

博士学位論文

スキャフォールド・ホッピングのためのトポロ
ジカル・ファーマコフォアグラフのマイニング

Mining of topological pharmacophore graph for scaffold-hopping

2021 年 12 月

奈良先端科学技術大学院大学

物質創成科学領域

データ駆動型化学研究室

中野 博史

目次

略語集	4
第 1 章 序論	6
1.1 背景	6
1.2 ファーマコフォア	7
1.3 ジオメトリカルな手法とトポロジカルな手法	9
1.4 トポロジカルなファーマコフォア	11
1.5 本研究の目的	12
1.6 参考文献	13
第 2 章 Scaffold hopping(SH)のための Pharmacophore Graph (PhG)の マイニング手法の提案	16
2.1 緒言	16
2.1.1 トポロジカルな手法を用いた SH	16
2.1.2 PhG とそのマイニング方法	17
2.1.3 SH のための PhG のマイニング手法の提案	19
2.2 研究方法	20
2.2.1 PhG の作成方法	20
2.2.2 PhG を用いた VS(Virtual Screening)	22
2.2.3 本研究で用いたターゲット生体高分子	23
2.2.4 検証のための化合物データセット	24
2.3 結果と考察	27
2.3.1 検証方法	27
2.3.2 スコアリング手法の比較結果	27
2.3.3 既存の類似度算出手法との比較	32
2.3.4 抽出された SPhG に関する考察	34
2.4 結論	36
2.5 参考文献	37
第 3 章 解釈可能性を持つ Sparse PhG (SPhG)の提案	40
3.1 緒言	40
3.2 研究方法	42
3.2.1 PhG と SPhG の違い	42
3.2.2 SPhG 作成方法の概要	43
3.2.3 PF の検出方法	44
3.2.4 SPhG 作成アルゴリズムの詳細	45

3.2.5 SPhG の評価方法	49
3.2.6 SPhG を用いた VS	49
3.2.7 SH 性能の評価指標	50
3.2.8 化合物データセット	50
3.3 結果と考察	52
3.3.1 PhG と SPhG の比較評価の詳細	52
3.3.2 スパース性とトポロジカル距離の再現性の評価	53
3.3.3 SH 性能の比較結果	56
3.3.4 抽出された SPhG に関する考察	58
3.4 結論	62
3.5 参考文献	63
3.6 補足	66
第 4 章 SPhG のクラスタリング解析	67
4.1 緒言	67
4.2 研究方法	68
4.2.1 化合物データセット	68
4.2.2 トポロジカル・ファーマコフォアの表現方法	69
4.2.3 トポロジカル・ファーマコフォアの類似度評価	71
4.2.4 トポロジカル・ファーマコフォアのクラスタリング	72
4.3 結果と考察	74
4.3.1 トポロジカルファーマコフォアマップ	74
4.3.2 PhFP マップと Mol-SPhP マップの比較	75
4.3.3 SPhG マップ	78
4.4 結論	84
4.5 参考文献	85
4.6 補足	87
第 5 章 総括	94
5.1 結論	94
5.2 今後の予定	94
謝辞	97
研究業績	98

略語集

ABL1	Tyrosine Kinase ABL1
AMDET	absorption, metabolism, distribution, excretion, and toxicity
AR	aromatic ring
BM	Bemis-Murcko
CADD	computer-aided drug design
CCR	compound core relationships
CoMFA	comparable molecular field analysis
ErG	extended reduced graph
FN	false negative
FT	feature tree
fX.	Coagulation factor X
GED	graph edit distance
GNN	graph neural network
GPCR	G protein-coupled receptor 44
HBA	hydrogen bond acceptor
HBD	hydrogen bond donor
His.	Histamine H3 receptor
Kop.	κ -opioid receptor
LBDD	ligand-based drug design
MFP	Morgan fingerprint
Mol-SPhG	molecule-sparse pharmacophore graphy
NI	negatively ionizable
PDB	protein data bank
PF	pharmacophoric feature
PhFP	pharmacophore fingerprint
PhG	pharmacophore graph
PI	positively ionizable
PI3	PI3-kinase p110-alpha subunit
POC	proof of concept
RECAP	retrosynthetic combinatorial analysis procedure
RF	random forest
RG	reduced graph
SBDD	structure-based drug design

SH	scaffold-hopping
SPhG	sparse pharmacophore graph
SVM	support vector machine
Thr.	Thrombin
TN	true negative
TP	true positive
VS	virtual screening

第1章 序論

1.1 背景

近年、コンピュータの性能の飛躍的な向上や、データサイエンスの発展により、あらゆる研究開発分野で、コンピュータを用いた手法が適用されている。医薬品においても、コンピュータ支援型医薬品設計(Computer-Aided Drug Design, CADD)が、開発効率化に必須となっている[1]。

CADD は、リガンドベース創薬(Ligand-Based Drug Design: LBDD)と構造ベース創薬(Structure-Based Drug Design: SBDD)の2つがある [2]。図 1-1a に示すとおり、SBDD は、ターゲットとなる生体高分子の構造を X 線結晶構造解析や NMR(Nuclear Magnetic Resonance)などで同定した後、その構造を用いて化合物設計を行う。例えば、分子動力学法や量子化学計算などを用いてターゲットに化合物が結合した状態をシミュレーションする方法が用いられる[3-5]。一方で、LBDD は、図 1-1b に示すとおり、ターゲットとなる生体高分子の構造を用いずに、活性が既知の低分子化合物のデータのみを用いて新しい化合物の設計を行う[6]。

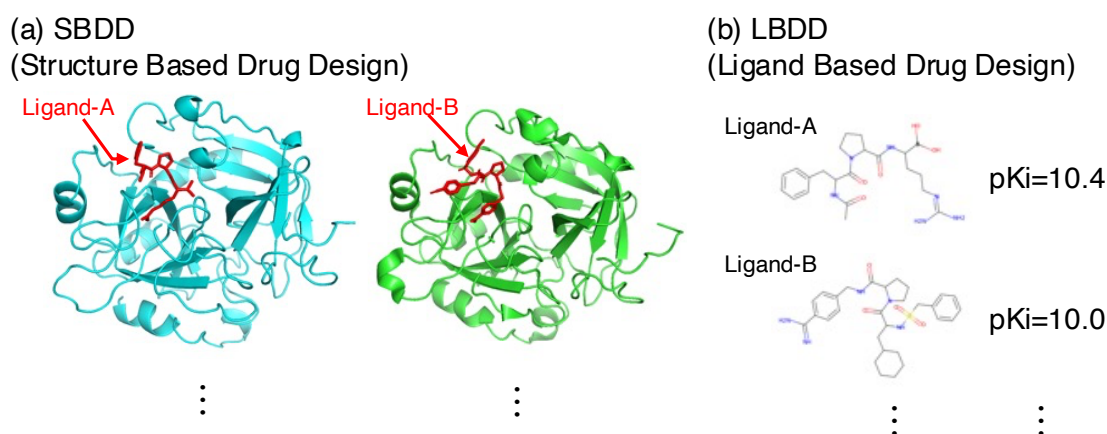


図 1-1 (a)SBDD と(b)LBDD

創薬の初期段階の Hit-to-Lead の段階やリード最適化の段階においては、候補化合物の多様性を高めることが必要となることが多い。候補となる活性化合物を多様化することで、図 1-2 に示したような、活性(activity)以外にも求められる特性、例えば合成(synthesis)が容易で、吸収(absorption)・分布(distribution)・代謝(metabolism)・排泄(excretion)・毒性(toxicity)といった

ADMET などの他の要件を満たす化合物を発見できる可能性が高くなる [7,8].
 その際に、既知の活性化合物と同等の活性を持ち、別の骨格(scaffold)に属する化合物を探索する Scaffold-Hopping (SH)を行う手法も必要となる [7,8].

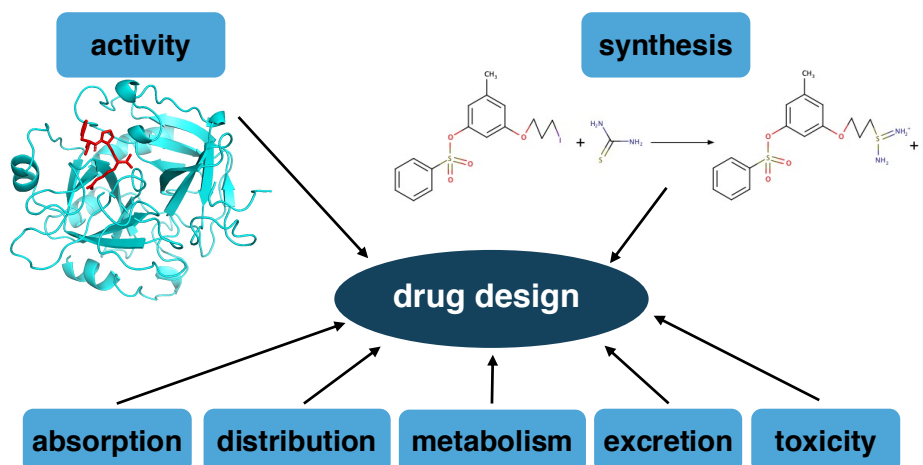


図 1-2 創薬において候補化合物に要求される特性

1.2 ファーマコフォア

SH を成功させるためには、scaffold に依存せず、活性を持つために必要な「本質的特徴」を抽出する必要がある。ここでは、まず、「本質的特徴」とは何か説明する。活性の発現は、図 1-3 に示すとおり、ターゲットの結合サイト(ポケット)に合った、化合物が、「鍵」と「鍵穴」の関係のように、結合することによって起きる。

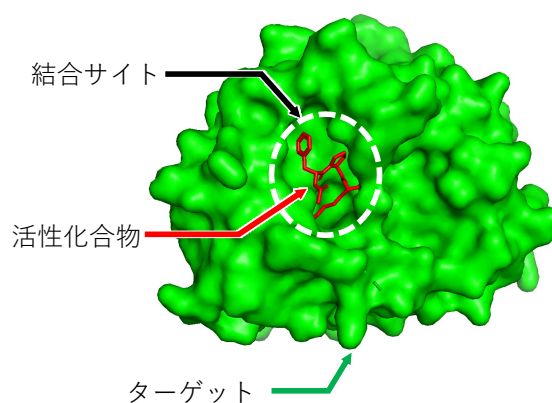


図 1-3 ターゲットの生体高分子の結合サイトと活性化合物の立体構造関係

活性化合物は、その分子構造中で、ターゲットと相互作用する可能性のある水素結合受容体(Hydrogen Bond Acceptor, HBA)・供与体(Hydrogen Bond Donor, HBD)・芳香環(Aromatic Ring, AR)・親油性(疎水性)特徴(Lipophilic, L)といった特徴を PF(Pharmacophoric Feature)と言い、対応する部分構造を PPP(Potential Pharmacophoric Point)と言う。ターゲットの結合サイトに化合物が結合するためには、PPP oughし、ターゲットの結合サイトの「鍵穴」に入るような「鍵」として特定の立体構造関係を有する必要がある。鍵の形状に相当する、PF とその物理的な位置関係が、活性を有するための本質的特徴である。そして、このような特徴を表現する方法の一つが、「ファーマコフォア」である。その一例を、図 1-4 に示す。

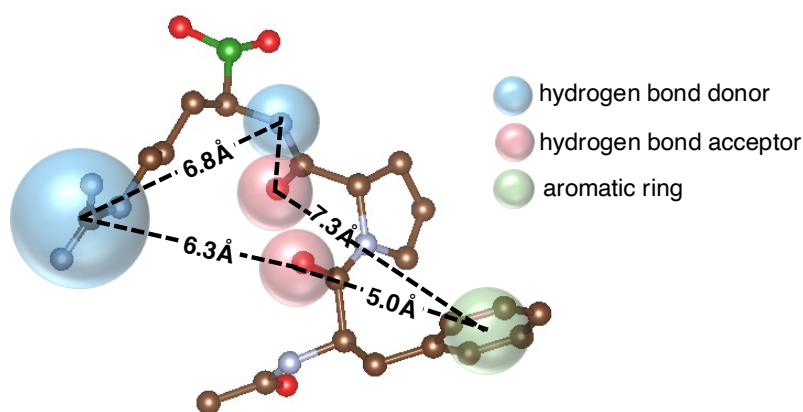


図 1-4 ジオメトリカルなファーマコフォア

ファーマコフォアは、PF とその相互位置関係を表現したものであり、図 1-4 のようにターゲットに結合した化合物の立体構造を基にした表現(ジオメトリカルな表現)が用いられることが多い。ジオメトリカルなファーマコフォアは、これらの PF とその立体的な位置関係とで表現される。つまり、図 1-4 のような「鍵」の立体構造そのものを表現していると言える。また、同図から明らかな通り、ファーマコフォアによる表現は、ターゲットと相互作用する PF とその位置関係のみで表現しているため、scaffold に依存していない。従って、目的とするターゲットについて、このよう scaffold に依存しないファーマコフォアを導出することができたら、図 1-5 に示すとおり、そのファーマコフォアに合致する化合物を、既知の活性化合物とは異なる scaffold を持つ化合物から見つける SH が可能であると考えた。

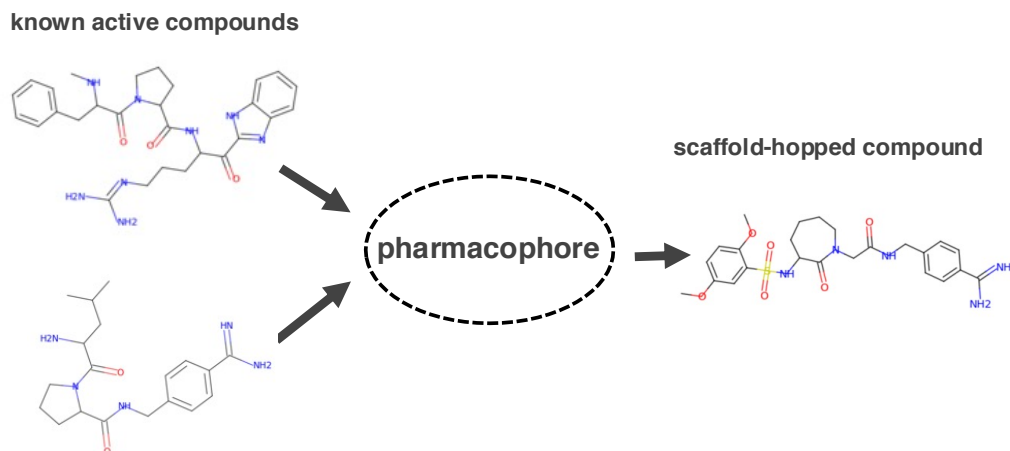


図 1-5 ファーマコフォアと Scaffold-Hopping

1.3 ジオメトリカルな手法とトポロジカルな手法

しかしながら，LBDD においては，図 1-4 に示すようなジオメトリカルなファーマコフォアを得ることは単純ではない．SBDD と異なり，ターゲットとなる生体高分子の立体構造を利用できないので，その結合サイトに合う立体構造も未知であるからである．

そこで，LBDD の手法はジオメトリカルな手法とトポロジカルな手法の 2 つの系統に分かれる．LBDD におけるジオメトリカルな手法は，ターゲットの立体構造が未知であるため，複数の活性化合物の構造のみから結合状態の立体構造を予測する必要がある．そしてその予測した立体構造に基づいて 3 次元空間上で活性化合物の特徴抽出を行う．この方法では，結合状態の立体構造を正しく予測できれば，立体構造に基づき活性予測ができる．例えば，その分子の立体構造から得られる静電ポテンシャルの分布を化合物間の類似性評価に用いる研究がある [9]．図 1-6 に示すように，CoMFA (Comparable Molecular Field Analysis) では，化合物の立体構造を再現し，その化合物の周りにプローブと言われる原子をおいたときにそのエネルギー変化を 3 次元のマップとして表すことで，静電ポテンシャルとは異なる特徴を抽出している [10]．CoMFA などジオメトリカルな手法の中で，分子構造そのものではなく，その化合物が作る立体的な場の情報を活用する手法では，scaffold に依存しない活性予測が可能となることが期待される．ジオメトリカルな手法を用いた SH 手法としては，CATS3D (Chemically Advanced Template Search 3D) や WHALES (Weighted Holistic Atom Localization and Entry Shape) といった手法が提案されている [10-14]．

しかし、LBDD におけるジオメトリカルな方法の欠点は、予測した化合物の結合時の立体構造が、実際の構造と異なっていた場合、正しい特徴抽出ができないことである。さらに、ターゲットの構造が未知であるため、予測したとしても実際にその構造が正しいか否か判断することが難しい。

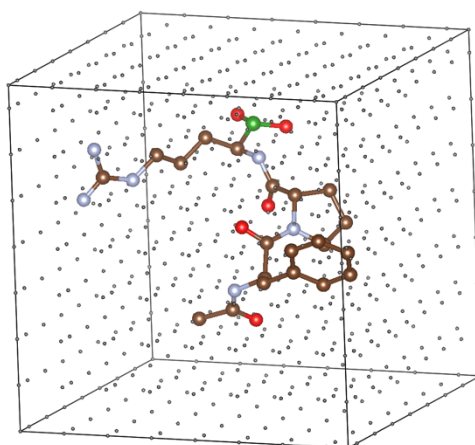


図 1-6 CoMFA

一方、トポロジカルな手法とは、分子構造をグラフとして捉えて特徴抽出を行う手法である。ECFP (Extended Connectivity Fingerprint) [15,16] のように、様々な部分構造を持つか否かという情報を 0/1 のビットとして表現してビット列として化合物を表現する手法や、溶解度などに関する特徴を記述子と呼ばれる数値として表現する手法がある。これらの手法は、固定長のベクトルに変換することで、SVM(Support Vector Machine)や RF(Random Forest)などの一般的な機械学習手法を適用できるメリットがある [17,18]。特に近年では、GNN(Graph Neural Network)を用いた手法が大きな注目を浴び[19,20]、GNNを用いた SH 手法も提案されつつある。GNN に自然言語処理分野で開発された Attention を組み合わせた用いた SH [19]、Multimodal Transfer NN [20]を用いた手法も提案されている。

また、LBDD で利用可能な情報はあくまでも、トポロジカルに表現可能な候補化合物の分子構造とその活性情報のみである。前述の CoMFA のように立体構造を介した活性予測手法は存在するが、それも、結局は、分子構造を基にして予測した立体構造を用いているに過ぎない。そのため、LBDD においては活用可能な情報がトポロジカルな情報に限定されているため、同じトポロジカルな表現も用いた手法でも、ジオメトリカルな手法と同等の予測性能を持つ可能性がある。

1.4 トポロジカルなファーマコフォア

トポロジカルな表現を用いて、「本質的特徴」を表現しようとした研究例は過去に存在する [21]. その代表的例は図 1-6a に示す Pharmacophore Graph (PhG) である. PhG は, 活性化合物の中の PF を頂点とし, その PF 間の辺の長さを, その PF が割り当てられた部分構造間の化学結合の数(トポロジカル距離)で表した完全グラフである. 図 1-7b 橙色の矢印では, 2つの HBD 間の距離が2つの結合を隔てているのでトポロジカル距離は2である. 図 1-7a にも2つの HBD 間の距離が2となっている.

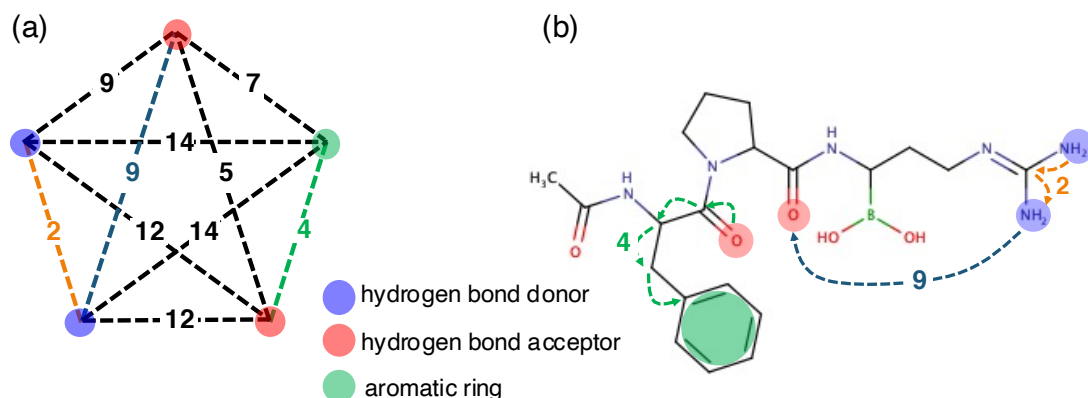


図 1-7 (a) Pharmacophore Graph(PhG) (b) 対応する PPP とトポロジカル距離

このように, 化合物がターゲットに結合したときの PF 間の物理的な距離は不明であっても, その PF に対応する PPP 間のトポロジカル距離によって代替することで, LBDD という条件の中で, 活性に必要な特徴をファーマコフォアという形で表現できる. なお, 2つの PF に対応する PPP 間のトポロジカル距離を簡単に「PF 間トポロジカル距離」と呼ぶこととする.

活性化合物の中で, 部分構造から検出した PF は, 実際に結合サイトで相互作用をする PF よりも多い. 例えば, 図 1-7a の PhG の各頂点が実際にターゲットと相互作用する PF とすると, ボロン(B)原子に結合した OH 基は水素結合し得る部分構造であるが, 相互作用はしていない. このように, 相互作用する可能性のある PF を適切に選択して, 活性化合物の中からできるだけ多く共通する PhG を選ぶ必要がある. つまり, 既知の活性化合物の情報を基に, 活性を得るために必要な PhG をマイニングする手法が必要である.

Métivier らの検討では, この PhG をマイニングする手法として, *Coverage* と *Growth-Rate* という基準を用いた [21]. *Coverage* は, 既知の活性化合物に該当する数によって PhG をスコアリングする手法, *Growth-Rate* は該当する

活性と不活性化化合物の割合でスコアリングする手法である。しかし、両者とも、scaffold の多様性を考慮していない。そのため、特定の scaffold に属する化合物のデータが多い場合など、scaffold に依存した PhG を高くスコアリングする可能性がある。つまり、SH のために、scaffold に依存しない本質的特徴を抽出するためのスコアリング手法の開発することが課題である。

また、図 1-7a に示すとおり、PhG 自体は各頂点が他のすべての頂点と、辺を持って繋がっている完全グラフである。図 1-7b の分子構造をグラフと見ると、当然、各原子は他の全原子と繋がっているのではなく、省略している水素原子を入れても、隣接した 2-4 個の原子としか接続していない。つまり点の数に対して辺の数が少ない「スパース」なグラフとなっている。したがって、完全グラフを用いる PhG は、活性化化合物に共通する特徴を抽出した表現、いわば活性化化合物のモチーフとしては、「解釈性」に改善の余地がある。

1.5 本研究の目的

前節で示したとおり、トポロジカルなファーマコフォア表現として唯一挙げられる PhG は、SH のような scaffold に依存しない特徴抽出が必要な予測への適用が検討されていないという課題がある。また、完全グラフである PhG 自体、解釈性が低いという課題もある。そこで、本研究では、下記の 2 点を目的とした検討を行った。

目的① 解釈性の高いトポロジカルなファーマコフォアの表現を見つけること

目的② その表現に基づいて SH を行う手法を提案すること

第 2 章では、活性化化合物のデータセットから、SH のために重要な PhG を見つけるスコアリング手法を提案する。活性化化合物群からは様々な PhG を抽出することができるが、「本質的特徴」に合致する PhG を同定する手法が求められる。既知の活性化化合物に含まれる割合によって PhG をスコアリングする *Coverage* と、含まれる活性化化合物数と不活性化化合物数の比でスコアリングする *Growth-Rate* という手法があるが、本研究では、既存の活性化化合物の中で、該当する scaffold 数によってスコアリングする *NScaffold* という手法を提案する。なお、各々のスコアリング手法の詳細な定義は第 2 章で説明する。

第 3 章では、PhG の課題であった「グラフの解釈性が低い」という弱点を克服するため、頂点数に対して辺の数が少ないスパースなグラフ表現 Sparse Pharmacophore Graph (SPhG) を提案する。また、複数の活性化化合物から共通

する特徴を抽出して SPhG として提示するアルゴリズムについて説明し, PhG より解釈性の高い SPhG が, PhG と同等の SH 性能を持つことを示す.

第 4 章では, Graph Edit Distance (GED) というグラフの類似度指標に基づいて, 複数の SPhG を平面上に配置するクラスタリング解析をした結果について説明する. そこで, 6 つのターゲットとなる生体高分子に対して, SPhG を用いた解析を行い, SPhG が, scaffold に依存しない活性化合物の特徴を表現できていることを示す. また, GED を用いたクラスタリング手法が複数の SPhG を把握するために適したものであることを示す [22,23]. 第 5 章で, 本研究のまとめと, SH を中心とした今後の展開について述べる.

