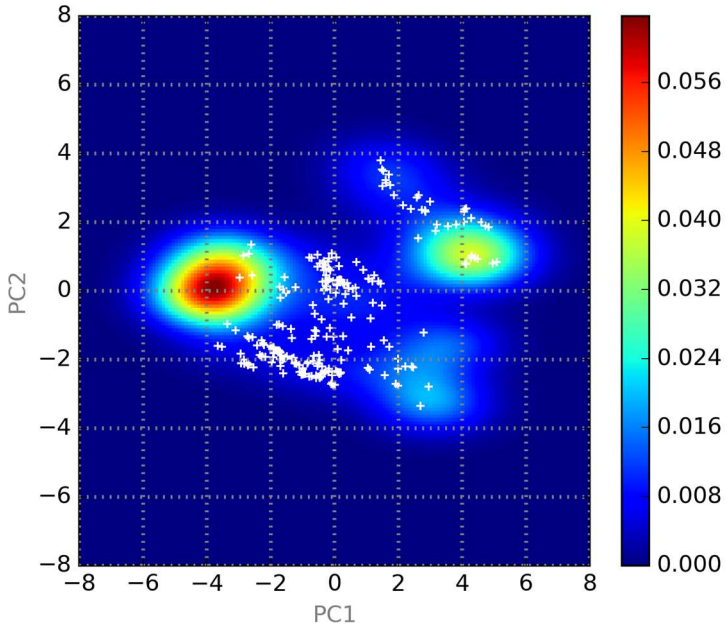

図 3.13　非特異的リガンド化合物群の主成分の、確率密度マップ上の分布
*確率密度マップは、学習用データセットから生成したもの（図 3.10）。

これらの化合物についても、それぞれのタンパク質リガンド群の Chemical Space には局在しないことが示された。

以上の結果から、各リガンドの Chemical Space が正しく特定されていることを確認した。

3.4.4　特徴ベクトルの主成分マップを用いた、BACE1 と cathepsin D それぞれのリガンドの化学構造的差異に関する考察

化合物スクリーニングの対象となる、天然化合物ライブラリ内の化合物について、同じ分類器／次元圧縮器を適用し得られた主成分の分布を見ると、化合物が第一主成分と第二主成分の両方によって表される領域によって二分されることが確認された（図 3.14）。

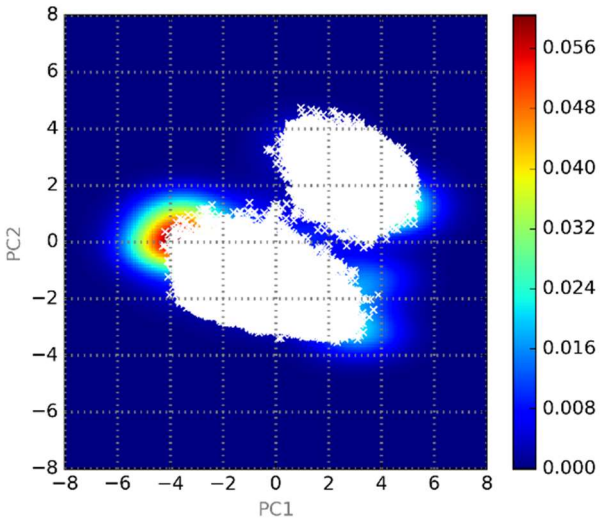

図 3.14　探索対象ライブラリ内の天然化合物における、主成分の確率密度マップ上の分布
*確率密度マップは、学習用データセットから生成したもの（図 3.10）。

このように化合物群が明確に二分されることが、どのような化学構造的特徴に由来するのかを明らかにするために、以下の手順による評価を試みた。

1. 図 3.14 の主成分マップ上において、以下の式で表される領域を用いて、化合物をクラス分けする。

$$\text{PC2a 群 : PC2} \geq -0.65 * \text{PC1} + 1.40$$
$$\text{PC2b 群 : PC2} < -0.65 * \text{PC1} + 1.40$$

2. 天然化合物群から無作為に抽出した 1000 例の化合物を対象に、PC2a 群と PC2b 群を判別する分類器を、GCNN により構築する。
3. 分類器が化合物を判別するうえで、判別の根拠としている化学構造を Grad-CAM により可視化する。

PC2a 群と PC2b 群のそれぞれに属する化合物について、Grad-CAM により算出された各原子の寄与率を色付き円によってハイライトした例を図 3.15 に示した。PC2a 群の化合物については、主にベンゼン環やカルボニル基を構成する、sp2 混成軌道を持つ炭素原子が強調されている。しかしそれ以上に顕著なのは、PC2b 群の化合物において、第四級炭素原子（4 つの炭素原子が結合する炭素原子）や、酸素原子が結合する第三級炭素原子（3 つの炭素原子が結合する炭素原子）が強調されている点である。これらの原子は、化合物の厚みの方向に嵩高い分子構造を作る。改めて図 3.15 の化合物の化学構造を見比べると、PC2a 群の化合物では全ての原子が同一平面付近に位置するような立体配座が可能であるのに対し、PC2b 群の化合物においてはそのような立体配座が取りづらい構造が確認できる。化合物内の全ての炭素原子に対する、sp2 混成軌道を持つ炭素原子の割合と、水素原子が結合していない炭素原子の数を、PC2a 群と PC2b 群で集計比較した結果を図 3.16 に示した。これらの結果から、両化合物群の構造的差異は、化合物の厚みの方向の嵩高さに起因すると考えられる。