1.5 参考文献

1. Schneider, G.; Baringhaus, K. H. *Molecular Design. Concepts and Applications*; Wiley-VCH, **2008**.
2. Stumpfe, D.; Bajorath, J. Current Trends, Overlooked Issues, and Unmet Challenges in Virtual Screening. *J. Chem. Inf. Model.* **2020**, 60 (9), 4112–4115.
3. Ślędź, P.; Caflisch, A. Protein Structure-Based Drug Design: From Docking to Molecular Dynamics. *Curr. Opin. Struct. Biol.* **2018**, 48, 93–102.
4. Ono, F.; Chiba, S.; Isaka, Y.; Matsumoto, S.; Ma, B.; Katayama, R.; Araki, M.; Okuno, Y. Improvement in Predicting Drug Sensitivity Changes Associated with Protein Mutations Using a Molecular Dynamics Based Alchemical Mutation Method. *Sci. Rep.* **2020**, 10 (1), 1–10.
5. Heifetz, A.; Morao, I.; Babu, M. M.; James, T.; Southey, M. W. Y.; Fedorov, D. G.; Aldeghi, M.; Bodkin, M. J.; Townsend-Nicholson, A. Characterizing Interhelical Interactions of G-Protein Coupled Receptors with the Fragment Molecular Orbital Method. *J. Chem. Theory Comput.* **2020**, 16 (4), 2814–2824.
6. Acharya, C.; Coop, A.; E. Polli, J.; D. MacKerell, A. Recent Advances in Ligand-Based Drug Design: Relevance and Utility of the Conformationally Sampled Pharmacophore Approach. *Curr. Comput. Aided-Drug Des.* **2010**, 7 (1), 10–22.
7. Schneider, G.; Schneider, P.; Renner, S. Scaffold-Hopping: How Far Can You Jump? *QSAR Comb. Sci.* **2006**, 25, 1162–1171.
8. Böhm, H. J.; Flohr, A.; Stahl, M. Scaffold Hopping. *Drug Discov. Today Technol.* **2004**, 1, 217–224.

9. Leach, A. R.; Gillet, V. J.; Lewis, R. A.; Taylor, R. Three-Dimensional Pharmacophore Methods in Drug Discovery. *J. Med. Chem.* **2010**, *53*, 539–558.
10. Kubinyi, H. *Handbook of Chemoinformatics* **2008**.
11. Renner, S.; Noeske, T.; Parsons, C. G.; Schneider, P.; Weil, T.; Schneider, G. New Allosteric Modulators of Metabotropic Glutamate Receptor 5 (mGluR5) Found by Ligand-Based Virtual Screening. *ChemBioChem* **2005**, *6*, 620–625.
12. Renner, S.; Schneider, G. Scaffold-Hopping Potential of Ligand-Based Similarity Concepts. *ChemMedChem* **2006**, *1*, 181–185.
13. Grisoni, F.; Merk, D.; Consonni, V.; Hiss, J. A.; Tagliabue, S. G.; Todeschini, R.; Schneider, G. Scaffold Hopping from Natural Products to Synthetic Mimetics by Holistic Molecular Similarity. *Commun. Chem.* **2018**, *1*, 44.
14. Grisoni, F.; Merk, D.; Byrne, R.; Schneider, G. Scaffold-Hopping from Synthetic Drugs by Holistic Molecular Representation. *Sci. Rep.* **2018**, *8*, 16469.
15. Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints *J. Chem. Inf. Model.* **2010**, 742–754.
16. Laufkötter, O.; Sturm, N.; Bajorath, J.; Chen, H.; Engkvist, O. Combining Structural and Bioactivity-Based Fingerprints Improves Prediction Performance and Scaffold Hopping Capability. *J. Cheminf.* **2019**, *11*, 54.
17. Svetnik, V.; Liaw, A.; Tong, C.; Christopher Culberson, J.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947–1958.
18. Doniger, S.; Hofmann, T.; Yeh, J. Predicting CNS permeability of drug molecules: comparison of neural network and support vector machine algorithms. *J. Comput. Biol.* **2002**, *9*, 849–864.
19. Stojanovic, L.; Popovic, M.; Tijanic, N.; Rakocëvic, G.; Kalinic, M. Improved Scaffold Hopping in Ligand-Based Virtual Screening Using Neural Representation Learning. *J. Chem. Inf. Model.* **2020**, *60*, 4629–4639.
20. Zheng, S.; Lei, Z.; Ai, H.; Chen, H.; Deng, D.; Yang, Y. Deep Scaffold Hopping with Multi-Modal Transformer Neural Networks **2020**, *ChemRxiv preprint* chemrxiv.13011767.v1.
21. Métivier, J. P.; Cuissart, B.; Bureau, R.; Lepailleur, A. The Pharmacophore Network: A Computational Method for Exploring Structure-Activity

- Relationships from a Large Chemical Data Set. *J. Med. Chem.* **2018**, *61*, 3551–3564.
22. Abu-Aisheh, Z.; Raveaux, R.; Ramel, J. Y.; Martineau, P. An Exact Graph Edit Distance Algorithm for Solving Pattern Recognition Problems. *ICPRAM 2015 - 4th Int. Conf. Pattern Recognit. Appl. Methods, Proc.* **2015**, *1*, 271–278.
23. Garcia-Hernandez, C.; Fernández, A.; Serratos, F. Ligand-Based Virtual Screening Using Graph Edit Distance as Molecular Similarity Measure. *J. Chem. Inf. Model.* **2019**, *59*, 1410–1421.

第2章 Scaffold hopping (SH)のための Pharmacophore

Graph (PhG)のマイニング手法の提案

2.1 緒言

2.1.1 トポロジカルな手法を用いた SH

前章で述べたとおり，創薬の初期段階では，候補化合物の多様性を高めることが必要となる．そのために，既知の活性化合物とは異なる scaffold の活性化合物を見つける SH が必要となることがある．

SH という概念を提唱した G.Schneider らは，PF とその PF 間トポロジカル距離を用いた Chemically Advanced Template Search (CATS) という手法により SH が可能であることを示した [1]．図 2-1 に示すとおり，CATS では，HBD, HBA に加え positively ionizable (PI), negatively ionizable (NI), および Lipophilic (L) という 5 種類の PF を扱い，2 つの PF ペアとその PF 間トポロジカル距離との組み合わせの頻度を，ヒストグラムで表す．

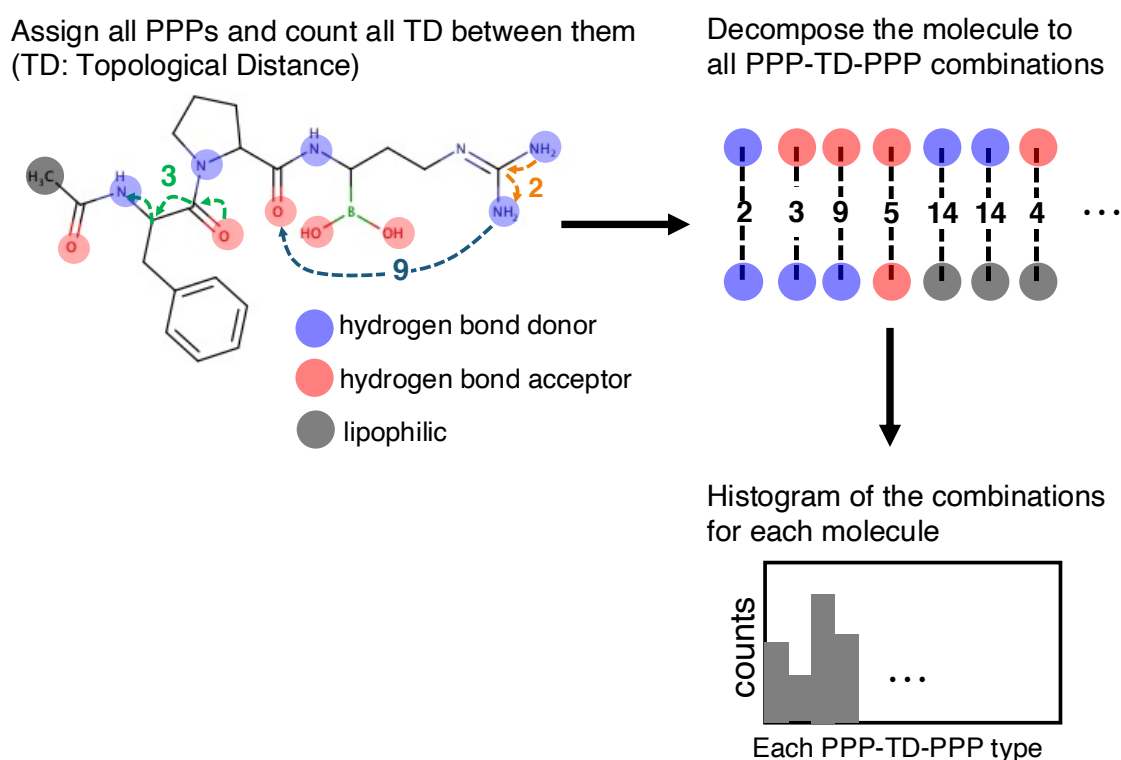


図 2-1 CATS (Chemically Advanced Template Search) [1]

同種の PF からなるペアを含む 5 種類の PF のペアは 15 種類あり、PF 間トポロジカル距離 1~10 までを考慮するため、150 次元の特徴ベクトルで分子を表現できる。このとき PF の種類とその PF 間トポロジカル距離しか用いていないため、scaffold 依存性の少ない類似度の表現が期待される。Bonachéra らは、CATS を発展させ、2 つの PF ではなく、3 つの PF と各 PF 間トポロジカル距離の組み合わせを用いたヒストグラムを化合物の特徴ベクトルとし、複雑な PF の位置関係を表現した [2]。このように、トポロジカルな手法においては、PF の物理的な距離関係ではなく、トポロジカル距離(化学結合の数)を用いられることが多い。

しかし、これらのヒストグラムを用いた手法では、scaffold を超えた活性予測が可能であったとしても、前章で説明した「本質的特徴」やその表現としてのファーマコフォアを導出しているわけではない。あくまで、ヒストグラムは、個々の化合物の特徴ベクトルであり、複数の活性分子に共通する特徴を抽出しているわけではない。

2.1.2 PhG とそのマイニング方法

Métivier らは、前章でも触れたとおり、CATS などの PF とその PF 間トポロジカル距離を用いる特徴抽出手法を発展させて、活性化合物に共通する特徴を表現する PhG を導出する手法を提案した[3]。CATS で用いられた 2 点の PF とそのトポロジカル距離で構成される単純な 2 頂点のグラフから、PF 数を 5 つまたは 6 つまで増やし、そのトポロジカル距離で構成される完全グラフを導出した。その例は図 2-2 の下半分の 3 つのグラフに示すとおりである。そして、そのグラフを、既知の活性化合物に「含まれる」割合によってスコアリングして、それが高いものを PhG として抽出(マイニング)した。このグラフ順位付けするスコアリング手法を *Coverage* と呼ぶ。

まず、図 2-2 を用いてこの「含まれる」ということを説明する。同図 2-2 の分子構造の中で、赤色矢印で示した PF を選択し、その PF ペア 15 個のトポロジカル距離をすべて算出した後、5 つの頂点が PF の種類(HBA や AR)で表され、その頂点間の辺の長さが各トポロジカル距離で表される PhG を作る。そのグラフは、同図左下の PhG と完全に一致する。このように、ある分子が持つ PF を適切に選択することで、クエリとなる PhG と完全一致する場合、同図上部の分子構造にクエリとなる PhG は「含まれる」と言う。一方で、他の 2 つのグラフは、どの PPP の組み合わせを選択しても、一致するグラフを作ることはいできない。例えば、同図右下の PhG では、AR が 3 つあるが、例示した分子構造には AR が 1 つしかないので、どのように PF の組み合わせを選んで

もこのグラフに一致するものは得られない。したがって、右下のクエリは、上部の分子に「含まれない」。

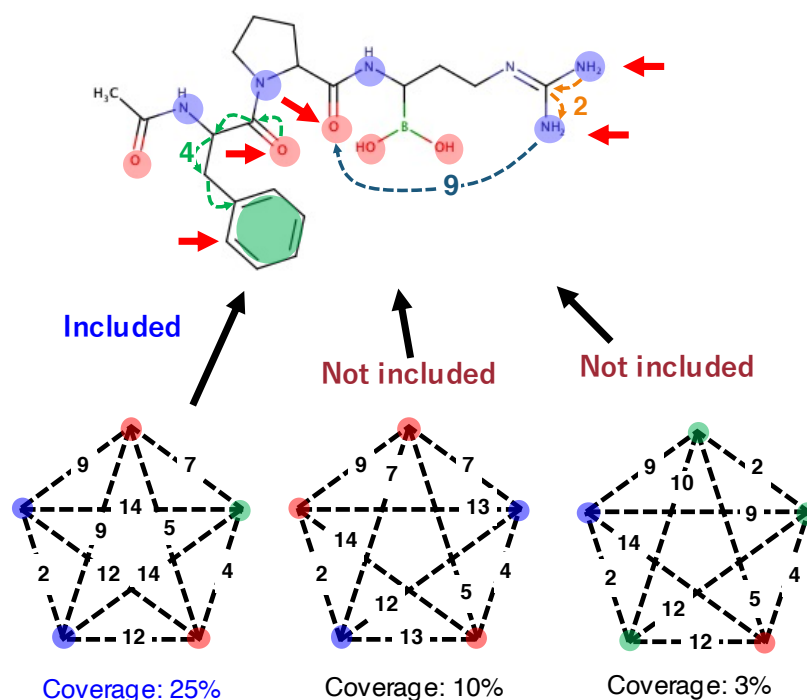


図 2-2 化合物に含まれるグラフと含まれない PhG [3]

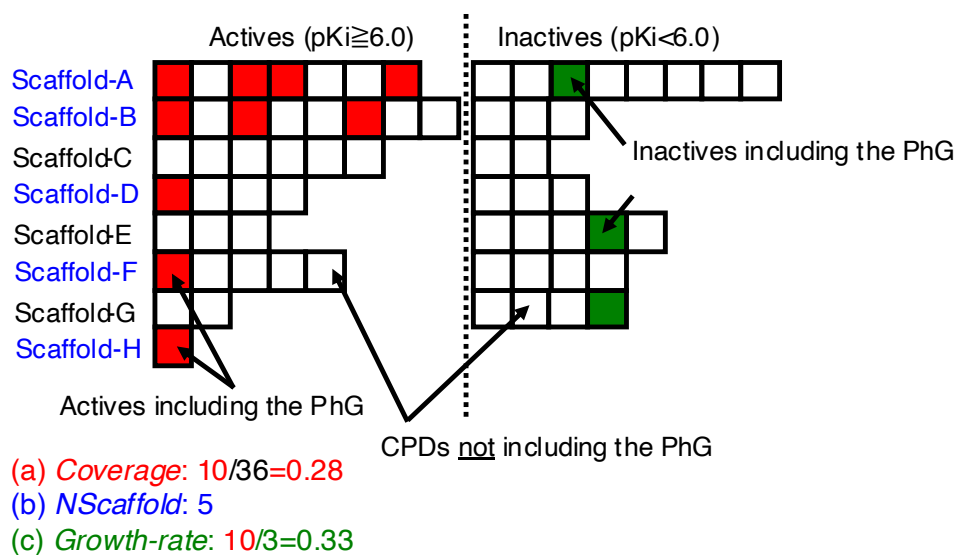


図 2-3 重要度の高い PhG をマイニングするためのスコアリング手法¹ [4]

(a) Coverage (b) NScaffold (c) Growth-rate

¹ Reprinted with permission from J. Chem. Inf. Model. 2020, 60, 2073–2081. Copyright 2020 American Chemical Society.

次に、図 2-3 を用いて *Coverage* と、含まれる活性化合物数と不活性化合物数の比でスコアリングする *Growth-Rate* という 2 つ手法を説明する [3]。これらの 2 つのスコアリング手法を、図 2-3 を用いて説明する。各マスは 1 つの化合物に対応しており、同じ行のマスは同じ scaffold に属する。活性のある化合物は左に、活性のない化合物は右に配置されている。各マスは 1 つの化合物に対応しており、同じ行のマスは同じ scaffold に属する。活性のある化合物は左に、不活性化合物は右に配置されている。まず、*Coverage* は、

$$Coverage(PhG) = \frac{N_{active,PhG}}{N_{all}}$$

と表される。ここで、 $N_{active,PhG}$ は PhG が含まれる活性化合物の数であり、 N_{all} は既知の活性化合物の数である。*Coverage* 法は、既知の活性化合物のうち、PhG が含まれる割合に基づいて PhG をランク付けする。図 2-3 では、 $N_{active,PhG}$ は 10 で N_{all} は 36 であるため、*Coverage* は $10/36 \div 0.28$ となる。*Growth-rate* は、PhG がカバーする非活性化合物の数に対する *Coverage* スコアの比率であり、次のように定義される。

$$Growth-rate(PhG) = \frac{N_{active,PhG}}{N_{inactive,PhG}}$$

ここで、 $N_{inactive,PhG}$ は PhG を含む不活性化合物の数でありここでは 3 である。つまり、*Growth-rate* のスコアは $10/3 \div 3.3$ となる。

2.1.3 SH のための PhG のマイニング方法の提案

図 2-3 で説明したスコアリング手法の *Coverage* を用いた場合、データセットに一つの scaffold に属する活性分子が存在する場合に、その scaffold の特徴に捕われて、scaffold に依存しない本質的特徴を、抽出できない可能性がある。また、*Growth-Rate* も多くの化合物に共通するのではなく比で表しているため、活性化合物の中にのみ存在する scaffold が存在した場合、*Coverage* と同様にその scaffold のみを高くスコアリングしてしまう。

そこで、本章の研究目的を、SH に適した PhG のマイニング方法の提案とした。具体的には、該当する分子の属する scaffold 数に基づいて PhG をスコアリングする *NScaffold* という手法を考案した。図 2-3 に示すとおり、*NScaffold* は、活性化合物の数の代わりに、PhG でカバーされる Scaffold 数 $N_{scaffold,PhG}$ を指標として用いる。*NScaffold* の定義は

$$NScaffold(PhG) = N_{scaffold,PhG}$$

である。本章では、*Coverage* など既存の PhG のスコアリング手法よりも SH 性能が優れていることを示した [4]。

まず、PhG をクエリとして使用した VS を実施することで、PhG の SH 性能を評価した。トレーニングデータとテストデータは、共通する scaffold に属する化合物が存在しないように分割し、トレーニングデータから抽出した PhG を用いてテストデータの活性を予測した。その結果、PhG の選択方法によって、VS における SH の性能が大きく変わることがわかった。特に PhG が該当する分子の数で優先順位付けする *Coverage* 法ではなく、*NScaffold* 法で PhG を選択すると、高い SH 性能を有することがわかった。さらに、トロンビン阻害剤のデータセットについてこの *NScaffold* 法を適用して得た PhG が持つ意味の解釈を試みた。すると、実験的に検証済みのトロンビンと阻害剤の結合点に相当する特徴を、抽出できていることがわかった。

2.2 研究方法

2.2.1 PhG の作成方法

本章の研究に用いた PhG の作成方法について、ここで正確に定義する。前述の通り、PhG とは、分子構造をグラフで表現し抽象化したもので、HBD・HBA などの PF が割り当てられた PPP を頂点、その PPP 間のトポロジカル距離を辺の長さとして表現している。なお、2つの PF に対応する PPP 間のトポロジカル距離を簡単に「トポロジカル距離」と呼ぶこととする。PF が割り当てられる PPP は、SMARTS [5]や SLN [6]などの線形表記手法を用いて特定できる。本研究では、HBD、HBA の他に、AR、PI、NI、亜鉛結合剤(Zinc binder, ZB)、親油性(疎水性)部分 (L)を PF として採用した。定義としては、疎水性の PF を除き、RDKit 付属のファイル名「BaseFeatures.fdef」[7,8]に記載された定義を採用した。疎水性部分は、Métivier らの定義を用いた [3]。また、PF のペアの少なくとも 1 つに PF に複数の原子が含まれる PPP が割り当てられる場合、PPP 間のトポロジカル距離が複数存在するときがある。その場合、その中の最短距離を、「トポロジカル距離」として定義した。

図 2-4 に PhG の構築の例を示す。まず、分子構造に、SMART ベースのマッチングアルゴリズムを適用して、全ての PF を割り当てた (図 2-4a)。次に、特定の数の PF の組み合わせをすべて個別に抽出し、PhG に変換した (図 2-4b および図 2-4c)。そして、グラフの重複を除去し、各分子が持つ PhG を算出した。本章では、計算コストを考慮して、1つのグラフに含まれる PF の数を 6

つに設定した。なお、2つのグラフの一致・不一致の判定を高速化するために、グラフはすべてハッシュ化した。

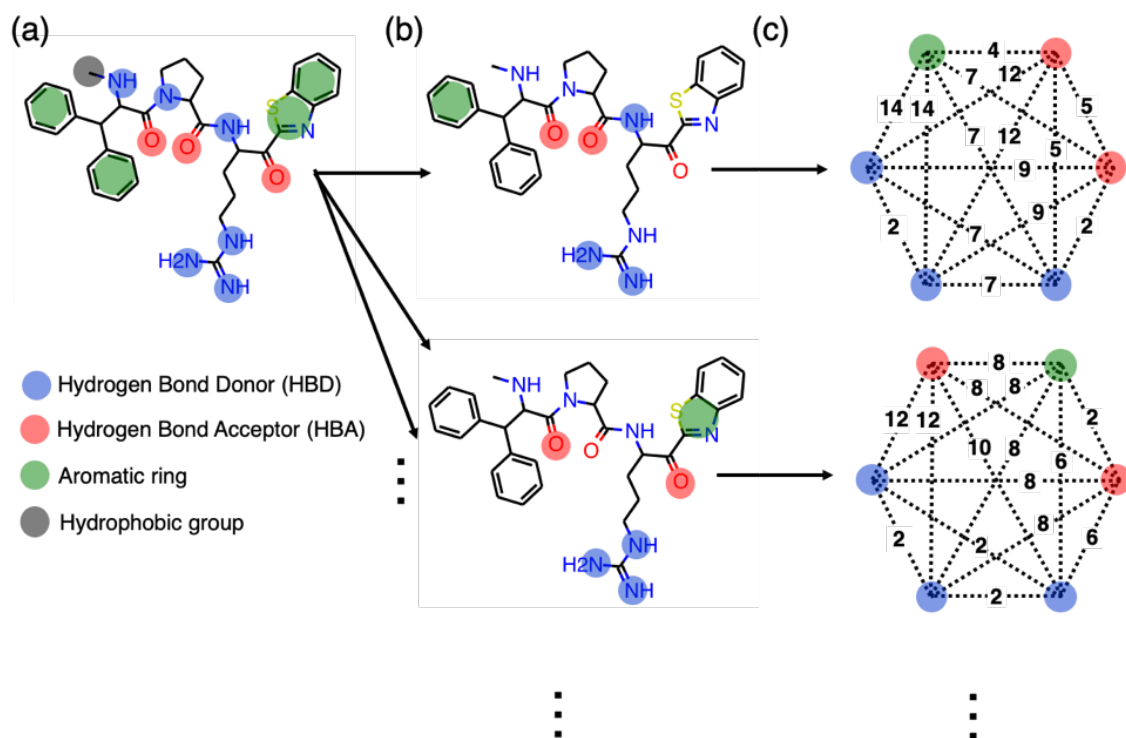


図 2-4 PhG の網羅的な作成¹[4]

(a) 分子構造の中の PPP を全て抽出する。(b) 抽出した PPP の中から、所定数の PF のすべての組み合わせを各々選択する。(c) 各々の組み合わせごとに網羅的に PhG の候補グラフを作成する。このグラフは、PPP が頂点に相当し、2つの PPP 間トポロジカル距離が、頂点間の辺の長さとする完全グラフである。

2.2.2 PhG を用いた VS(Virtual Screening)

VS で PhG を直接利用する最もシンプルな方法は、PhG をクエリとして、候補化合物のデータベースをスクリーニングすることである。図 2-5 はその VS の全体像を示している。

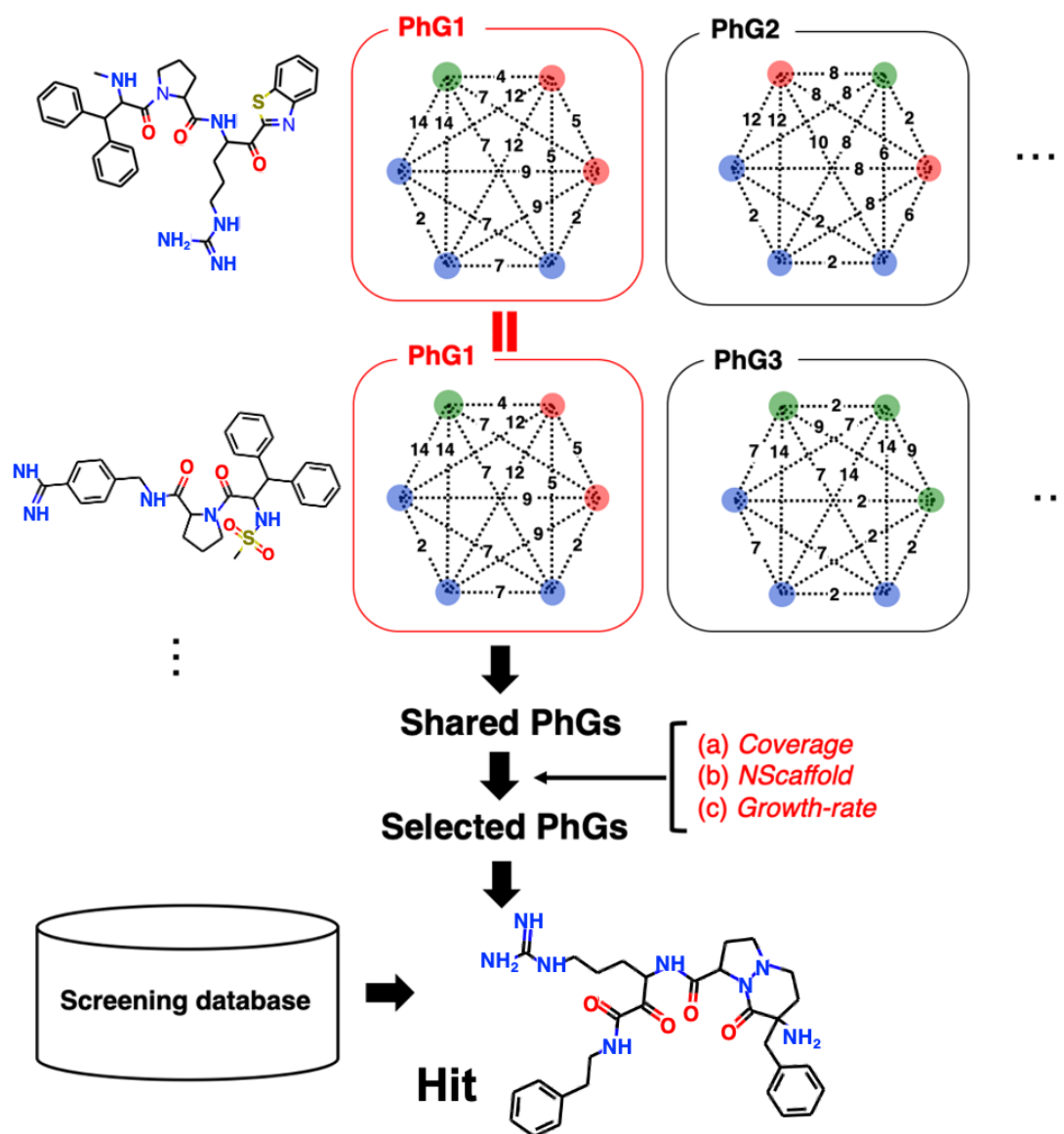


図 2-5 PhG を用いた VS のフロー概念図¹[4]

まず、既知の活性化合物のデータから、図 2-4 で説明した方法で、複数の PhG に変換する。ここで、前出の通り、グラフの持つ PPP 数を 6 と設定しているの
で、ある分子が 10 個の PPP を持つ場合、10 個の中から 6 個の組み合わせを選

ぶため、最大で 210 通りの PhG が存在する。210 個の中で完全に一致する PhG があるときは 1 つのみを残すので実際は 210 個より少なくなる可能性がある。複数の化合物に対してグラフを網羅的に作成すると、異なる化合物で共通している PhG が多く存在する。これを図 2-5 では Shared PhG と表現した。既知の活性化合物から生成される PhG は多数存在するため、クエリセットの PhG の選択は重要である。

次に、これらの PhG について、(a) *Coverage* (b) *NScaffold* (c) *Growth-rate* の 3 つのスコアリング手法を用いて優先順位付けして、上位 1~100 個の selected PhG を抽出し、最後に、その selected PhG の組み合わせをクエリとした活性予測を行う。既知の活性化合物と異なる scaffold に属する化合物で構成されるテスト化合物の中で、1 つでもクエリの PhG を含むものがあれば、そのテスト化合物が活性を持つとする。

本研究では、トレーニングデータの活性化合物から抽出した全ての PhG をスコアでランク付けし、上位 N 個の PhG をクエリセットとして使用した。本研究では、N の数を 1, 3, 10, 30, 50, 100 に設定した。

2.2.4 本研究で用いたターゲット生体高分子

ここで、2 章から 4 章で用いるターゲットの特徴を述べる。各章で全ターゲットを使っているわけではないがここにまとめて記載する。

Thrombin (Thr.)

Thr. は血液凝固プロセスに主要な役割を果たす酵素である [14]。血液凝固を阻害するためには、このタンパク質の活性部位に、本来結合するフェブリノゲン以外の分子を結合させればよい。血液凝固の原因となるフィブリンの発生を阻害することで、脳梗塞などの血栓症を抑制する。後述する TPS6 と同様のセリン・プロテアーゼである。

Coagulation factor X (fX.)

Thr. と同様に血液凝固カスケードに関与する酵素である [15]。fX. の機能阻害により、静脈血栓症・動脈血栓症など種々の血栓症に効果がある。

Tyrosine Kinase ABL1 (ABL1)

ABL1 は、生体内でエネルギーを媒介する adenosine triphosphate (ATP) を特定のタンパク質内のチロシン残基に結合させる機能を持つ。ATP の結合部位

などに、他の分子を結合させることで、この機能を阻害できる。阻害剤は、主に抗がん剤として利用できる [16].

PI3-kinase p110-alpha subunit (PI3)

PI3 は、キナーゼの一種であり、ホスファチジルイノシトールをリン酸化するシグナル伝達タンパク質である PI3K の変異種 PI3KCA のサブユニットである。PI3 は、がんとの関連性が指摘されているため、創薬上のターゲットとして適している可能性がある [17].

κ -opioid receptor (Kop.)

オピオイド系化合物と結合する G タンパク質共役受容体 (GPCR) の一種であり、視床下部・脊髄に多く存在する。モルヒネなどと結合し鎮痛作用を発現する [18].

G protein-coupled receptor 44 (GPCR44)

GPCR は、細胞外に伝達物質の結合部位を、細胞内に G タンパク質の結合部位を有し、細胞外の情報を細胞内に伝達する役割を持つ。多数発見されている GPCR の一つを、VS の対象として用いた [9].

Histamine H3 receptor (His.)

His. はアレルギー・胃酸分泌・中枢神経系に働くヒスタミンの受容体の一つであり GPCR ファミリーに属する。脳の神経系伝達に関与している [19].

Transmembrane protease serine 6 (TPS6)

プロテアーゼは、タンパク質のペプチド結合を切断し、ポリペプチドやアミノ酸等のより小さな単位に変換する役割を持つ。TPS6 は、鉄欠乏性貧血に関与していると考えられている [20].

なお、本章では、ABL1, Thr., Kop., fX, PI3, および His. の 6 つのターゲットに関するデータセットを用いた。

2.2.4 検証のための化合物データセット

表 2-1 に今回の VS に用いたデータセットの詳細を示す。活性化合物は、ChEMBL ver. 24 [9] から、ターゲットに掲載されている化合物の数と、高分子の種類（キナーゼ 2 種、プロテアーゼ 2 種、G タンパク質共役型受容体 2 種）に基づいて、6 つの高分子ターゲットを選択した。1 つの化合物に対して複数の

Ki 値が得られた場合、すべての値が同じオーダーに収まる限り、それらの幾何平均を計算して最終的な活性値とした。また、ChEMBL から、「not active」または「inactive」と記載された報告された化合物についても取得した。

本章では、先行研究に基づき、pKi 値が 6.0 以上の化合物を活性化合物と呼び、6.0 未満の化合物を不活性化合物と呼ぶこととする [3]。しかし、ChEMBL から得た不活性化合物数が少なく、6 ターゲット全てで活性化合物より少ない。そこで、VS の Precision 評価では、ネガティブな化合物数を増やすために、ChEMBL の不活性化合物ではなく、ZINC ver. 15 からランダムに選択した 250,000 個の化合物を不活性化合物とした [10]。

また、scaffold のタイプによる依存性を排除するために、BM (Bemis-Murcko scaffold) [11]、CCR (compound-core relationship) single [12] 及び CCR RECAP (CCR retrosynthetic combinatorial analysis procedure) [12,13] の 3 種類の scaffold タイプを用いた。6 つのターゲットの各 scaffold 数は、BM scaffold では 209 から 1380、CCR Single では 101 から 1041、CCR RECAP では 262 から 1737 であった。

なお、約 1,600 個の分子について PhG を生成し、3 つのスコアリング手法のいずれかで PhG を選択するまでの実行時間は、CPU に Intel Core-i9-7960X 2.80GHZ を搭載したデスクトップ PC のシングルコアを使用した場合、約 3 時間であった。

表 2-1 化合物データセットのプロファイル

ChEMBL ID ^a	Target	Code	#CPDs ^a (Active CPDs ^b)	#Scaffolds		
				Bemis Murko ^c	CCR Single ^d	CCR RECAP ^e
CHEMBL1862	Tyrosine kinase ABL1	ABL1	634 (544)	209	101	262
CHEMBL204	Thrombin	Thr.	1643 (600)	884	659	1045
CHEMBL237	κ -opioid receptor	Kop.	3176 (1745)	1380	1041	1737
CHEMBL244	Coagulation factor X	fX.	1758 (1209)	735	332	826
CHEMBL4005	PI3-kinase p110- alpha subunit	PI3	945 (864)	365	175	425
CHEMBL264	Histamine H3 receptor	His.	2627 (2440)	1268	667	1289

^a CPDs: ChEMBL から取得した化合物数 [9].

^b 活性化合物: ChEMBL から得た化合物の中で, pK_i が 6 以上のもの. 本章では, 単に「活性化合物」と呼ぶ.

^c Bemis Murko: RDKit により作成した Bemis Murko scaffold [11]

^d CCR Single: Compound core-relationship (CCR)ベースの scaffold.³³

^e CCR RECAP: RECAP(Retrosynthetic Combinatorial Analysis Procedure)ルールにより結合を切断した CCR ベースの scaffolds [12,13]

2.3 結果と考察

2.3.1 検証方法

候補分子には、実際にターゲットとの結合に寄与する PF 以外にも、一般的に PF となり得る部分構造を持つことが多く、PF 数を 6 個程度に限定しても PhG の候補を複数作ることができる。そこで、2.1.3 節に示した 3 つのスコアリング手法を用いて、SH のために必要な PhG をマイニングし、トレーニングデータとは異なる scaffold を持つ活性化合物を同定する能力について比較した。

SH 性能の評価は、トレーニングデータに無い、新しい scaffold の活性の正しく識別できる割合を指標として用いた。まず、共通の scaffold を持たないようにトレーニングデータセットとテストデータセットを作成した。このとき、3 つの scaffold タイプそれぞれについて 10 回行い、各回でトレーニングデータとテストデータに含まれる scaffold が 1:1 または 1:4 になるように、ランダムに選んだ。合計 30 回の VS を平均して、SH 性能の評価を行った。特に、トレーニングデータがテストデータの 1/4 の scaffold しか持たない条件は、創薬プロセスの初期段階を想定している。SH 性能は、正しく活性であると判別された活性化合物数 (True Positive, TP), 誤って不活性であると予測された活性化合物数 (False Negative, FN), ZINC から得た不活性化合物の中で、VS により活性であると判別された化合物数 FP を用いて評価した。これらの値を用いて、precision ($TP/(TP+FP)$) と recall ($TP/(TP+FN)$) の 2 つの指標を計算しグラフにプロットした。一般的に、この 2 つの指標はトレードオフの関係にあり、クエリセットに含まれる PhG (selected PhG) の数がトレードオフのバランスを左右するパラメータとなる。1 つの PhG のみを使用すると、recall を犠牲にして高い precision を示すことができる。一方、100 位までの PhG を使用すると、FP が増加して precision が増加するリスクはあるものの、より多くの活性化合物を特定できる。したがって、クエリに用いる PhG 数を 1~100 まで変化させながら、precision と recall の関係をグラフとして可視化し、その precision-recall 曲線下の面積の大きさが SH 性能と比例するとして、SH 性能の議論を行った。

2.3.2 スコアリング手法の比較結果

図 2-6 に、トレーニングデータとテストデータを、同一の scaffold 数 (1:1) で分割した場合の SH 性能を示した。各グラフは、それぞれ 6 つのターゲットに相当し、各グラフで 3 つのスコアリング手法の precision-recall 曲線を比較した。図 2-6 では、最も Recall の低い点 (四角、丸、ダイヤモンドの各プロットの左端) は、単一のクエリ $N=1$ を使用した場合に対応しており、 N が増加するにつれて回収率の値が増加する。各線上の 6 つのマークは、それぞれ $N=1, 3,$

10, 30, 50, 100 に対応している. VS のパフォーマンスは, recall と precision のトレードオフの関係を示している. Kop., fX., PI3, 及び HI3 の 4 つのターゲットでは, *NScaffold* 法が他のスコアリング法を常に上回っていた. これは, *NScaffold* 法が他の 2 つの手法よりも, scaffold に依存しない特徴をより多く抽出できることを示している. つまり, *Coverage* と *Growth-rate* 法で選択された PhG は, scaffold に依存した特徴を優先しやすい傾向がある. しかし, ABL1 でと Thr.では, *Coverage* と *NScaffold* に大きな差は見られなかった.

次に, 図 2-7 に, トレーニングデータとテストデータの scaffold 数が 1:4 の場合の precision-recall 曲線を示す. この条件では, トレーニングデータに含まれる scaffold の数は, 図 2-4 の scaffold 数の 5 分の 2 に減少した. 図 2-7 は, *NScaffold* が他のスコアリング手法よりも優れていることを, 図 2-6 よりも明確に示している. 特に, ABL1 と Thr.では, 図 2-6 よりも *NScaffold* の *Coverage* に対する優位性がより顕著になった. これらの結果から, *NScaffold* 法は, 他の従来の手法と比較して, トレーニングデータの scaffold 数が少ない場合に, その有効性が高くなることがわかった. *Coverage* や *Growth-rate* 基準では, トレーニングデータの中に, 非常に多くの活性化化合物が属する Scaffold が含まれていた場合, その Scaffold のみに含まれる PhG が上位選択され, 他の Scaffold をもつ活性化化合物を予測できない. そのため特にトレーニングデータが少ない場合に, *NScaffold* との Recall の差が広がったと考えることができる.

また, Kop.では *NScaffold* 基準と *Coverage* 基準の Precision-Recall 曲線の傾きが異なるが, fX.では同じ傾きを持つ. これは, 入手したデータセットごとの特徴が反映されたと考えている. (そのため今回は複数のターゲットを用いて検証を行った.) 今回の VS では, クエリとした PhG のうちどれか一個でも含めば活性があると判定する方法を採用している. fX.のように傾きが同じで Recall が 0.1 程度異なる現象は, *NScaffold* 基準で抽出できたが, *Coverage* では抽出できなかった PhG が存在したためと考える. *Coverage* 基準では, その PhG を含むテスト活性化化合物を抽出できなかった. Kop.では, *NScaffold* 基準でのみ抽出された PhG が, クエリ数 10-50 の間で多い. その際, Recall が上昇する(右側に点が移動する)ため傾きが異なって見えると推察する.

なお, 6 つのターゲットのうち, His.と Kop.については, 3 つの手法の Recall が最も低かった (図 2-6c,e および図 2-7c,e). この 2 つのターゲットでは, 他のターゲットに比べて scaffold 数が多かった. ヒスタミン H3 の場合, BM (CCR Single) のスカフォールド数は 1268 (667), κ -オピオイドレセプターの場合は 1380 (1041) であった. 一方, BM(CCR Single)スカフォールド数が 884(659) と 3 番目に多かった Thr.は, 比較的高い recall を示した.

図 2-6 と図 2-7 から、クエリとする PhG 数が多いほど、Recall が上昇する代わりに、Precision が低下する傾向がある。したがって、実際に今回提案する手法を用いるときには、既知のデータで検証して Precision-Recall 曲線を算出し、PhG 数を調整することで、目的の Precision または Recall が得られるようになると思う。また、 $NScaffold$ における $N_{scaffold,PhG}$ の値にしきい値を設ける方法も有効である可能性があるが、現段階では検証できてない。対象とするターゲットやデータセットに依存せず、骨格に依存しない特徴を抽出できる $N_{scaffold,PhG}$ の値が規定できるか検証することも、創薬として本質的な課題であると考えている。

さらに、今回は $NScaffold$ 基準を用いてスコアが高い PhG を 1~100 個抽出したが、1 個目の PhG を選択したあと、2 個目以降の PhG では、先に選んだ PhG と異なる Scaffold の活性分子に含まれるものを優先して選択するという方法も、今後の改良案として考えている。既知の活性分子を網羅する最小数の PhG の組み合わせを探索する、もしくはそれに準ずるような組み合わせを選択できたらより高いパフォーマンスを発揮できる可能性があると考えている。

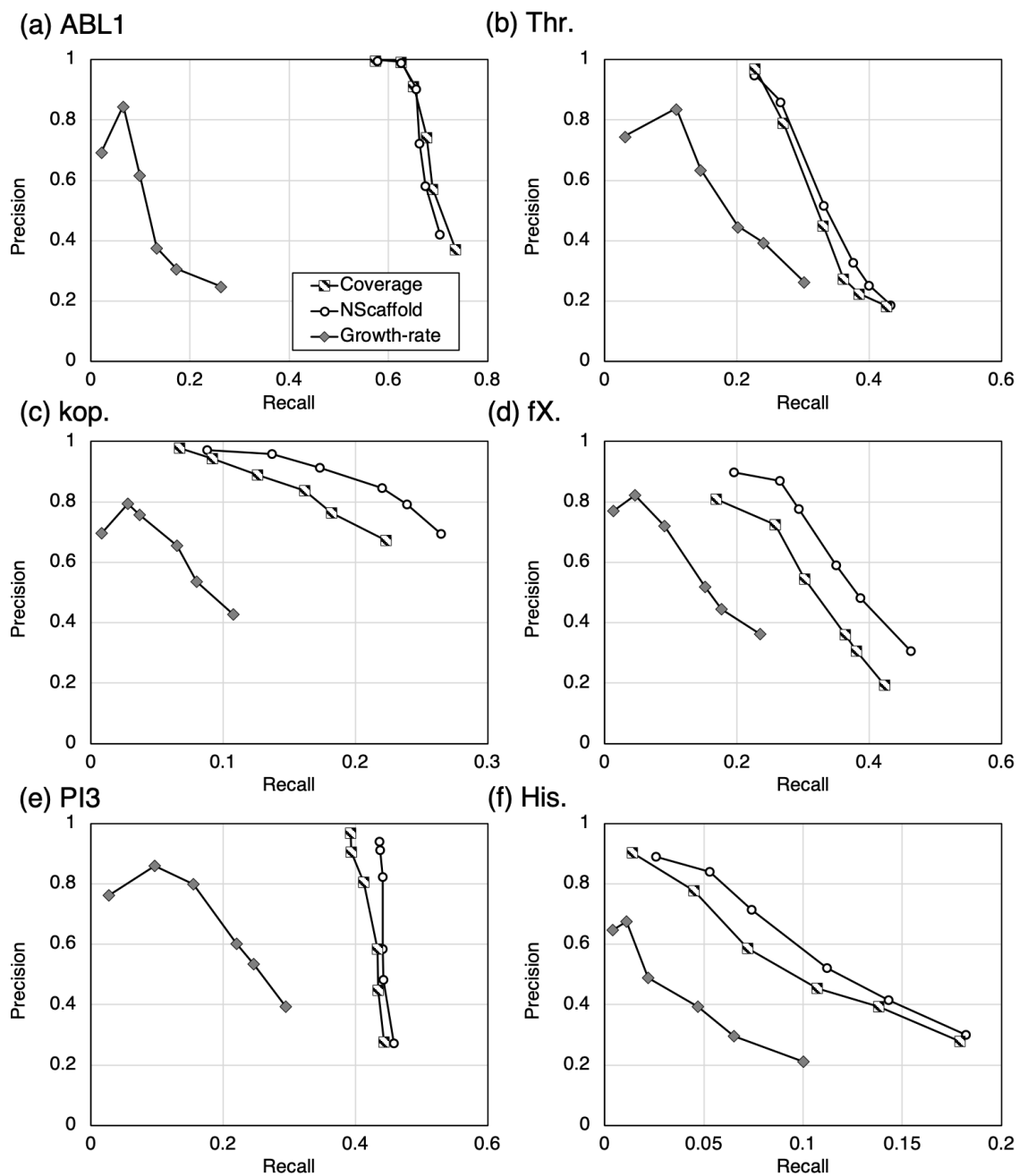


図 2-6 (a)ABL1, (b) Thr., (c) Kop. (d) fX. (e) PI3 (f) His.における, 3つのスコアリング手法の VS (トレーニングデータ数: テストデータ数= 1:1)¹[4]

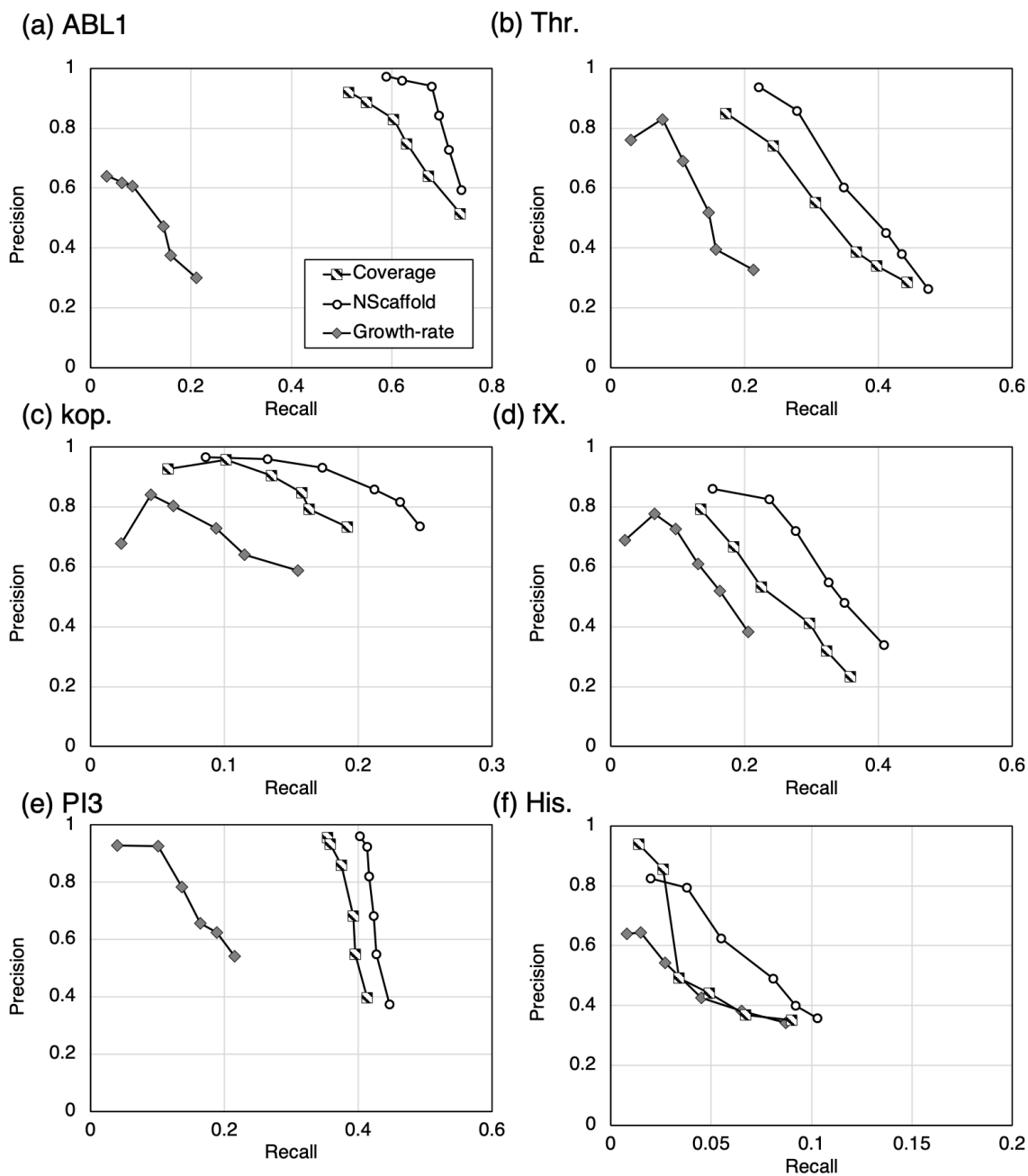


図 2-7 (a)ABL1, (b) Thr., (c) Kop. (d) fX. (e) PI3 (f) His.における, 3つのスコアリング手法の VS (トレーニングデータ数: テストデータ数= 1:4)¹[4]

2.3.3 既存の類似度算出手法との比較

前節までは、PhG どのスコアリング手法の比較であったが、既存の類似度算出手法との比較も実施した。分子表現として 2048 ビットのベクトル形式の Morgan Fingerprint (MFP)を用い[7]，類似度計算の指標として Tanimoto Similarity を用いた。このベンチマーク計算では、トレーニングデータとテストデータの scaffold 数の比が 1:4 (CCR Single) になるように、トレーニングとテストの活性化合物データセットを 1 組ずつランダムに作成した。VS 用の不活性化合物として、6 つのターゲットそれぞれについて、ZINC データセットまたは不活性化合物データセットのいずれかを用意した。

表 2-2 は、*NScaffold*でスコアリングした PhG の最も優先度の高い PhG を使用した場合(N=1 の場合)の、各ターゲットに関する Precision と Recall を示す。また、比較として、PhG と同じリコール値での MFP の Precision を表す。*NScaffold*法で選択された最良の PhG を使用した場合、従来の MFP を使用した場合と同等の精度値が得られた (表 2-2a)。一方で、しかし、ChEMBL の不活性化合物と活性化合物からなるデータセットでは、最適な PhG をクエリとして使用した場合が、MFP よりも高い Precision を示した (表 2-2b)。

表 2-2 PhG と従来の MFP を用いた SH 性能の比較

(a) ZINC の 250,000 個の化合物を不活性化合物として用いた場合

target	recall	precision	
		PhG	MFP
ABL1	0.66	1.00	1.00
Thr.	0.25	0.99	1.00
Kop.	0.10	1.00	0.99
fX.	0.05	0.91	1.00
PI3	0.51	1.00	1.00
His.	0.01	1.00	1.00

(b) ChEMBL の不活性化合物を用いた場合

target	recall	precision	
		PhG	MFP
ABL1	0.66	1.00	0.94
Thr.	0.25	0.68	0.74
Kop.	0.10	1.00	0.84
fX.	0.05	1.00	0.83
PI3	0.51	1.00	0.92
His.	0.01	1.00	0.96

2.4.4 抽出された SPhG に関する考察

PhG 用いれば, VS のクエリとして SH が可能となるだけでなく, 活性化化合物に共通して存在する特徴を解釈することもできる. 図 2-8 に, *NScaffold*法 (図 2-8a) と *Coverage*法 (図 2-8b) を用いて得た PhG の例を示した. Thr.の活性化化合物を, CCR Single に基づき, データセットを 1:4 に分割して得た上位の PhG であり. 図中には, トレーニングデータに含まれる化合物 (左) とテストデータに含まれる化合物 (右) の例も示した.