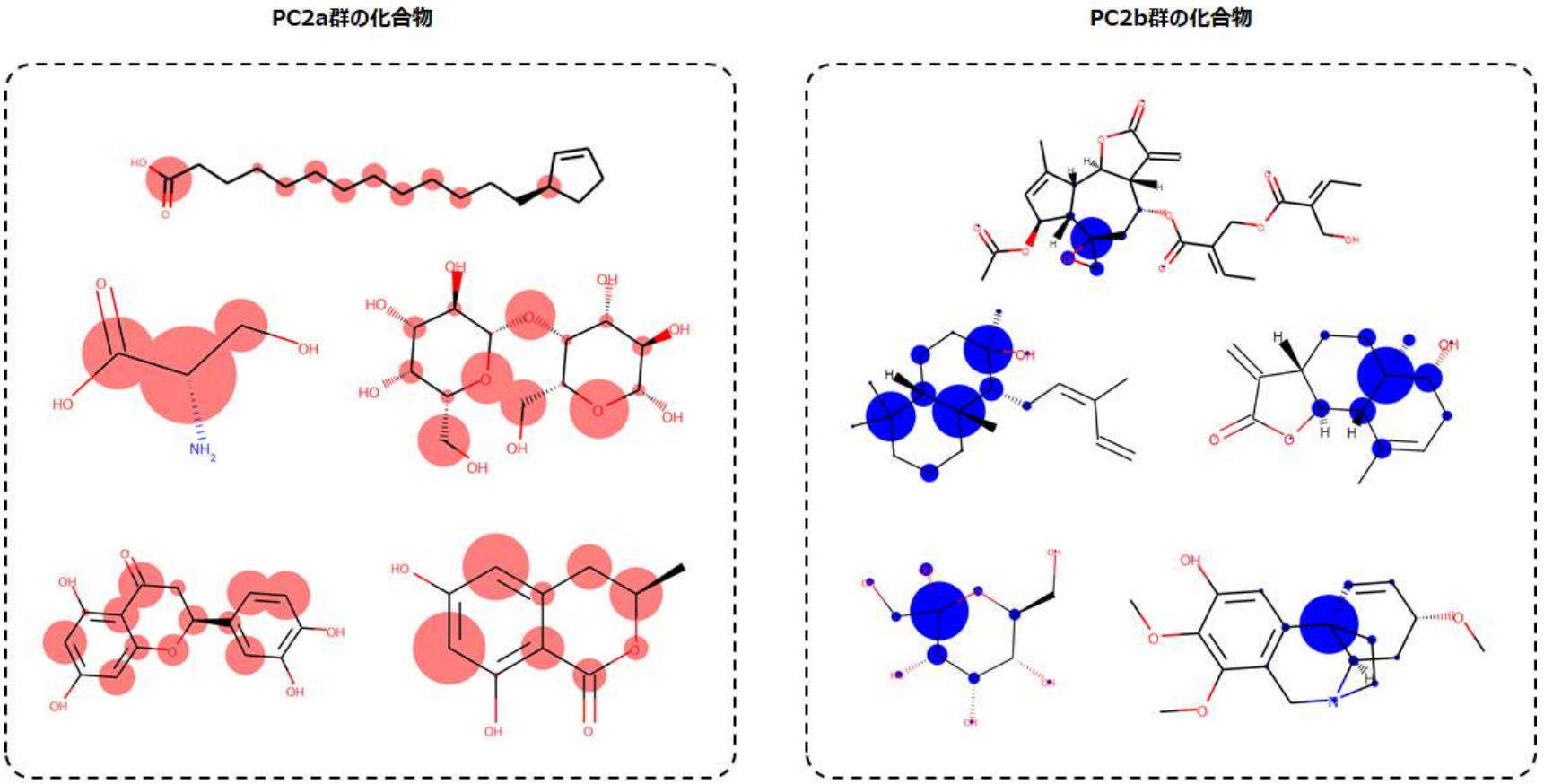

図 3.15　PC2a 群と PC2b 群のそれぞれの化合物における、分類に寄与する構造の可視化

*PC2a 群への分類に寄与する原子を赤色、PC2b 群への分類に寄与する原子を青色で示しており、円の大きさは寄与の度合いを表す。

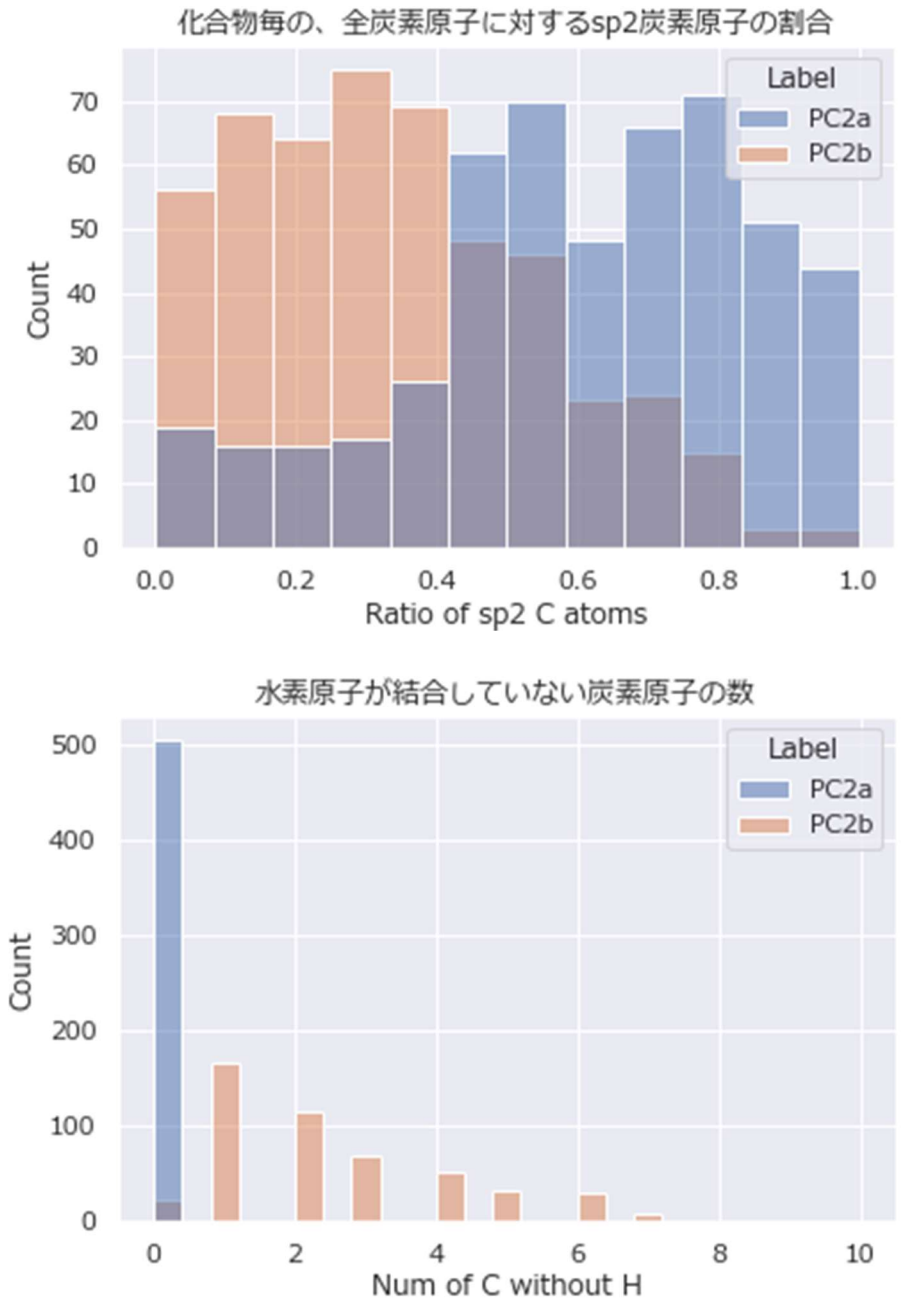

図 3.16　PC2a 群と PC2b 群のそれぞれにおける、sp2 混成軌道を持つ炭素原子の割合と、水素原子が結合していない炭素原子の数の比較集計結果

ここで改めて BACE1 と cathepsin D それぞれのリガンドの Chemical Space を確認すると、BACE1 リガンドが最も高密度に分布する領域は PC2b 群に属し、cathepsin D リガンドの場合は PC2a 群に属していることが分かる。この領域に分布する BACE1 リガンドと cathepsin D リガンドの化学構造を比較した結果を、図 3.17-18 に示した。BACE1 リガンドには、水素原子が結合していない炭素原子が 1 つ以上確認できる。また、そのような炭素原子が環構造を形成していることが、該当領域の全ての BACE1 リガンドにおいて確認できた。これに対し、cathepsin D リガンドにはそのような部分構造が確認されなかった。以上から、BACE1 に特異的に結合するリガンドの化学構造的特徴の 1 つとして、化合物の厚みを作る炭素原子の存在が重要である可能性が示唆された。