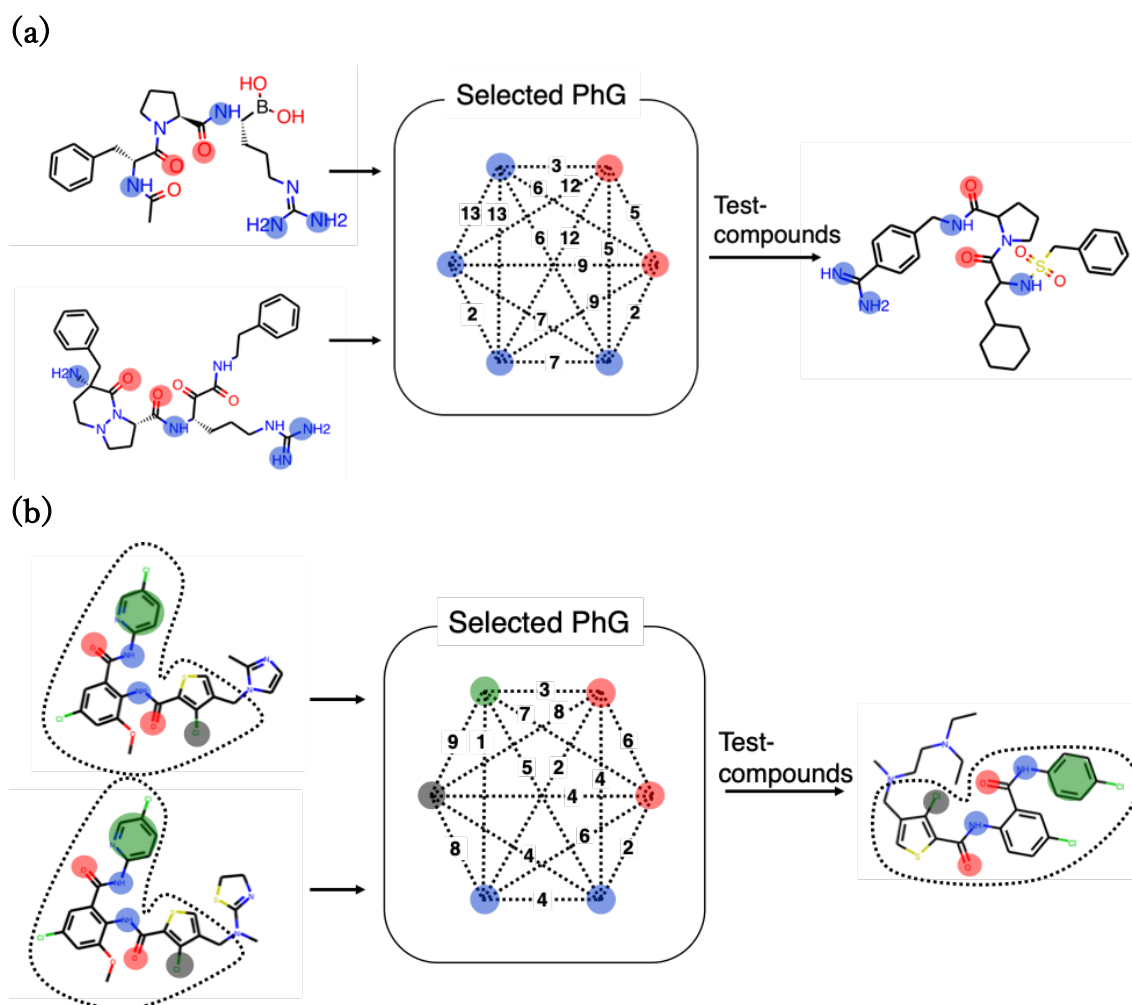


図 2-8 スコアリング手法の違いで得られる PhG の違いと対応化合物の例¹[4]

(a) *NScaffold* 法 (b) *Coverage* 法

図 2-8a に示す *NScaffold* 法の PhG の 6 つの PF (4 つの水素結合ドナーと 2 つのアクセプター) は, 既に構造解析によって知られている Thr. の結合点と一致した. 構造に基づいた PF の解釈によると, 距離 2 の水素結合ドナーは, トロンビン阻害剤のアルギニン残基に対応し, Asp189 のカルボニルオキシゲンと相互

作用することがわかっている [21-25]. また, アルギニン残基の NH₂ から 7 トポロジカル距離離れた HBD とさらに 2 トポロジカル距離離れた HBA のペアは, Ser214 の残基と相互作用し [21], 他のカルボニル部分は Gly216 と相互作用することが報告されている [22], また, これらの特徴は多くの高活性のトロンビン阻害剤に共通することが証明されている [21-25].

一方, *Coverage* 法では, ターゲットーリガンド相互作用に共通する特徴を抽出することができず, 限られた scaffold に属する化合物の特徴を抽出する傾向がある. 図 2-6b に示す PhG は, 1 つの芳香環, 1 つの親油性部分, 2 つの水素結合のドナーとアクセプターを含んでいる. この PhG は, 対応する scaffold は, 少ないが, 1 つの scaffold に対して多くの化合物が属しているため, *Coverage* 法では, 優先度が高く評価されていた. 図 2-6b 内のトレーニング化合物(左側)とテスト化合物(右側)が部分構造を共有していることからそれが裏付けられた. また, 図 2-6b の PhG は, 2 つの PF が水素結合に関係していなかった. 1 つは通常弱い相互作用をもたらす芳香環であり, もう 1 つはエントロピー効果に加えて分散的な相互作用をもたらす疎水性の PF であった [26,27].

もっとも, 本章で扱う PhG は完全グラフであり, 頂点の数に対して辺の数が多い. 分子構造を原子を頂点とし, 化学結合を辺とするグラフとみなすと頂点数に対して辺の数が少ないスパースなグラフとなっている. PhG はここで示したとおり, 一定の解釈可能性はあるが, 改善の余地がある. 次章ではこの点を主題として扱う.

最後に, 3 つのスコアリング手法で得られたテストデータの scaffold ごとの分布を図 2-9 に示した.

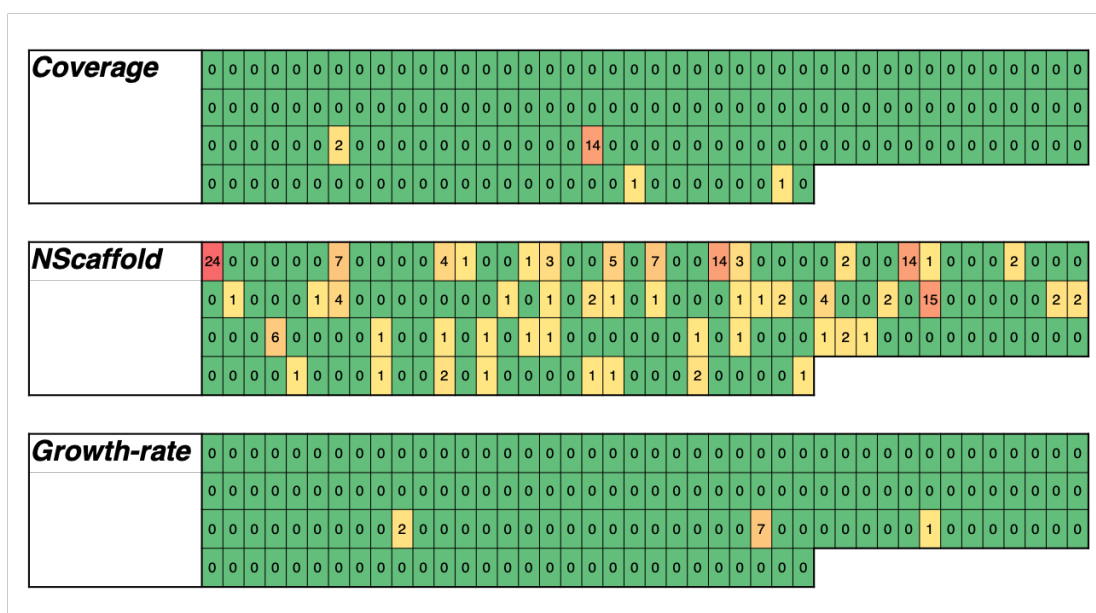


図 2-9 PhG のカバーする scaffold の化合物数のヒートマップ¹ [4]

同図では、図 2-8 と同じ Thr. のデータセットから得られた優先順位の高い 10 個の PhG をクエリとして用いた。各セルはテストデータの各 scaffold に相当し、ヒットした化合物数をセル内の数字と色で表している。テストデータの scaffold の数は 155、化合物の数は 512 であった。Coverage 法で同定された PhG は、トレーニングデータの活性化合物の 31.8% をカバーしていたが、テストデータ内の 18 の活性化合物にマッチする 4 つの scaffold しか同定できなかった。Growth-rate 法で決定された PhG は、トレーニングデータの活性化合物の 25.0% をカバーしたが、テストデータにおいて 10 個の活性化合物にマッチして 3 個の scaffold しか識別できなかった。一方、NScaffold 法で決定された PhG は、トレーニング活性化合物のカバー率は 22.7% であるが、テストデータにある 49 個の scaffold を検出できた。つまり、NScaffold 法で得た最も高い優先順位の PhG は、scaffold に依存せずターゲット-リガンド相互作用に共通する本質的な特徴を取り出すことに成功している。

2.5 結論

本章では、トレーニングデータとは異なる scaffold を持つ化合物の活性を予測する SH 手法の提案を目的として、ファーマコフォアをグラフで表現した PhG を、SH 向けにスコアリングする新たな手法 NScaffold 法を提案した。NScaffold 法は、活性化合物の数ではなく、scaffold 数で PhG を優先順位付けする手法である。6 つのターゲットに対して、従来の Coverage 法や Growth-rate 法に加え NScaffold 法で PhG を抽出し、その PhG クエリとして VS を実施した。その結果、本手法は、従来法に比べて、Precision-Recall 評価により、より高い SH 性能を示した。また、創薬の初期段階を想定して実施したトレーニングデータの scaffold 数が少ない場合については、NScaffold 法は他の方法よりも Precision-Recall 評価の観点から優れていた。Thr. を例にとると、提案した手法は、残基 Asp189, Ser214, および Gly216 に結合する PF とその位置関係を示す PhG が得られることがわかった。この結果は、NScaffold 法によって選択された PhG が、scaffold に依存せず、かつ解釈性のあるファーマコフォアを導出できることを示している。この方法は非常にシンプルで、余分なパラメータは必要無い。NScaffold 法を用いて得た PhG を用いれば、今後の創薬プロジェクトで、SH の成功例がさらに増えると期待する。

2.6 参考文献

1. Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. "Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chemie - Int. Ed.* **1999**, *38*, 2894–2896.
2. Bonachéra, F.; Parent, B.; Barbosa, F.; Froloff, N.; Horvath, D. Fuzzy Tricentric Pharmacophore Fingerprints. 1. Topological Fuzzy Pharmacophore Triplets and Adapted Molecular Similarity Scoring Schemes. *J. Chem. Inf. Model.* **2006**, *46*, 2457–2477.
3. Métivier, J. P.; Cuissart, B.; Bureau, R.; Lepailleur, A. The Pharmacophore Network: A Computational Method for Exploring Structure-Activity Relationships from a Large Chemical Data Set. *J. Med. Chem.* **2018**, *61*, 3551–3564.
4. Nakano, H.; Miyao, T.; Funatsu, K. Exploring Topological Pharmacophore Graphs for Scaffold Hopping. *J. Chem. Inf. Model.* **2020**, *60*, 2073–2081.
5. Daylight Chemical Information Systems, Inc. Daylight Theory Manual. <http://www.daylight.com/dayhtml/doc/theory/index.html> (accessed Dec 28, 2019)
6. Ash, S.; Cline, M. A.; Homer, R. W.; Hurst, T.; Smith, G. B. SYBYL Line Notation (SLN): A Versatile Language for Chemical Structure Representation. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 71–79.
7. Landrum, G., RDKit: open-source cheminformatics software <http://www.rdkit.org> (accessed Dec 28, 2019)
8. <https://github.com/rdkit/rdkit/blob/master/Data/BaseFeatures.fdef> (accessed Dec 28, 2019)
9. Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, Mutowo P, Atkinson F, Bellis LJ, Cibrián-Uhalte E, Davies M, Dedman N, Karlsson A, Magariños MP, Overington JP, Papadatos G, Smit I, Leach AR. 'The ChEMBL database in 2017.' *Nucleic Acids Res.*, **2017**, *45*, D945–D954.
10. Sterling, T.; Irwin, J. J. ZINC 15 – Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324–2337.
11. Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.
12. Naveja, J. J.; Vogt, M.; Stumpfe, D.; Medina-franco, J. L.; Jesu, J.; Bajorath, J. Systematic Extraction of Analogue Series from Large Compound Collections

Using a New Computational Compound – Core Relationship Method. *ACS Omega* **2019**, *4*, 1027–1032.

13. Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP–Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511–522.
14. Royle, N. J.; Irwin, D. M.; Koschinsky, M. L.; MacGillivray, R. T. A.; Hamerton, J. L.; Human genes encoding prothrombin and ceruloplasmin map to 11p11–q12 and 3q21–24, respectively. *Somat Cell Mol Genet*, **1987**, *13*, 285–292.
15. 小西典子; 廣江克彦; 川村正起; Factor Xa(FXa)阻害薬の次世代経口抗凝固薬としての可能性 –前臨床成績から見た将来展望–, *日薬理誌*, **2010**, *136*, 88-92.
16. Shah, N. P.; Tran, C.; Lee, F. Y.; Chen, P.; Norris, D.; Sawyers, C. L. Overriding Imatinib Resistance with a Novel ABL Kinase Inhibitor. *Science* **2004**, *305*, 399–401.
17. Samuels, Y.; Wang, Z.; Bardelli, A.; Silliman, N.; Ptak, J.; Szabo, S.; Yan, H.; Gazdar, A.; Powell, S. M.; Riggins, G. J.; et al. High Frequency of Mutations of the PIK3CA Gene in Human Cancers. *Science* **2004**, *304*, 554.
18. Anderson, R. I.; Becker, H. C. Role of the Dynorphin/Kappa Opioid Receptor System in the Motivational Effects of Ethanol. *Alcohol. Clin. Exp. Res.* **2017**, *41*, 1402–1418.
19. Arrang, J.-M.; Garbarg, M.; Schwartz, J.-C. Auto-Inhibition of Brain Histamine Release Mediated by a Novel Class (H3) of Histamine Receptor. *Nature*, **1983**, *302*, 832–837.
20. Gichohi-Wainaina, W. N.; Towers, G. W.; Swinkels, D. W.; Zimmermann, M. B.; Feskens, E. J.; Melse-Boonstra, A. Inter-Ethnic Differences in Genetic Variants within the Transmembrane Protease, Serine 6 (TMPRSS6) Gene Associated with Iron Status Indicators: A Systematic Review with Meta-Analyses. *Genes Nutr.* **2015**, *10*, 1–15.
21. Friedrich, R.; Steinmetzer, T.; Huber, R.; Stürzebecher, J.; Bode, W. The Methyl Group of N^α(Me)Arg-Containing Peptides Disturbs the Active-Site Geometry of Thrombin, Impairing Efficient Cleavage. *J. Mol. Biol.* **2002**, *316*, 869–874.

22. Steinmetzer, T.; Baum, B.; Biela, A.; Klebe, G.; Nowak, G.; Bucha, E. Beyond Heparinization: Design of Highly Potent Thrombin Inhibitors Suitable for Surface Coupling. *ChemMedChem* **2012**, *7*, 1965–1973.
23. Berman, H. M.; Westbrook, J. D.; Feng, Z.; Gilliland, G. L.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
24. Krishnan, R.; Mochalkin, I.; Arni, R.; Tulinsky, A. Structure of Thrombin Complexed with Selective Non-Electrophilic Inhibitors Having Cyclohexyl Moieties at P1. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2000**, *56*, 294–303.
25. Weber, P. C.; Lee, S. L.; Lewandowski, F. A.; Schadt, M. C.; Chang, C. H.; Kettner, C. A. Kinetic and Crystallographic Studies of Thrombin with Ac-(D)Phe-Pro-BoroArg-OH and Its Lysine, Amidine, Homolysine, and Ornithine Analogs. *Biochemistry* **1995**, *34*, 3750–3757.
26. Desiraju, G. R. Hydrogen Bridges in Crystal Engineering: Interactions without Borders. *Acc. Chem. Res.* **2002**, *35*, 565–573.
27. Schneider, G.; Baringhaus, K. H. *Molecular Design Concepts and Applications*; Wiley-VCH, **2008**; pp. 54–57.

第 3 章 解釈可能性を持つ Sparse PhG (SPhG) の提案

3.1 緒言

第 2 章で検討した PhG は、全ての頂点が、他の全ての頂点と辺で結ばれたグラフ、つまり完全グラフである [1]. しかし、化合物に含まれるほとんどの原子が持つ化学結合の数は最大でも 4 であるため、分子構造をグラフとして見ると、頂点数に比べて辺が少ないスパースなグラフとなる. そのため、完全グラフである PhG を、分子構造と同じスパースなグラフに変換できれば、解釈性が高くなると考えた.

分子構造をスパースに保ったまま PF を頂点として抽象化した縮約グラフ (Reduced Graph) を作る方法は、過去に研究例が存在する. Rarey らは、部分構造を PPP の種類に応じて数種類のノードに置き換え、リングをノードに折り畳むことで、分子から生成される「Feature Tree (FT)」とその効率的なマッチングアルゴリズムを提案した [2]. Barker らは、段階的に縮約可能な分子構造の抽象表現を提案した [3]. (以下 Barker の縮約表現を Reduced Graph と呼ぶ.) Stiefl らは、「Extended eeduced Graph (ErG)」と呼ばれる別のタイプの縮約グラフを提案した [4]. ErG では、芳香環の中心に仮想的な原子があると仮定して、その仮想的な原子に芳香環の PF を割り当てることなどによりグラフを分子構造から縮約する. これらの手法は、異なる分子の縮約グラフどうしの類似性を算出することで、分子構造どうしの類似性を直接比較するよりも分子の scaffold に依存しにくい、活性の推論が可能になる [2-4].

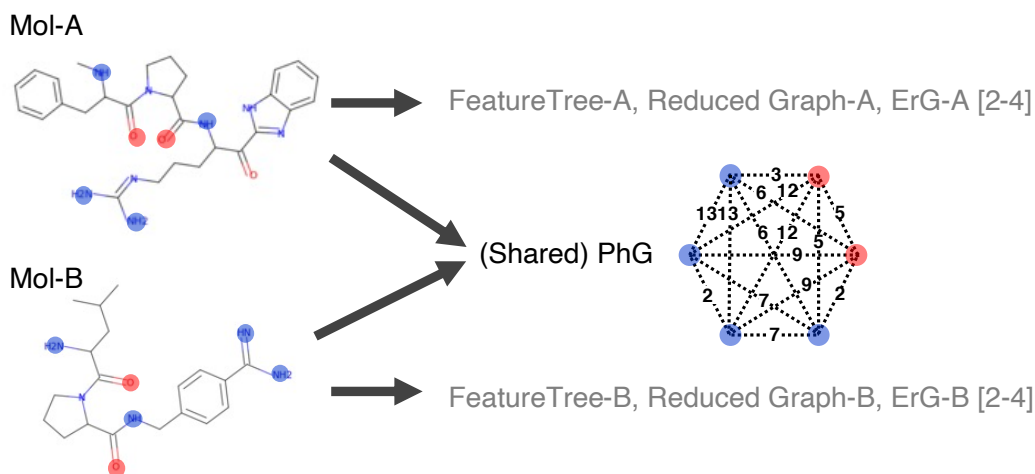


図 3-1 PhG と他の縮約グラフの違い[2-4]

Rarey らの Feature Tree, Barker らの Reduced Graph, および Stiefl らの ErG はすべて化合物それぞれに対応するものであり、複数の化合物に共通する特徴を表す PhG とは異なる。

しかしながら、図 3-1 に示すとおり、上述の既往研究の縮約グラフは、個々の分子を、抽象化したグラフである [1-4]。一方で、(Shared) PhG は、異なる scaffold に属する複数の化合物から抽出している通り、scaffold を超えて活性分子間に共通して存在する特徴・活性を発現するための本質的な特徴を抽出している。実際に、*NScaffold*法で抽出することで、高い SH 性能を実現した。これは PhG が、scaffold を超えて複数の活性分子に共通する特徴を暗に含んでいることを意味する。

そこで、本章の目的を、SH 性能を維持しながら、PhG の解釈性を高めるように改良することとした。本章で提案する Sparse Pharmacophore Graph (SPhG) は、分子構造と同様のスパースなグラフでありながら、完全グラフである PhG と同等の高い SH 性能を実現できる [5]。具体的には、ジャンクションノード (junction node) の導入と、独自のグラフ縮約アルゴリズムを開発することで、トポロジカル距離を可能な限り維持しつつ、PhG をスパースな表現に変換した。

3.2 節では、活性分子のデータセットから SPhG を作成するための詳細なアルゴリズムを説明した。3.3 節では、SPhG を評価するために、ChEMBL [6] と PubChem [7] のデータベースから抽出した約 90 万化合物を用いて、スパース性とトポロジカル距離の再現性に関してベンチマーク計算を行った。さらに、3 つのターゲット (Thr., ABL1, および Kop.) について、PhG と SPhG を用いた VS を実施し、SH の性能を比較した [6]。

3.2 研究方法

3.2.1 PhG と SPhG の違い

SPhG は、前章で説明した PhG をスパース化したものである [1,5]. PhG は PF を表す頂点と、PF 間トポロジカル距離を表す辺で構成される完全なグラフである. 図 3-2a は, Thr. に対して活性を持つ分子を示す. 青と赤でハイライトした原子は, 着目している PF を表し, 各々 HBD と HBA を表す. 図 3-2a に示した PF の組み合わせに着目して PhG を作成すると, 図 3-2b に示した通りの完全グラフとなる. 一方, 同じ PF の組み合わせに着目して作成した SPhG は, 図 3-2c のようになる. ここでジャンクションノードは, 元の化学構造のノード内距離をできる限り維持するために導入されたもので, 3 つ以上のエッジを持つグラフの分岐点の役割を持つ. このジャンクションノードのおかげで, SPhG は, 冗長な辺を削除できるため, PhG (図 3-2b) よりもスパースになり, 解釈性が向上する (図 3-2c). 以下に, この SPhG の作成方法を説明する.