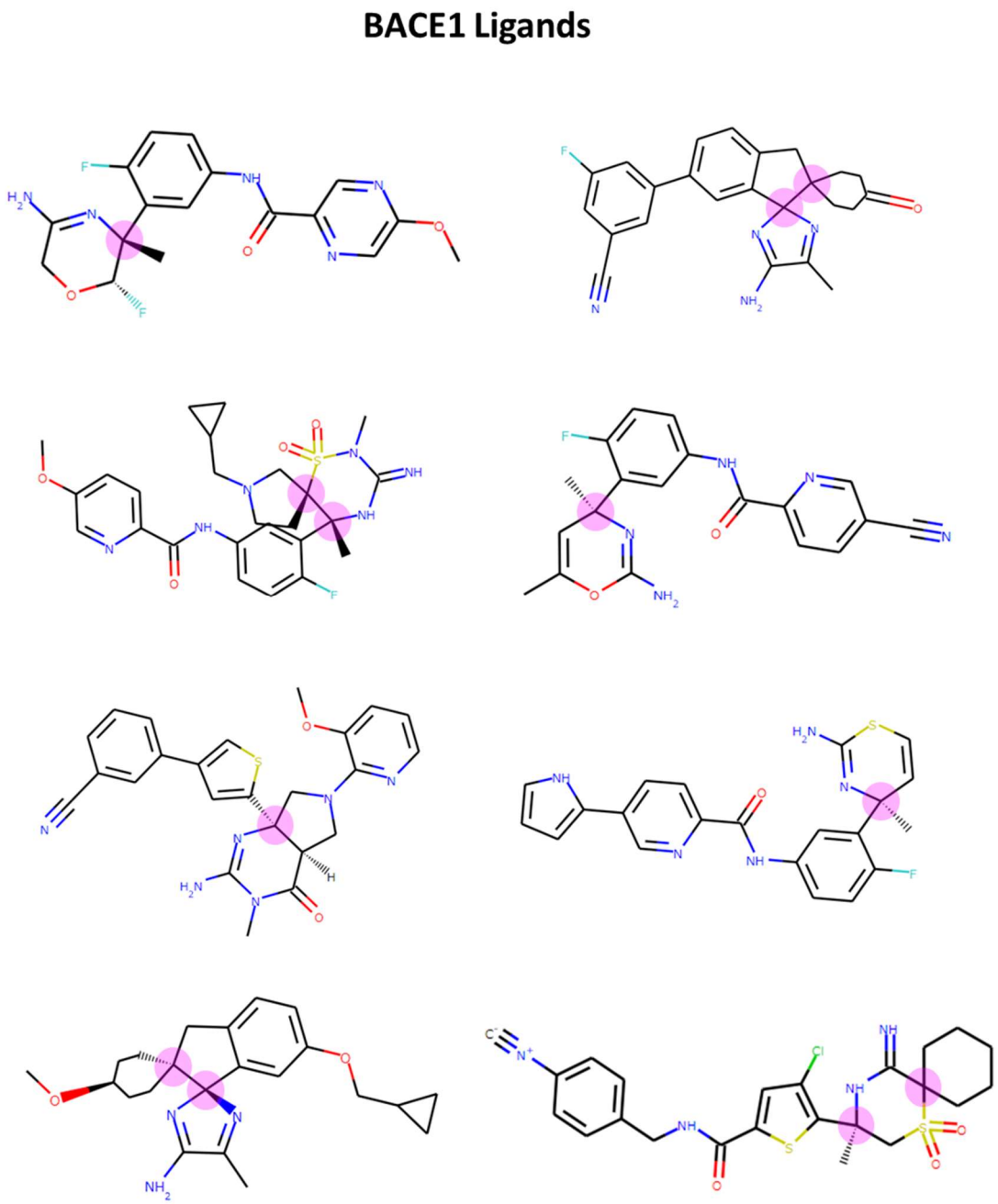

図 3.17　BACE1 リガンドの Chemical Space に分布する化合物の例

*桃色の円は、厚み方向の嵩高さを作る「水素原子が結合していない炭素原子」を表す。

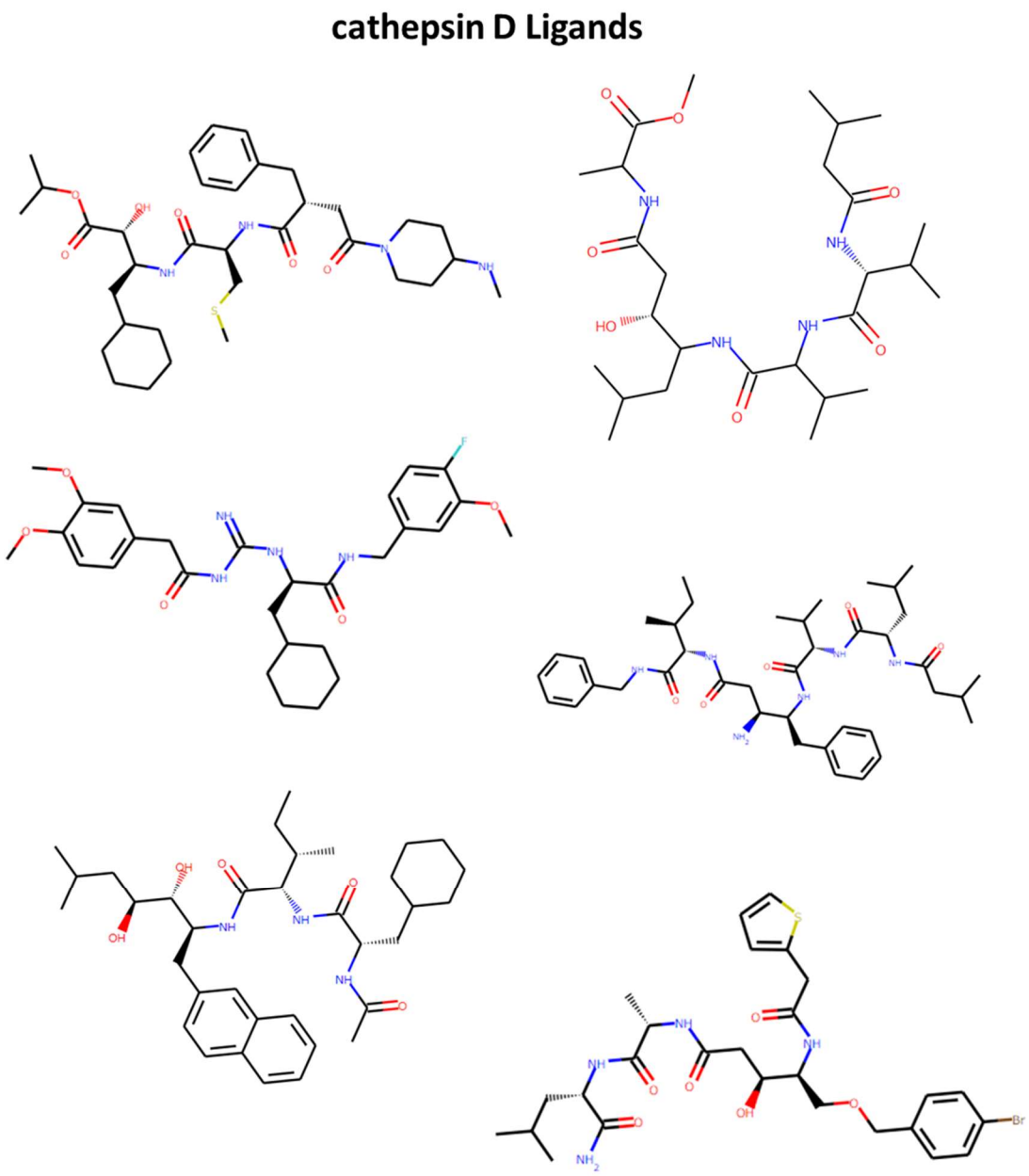

図 3.18　cathepsin D リガンドの Chemical Space に分布する化合物の例

3.5　開発手法を用いた、BACE1 リガンド候補化合物の探索

今回の研究で構築した GCNN 分類器を用いることで、BACE1 又は cathepsin D に選択的に結合するリガンドの候補化合物を探索することができると考えられる。アルツハイマー病の治療薬として有用と考えられているのは BACE1 のリガンドであるため、BACE1 の特異的リガンドの候補化合物を、以下のプロセスにより探索した。KNApSAcK Core Database に蓄積されている約 50,000 種類の天然物由来の化合物のうち、主成分マップ上で BACE1 リガンドの Chemical Space に分布する 250 の化合物について、GCNN における Softmax 化前の出力値を調べた。これらの化合物は、楕円の方程式を用いて凡そ以下の不等式によって記述される領域内に存在するものである。

$$(x_{PC1} + 3.6)^2 + \frac{(x_{PC2} - 0.4)^2}{0.6^2} \leq 1 \qquad (7)$$

これらの化合物について、GCNN の Softmax 化前の二次元出力ベクトルをプロットした結果を図 3.19 に示した。

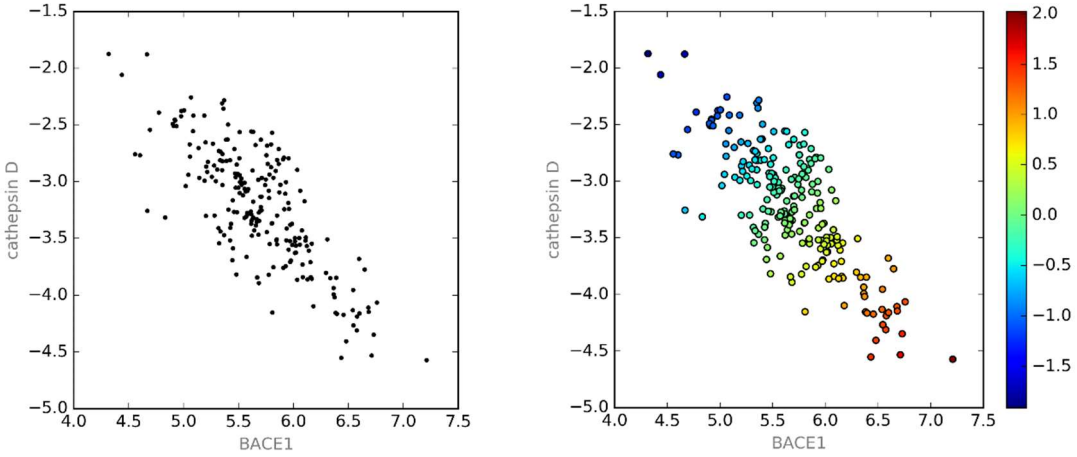

図 3.19　BACE1 リガンドの Chemical Space に位置した 250 の天然化合物の、
Softmax 化前の二次元出力ベクトル値のプロット（左）と、
二次元出力ベクトルの第一主成分の値に応じてプロットを色付けした図（右）

この出力ベクトルは、1 番目の要素の値（横軸）が大きければ BACE1 のリガンドに、2 番目の要素の値（縦軸）が大きければ cathepsin D のリガンドに、その後の Softmax 化を経て分類されることを表す。すなわち、横軸の値が大きく、同時に縦軸の値が小さいほど、BACE1 への選択的結合性の高い化合物である可能性が高いと考えられる。その度合いは、本例においては、二次元プロットから主成分分析によって抽出された第一主成分の

値により評価できる。第一主成分の値が上位 20%に属する 50 の天然化合物の一覧を、表 3.2 に示した。

CID	Compound Name	Kingdom	Family	Species	PC1 value
C00027630	(+)-Graciline	Plantae	Amaryllidaceae	Galanthus gracilis	2.04
C00018667	Angustomycin A	Bacteria	Streptomycetaceae	Streptomyces hygroscopicus var. angustmyceticus	1.682
C00012553	Aplysinol	Animalia	Aplysiidae	Aplysia kurodai	1.607
C00039568	Kopsamidine A	Plantae	Apocynaceae	Kopsia arborea	1.577
C00039956	Periglaucine C	Plantae	Menispermaceae	Pericampylus glaucus	1.555
C00028441	Labrandine	Plantae	Papaveraceae	Roemeria hybrida	1.512
C00012555	Isoaplysin	-	-	Laurencia nipponica	1.433
C00003317	Linderane	Plantae	Lauraceae	Lindera aggregata SIMS KOSTERM.	1.425
C00027621	(+)-3-epi-3,4-Dihydro-3-hydroxygraciline	Plantae	Amaryllidaceae	Galanthus gracilis	1.373
C00027620	(+)-3,4-Dihydro-3-hydroxygraciline	Plantae	Amaryllidaceae	Galanthus plicatus	1.373
C00026045	Stephanaberrine	Plantae	Menispermaceae	Stephania japonica Miers	1.373
C00039592	Kopsofinone	Plantae	Apocynaceae	Kopsia singapurensis	1.362
C00024449	Kopsinine H	Plantae	Apocynaceae	Kopsia officinalis	1.342
C00036916	Clavilactone D	Fungi	Tricholomataceae	Clitocybe clavipes	1.336
C00039955	Periglaucine B	Plantae	Menispermaceae	Pericampylus glaucus	1.329
C00039954	Periglaucine A	Plantae	Menispermaceae	Pericampylus glaucus	1.329
C00025949	N,O-Dimethyloxostephine	Plantae	Menispermaceae	Stephania japonica Miers	1.267
C00042100	3',4',4a',9a'-Tetrahydro-6,7'-dimethylspiro[benzofuran-3(2H),2'-pyrano[2,3-b]benzofuran]-2,4a',diol	Plantae	Asteraceae	Hofmeisteria schaffneri	1.243