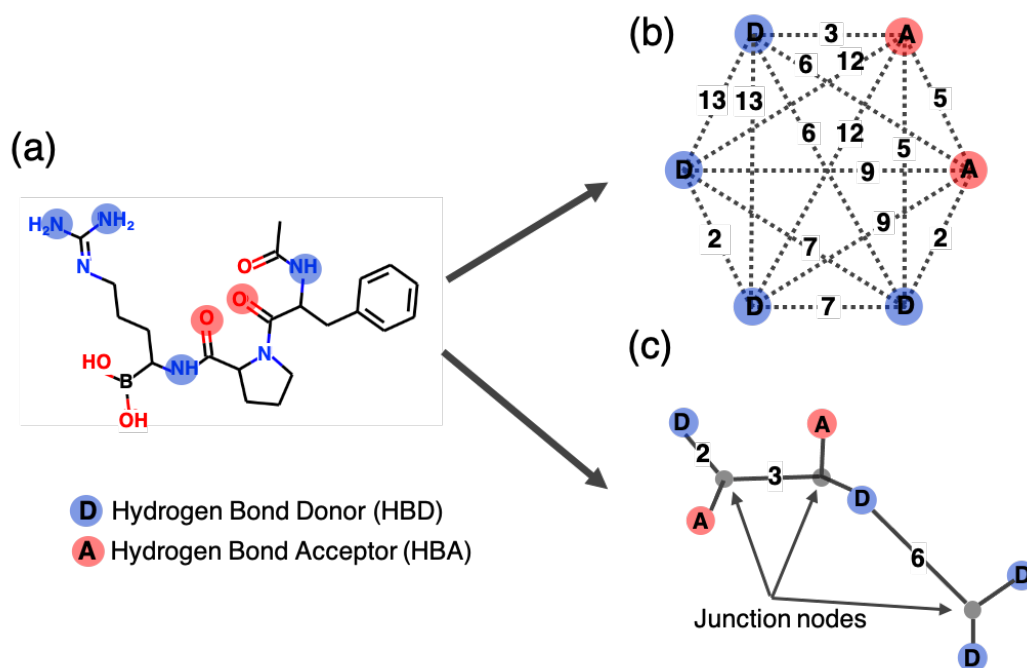


図 3-2 PhG と SPhG の違い² [5]

(a)Thr.の活性化化合物の分子構造 (b) PhG の一例 (c) SPhG の一例

² Reprinted with permission from J. Chem. Inf. Model. 2021, 61, 3348–3360. Copyright 2021 American Chemical Society.

3.2.2 SPhG 作成方法の概要

図 3-3 に用いて段階的に SPhG を作成する方法の概要を示す．まず，PF を割り当てられる PPP を全て検出する．図 3-3a の化合物は，6 つの HBD，4 つの HBA，2 つの PI，4 つの AR を持つ．全化学結合は，種類に関係なく同じ辺に変換される．次に，すべての PPP を頂点として残したまま，グラフを「Mol-SPhG」と呼ばれる縮約グラフに変換する（図 3-3b）．この Mol-SPhG は，図 3-1 で示した FT などと同様に各化合物に対応する．その次に，あらかじめ定義された数の PF を選択し，残りの頂点を図 3-3c および図 3-3d の黒丸で示す非 PF 頂点に変換する．この段階で，あらかじめ決められた数の PF の組み合わせを，網羅的に選択するが，図 3-3 には 2 つを例示している．本章は，第 2 章との整合性をとり，また組み合わせ爆発を避けるため，選択する PF の数を 6 つにした [1]．最後のステップでは，得られたグラフに，別のグラフ縮約アルゴリズムが適用され（図 3-3e，3-3f）．芳香環の PF を持つ頂点を芳香族結合に変換し，ジャンクションノート（図 3-3 内の灰色の丸）を用いて，トポロジカル距離を保存しながら冗長な辺を除去することで SPhG の候補を得る（図 3-3f）．この後，前章で提案した *NScaffold* 手法，比較として用いた *Coverage* 手法，及び *Growth-Rate* 手法を用いて，SPhG をスコアリングする．

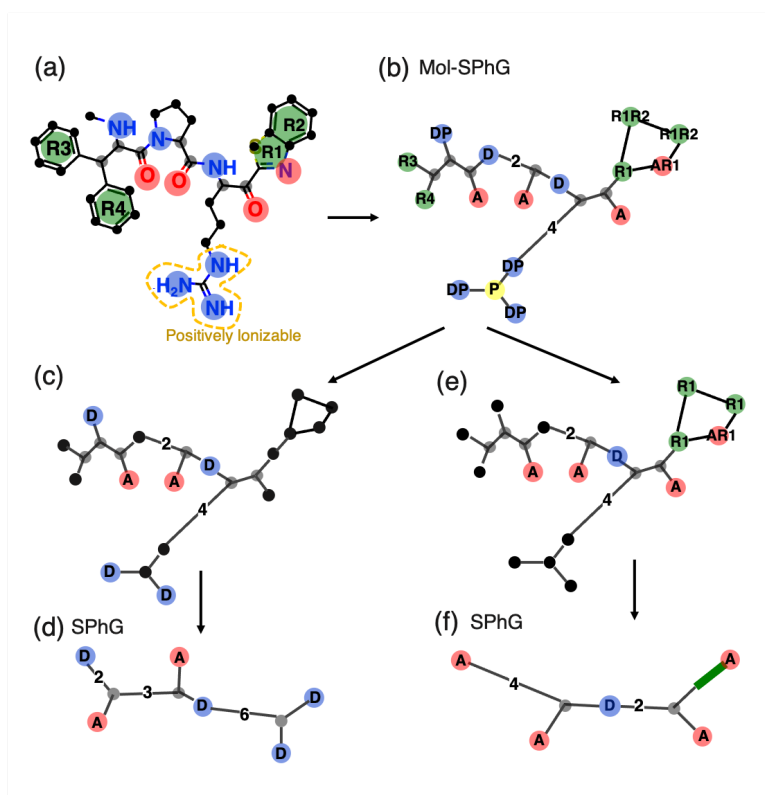


図 3-3 SPhG 作成アルゴリズムの概要² [5]

(a) PPP の割り当て (b) Mol-SPhG (c)(e) 選択された PPP (d)(f) SPhG

3.2.3 PF の検出方法

PF の検出は、前章で説明した方法と同じ方法を用いた。一般的に、PF の定義には SMARTS [8] や SLN [9] などの線形表記を用いた手法が一般的に用いられるが、本章でも RDKit の実装である "BaseFeatures.fdef" というファイルを採用した [10,11]。SPhG については、PF の中で、HBD, HBA, PI, NI, AR を採用した。また 1 つの PF は、複数の原子の集合に割り当てることができる。例えば、芳香環の特徴は通常 5~6 個の原子に割り当てられる。図 3-3a に示すグアジニウムでは、正にイオン化可能な特徴は 3 つの窒素原子と中心の炭素原子に割り当てられる。このとき、PF に対応する PPP 間のトポロジカル距離が複数存在するときがあるが、その場合、その中の最短距離を、「PF 間トポロジカル距離」として定義した。

3.2.4 SPhG 作成アルゴリズムの詳細

図 3-4 に、SPhG の作成アルゴリズムのフローチャートを示す。全体のプロセスは、図 3-3a から図 3-3b に対応する化学構造からの Mol-SPhG の生成と、図 3-3b から図 3-3d または f のプロセスに対応する Mol-SPhG からの SPhG の生成の 2 つの部分に分けられる。前者のアルゴリズムを図 3-4a に、後者のアルゴリズムを図 3-4b に示す。

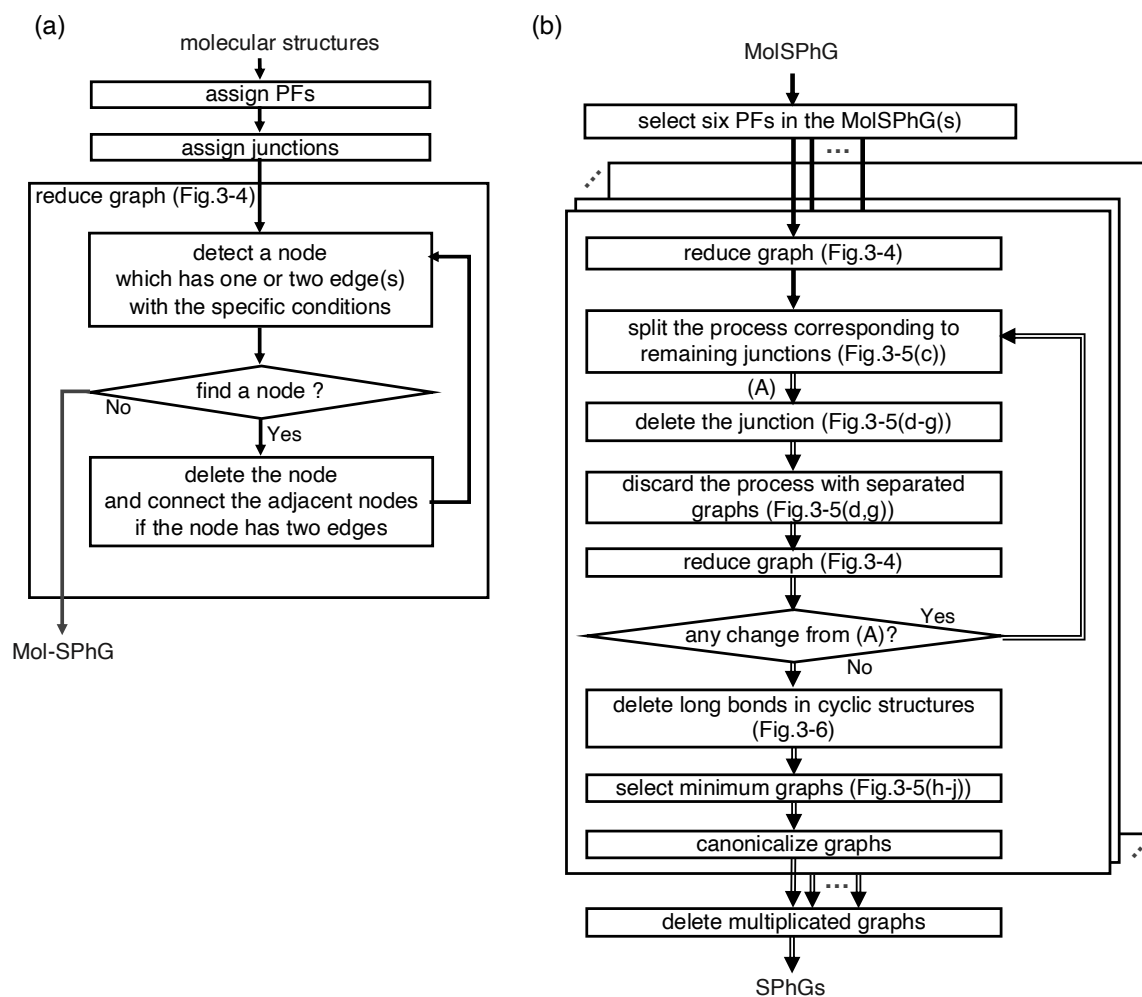


図 3-4 SPhG 作成アルゴリズムのフローチャート² [5]

(a)は、分子構造を Mol-SPhG に変換する手順を示す。このプロセスでは、元の分子と同じトポロジカル距離関係を維持しながら、図 3-5 に説明する「reduce graph」という手法で、2つ以下のエッジを持つノードを削除する。(b)は、Mol-SPhG を、6つのPFを持つSPhGに変換する手順を示す。(b)の二重線付きの矢印は、そのステップを処理した結果、複数のグラフに分岐する可能性があることを意味する。

分子構造から Mol-SPhG の作成

Mol-SPhG は、化合物が持つ全ての PF に対応する頂点を含む縮約グラフである。一方、SPhG は、これらの PF の一部のみを含み、PF 間トポロジカル距離をできるだけ維持する最低限の辺や頂点で構成されている。Mol-SPhG を作成するためには、分子構造中の原子全てを、少なくとも 1 つの PF に対応する頂点か、PF に対応しない頂点(非 PF 頂点)に変換する。これらの処理に続いて、PF やジャンクションノード(=3 つ以上の辺を持つ頂点)以外の頂点を削除したり結合したりすることで、グラフをよりシンプルなものに繰り返し改良する「reduce graph」処理が行われる。この「reduce graph」という処理を、図 3-5 を用いて次に説明する。

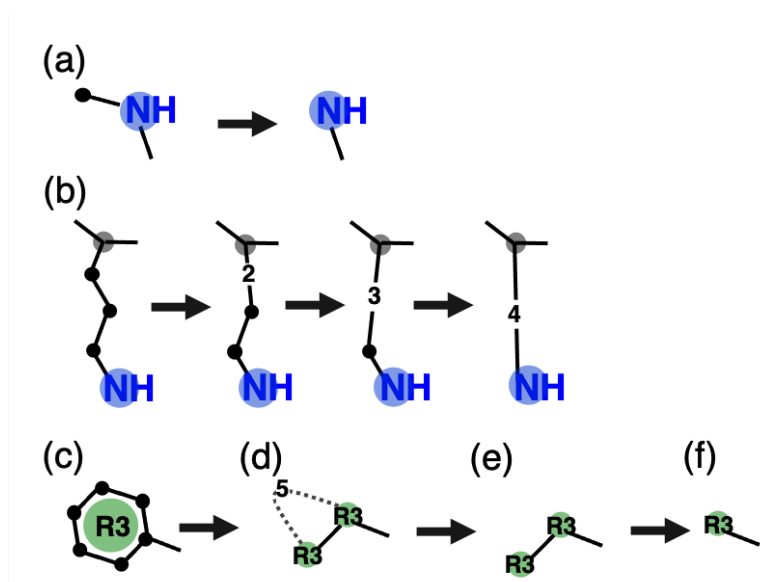


図 3-5 reduce graph アルゴリズム² [5]

(a) 1 辺と接続する頂点 (b) 2 辺と接続する頂点 (c)-(f) 端にある芳香環の簡略化

reduce graph プロセスは、非 PF 頂点と芳香環の PF を持つ頂点を削減するための操作である。非 PF 頂点は PF が割り当てられていないノードであるが、図 3-5a に示すように、グラフ上の末端にある場合は、PF 間のトポロジカル距離には無関係であるので削除する (図 3-5a)。同様に、2 つのエッジを持つノードは、隣接するノードを直接接続して削除する。これらの隣接ノード間の距離は、削除前の隣接ノード間のトポロジカル距離の合計のトポロジカル距離を持つ。図 3-5b では、3 つの 2 つのエッジを持つノードが削除され、灰色のジャンクションノードと NH (HBD) が 4 つの長さのエッジで直接接続されている。

また、末端の芳香環は次のようにして 1 つのノードに統合される (図 3-5c)。図 3-5b で説明した、2 つのエッジを持つノードの削除ルールを、同じ PF 内の

頂点に対して適用する．すると，長さの異なる 2 つのエッジで直接接続された 2 つの隣接ノードが現れる (図 3-5d)．隣接する 2 つのノードは 1 つのエッジでしか接続できないため，最も短い長さのエッジのみを維持する (図 3-5e)．図 3-5e に存在する末端の頂点は，同じ PF「R3」が割り当てられている頂点に統合される (図 3-5f)．(なお，アゾ化合物の窒素原子のように，同じ種類の PF を持つ 2 つの隣接する頂点は，異なる PF が割り当てられているため，統合されない．) 結果的に，図 3-5f に示す，「R3」というラベルで示した芳香環の PF が割り当てられた頂点が 1 つ生き残る．「reduce graph」プロセスでは，ここで示した通り，頂点を削除する操作を繰り返し適用して，これ以上削除できなくなった段階で，最終的なグラフを Mol-SPhG として出力する．

Mol-SPhG から SPhG の作成

Mol-SPhG (図 3-4b) から SPhG を作る後半部分を図 3-6 を用いて説明する．

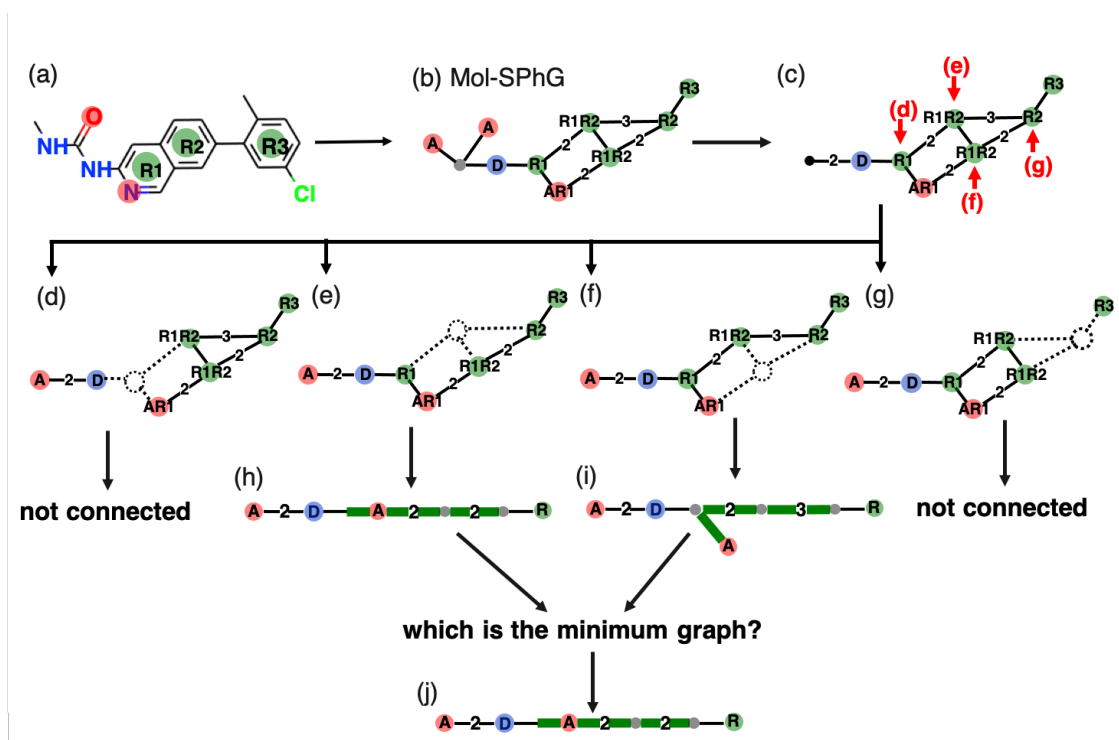


図 3-6 Mol-SPhG から SPhG の作成方法² [5]

(a) 分子構造 (b) Mol-SPhG (c) 選択された PPP とジャンクションノード (d)-(g) ジャンクションノードの除去 (h)(i) グラフの単純化 (j) SPhG

図 3-6a に示す分子を Mol-SPhG に変換後 (図 3-6b)，PF の組み合わせを選択する (図 3-6c)．グラフをさらに単純化するために，図 3-6c に赤で(d)～(g)と記

した通りに、ジャンクションノードを検出する。次に、検出された各ジャンクションノードを削除した後（図 3-6d～3-6g）、連結を維持しているグラフのみを残す。図 3-6e と 3-6f のグラフは単一の連結グラフを形成しているが、図 3-6d と 3-6g のグラフは 2 つに分離している。したがって、前者のグラフのみに「reduce graph」処理が適用される。新たに作成されたグラフに、再度 3 つ以上のエッジを持つ頂点が検出された場合は、この頂点を削除して「reduce graph」処理を適用するというプロセスを、グラフの変化が起こらなくなるまで繰り返し適用する。例えば図 3-6f は、芳香環の PF である「R1」が割り当てられた頂点の一つがまだ 3 つ以上の辺を持つ。この頂点を削除すると、グラフは 3 つに分解されるため、図 3-6f はそのまま次のステップに移行する。

ここで、SPhG をよりシンプルで直感的なものにするため、芳香環の特徴を辺ベースの記述に変換する（図 3-6h, 図 3-6i）。同じ芳香環特徴を持つ 2 つの頂点を持つ辺は、芳香環を示す頂点を使わずに、太い線と緑の線に変更する。例えば、図 3-6h では、芳香環の PF である「R1」と「R2」は辺ベースの緑の太線に変更しているが、同じ芳香環の PF である「R3」は 1 つの頂点しか持っていないため、頂点ベースの表現のままである。なお、この処理は、グラフの可視化表現上の変更であり、グラフ構造には変化がないため、図 3-4 では省略している。

次に、必要に応じて、図 3-7a に示すような複合環状構造（図 3-7a）とマクロ環状構造（図 3-7b）に対する処理を行う。

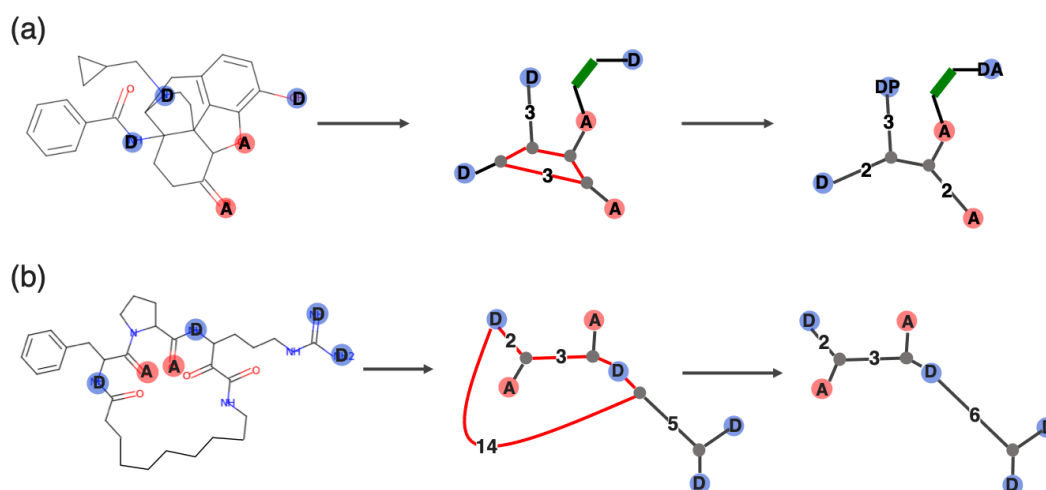


図 3-7 SPhG のループ構造の除去方法² [5]

(a) 複合環状構造 (b) マクロ環構造

図 3-4b のフローチャート内の、「delete long bonds in cyclic structures」に相当するが、図 3-6 に例として示した分子においては不要である。この処理では、ここまでのプロセスで得たグラフに、「ループ構造があり、かつ、そのループを構

成する最も長いエッジが、他のループのエッジの合計よりも長い」場合、その最も長いエッジを除去する。最後に辺の長さの合計が最小となるグラフが選択される (図 3-6j)。複数のグラフが同じサイズを持つ場合は、複数のグラフが出力される。残ったグラフを、N. Schneider らが提案したアルゴリズムを用いてグラフを正準化する [12]。この正準化されたグラフが SPhG である。

3.2.5 SPhG の評価方法

SPhG を評価する上で重要な指標は、スパース性と PF 間トポロジカル距離の情報である。PF 間トポロジカルな距離が、薬物候補の活性を得るための重要な指標であることが示されている [1, 13-15]。一方、SPhG の解釈性を高めるためには、スパース性が必要である。分子構造は、前述の通りスパースなグラフであるため、スパースであることは、合成学者や創薬の研究者がグラフをモデル化する際に、可能な分子構造を考えるのにも役立つ。

そこで、SPhG がこの 2 つを両立できていることを示すために、SPhG のスパース性と、トポロジカル距離の再現性を、定量的に評価した。まず、SPhG のスパース性は以下の Sparse Index で評価する。

$$\text{Sparse Index}(\text{SPhG}) = \frac{N_E}{N_N - 1}$$

ここで N_E はエッジの数、 N_N はノードの数である。ループの無いグラフである木構造は $N_N - 1$ 個のエッジを持つことが証明されており、言い換えれば、木構造の Sparse Index は常に 1 である。また、PF 間のトポロジカル距離の再現性は、ある SPhG 内の 2 つの PF の最短距離と、同じ条件で作成した PhG における同じ PF 間の距離が、一致する割合によって評価する。つまり各 SPhG 毎に、0 から 1 の値を持ち、1 に近いほど再現性が高いことを意味する。

3.2.6 SPhG を用いた VS

上述の 2 指標に加えて、SPhG の SH 性能評価に関する VS を実施した。前章の PhG と同様に、SPhG は VS のクエリとして使用できる。つまり、優先順位の高い SPhG を 1~100 個程度をクエリとして、そのクエリセットの少なくとも 1 つの SPhG を持つ候補化合物をヒット化合物とみなすことで VS が可能である [1]。なお、このときファジー性は認めず、2 つの SPhG 間の完全一致か否かのみによって活性を評価する。SPhG のスコアリングには、前章と同じ (a) *Coverage*, (b) *NScaffold*, および (c) *Growth-rate* の 3 つのスコアリング法を適用した。

3.2.7 SH 性能の評価指標

SPhG は, PhG と同様に, scaffold に依存しない, 活性に関する本質的特徴を抽出することを目的としている. したがって, SPhG がこの目的を果たしていることは, データセットにない, 新しい scaffold に属する化合物の中から, 活性を持つ化合物を特定する能力, つまり SH 性能によって評価できる. scaffold の定義には前章と同様の BM scaffold, CCR Single, CCR RECAP の 3 つのタイプを用いた [1,16-18]. 対象とするターゲットは, Thr., ABL1, および Kop. の 3 つであり, データセットは, ChEMBL に活性情報が登録されている化合物から抽出した. 前章と同様に, ChEMBL に登録されている各標的の全化合物を, トレーニングデータセットとテストデータセットに分けた. このとき, 2 つのデータセットで共通の scaffold に属する化合物が無く, 比率が 1:4 になるようにランダムに分割を行った [1]. 統計的な妥当性確保のために, 3 つの scaffold のタイプごとに 10 回ずつ, ランダムな分割を実施し, それぞれのデータセットで SH 性能を評価して平均化した.