C00024789	Gliovictin	-	-	Astromyces cruciatus	1.197
C00018666	Angustmycin C	Bacteria	Streptomycetaceae	Streptomyces hygroscopicus var. angustmyceticus	1.18
C00024557	Kopsinginol	Plantae	Apocynaceae	Kopsia teoi	1.136
C00039957	Periglaucine D	Plantae	Menispermaceae	Pericampylus glaucus	1.072
C00024540	Kitraline	Plantae	Apocynaceae	Catharanthus ovalis	1.07
C00024541	Kitramine	Plantae	Apocynaceae	Catharanthus ovalis	1.07
C00035856	Noraugustamine	Plantae	Amaryllidaceae	Crinum kirkii	1.047
C00025060	(+)-Andrangine	Plantae	Apocynaceae	Craspidospermum verticillatum	1.004
C00001682	Akuammine	Plantae	Apocynaceae	Picralima klaineana	1.004
C00024584	N-Methyl-14,15-didehydro-aspidofractinine	Plantae	Apocynaceae	Vinca sardoa	0.967
C00026625	Chetoseminudin A	Fungi	Chaetomiaceae	Chaetomium seminudum	0.955
C00027604	(-)-Digracine	Plantae	Amaryllidaceae	Galanthus gracilis	0.92
C00027347	Drupacine	Plantae	Cephalotaxaceae	Cephalotaxus harringtonia	0.86
C00046555	12-epi-Montanin D	Plantae	Labiatae	Teucrium maghrebinum	0.856
C00036147	Montanin D	Plantae	Labiatae	Teucrium botrys L.	0.856
C00026622	Perophoramidine	-	-	Perophora namei	0.814
C00048678	Cylindradine A	Animalia	Axinellidae	Axinella cylindratus	0.811
C00028179	Dibromoisophakellin	Animalia	Axinellidae	Axinella brevistyla	0.799
C00012367	Litseaculane	Plantae	Lauraceae	Neolitsea aciculata	0.795
C00042159	6-Bromo-1'-ethoxy-1',8-dihydroaplysinopsin	-	-	Smenospongia sp. specimen (UCSC coll.no.91111)	0.79
C00023530	Teucrin H319-Acetylgnaphalin (diterpene)	Plantae	Labiatae	Teucrium hyrcanicum	0.747
C00024367	Augustamine	Plantae	Amaryllidaceae	Crinum angustum Rox.	0.732

C00025066	(+)-Schizozygine	Plantae	Apocynaceae	Schizozygia caffaeoidea	0.694
C00047889	Flustramine N	-	-	Flustra foliacea	0.685
C00012365	Zeylanan	Plantae	Lauraceae	Neolitsea aciculata	0.662
C00038949	Dasyrachine	Plantae	Apocynaceae	Kopsia dasyrachis	0.645
C00023552	Teulepicin	Plantae	Labiatae	Teucrium lepicephalum	0.61
C00023543	2-Deoxychamaedroxide	Plantae	Labiatae	Teucrium divaricatum	0.58
C00011745	(+)-Juricanolide	Plantae	Asteraceae	Jurinea albicaulis	0.573
C00017979	A26771A	Fungi	Trichocomaceae	Penicillium turbatum	0.572
C00025540	Hyalodendrin	-	-	Hyalodendron sp.	0.572
C00024417	O-Demethyllycoramine	Plantae	Amaryllidaceae	Lycoris radiata Herb.	0.57

表 3.2　BACE1 への特異的結合性が高い化合物として抽出された、50 の天然化合物

表 3.2 の化合物群は、植物（*Plantae*）由来のものが全体の 72%を占めた。植物由来の化合物群の中では、キョウチクトウ科（*Apocynaceae*）由来のものが 30%と最も多く、次いでヒガンバナ科（*Amaryllidaceae*）が 19%、ツヅラフジ科（*Menispermaceae*）が 17%を占めた。それぞれの科に属する植物由来の化合物から、最も BACE1 への特異的結合性が高いと考えられる（表 3.2 中の第一主成分の値が高い）ものを 1 つずつ選び、化学構造を図 3.20 に示した。

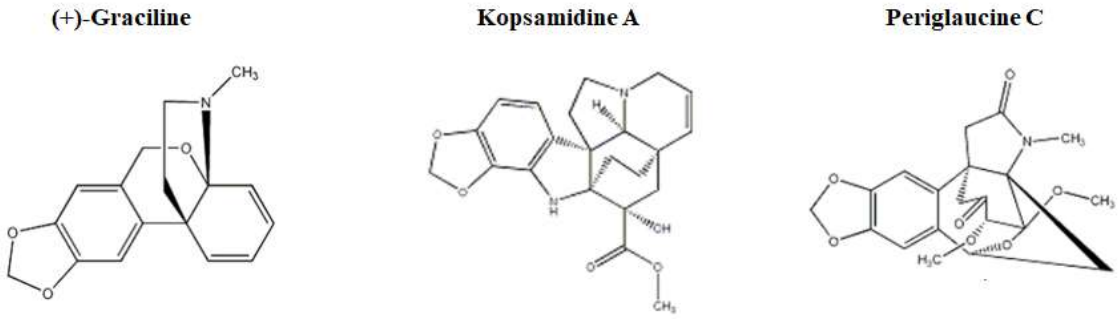

図 3.20　(+)-Graciline（キョウチクトウ科）、Kopsamidine A（ヒガンバナ科）、Periglaucine C（ツヅラフジ科）の化学構造

これらの化合物に共通するのはヘテロ 5 員環が 6 員芳香環に結合した部分構造であり、このような構造的特徴が BACE1 への結合に何等かの役割を果たしている可能性も示唆される。また Graciline については、近年の *in silico* 解析により BACE1 への結合力を有する可能性が指摘されており、その中ではヘテロ 5 員環内の酸素原子が BACE1 中のアルギニン残基と相互作用する様子が示されている[17]。一方で、Graciline が cathepsin D に対し結合力を有することを示す研究結果は見当たらなかった。また、動物（*Animalia*）由来の化合物に分類されている化合物は 3 件のみであったが、これらは全て海洋生物に由来するものであった（*Aplysiidae*; アメフラシ上科、*Axinellidae*; アキシネリダ科（海綿の一種））。また、Family が未分類の化合物の中にも、海綿の一種（*Smenospongia*）に由来するとみられる化合物が確認された。その化合物は 6-Bromo-1'-ethoxy-1',8-dihydroaplysinopsin であり、名称から Aplysinopsin の類似化合物であると推測される。図 3.21 に示すとおり、両化合物は実際に化学構造が極めて類似している。

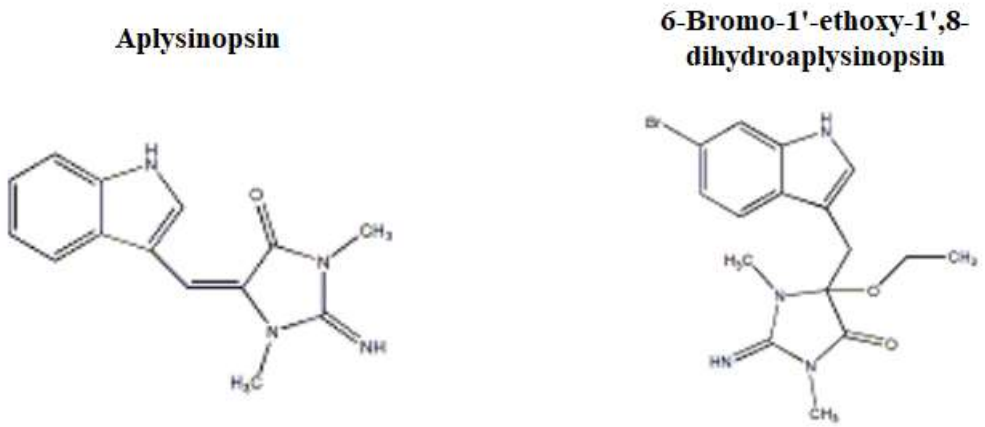

図 3.21　Aplysinopsin、及び 6-Bromo-1'-ethoxy-1',8-dihydroaplysinopsin の化学構造

Aplysinopsin については、*in vitro* 試験で BACE1 に対する弱い阻害活性を示したことが最近報告されている[18]。このことから 6-Bromo-1'-ethoxy-1',8-dihydroaplysinopsin も BACE1 に対する阻害活性を有すると結論付けることはできないが、骨格構造の類似性に加え、3.4.4 において BACE1 リガンドの構造的特徴として特定した部分構造（水素原子が結合していない炭素原子）を 6-Bromo-1'-ethoxy-1',8-dihydroaplysinopsin のみが有していることを鑑みると、その可能性は十分にあると考えられる。以上のことからも、本論で提案した特異的リガンドの探索手法の有用性が示唆された。

3.6 次元圧縮器の評価

GCNN 分類器を経て得られた特徴ベクトルの主成分について確率密度分布を可視化することで、BACE1 と cathepsin D それぞれの特異的リガンド群の Chemical Space を分離して明示することができた。主成分については第一主成分と第二主成分のみを用いており、これら 2 成分による累積寄与率は 55%であった。この値が有意に高いものであることを示すために、トレーニング用データセットに含まれる化合物をフィンガープリント法（フラグメント半径=2）で表現したベクトルに変換し、第二主成分までの累積寄与率を計算したところ、31%であった。なお、機械学習の入力データとして用いる場合にはベクトルの次元を揃えるためにハッシュ化を行うが、ハッシュ化ベクトルはそれぞれの化合物を区別するものの構造の類似性や多様性を反映できないため、ここではハッシュ化を行わなかった。ここで算出された累積寄与率と比べても、GCNN 分類器を経て得られた特徴ベクトルの第二主成分までの累積寄与率は有意に高く、これらの主成分で化合物の多様性をある程度反映できていることがわかる。

特徴ベクトルの多様性の全てを二次元平面上に反映するために利用可能な手法としては、多次元超空間上のベクトル群がより低次元の超平面上に分布していると仮定したうえで、最適な分布を推定するというものも存在する。このような手法は、多様体学習（manifold learning）と総称される。本研究において GCNN 分類器を経て得られた特徴ベクトルに対し、多様体学習による二次元マッピングを行えば、多次元ベクトルの超空間上の相対的位置関係を二次元平面上でも反映することができ、より効率的な Chemical Space の特定に役立つかもしれない。多様体学習の代表的な手法としては、t-SNE（t-distributed Stochastic Neighbor Embedding）が知られている。t-SNE は、入力となる多次元ベクトルの低次元（例えば二次元）空間上の分布に確率分布（t-分布）を仮定し、最尤のマッピング結果を推定する手法である。t-SNE は、元の多次元空間における各ベクトルの位置関係を、マッピング後の低次元空間でもよく保存できるという意味で性能が優れている[19]一方、未知の入力ベクトルが低次元マップ上のどこに位置するかを算出することができないという問題点があった。そこで近年では、リーマン幾何学に基づき、t-SNE と同等の性能を有しながら未知の入力に対しても利用可能な手法として、UMAP（Uniform Manifold Approximation and Projection）という手法が開発された[20]。以上を鑑み、多様体学習の手法として、以下では UMAP を考える。本研究において GCNN 分類器を経て得られた特徴ベクトルについて、UMAP を用いて二次元平面上にマッピングした結果を図 3.22 に示した。

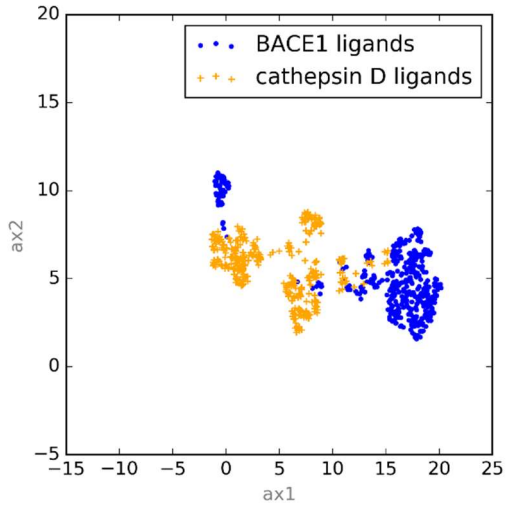

図 3.22　UMAP を用いた、学習用データにおける特徴ベクトルの二次元プロット

主成分分析では、多次元データ内の有意な分散を主成分によって表現することができる。そのため、主成分マッピングにおいては BACE1 と cathepsin D それぞれのリガンドの Chemical Space を第一主成分軸に沿って分離することができた。しかし UMAP のような多様体学習では、次元圧縮後の空間（平面）を構成する軸が何らかの意味を持っているとは限らない。本例においても、縦軸と横軸のいずれに沿っても両リガンドの Chemical Space が明確には分離されていない。そのため、UMAP を用いた二次元プロットから確率密度マップを生成する場合には、各リガンドについて別々に可視化する方が望ましいと考えられる。学習用データから生成した各リガンドの確率密度マップ上における、テスト用データ、KNApSAcK Core Database から無作為抽出した 200 例の天然化合物、非特異的リガンドの UMAP による二次元プロットの分布を図 3.23 に示した。主成分マップを用いた場合と比べると、cathepsin D リガンドの Chemical Space が、3 つの領域により明確に分離したことが分かる（図 3.23(a)）。これらの領域の間には、天然化合物や非特異的リガンドの分布に差異が見られた。例えば、領域①及び②には非特異的リガンドも分布している一方、領域③には分布が見られなかった。このような複数の Chemical Space 間での化合物の分布の違いは、GCNN 分類器の Softmax 化前の二次元出力ベクトルからは特定できなかった。以上の結果からも、GCNN 分類器を経て得られた特徴ベクトルのマッピングを用いた化合物スクリーニングの有用性が支持された。