SH 性能は, 正しく予測された活性化合物の値 (TP), 誤って予測された活性化合物の値 (FN), 誤って予測された不活性化合物の値 (FP) を用いて評価した. ここで, テストデータセットに含まれる不活性化合物は, データセットのサイズを大きくするために, PubChem データベースに不活性として登録されている化合物に置き換えた. Precision ($TP/(TP+FP)$) と recall ($TP/(TP+FN)$) をグラフにプロットした. また, scaffold ベースの recall と precision も算出した. つまり, TP, FP, FN を, 化合物数ではなく, TP, FP, FN の化合物に含まれる scaffold の種類数でそれぞれ計算した.

3.2.8 化合物データセット

ChEMBL ver. 24 から, 登録化合物の数, ターゲットの種類に基づいて, 表 3-1 に示す 3 つの高分子ターゲットを選択した [6]. 前章と同様の方法で, 信頼度の高い pKi の値を抽出した. また, ChEMBL 抽出した化合物のうち, pKi 値が 6.0 以上のものを(高い)活性があるとみなし, pKi 値が 6.0 未満の化合物は, ChEMBL の不活性化合物と同様に不活性とみなした. 前章と異なり, 分子量が 200 以下, 600 以上の化合物は除外した. scaffold タイプによって, BM scaffold では 200 から 1237, CCR Single では 97 から 883, CCR RECAP では 254 から 1506 の範囲の活性化合物が得られた. 不活性化合物としては, PubChem から得た 各ターゲットに関するものをそれぞれ 346990, 211784, 284546 個使用した [7].

表 3-1 化合物データセットのプロファイル

ChEMBL ID ⁴	Target	Code	#CPDs ^a (Active CPDs ^b)	#Inactive CPDs	#Scaffolds		
					Bemis Murko ^d	CCR Single ^e	CCR RECAP ^f
CHEMBL1862	Tyrosine kinase ABL1	ABL1	601 (511)	346,990	200	97	254
CHEMBL204	Thrombin	Thr.	1387 (514)	211,784	785	575	915
CHEMBL237	κ -opioid receptor	Kop.	2516 (1425)	284,546	1237	883	1506

^a CPDs: ChEMBL から取得した化合物数 [6].

^b 活性化化合物: ChEMBL から得た化合物の中で、pK_i が 6 以上のもの.

^c PubChem から得た不活性化化合物[7].

^d Bemis Murko: RdKit により作成した Bemis Murko scaffold [16]

^e CCR Single: Compound core-relationship (CCR)ベースの scaffold [17].

^f CCR RECAP: RECAP(Retrosynthetic Combinatorial Analysis Procedure)ルールにより結合を切断した CCR ベースの scaffold [17,18]

3.3 結果と考察

3.3.1 PhG と SPhG の比較評価の詳細

前章で説明した PhG は完全グラフであるため、PhG が表す特徴を解釈する際には、対応する分子構造が必要となる (図 3-1). 一方、スパースなグラフは、エッジの数が少ないために解釈しやすいが、PhG をスパース形式に変換するためには、PF 間の最短距離をできるだけ保持する必要がある。本章で提案した SPhG は、このトレードオフを打破して、SH 性能とグラフの解釈性という 2 つの要求を可能な限り満たすように設計されている。

まずは、スパース性と、PF 間トポロジカル距離の再現性から SPhG を評価した。データセットとしては、ChEMBL と PubChem の 3 つの異なるターゲットに対する活性化合物と不活性化合物の SPhG を使用した (表 3-1) [6,7]. また、SPhG の中には、分子構造から生成される過程で、PF 間トポロジカル距離が保存されないものがある。そのため、SPhG を用いた SH 性能は、スパース性を獲得する一方で、PhG を用いた場合よりも悪くなることが予想される。そこで、PF トポロジカル距離を再現できた SPhG の比率を算出した。

次に、SH 性能のベンチマーク計算を、3.2.6 節に示した方法で実施した。NScaffold 法を用いてスコアリングして得た SPhG をクエリーとして用いて、SH 性能評価のための VS を行った。比較対象として、前章で説明した完全グラフを用いる PhG をクエリーとして用いる方法も評価した。ChEMBL に登録されている活性化合物と不活性化合物をトレーニングデータセットとテストデータセットに分け、2 つのデータセットで scaffold が共有されないようにした。さらに、トレーニングデータセットとテストデータセットのスcaffold の数の比率を 1:4 にし、このデータ分割を、前章と同じ 3 つタイプの scaffold ごとにランダムに 10 回行い、合計 30 個のデータセット全体について、precision と recall の値を平均した。クエリーの SPhG (または PhG) は、3 つのスコアリング手法 (*Coverage*, *NScaffold*, *Growth-rate*) に基づいてスコアリングして SPhG (または PhG) を選択した、このときクエリサイズを 1 個 (単一グラフ) から 100 個まで設定し、テストデータの化合物に、クエリーの SPhG (または PhG) を 1 つでも含めば、そのテストデータの化合物は活性があると見なした。

また、本章の VS では、テストデータセット内の不活性化合物の数を増やすために、PubChem の不活性化合物を使用した。ここで用いたターゲットの不活性化合物は、表 3-1 の通り 20 万個以上あり、今回の VS は、これらの大量の不活性化合物から数百個から千個の活性化合物を見つけ出す能力を評価していることになる。また、前章の研究と同様に PF 数を 6 に固定した。今回の評価では、

PhG については、前章と同条件になるように、さらに 2 種類の疎水性結合と亜鉛結合の 2 種類の PF を追加している。

3.3.2 スパース性とトポロジカル距離の再現性の評価

最初に SPhG の Sparse Index を比較した。図 3-8a は ChEMBL から取得した活性化合物、図 3-8b は PubChem から取得した不活性化合物に関する SPhG の Sparse Index をボックスプロットで示した。6 個のノードを持つ完全なグラフは 15 個のエッジを持つため、PhG の Sparse Index は常に 3 である。一方、活性化合物の SPhG では、PhG の Sparse Index は、ほとんど例外なく 1.0 から 1.2 の範囲であった (図 3-8a)。これは SPhG がツリー構造 (閉ループのないグラフ) に近いスパースなグラフであったことを意味する。また、PubChem の不活性化合物を用いた場合も同様の傾向が見られた (図 3-8b)。図 3-8a と 3-8b のすべてのボックスプロットでは、中央値が全て 1 になっている。これは、半数以上の SPhG が木構造であったからである (ABL1 で 53.2%, Thr. で 50.0%, Kop. で 52.3%)。

次に、ChEMBL の活性化合物に関するトポロジカル距離の再現性を図 3-8c に、PubChem から得た不活性化合物のトポロジカル距離の再現性を図 3-8d に表す。両図は、各化合物について、その化合物に属する SPhG のうち、全ての PF 間のトポロジカル距離を保存している SPhG の割合を表す。ChEMBL から得た化合物の距離の保存割合の平均は、ABL1, Thr., Kop. でそれぞれ 32.6%, 91.7%, 64.6% であった。このことから、ABL1 と Kop. に対して活性を持つ化合物は、一度 SPhG に変換されると、PPP 間のトポロジカル距離を維持することが難しいことがわかった。これは、各ターゲットに活性を持つ分子は、共通した特徴を持っていることから解釈できる。例えば ABL1 は adenine に類似の、複素芳香環を持つ活性分子が多く、複数の距離の保存が難しい。Thr. は、その機能からもペプチド鎖の類似構造を持つ分子が活性であることが多く、複雑な縮環構造は持ちにくい。Kop. はモルフィネのような立体的な閉ループが多く ABL1 に近く距離の保存しない割合が高い。

しかし、PubChem 不活性化合物については、図 3-8d に示すように、標的間の距離を保持する化合物の数には大きな差がなかった。実際、ABL1 では 79.3% の化合物、Thr. では 79.2% の化合物、Kop. では 79.2% の化合物が PF 間トポロジカル距離を維持していた。

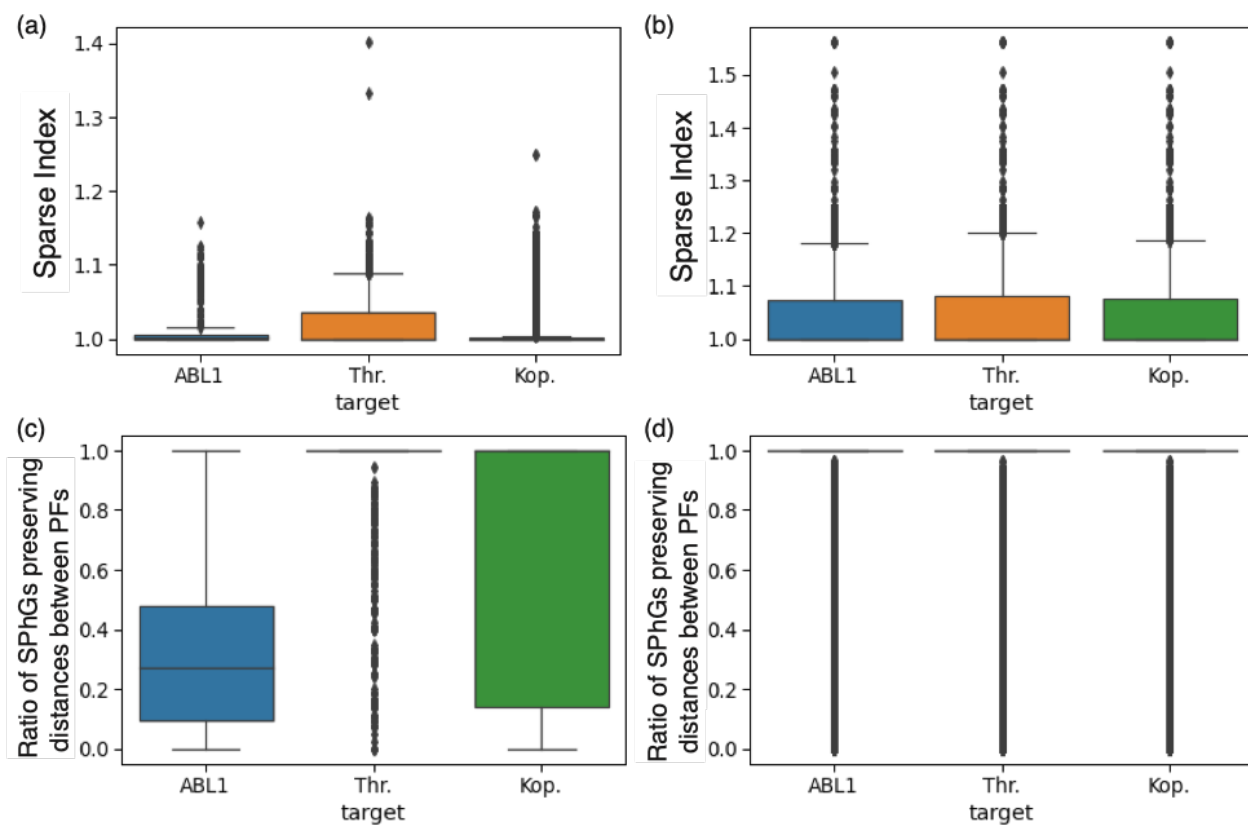


図 3-8 SPhG のスパース性と PF 間トポロジカル距離の再現性² [5]

(a) ChEMBL から得た化合物と (b) PubChem から得た不活性化化合物における Sparse Index のボックスプロット. (c) ChEMBL から得た化合物と (d) PubChem の不活性化化合物における, トポロジカル距離の保存割合のボックスプロット.

このボックスプロットでは, 上位 25% から下位 25% をボックスとして表し, ボックスの中の横線は中央値を, また短い横線は, 中央値からボックスの上下端までの幅の 1.5 倍上下に, ボックスから外れた値を表す.

図 3-9 に、PhG の距離関係を維持できなかった 2 つの SPhG の例を示す。図 3-9a は、ある ABL1 に活性を持つ分子構造と表し、選択済み PF をハイライトしている。赤い太線で描かれているように、芳香環「R1」から HBD「D1」までの最短経路長は 3 である。しかし、SPhG 作成アルゴリズムで生成した SPhG (図 3-9b) では、図 3-9a に灰色の太線で描かれている別のパスが選択されている。その結果、「R1」から「D1」の間のパスの長さは 4 となっている。これは、「A1」という別の PPP が存在することが原因で、すべての PPP を結ぶ辺の数が最小となるグラフを作成するため、敢えて赤いパスを選択せずに灰色のパスを選択しているからである。もう一つの距離が保存されない例は、図 3-8c と 3-8d に示す Kop.における SPhG である。「R1」から「D1」への最短パスは赤い太線で示されているが (パスの長さは 3)、対応する SPhG は、もう一つの PPP である「A1」が存在するため、灰色の線のパスが選択されている。このように、SPhG における PPP 間の距離は、その PF だけでなく、他の選択された全ての PF の組み合わせによって変動する。例えば図 3-8c で D1 を選ばずに他の PPP を選んでいた場合、最短距離を保存していた可能性もある。

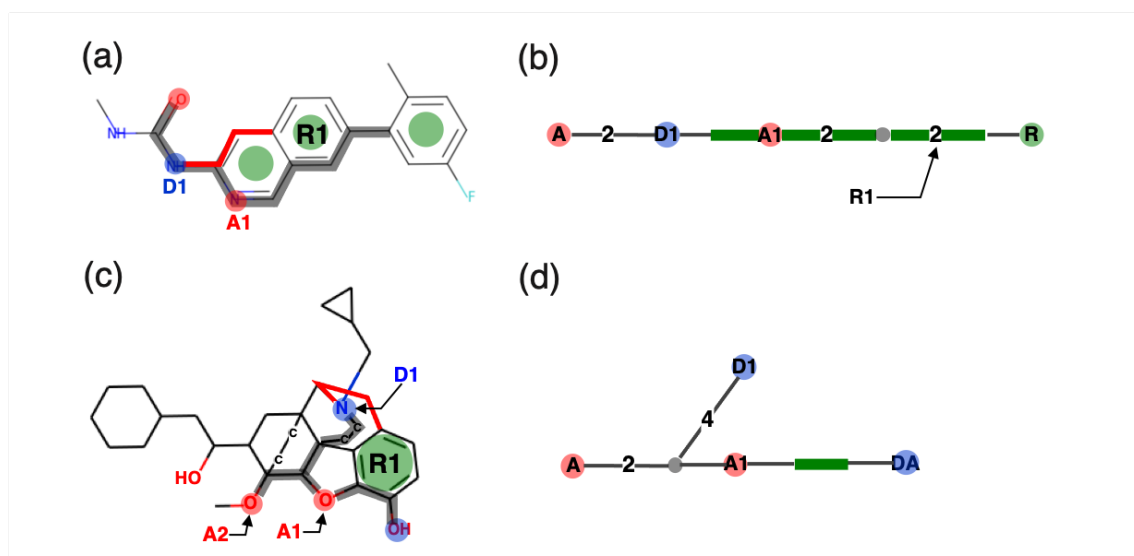


図 3-9 最短距離を保存しない SPhG の例² [5]

(a) ある PPP と別の PPP との間の最短経路を保存しない ABL1 の SPhG の一例。太い灰色の線は、(b)に示す SPhG に対応する経路を表す。赤線で示した芳香環「R1」と HBD「D1」の間の最短経路は、灰色の経路とは異なる。(c) (d) は Kop.での別の例を表す。「R1」と「D1」の間の最短パス (赤のパス) と SPhG のパス (太い灰色のパス) が異なる。

以上の解析結果から、ほとんどの化合物は、最短距離を保ったままよりスパースな SPhG を作成できることが分かった。ただし、6つの PF を繋ぐためのエッジの距離の合計を最小にすることを優先するために、ごく一部の PF 間の最短パスがその SPhG に含まれなくなる場合があることがわかった。

3.3.3 SH 性能の比較結果

図 3-9 は、3つの(S)PhG の優先度評価手法に基づいて、1~100 個選択した (S)PhG の Precision-Recall 曲線を示したものである。

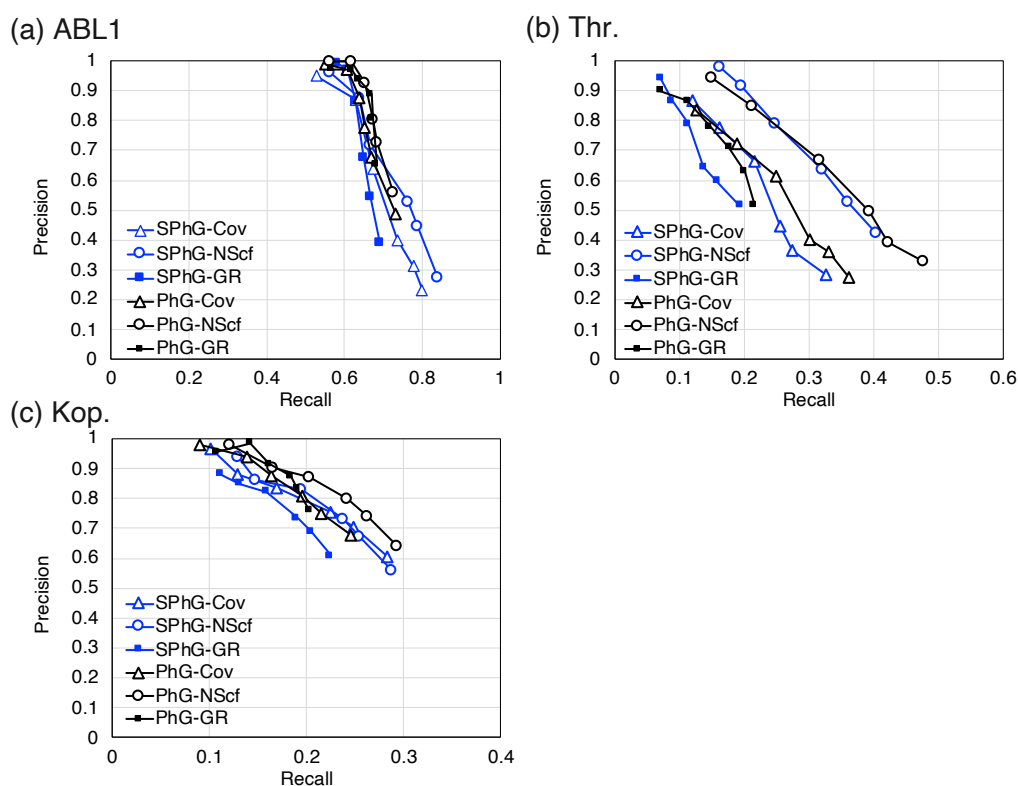


図 3-10 PhG と SPhG の VS による SH 性能比較 (化合物数ベース)² [5]

(a) ABL1, (b) Thr., および (c) Kop.の SH 性能の Precision-Recall 曲線

それぞれの曲線について、Recall が最も低いマーク (四角, 丸, ダイヤモンド) は、SPhG が 1 個 (N=1) に対応し、N の増加につれて Recall 値が増加する。各線上の 6 つのマークは、それぞれ N=1, 3, 10, 30, 50, 100 を示す。SPhG を使用した場合の性能は青で、PhG を使用した場合は黒でプロットした。同図から分かる通り、ほとんどの SPhG は PhG と同等の SH 性能を持つことがわかった。

図 3-10a に示した ABL1 の性能は、N=1, 3, 10, 30 では SPhGs と PhGs の間で同程度であるが、N=50, 100 では精度が約 20% 劣化している。Thr.については、図 3-10b に示すように、*NScaffold* 法で選択された SPhG を使用した場合 (SPhG-NScF) は、PhG(PhG-NScF)とほぼ同等の性能を示す。特に、N=1 または 3 では SPhG が PhG よりも優れていた。Kop.は、PhG と PhG の間の精度の差は 10%以下だった。ABL1 と Kop.についての SPhG と PhG の間のこのような性能差は、図 3-8c に示すように、PPP 間の最短距離を保存する比率が小さいことに起因すると考えている。

図 3-11 は、SH 性能の Precision-Recall 曲線を scaffold 数ベースの値に置き換えたものである。図 3-11a と 3-11b は、それぞれ図 3-10a と図 3-10b に示した化合物ベースの分割を用いた場合と顕著な違いは見られない。図 3-11c に示す Kop.の場合、*Coverage* 法を用いた場合には性能に差はないものの、*NScaffold* 法のクエリ選択を用いた場合には、PhG が SPhG よりも優れた性能を示したこの傾向は図 3-10 でも確認できるが、scaffold 数ベースで表した図 3-11c でより、顕著になっている。なお、より詳細の scaffold のタイプごとの SH 性能比較は、3.6 節の図 3-S1 に示した。

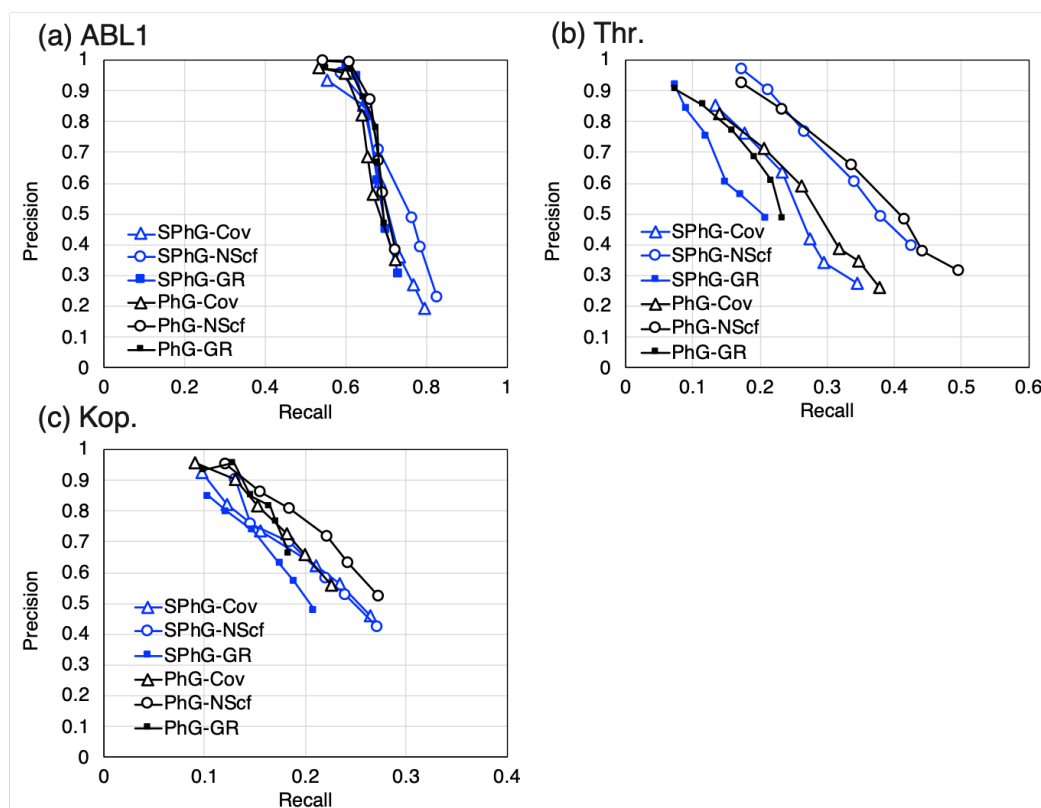


図 3-11 PhG と SPhG の VS による SH 性能比較 (scaffold 数ベース)² [5]

(a) ABL1, (b) Thr., および (c) Kop.の SH 性能の Precision-Recall 曲線

3.3.4 抽出された SPhG に関する考察

ABL1

図 3-12 に、ABL1 の活性分子から得られた SPhG の一例と、その SPhG に該当する活性化合物を示す。この SPhG は、3 つの芳香族結合(緑太線)を持ち、その中に、HBA を含む。これは、2 つ以上の芳香環が縮環していて、その環を構成する原子の少なくとも 1 つが、N 原子などの HBA となっていることを表す。この特徴は、ATP (Adenosine Tri-Phosphates) 自身が持つ縮環構造とも共通する。元々 ABL1 は ATP が結合するたんぱく質であり、その阻害剤として働く活性化合物が類似構造を有することは、薬理的にも正しい。このような、解釈性のある特徴が、自動的に導出できたことは、SPhG の有効性を示す一つの根拠である。

この図 3-12 に示す SPhG の妥当性を評価するために、この SPhG を含む化合物が ABL1 と結合している構造を PDB 用いて探索したが、目的の構造は登録されていなかった。しかし、複数の化合物が、この SPhG の示すとおり、縮環した芳香環を用ち、その構成原子に HBA となる部分を持っていた。そして、その HBA が ABL1 の残基 Phe90 と結合していることを確認できた [19,20]。さらに PDB に登録されている別の構造(2HZI, 1OPK)では、SPhG の HBD に相当する部分が、メチオニン残基と相互作用していることを確認できた[21,22]。