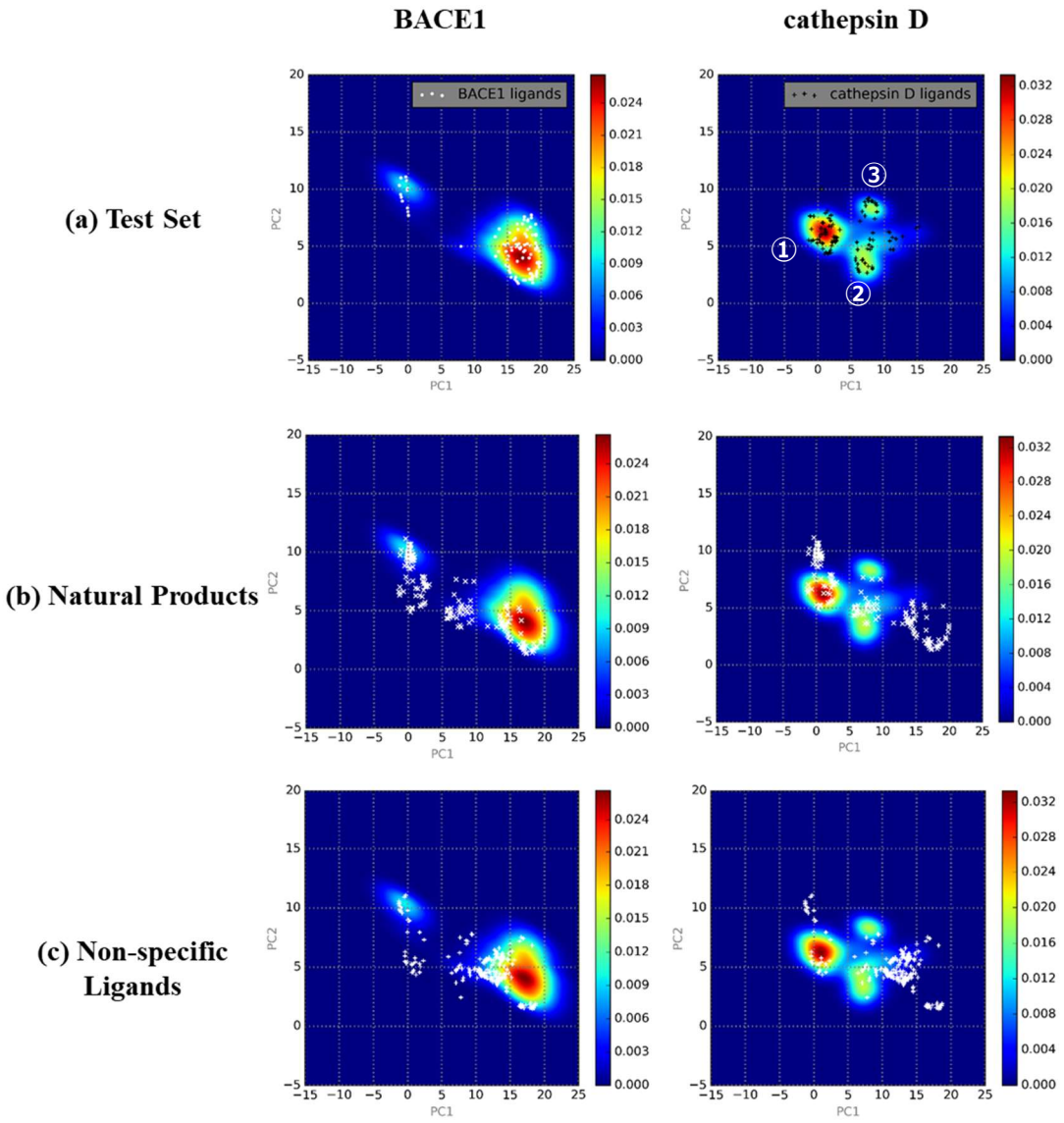

図 3.23　テスト用データ（a)、天然化合物（b)、非特異的リガンド（c)の特徴ベクトルの、UMAP による二次元プロット

*確率密度マップは、学習用データから生成されたもの。また、(a)の図中の付番は、3 つの領域に分離された cathepsin D リガンドの Chemical Space のそれぞれを区別するもの。

第 4 章　結論

本論文では、副作用を引き起こすタンパク質に比して、標的タンパク質への結合特異性が高いと考えられる化合物を網羅的に探索する手法を提案した。本手法は一般に入手困難な true-negative data を必要とせず、また GCNN を用いることで化合物の構造的特徴を効率的に抽出している点も含めて、機械学習的アプローチを用いた LBDD における技術的課題を克服していると考えられる。また必要な計算にかかる時間が分子動力学シミュレーション等と比べて少ないという LBDD の利点も保持されており、解析手法も Python や R 等の Open Source 解析ソフトウェアによって容易に実装できる。以上の点からも、本手法は簡便な新薬候補化合物の探索手段として幅広く普及することが期待できる。

本研究では、候補化合物の探索対象として天然化合物ライブラリ（KNApSAcK Core Database）を用いた。天然化合物を基に医薬品を開発することについては、化合物の合成にかかる手間が少ないことや、優れた生分解性が期待できる点から残留化合物による予期せぬ副作用の発現を回避しやすいこと等の利点が注目されている。それに伴い、今後天然化合物について一層の知見が集積されライブラリで利用可能になれば、本手法により有望な新薬候補化合物が見つかる可能性も高くなると考えられる。

本研究では、アルツハイマー病の治療薬探索を想定し、BACE1 を標的タンパク質に、cathepsin D を副作用発現タンパク質に採用した。今後、他の疾患の治療薬開発を前提として、種々の標的／副作用発現タンパク質の組み合わせについてリガンド化合物の分類器が構築されていけば、化合物のリガンド性に普遍的に関係する構造的特徴が見いだせるかもしれない。その特徴を抽出する Convolution 層を、異なるタンパク質の組み合わせに対する分類器の学習に転移させることで、標的／副作用発現タンパク質への結合性に関するデータが少ない場合でも本手法が適用できる可能性もある。このようにして、治療薬候補化合物の探索における本手法の適用範囲を広げていくことで、様々な疾患に対する有望な治療薬の開発が一段と加速することを期待する。

謝辞

本研究を進めるに当たり、奈良先端科学技術大学院大学　先端科学技術研究科　金谷重彦　教授には、多大なる御指導、御鞭撻を賜りました。特に、主体性を持って研究に取り組む姿勢と、困難な課題に根気強く向き合うことの大切さを、研究活動を通して学ばせていただきました。心より御礼申し上げます。本研究の副審査委員である奈良先端科学技術大学院大学　先端科学技術研究科　安本慶一　教授には、本研究の審査に当たり貴重な御助言を賜りました。心より御礼申し上げます。奈良先端科学技術大学院大学　先端科学技術研究科　MD.ALTAF-UL-AMIN　准教授には、日頃の研究活動に加え、本研究に係る学会発表、論文公表に当たり、多大なる御助言・御助力を賜りました。心より御礼申し上げます。奈良先端科学技術大学院大学　先端科学技術研究科　小野直亮　准教授には、本研究に係る、研究計画から結果の考察・公表に至るまで、並々ならぬ御助言・御助力を賜りました。心より御礼申し上げます。奈良先端科学技術大学院大学　先端科学技術研究科　黄銘　助教には、日頃の研究活動に加え、本研究の審査に当たり多大なる御助力を賜りました。心より御礼申し上げます。企業に勤めながら課程博士として活動することの困難を伴う中、奈良先端科学技術大学院大学　計算システムズ生物学研究室にご在籍の皆様、修了生の皆様に御理解と御助力をいただくことで、本論文を執筆するに至りましたこと、この場をお借りして深く感謝申し上げます。本研究、並びに本論文の執筆を後押ししてくださった全ての方々に、深く感謝申し上げます。

引用文献

[1] How to improve R&D productivity: the pharmaceutical industry's grand challenge, Steven M. Paul, Daniel S. Mytelka, Christopher T. Dunwiddie, Charles C. Persinger, Bernard H. Munos, Stacy R. Lindborg & Aaron L. Schacht, 19 February 2010, Nature Reviews Drug Discovery, 9, 203-214.

[2] Synthesis and Activity of New Aryl- and Heteroaryl-Substituted Pyrazole Inhibitors of the Transforming Growth Factor-β Type I Receptor Kinase Domain, J. Scott Sawyer, Bryan D. Anderson, Douglas W. Beight, Robert M. Campbell, Michael L. Jones, David K. Herron, John W.Lampe, Jefferson R. McCowan, William T. McMillen, Nicholas Mort, Stephen Parsons, Edward C. R. Smith, Michal Vieth, Leonard C. Weir, Lei Yan, Faming Zhang, & Jonathan M. Yingling, 12 August 2003, Journal of Medicinal Chemistry, 46 (19), 3953–3956.

[3] Successful shape-based virtual screening: the discovery of a potent inhibitor of the type I TGFbeta receptor kinase (TbetaRI), Juswinder Singh, Claudio E Chuaqui, P Ann Boriack-Sjodin, Wen Cherng Lee, Timothy Pontz, Michael J Corbley, H-Kam Cheung, Robert M Arduini, Jonathan N Mead, Miki N Newman, James L Papadatos, Scott Bowes, Serene Josiah & Leona E Ling, 15 Dec 2003, Bioorg Med Chem Lett., 13 (24), 4355-4359.

[4] FK506-Binding Protein Ligands: Structure-Based Design, Synthesis, and Neurotrophic/Neuroprotective Properties of Substituted 5,5-Dimethyl-2-(4-thiazolidine)carboxylates, Liqin Zhao, Wei Huang, Hongying Liu, Lili Wang, Wu Zhong, Junhai Xiao, Yuandong Hu & Song Li, 21 June 2006, Journal of Medicinal Chemistry, 49 (14), 4059-4071.

[5] Discovery of Plasmepsin Inhibitors by Fragment-Based Docking and Consensus Scoring, Friedman, R. and Caflisch, A., 30 July 2009, ChemMedChem, 4, 1317-1326.

[6] A small-molecule inhibitor of BCL6 kills DLBCL cells in vitro and in vivo, Leandro C Cerchietti, Alexandru F Ghetu, Xiao Zhu, Gustavo F Da Silva, Shijun Zhong, Marilyn Matthews, Karen L Bunting, Jose M Polo, Christophe Farès, Cheryl H Arrowsmith, Shao Ning Yang, Monica Garcia, Andrew Coop, Alexander D Mackerell Jr, Gilbert G Privé & Ari Melnick, 13 Apr 2010, Cancer Cell, 17 (4), 400-411.

[7] Virtual and biomolecular screening converge on a selective agonist for GPR30, Cristian G Bologa, Chetana M Revankar, Susan M Young, Bruce S Edwards, Jeffrey B Arterburn, Alexander S Kiselyov, Matthew A Parker, Sergey E Tkachenko, Nikolay P Savchuck, Larry A Sklar, Tudor I Oprea & Eric R Prossnitz, 05 Mar 2006, Nature Chemical Biology, 2, 207-212.

[8] Deep Neural Nets as a Method for Quantitative Structure–Activity Relationships, Junshui Ma, Robert P Sheridan, Andy Liaw, George E Dahl, Vladimir Svetnik, 23 Feb 2015, Journal

of Chemical Information and Modeling, 55 (2), 263-274.
[9] Applications of machine learning in drug discovery and development, Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, Li B, Madabhushi A, Shah P, Spitzer M, Zhao S., 11 April 2019, Nature Reviews Drug Discovery, 18, 463–477.
[10] A Systematic Prediction of Multiple Drug-Target Interactions from Chemical, Genomic, and Pharmacological Data, Hua Yu, Jianxin Chen, Xue Xu, Yan Li, Huihui Zhao, Yupeng Fang, Xiuxiu Li, Wei Zhou, Wei Wang & Yonghua Wang, 30 May 2012, PLoS ONE, 7 (5), e37608.
[11] Is donepezil effective for treating Alzheimer's disease?, L. S. Steele & R. H. Glazier, Apr 1999, Can Fam Physician, 45, 917–919.
[12] Targeting the β secretase BACE1 for Alzheimer's disease therapy, Riqiang Yan & Robert Vassar, Mar 2014, Lancet Neurol., 13 (3), 319-329.
[13] Graph Convolutional Neural Networks for Predicting Drug-Target Interactions, Wen Torng & Russ B Altman, 28 Oct 2019, Journal of Chemical Information and Modeling, 59 (10), 4131–4149.
[14] Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh & Dhruv Batra, 03 Dec 2019, arXiv, 1610.02391v4.
[15] KNApSAcK Family Databases: Integrated Metabolite–Plant Species Databases for Multifaceted Plant Research, Farit Mochamad Afendi, Taketo Okada, Mami Yamazaki, Aki Hirai-Morita, Yukiko Nakamura, Kensuke Nakamura, Shun Ikeda, Hiroki Takahashi, Md. Altaf-Ul-Amin, Latifah K. Darusman, Kazuki Saito & Shigehiko Kanaya, 28 Nov 2011, Plant and Cell Physiology, 53 (2), e1.
[16] Adam: A Method for Stochastic Optimization, Diederik P. Kingma & Jimmy Ba, 30 Jan 2017, arXiv, 1412.6980.
[17] Fucoidan serves a neuroprotective effect in an Alzheimer’s disease model, Subaraja M, Anantha Krishnan D, Edwin Hillary V, William Raja TR, Mathew P, Ravikumar S, Gabriel Paulraj M & Ignacimuthu S, Frontiers in Bioscience (Elite Edition), 01 Jan 2020, 12, 1-34.
[18] Identification of aplysinopsin as a blood-brain barrier permeable scaffold for anti-cholinesterase and anti-BACE-1 activity, Vijay K. Nuthakki, Rammohan R. Yadav Bheemanaboina & Sandip B. Bharate, 19 Dec 2020, Bioorganic Chemistry, 107, 104568.
[19] Visualizing Data using t-SNE, Laurens van der Maaten & Geoffrey Hinton, 2008, JMLR, 86, 2579-2605.
[20] UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, Leland McInnes, John Healy & James Melville, 18 Sep 2020, arXiv, 1802.03426.