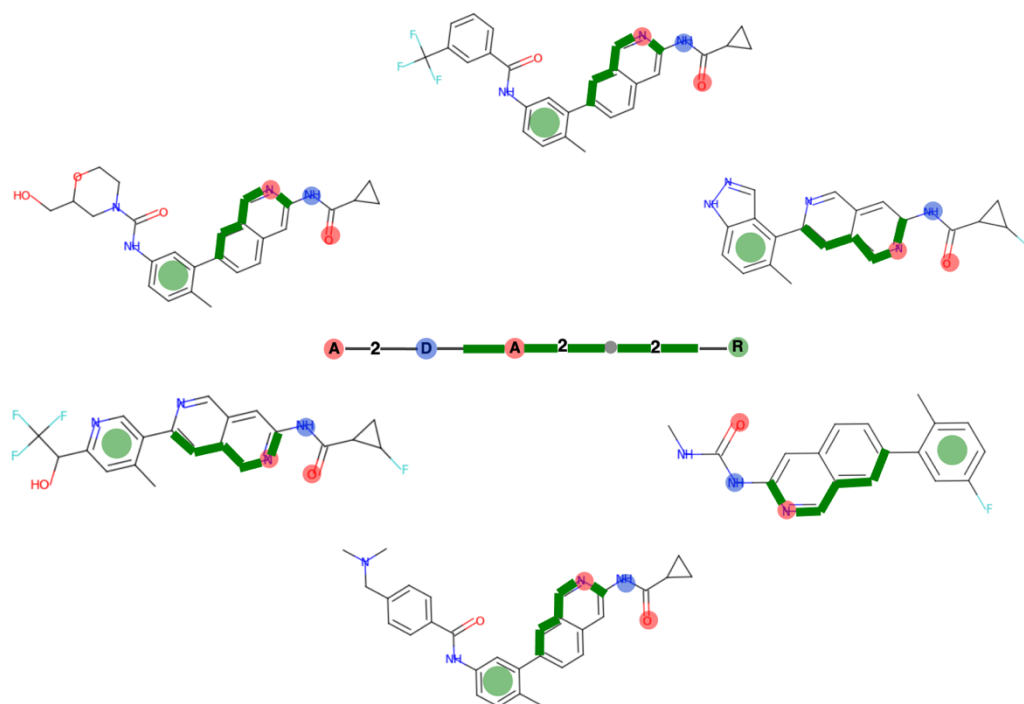


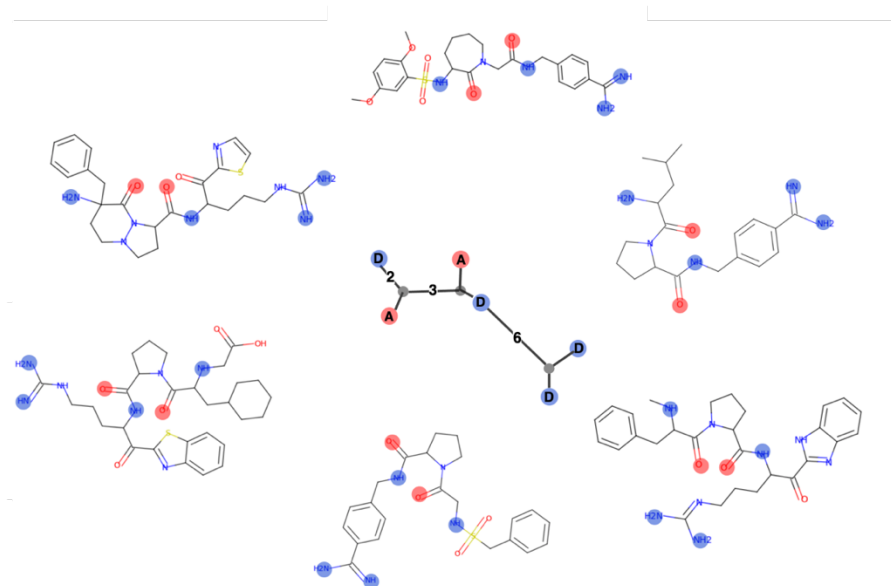
図 3-12 ABL1 に関する SPhG の例² [5]

Thr.

次に、Thr.の SPhG の例を図 3-13 に示す。図 3-13a の SPhG は、*NScaffold* 法を用いたときにスコアが最も高くなることが多い SPhG であり、4 つの HBD と 2 つの HBA を持っている。6 つの PF はすべて、実験的に構造解析により得られたトロンビン阻害剤の結合状態と一致した。2 つの HBD が離れた 1 つの HBD とジャンクションノードを共有しているのは、アルギニンに見られるグアニウム骨格やアミジン骨格に対応している。この部分構造は、Thr.残基 Asp189 と結合することが確認されている [19,23-26]。このようにジャンクションノードは必ず分子内のある原子に相当するようになっている。SPhG は減ったエッジの分だけ情報は減るが、ジャンクションノードもそれに変わる情報を持っている。結果として図 3-10,11 に示したとおり SH 性能は PhG と遜色ない結果となった。さらに、距離 2 の HBD と HBA のペア（アミド基）は、Ser214 と結合する部分である [23]。また、距離 3 の HBD と HBA のペアは、Gly216 と結合する [24]。

図 3-13b は、*NScaffold*法を用いた場合の 2 番目に良い SPhG の例を示している。黒の破線の円で示されているように、この SPhG は、端に HBA がある芳香族結合を持ち、その反対側の端と、ある別の HBA が、お互い距離 1 離れてジャンクションノードを共有している。このようなパターンは、図 3-13a には存在しない。この SPhG に該当する分子を、図 3-14b の周りに 6 つ示した。つまり、この破線で囲まれた構造はヘテロ五員環とカルボニル基を表すことがわかる。PDB [19]を調査したところ、1DOJ や 1A61 などが、Thr.に、これらと同様の特性を持つ分子が結合した構造であることがわかった [27,28]。これらの構造解析結果によると、カルボニル部分が Gly193 と相互作用し、芳香環の窒素原子が Hys57 の残基と相互作用する部分であることがわかった [27,28]。

(a)



(b)

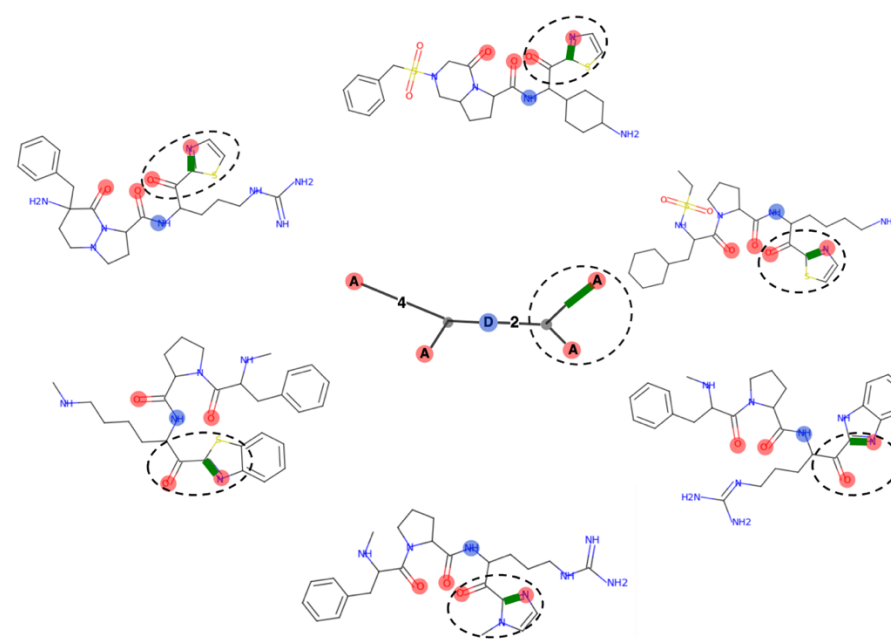


図 3-13 Thr.に関する SPhG の例² [5]

(a) *NScaffold* 法において優先度が最も高くなることが多い SPhG (b) 同様に 2 番目に優先度が高くなることが多い SPhG

Kop.

図 3-14 に、Kop. のデータセットから抽出した SPhG の例と該当する活性化化合物を示す。これらの化合物は、芳香環に隣接する環構造と立体的構造を持つ環を含む複雑な構造を有している。このような構造では、PF から別の PF に、複数の経路を持つことが多く、分子自体がスパースではない。そのため、スパースにするために、エッジを除去する処理を多く取り入れている SPhG では限界がある。そのため、PhG に近いパフォーマンスを得るためには、このようなスパース性が比較的低い分子にも対応できるように SPhG を改良する必要がある。これは今後の研究課題である。

この図 3-14 に示す SPhG を含む化合物が, Kop. と結合している構造を, ABL1・Thr. と同様に PDB で確認したところ, Kop. の残基 Asp198 と SPhG の「DP(HBD / Positively Ionizable)」の PF に相当する部分が相互作用していることがわかった [19,29]. ただ, 他の PF は Kop. と相互作用していることは確認できなかった.

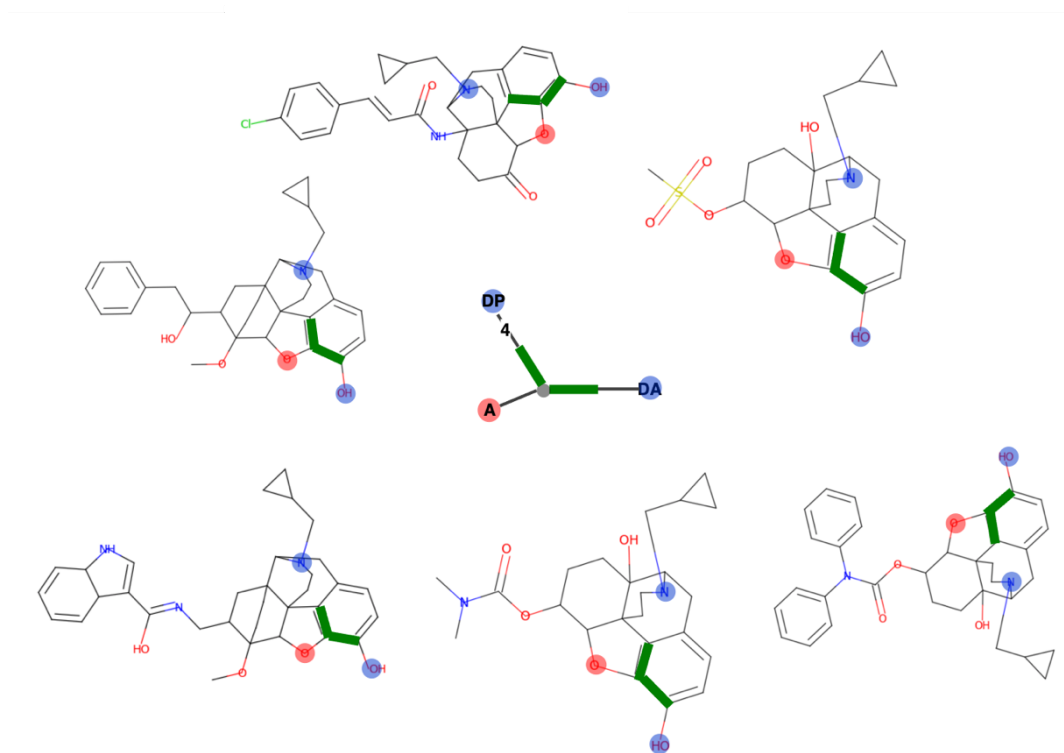


図 3-14 Kop.に関する SPhG の例² [5]

3.4 結論

本章では、既存の活性化合物と異なる scaffold を持つ化合物の活性を、解釈可能な形で予測するための方法として、SPhG を提案した。SPhG の頂点は PF とジャンクションノードで構成されている。また、トポロジカル距離をパラメータとして持つ辺によって繋がっている。完全グラフである PhG とは異なり、分子構造のグラフと同様のスパースなグラフであるため、グラフの解釈性が向上した。実際に、グラフのスパース性と、PF 間トポロジカル距離の再現性も十分高いことを定量的に確認した。

同時に、ABL1, Thr., および Kop. という 3 つのターゲットに対して、ChEMBL と PubChem からそれぞれ得た化合物を用いて、SH の VS 試験を行った。VS の性能は、完全グラフである PhG を用いた場合と同等であった。

これは、3 つ以上のエッジを持ち、特定の PF が割り当てられていない頂点として、ジャンクションノードを導入することで、頂点数を減らしながらも PF 間トポロジカルな距離を保つことができたからである。特に ABL1 や Thr. においては、抽出された SPhG が、構造解析に基づく解釈とも一致することを確認した。

3.5 参考文献

1. Nakano, H.; Miyao, T.; Funatsu, K. Exploring Topological Pharmacophore Graphs for Scaffold Hopping. *J. Chem. Inf. Model.* **2020**, *60*, 2073–2081.
2. Rarey, M.; Dixon, J. S. Feature Trees: A New Molecular Similarity Measure Based on Tree Matching. *J. Comput. Aided. Mol. Des.* **1998**, *12*, 471–490.
3. Barker, E. J.; Buttar, D.; Cosgrove, D. A.; Gardiner, E. J.; Kitts, P.; Willett, P.; Gillet, V. J. Scaffold Hopping Using Clique Detection Applied to Reduced Graphs. *J. Chem. Inf. Model.* **2006**, *46*, 503–511.
4. Stiefl, N.; Watson, I. A.; Baumann, K.; Zaliani, A. ErG: 2D Pharmacophore Descriptions for Scaffold Hopping. *J. Chem. Inf. Model.* **2006**, *46*, 208–220.
5. Nakano, H.; Miyao, T.; Swarit, J.; Funatsu, K. Sparse Topological Pharmacophore Graphs for Interpretable Scaffold Hopping. *J. Chem. Inf. Model.* **2021**, *61*, 3348–3360.
6. Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M.P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. 'The ChEMBL Database in 2017.' *Nucleic Acids Res.* **2017**, *45*, D945–D954.
7. Kim, S.; Chen J.; Cheng T.; Gindulyte A.; He J.; He S.; Li Q.; Shoemaker B.A.; Thiessen P.A.; Yu B.; Zaslavsky L.; Zhang J.; Bolton E. E. PubChem in 2021: New Data Content and Improved Web Interfaces. *Nucleic Acids Res.* **2021**, *49*, D1388–D1395.
8. Daylight Chemical Information Systems, Inc. Daylight Theory Manual. <http://www.daylight.com/dayhtml/doc/theory/index.html> (accessed Dec 28, 2019)
9. Ash, S.; Cline, M. A.; Homer, R. W.; Hurst, T.; Smith, G. B. SYBYL Line Notation (SLN): A Versatile Language for Chemical Structure Representation. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 71–79.
10. Landrum, G., RDKit: Open-Source Cheminformatics Software <http://www.rdkit.org> (accessed Dec 28, 2019)
11. <https://github.com/rdkit/rdkit/blob/master/Data/BaseFeatures.fdef> (accessed Dec 28, 2019)
12. Schneider, N.; Sayle, R. A.; Landrum, G. A. Get Your Atoms in Order-An Open-Source Implementation of a Novel and Robust Molecular Canonicalization Algorithm. *J. Chem. Inf. Model.* **2015**, *55*, 2111–2120.

13. Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. "Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chemie - Int. Ed.* **1999**, *38*, 2894–2896.
14. Bonachéra, F.; Parent, B.; Barbosa, F.; Froloff, N.; Horvath, D. Fuzzy Tricentric Pharmacophore Fingerprints. 1. Topological Fuzzy Pharmacophore Triplets and Adapted Molecular Similarity Scoring Schemes. *J. Chem. Inf. Model.* **2006**, *46*, 2457–2477.
15. Métivier, J. P.; Cuissart, B.; Bureau, R.; Lepailleur, A. The Pharmacophore Network: A Computational Method for Exploring Structure-Activity Relationships from a Large Chemical Data Set. *J. Med. Chem.* **2018**, *61*, 3551–3564.
16. Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.
17. Naveja, J. J.; Vogt, M.; Stumpfe, D.; Medina-franco, J. L.; Jesu, J.; Bajorath, J. Systematic Extraction of Analogue Series from Large Compound Collections Using a New Computational Compound – Core Relationship Method. *ACS Omega* **2019**, *4*, 1027–1032.
18. Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP–Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511–522.
19. Berman, H. M.; Westbrook, J. D.; Feng, Z.; Gilliland, G. L.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
20. Zhou, T.; Parillon, L.; Li, F.; Wang, Y.; Keats, J.; Lamore, S.; Xu, Q.; Shakespeare, W.; Dalgarno, D.; Zhu, X. Crystal Structure of the T315I Mutant of Abl Kinase, *Chem. Biol. Drug. Des.* **2007**, *70*, 171–181
21. Nagar, B.; Hantschel, O.; Young, M.A.; Scheffzek, K.; Veach, D.; Bornmann, W.; Clarkson, B.; Superti-Furga, G.; Kuriyan, J., Structural basis for the autoinhibition of c-Abl tyrosine kinase, *Cell*, **2003**, *112*, 859–871.
22. Cowan-Jacob, S.W.; Fendrich, G.; Floersheimer, A.; Furet, P.; Liebetanz, J.; Rummel, G.; Rheinberger, P.; Centeleghe, M.; Fabbro, D.; Manley, P.W., Structural biology contributions to the discovery of drugs to treat chronic myelogenous leukaemia, *Acta. Crystallogr. D Biol. Crystallogr.* **2007**, *63*, 80–93.

23. Friedrich, R.; Steinmetzer, T.; Huber, R.; Stürzebecher, J.; Bode, W. The Methyl Group of N^α(Me)Arg-Containing Peptides Disturbs the Active-Site Geometry of Thrombin, Impairing Efficient Cleavage. *J. Mol. Biol.* **2002**, *316*, 869–874.
24. Steinmetzer, T.; Baum, B.; Biela, A.; Klebe, G.; Nowak, G.; Bucha, E. Beyond Heparinization: Design of Highly Potent Thrombin Inhibitors Suitable for Surface Coupling. *ChemMedChem* **2012**, *7*, 1965–1973.
25. Krishnan, R.; Mochalkin, I.; Arni, R.; Tulinsky, A. Structure of Thrombin Complexed with Selective Non-Electrophilic Inhibitors Having Cyclohexyl Moieties at P1. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2000**, *56*, 294–303.
26. Weber, P. C.; Lee, S. L.; Lewandowski, F. A.; Schadt, M. C.; Chang, C. H.; Kettner, C. A. Kinetic and Crystallographic Studies of Thrombin with Ac-(D)Phe-Pro-BoroArg-OH and Its Lysine, Amidine, Homolysine, and Ornithine Analogs. *Biochemistry* **1995**, *34*, 3750–3757.
27. Recacha, R.; Costanzo, M. J.; Maryanoff, B. E.; Carson, M.; DeLucas, L.; Chattopadhyay, D. Structure of Human κ -Thrombin Complexed with RWJ-51438 at 1.7 Å: Unusual Perturbation of the 60A-60I Insertion Loop. *Acta Cryst.* **2000**, *D56*, 1395–1400.
28. St Charles, R.; Matthews, J. H.; Zhang, E.; Tulinsky, A. Bound Structures of Novel P3-P1' Beta-Strand Mimetic Inhibitors of Thrombin. *J. Med. Chem.* **1999**, *42*, 1376–1383.
29. Che, T.; Majumdar, S.; Zaidi, S.A.; Kormos, C.; McCorvy, J.D.; Wang, S.; Mosier, P.D.; Uprety, R.; Vardy, E.; Krumm, B.E.; Han, G.W.; Lee, M.Y.; Pardon, E.; Steyaert, J.; Huang, X.P.; Strachan, R.T.; Tribo, A.R.; Pasternak, G.W.; Carroll, I.F.; Stevens, R.C.; Cherezov, V.; Katritch, V.; Wacker, D.; Roth, B.L., Crystal Structure of a nanobody-stabilized active state of the kappa-opioid receptor, *Cell*, **2018**, *172*, 55-67.e15.

3.6 補足

図 3-S1 に scaffold のタイプ別の Precision-Recall 曲線(化合物数ベース)を全て示す. 図 3-S1g, 3-S1h, 3-S1i は, それぞれ Kop. の scaffold のタイプごとの precision-recall 曲線を示す. BM (図 3-S1g) と CCR RECAP (図 3-S1i) の scaffold の場合, PhG は SPhG よりも有意に優れていた.

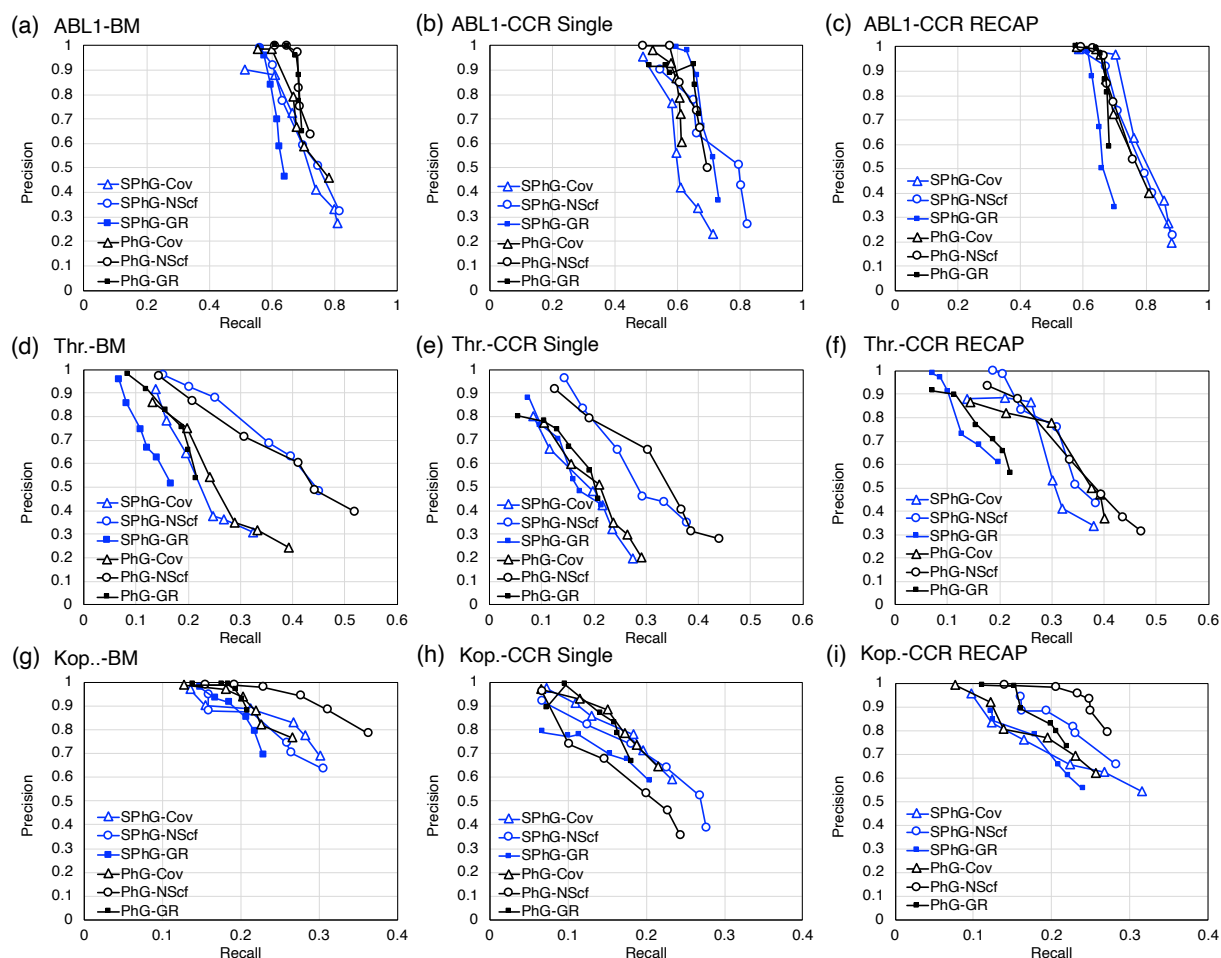


図 3-S1 scaffold タイプ別の SH 性能比較 (化合物数ベース)² [5]

(a)-(c) ABL1, (d)-(f) Thr., および (g-i) Kop. の SH 性能の precision-recall 曲線. 各列は scaffold タイプを表し, 一番右の列から順に, BM, CCR Single, CCR RECAP に対応する.

第 4 章 SPhG のクラスタリング解析

4.1 緒言

第 2 章では、PF とその PF 間の距離を表すグラフ PhG を用いて、SH を実施するための PhG のマイニング法を提案した [1]. 特定の scaffold に依存した PhG ではなく、複数の scaffold に共通して存在する特徴を表現した PhG を優先的に抽出できた. 第 3 章では、PhG を実際の化合物にのモチーフとして捉えられるように、辺を減らしてスパースに改良した SPhG を用いて、解釈性を向上した表現方法を開発した [2]. このマイニング法では、あるターゲットに対する活性化合物のデータセットから、一つの SPhG ではなく複数の SPhG を得ることができる. 実際、第 3 章では、100 個までの SPhG をクエリとして、各 SPhG に該当する化合物をスクリーニングしたところ、トップスコアの SPhG 以外の SPhG も、precision を大きく損なうことなく recall の向上に寄与するものがあることが示された.

しかし、これは、スコアリングにより SPhG を順位付けしたとしても、トップの SPhG だけでは、ターゲットと化合物の相互作用の特徴の全てを表現できていないことを示している. もちろん、前章に示した Thr. の例の通り、トップの SPhG だけでも、X 線構造解析などにより実験的に解明された特徴を抽出して表現できた. しかし、複数の SPhG には、1 つの SPhG 以上の情報が含まれていることは確かである. 一方で、100 個オーダーの重要な SPhG を一度に解釈し、SPhG 群の中から最適な化合物の設計コンセプトを選択することは困難であった.

そこで、本章では、ターゲットとの相互作用に関与する抽出すべき特徴を有する複数の SPhG の全体を捉え、候補化合物の設計指針を得ることができる手法を提案することを目的として、Graph Edit Distance (GED) というグラフの類似度指標に基づいて、複数の SPhG を平面上に配置するクラスタリング解析を実施した.

4.2 研究方法

4.2.1 化合物データセット

ChEMBL ver.24 から 6 つのターゲットに対する活性化合物を抽出した [3]. 選択したターゲットは, Thr., ABL1, Kop., PI3, GPCR, および TPS6 であり, タンパク質の種類を考慮して選択した. 各データセットに含まれる高活性化合物の数を, 略語と ChEMBL ID とともに表 1 に示す. 6 つのターゲットのうち TPS6 を除く 5 つは, pKi が 6 以上の化合物を高活性化合物として抽出した. TPS6 については, 対象となる化合物数が限られていたため pKi の閾値を 5.0 に下げた. 2 章・3 章と同様の前処理を行って, pKi 値を算出し, 分子量が 200~600 の化合物を以降の分析に使用した. TPS6 については, pKi 値とともに全化合物の SMILES (Simplified Molecular-Input Line-Entry System) を, 4.6 節の表 4-S1 に示した.

表 4-1 化合物データセット

ChEMBL ID	Target	Code	#Active CPDs ^a
CHEMBL204	Thrombin	Thr.	514
CHEMBL1862	Tyrosine kinase ABL1	ABL1	511
CHEMBL237	κ -opioid receptor	Kop.	1425
CHEMBL2498	PI3-kinase p110-alpha subunit	PI3	812
CHEMBL5701	G protein-coupled receptor 44	GPCR44	686
CHEMBL1795139	Transmembrane protease serine 6	TPS6	21

^a #ActiveCPDs: pKi 値が 6.0 以上の化合物数. ただし, TPS6 は, 5.0 以上の化合物.

4.2.2 トポロジカル・ファーマコフォアの表現方法

本章では、グラフなどのトポロジカルな情報を用いてファーマコフォアを表現する方法として、従来から存在する Pharmacophore Fingerprint (PhFP)と、前章で説明した Mol-SPhGs, および SPhGs の3種類で比較した [2,4]. PhFPは、各要素が0か1の値を持つビットベクトルであり、Mol-SPhGとSPhGはグラフ表現である。この3つの表現方法について図4-2に示し、以下に詳細に説明する。

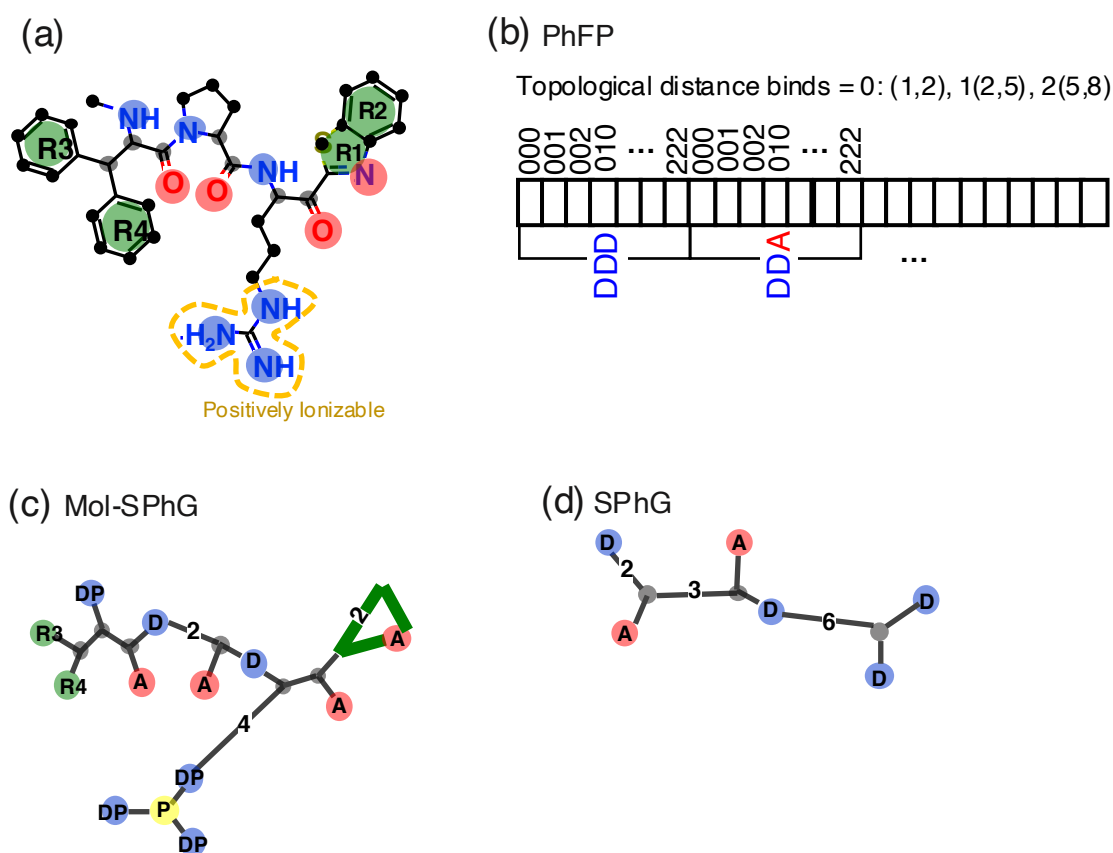


図4-2 トポロジカル・ファーマコフォアの表現方法

(a) PPPがハイライトされたサンプル分子。青色の丸はHBD、赤色の丸はHBA、緑色の丸はARを表す。(b)PhFPの一例。各ボックスは0か1の値しか持たない。箱の中の値が1であれば、2つまたは3つのPFの組み合わせと、それらの間の位相的な距離によって定義される特定のパターンを意味する。距離は3つのカテゴリーに分類される。DDDは、3つのPFが全てHBDの場合、DDAは2つがHBDでもうち1つがHBAの場合を表す。(c) 図4-2aの分子をMol-SPhGに変換したもの。(d) 図4-2cのMol-SPhGを含む高活性化合物のMol-SPhGから第3章のアルゴリズムでマイニングして得られたSPhG。

いずれの表現においても、まず、分子構造からターゲットとの相互作用に寄与し得る部位(PPP)を検出する。今回は、第3章と同様に HBD, HBA, AR, PI, および NI という 5 つの PF を用いた。PF の定義としては、RDKit 内で使用されている "BaseFeatures.fdef" というファイル内の定義を用いた [4,5]。

PhFP

PhFP は、PF とその PF 間トポロジカル距離の情報を、フィンガープリント形式で表す方法である。化合物内に、通常 2-3 個の PF とその間のトポロジカル距離のパターンが存在するか否かを 0 と 1 で表し、そのパターンを、フィンガープリントの 1 ビットに対応させる (図 4-2b)。本研究では、RDKit に実装されている方法をデフォルトのパラメータ値で利用した [4]。PF の数は 3 個まで、距離の最大値は 8 としている。また、距離情報圧縮のため、距離 8 までを 3 つの範囲に分割し、その 3 つのうちのいずれに入るかの情報を 0,1,2 の 3 値で保持する。例えば、ある 3 つの PF を持つパターンは、その PF 間の 3 つトポロジカル距離が各々 3 カテゴリーを持つので 27 個のビットで表現される。図 4-2b に示す 000, 001, 002, 010, ..., 222 のラベルに対応する。3 個の PF の組み合わせとして、全て HBD を保つ場合(DDD)に 27 個、さらに 2 つの HBD と 1 つの HBA の場合に 27 個というように、PhFP は、3 つの PF とそのトポロジカル距離の組み合わせの数だけ、要素を持つビット列である。

Mol-SPhG

図 4-2b に例示した Mol-SPhG は、後述の SPhG を作成する過程に必要な、分子構造の縮約グラフである [2]。Mol-SPhG の頂点は、PF に割り当てられた頂点またはジャンクションノードに相当する。ジャンクションノードは、ターゲットと相互作用する特徴を持たない頂点であり、3 つの辺と接続されている。Mol-SPhG は、全ての PF 間トポロジカル距離を保持している。

SPhG

SPhG は、PhFP・Mol-SPhG と異なり、それぞれが、単独の分子構造に対応するわけではない。活性化合物の分子構造を Mol-SPhG に変換した後、その Mol-SPhG の中から、第 2 章で説明した方法により、共通する部分グラフを抽出することで得られる [1,2]。SPhG は、トポロジカル・ファーマコフォアの直感的な理解を促すためにスパースな形を採っている。同時に、PF 間トポロジカルな距離も可能な限り維持することで、解釈しやすさと SH 性能とのバランスをとっている。SPhG を得るために、まず、あらかじめ定義された数の PF を選択した後、不要な頂点と辺を削除することで、Mol-SPhG から、候補

となる SPhG 群を生成する。これらの候補 SPhG はさらに、その候補 SPhG を含む活性化合物の scaffold の種類数(*NScaffold*法)によってスコアリングされる。この中でスコアが比較的高い候補 SPhG を、真の SPhG とする。本章では、6 つの PF を選択して候補 SPhG を形成した。各ターゲットについて、BM scaffold の数 (*NScaffolds*法) でスコアリングしたの観点からトップ 300 の SPhG を選択した [1,2,6]。前章の上位 100 個までの Precision-Recall 曲線において、100 個まで SPhG を増やすといずれのターゲットにおいても Precision が 0.5 近くまで減少したことから、その 3 倍の個数を含めれば、十分に抽出すべき特徴を捉えられていると考え、各ターゲット上位 300 個の SPhG を対象としたクラスタリング解析を行った。

4.2.3 トポロジカル・ファーマコフォアの類似度評価

3 種類のトポロジカル・ファーマコフォア表現を用いた類似度評価に基づいてクラスタリング解析を行うため、各表現方法について、類似度評価手法が必要である。PhFP は、0 と 1 のビットベクトルであるため、Jaccard 距離を用いた類似度評価を用いる。また、グラフで表現される Mol-SPhG と SPhG は、以下に説明する Graph Edit Distance (GED) によって定量的に類似度評価を行う [7]。

GED

GED は、あるグラフ A の頂点と辺を編集して、グラフ A を、もう一方のグラフ B に変換する際の総コストとして定義される。つまり、グラフの類似性は、グラフ A がどれだけ「簡単に」グラフ B に変換できるかという指標である。ここでは、頂点の置換、頂点の挿入、頂点の削除、辺の置換、辺の挿入、辺の削除の 6 つの編集操作を用い、それぞれの操作に対してコストが決められる。Garcia-Hernandez らの研究に基づいて、すべての頂点操作のコストを表 4-3 に、辺の操作のコストを表 4-4 に示した [8]。これらのコストは、3 つの点で SPhGs または Mol-SPhGs に合わせるために、オリジナルの研究の表から修正を行った。まず、頂点の変換表 (表 4-3) に J と表記されるジャンクションノードの定義を追加し、J は他の頂点の種類と同様に修正コストをすべて 2 と定義した。次に、辺の距離演算のコスト定義を表 4-4 のように新たに定義した。長さ n と m ($n > m$) の 2 つの辺 (芳香族の属性を持つ辺を除く) を置換する際のコストは次のように定義した。

$$\sum_{k=m+1}^n \frac{1}{k}$$

このように、辺の長さに応じてコストが単調に減少することは、直感的な化学的理解と一致する。例えば、PF 間トポロジカル距離を、1 から 2 へ伸ばす処理は、6 から 7 へ伸ばすよりも大きな影響を与える。また、同じ長さの、非芳香族の辺から芳香族の辺への置換コストは、文献[8]に基づいて、それらのエッジの長さの 3 倍として定義した。同様に、長さの異なる 2 つの芳香族エッジ間の置換は、対応する非芳香族エッジの 10 倍のコストを要した。これは、同文献で単結合、二重結合、三重結合の挿入と削除に対して定義されたコストに基づいている。

また、Mol-SPhG の GED を計算する際には、SPhG に比べて頂点 (PPP) 数が多いことから、NetworkX に実装されている時間制限アルゴリズムを適用した。このオプションは、あらかじめ設定された時間後に最小グラフ編集パスの探索を終了し、現在の最小距離を出力する。本研究では、10 秒を設定した。例えば、1425 化合物を含む Kop. データセットの GED の計算に 283 時間の CPU 時間を要する。なお、GED の計算には、NetworkX ver. 2.5 に実装されている近似 GED を使用した [9]。

なお、3 章で提案した SPhG でなく 2 章で説明した PhG でも同様に GED を用いたクラスタリングは可能であるが、3 章で実証したように、SH 性能に差が小さいことから優位性は小さいと考える。また、PhG で GED を計算する場合エッジの数が増える分だけ計算量が増加する。従って、今回は PhG ではなく、SPhG を用いて解析を行った。

4.2.4 トポロジカル・ファーマコフォアのクラスタリング

4.2.3 で定義した類似度評価手法を用いることで、PhFP・Mol-SPhG を用いた化合物のクラスタリングと、SPhG のクラスタリングが可能になる。本章では、t-distributed stochastic neighbor embedding (t-SNE) [10], isomap [11], 多次元スケーリング (multidimensional scaling, MDS) [12] の 3 つのクラスタリング方法を用いた。(全て scikit-learn ライブラリバージョン 0.23.2 に実装されているプログラムを用いた [13]) これらのマッピング手法を、前述の 3 つ表現 PhFP, Mol-SPhG, および SPhG に適用する。以下、3 つのマップはそれぞれ、「PhFP マップ」、「Mol-SPhG マップ」、「SPhG マップ」と呼ぶ。

PhFP マップと Mol-SPhG マップは、マップ内の各点が化合物 (CPD) に対応し、各ポイントは、その活性値 (pKi) によって色分けした。一方、SPhG マップ

プでは、各点は、活性化合物間の共通グラフから抽出された SPhG に対応する。各点の色は、その SPhG を持つ化合物数をデータセット内の全化合物数で割ったカバレッジを示す。

表 4-3 Graph Edit Distance (GED)における頂点の編集コスト一覧表

	D	A	P	N	R	J	DA	DP	DN	AP	AN	PN	DAP	DAN
D ^a	0	2	2	2	2	2	1	1	1	2	2	2	1	1
A ^b	2	0	2	2	2	2	1	2	2	1	1	2	1	1
P ^c	2	2	0	2	2	2	2	1	2	1	2	1	1	2
N ^d	2	2	2	0	2	2	2	2	1	2	1	1	2	1
R ^e	2	2	2	2	0	2	2	2	2	2	2	2	2	2
J ^f	2	2	2	2	2	0	2	2	2	2	2	2	2	2
DA ^g	1	1	2	2	2	2	0	2	2	2	2	2	2	2
DP ^g	1	2	1	2	2	2	2	0	2	2	2	2	2	2
DN ^g	1	2	2	1	2	2	2	2	0	2	2	2	2	2
AP ^g	2	1	1	2	2	2	2	2	2	0	2	2	2	2
AN ^g	2	1	2	1	2	2	2	2	2	2	0	2	2	2
PN ^g	2	2	1	1	2	2	2	2	2	2	2	0	2	2
DAP ^g	1	1	1	2	2	2	2	2	2	2	2	2	0	2
DAN ^g	1	1	2	1	2	2	2	2	2	2	2	2	2	0
insertion	1	1	1	1	1	0.5	1	1	1	1	1	1	1	1
deletion	1	1	1	1	1	0.5	1	1	1	1	1	1	1	1

^aD: hydrogen bond donor, ^bA: hydrogen bond acceptor, ^cP: positively ionizable, ^dN: negatively ionizable, ^eR: aromatic ring, ^fJ: junction. ^gDA, DAP など複数のシンボルは、一つのノードに複数の PPP が定義されているケースを表す。

表 4-4 GED における辺の編集コスト一覧表

	長さ n の非芳香族の辺 ^a	長さ n の芳香族野編 ^a
長さ m の非芳香族の 辺 ^a	$\sum_{k=m+1}^n \frac{1}{k}$	$3n \quad (n = m)$ $3n + \sum_{k=m+1}^n \frac{1}{k} \quad (n > m)$
長さ m の芳香族の辺 ^a	$3m \quad (n = m)$ $3m + \sum_{k=m+1}^n \frac{1}{k} \quad (n > m)$	$\sum_{k=m+1}^n \frac{10}{k}$
辺の挿入	0.1	1.0
辺の削除	0.1	1.0

^a ここでは $n > m$ と仮定する（一般性を失うことはない）。

4.3 結果と考察

4.3.1 トポロジカルファーマコフォアマップ

本研究で扱うマップは大きく分けて2つに分類できる。一つは、各化合物がマップ上の点に対応していて、化合物間の類似性を評価することでクラスタリングしたものである。もう一方は、一つのデータセットから得られた複数の SPhG をクラスタリングしたものである。SPhG をクラスタリングすることで、前者の化合物ベースのマップでは理解することができない、SPhG 間の関係を直感的に理解することができる。

我々は、Isomap, MDS, および t-SNE の3つのマッピングアルゴリズムを適用して、PhFP, Mol-SPhG, SPhG で表される活性化化合物のデータセットに対するマップを作成した。6つのターゲットのすべてのマップは、補足資料の図 4-S1 から 4-S6 に示した。t-SNE アルゴリズムによるマップは、3つのアルゴリズムの中で最も優れていた。これは、MDS によるほとんどのマップでは、クラスター化された領域が全く作成されなかったためであり、Isomap によるいくつかのマップでは、このアルゴリズムではクラスター化された領域を作成できるにもかかわらず、点（化合物または SPhG）が互いに重なっていたためである。そこで、t-SNE アルゴリズムによるマップをもとに議論を進めることにした。以下では、まず化合物が各点に対応するマップである Mol-SPhG マップと PhFP マップの違いを明らかにした。次に、活性化化合物のデータセットから SPhG マップによって抽出できる情報について議論する。

4.3.2 PhFP マップと Mol-SPhG マップの比較

図 4-3 は、Thr.の活性化合物のデータセットに基づいた PhFP マップと Mol-SPhG マップである。図 4-3a は PhFP マップである。CPD1, CPD2, CPD3 はグアニジウム構造を共有しているが、CPD4 はグアニジウムがアミジンとチオフエンに置換されており、CPD5 には類似の部位がないことが読み取れる。

Mol-SPhG マップを図 4-3b に示す。Mol-SPhG マップの各点は化合物を表しているのので、PhFP マップと同様である。しかし、scaffold に対する依存性は大きく異なっている。例えば、CPD1 と CPD3 は、図 4-3a ではお互いに離れているが、図 4-3b では近接している。CPD1 と CPD3 は全く異なる scaffold に属しているように見えるが、HBD や HBA の PF を持つ含むグアニジウム構造とアミド結合を共有している。図 4-3a では、CPD1 と CPD3 の間の距離は、グアニジウム構造を持たない全く異なる scaffold を持つ CPD5 と CPD3 の間の距離よりも長い。図 4-3b では、CPD1 と CPD3 の間の距離は、CPD5 の間の距離よりも短い。これは、図 4-3b は、scaffold から離れて、ファーマコフォアの特徴を優先した機能的な類似性に基づくマッピングがとなっていることを示している。

次に、図 4-4 に ABL1 の活性化合物のデータセットを用いて作成した PhFP マップと Mol-SPhG マップを示した。どちらのマップも、比較的高い pKi (約 9) を持つ大きなクラスターと、比較的低い pKi (約 7) を持つ小さなクラスターを分離することに成功した。図 4-4b を見ると、大きなクラスターに含まれる化合物は縮環構造を共通して持っていることがわかる。そのうちの 1 つは、環構造上に HBA を持つヘテロ芳香族環であり、もう 1 つは PF を持たない芳香環でなければならない。図 4-4b の Mol-SPhG マップでは、PF がハイライトされ、他の冗長な構造が省略されているので、これらの共通の特徴を見つけるのは図 4-4a よりも容易である。

このように、Mol-SPhG マップは、PF とその PF 間トポロジカル距離のみのグラフ表現に縮約することで、scaffold 依存性を排し、ターゲットとの相互作用に関わらない部分を省くことで、PhFP マップよりも、本質的特徴に基づくマップとなっている。

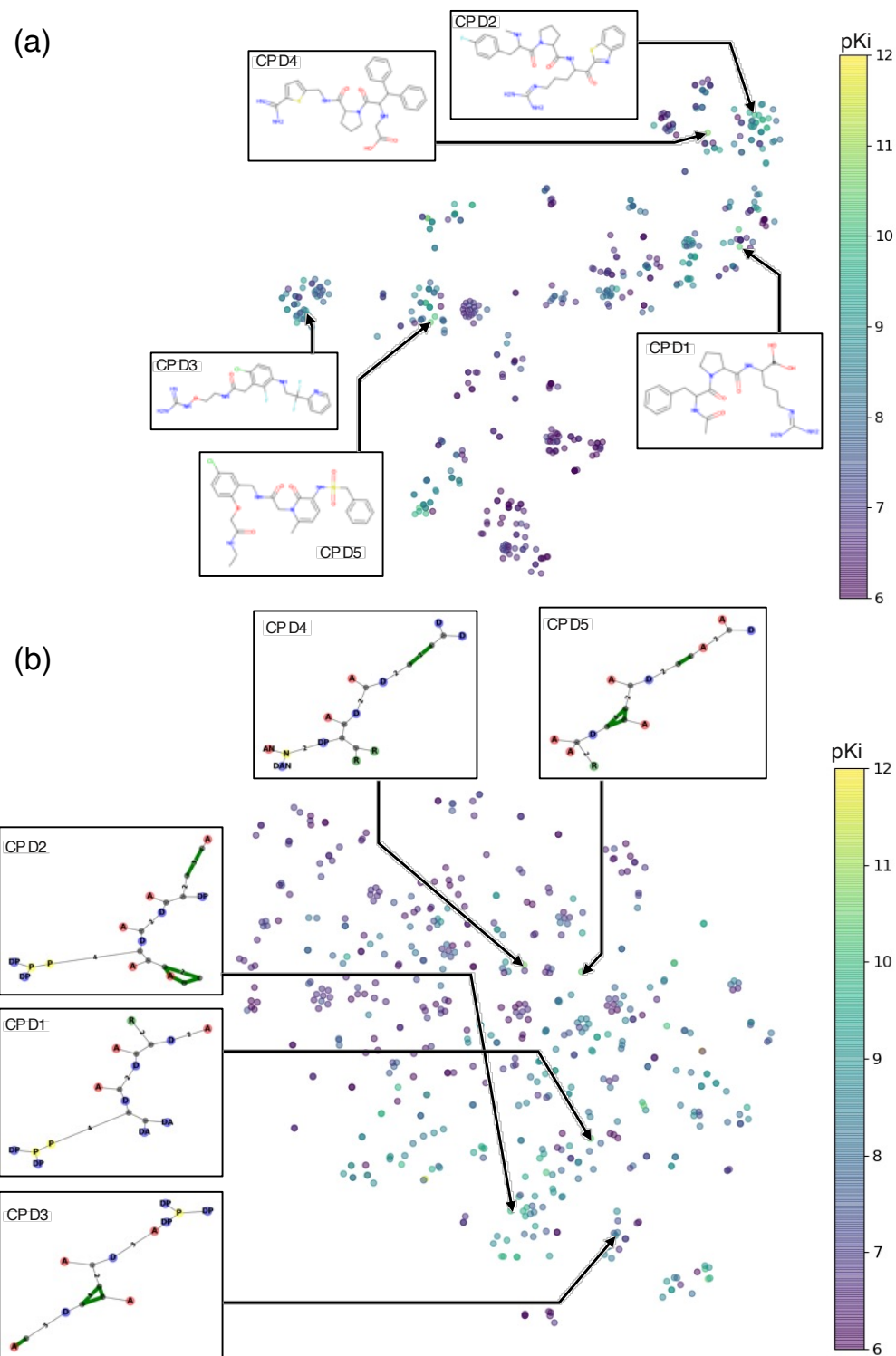


図 4-3 Thr.の PhFP マップと Mol-SPhG マップ

(a)各点は、それぞれ PhFP を用いてクラスタリングした化合物を表し、点の色は pKi を表す。また代表的な化合物をマップ上に示した。(b) Mol-SPhG を GED による類似度によりクラスタリングした。(a)と同様に点の色は pKi を表す。

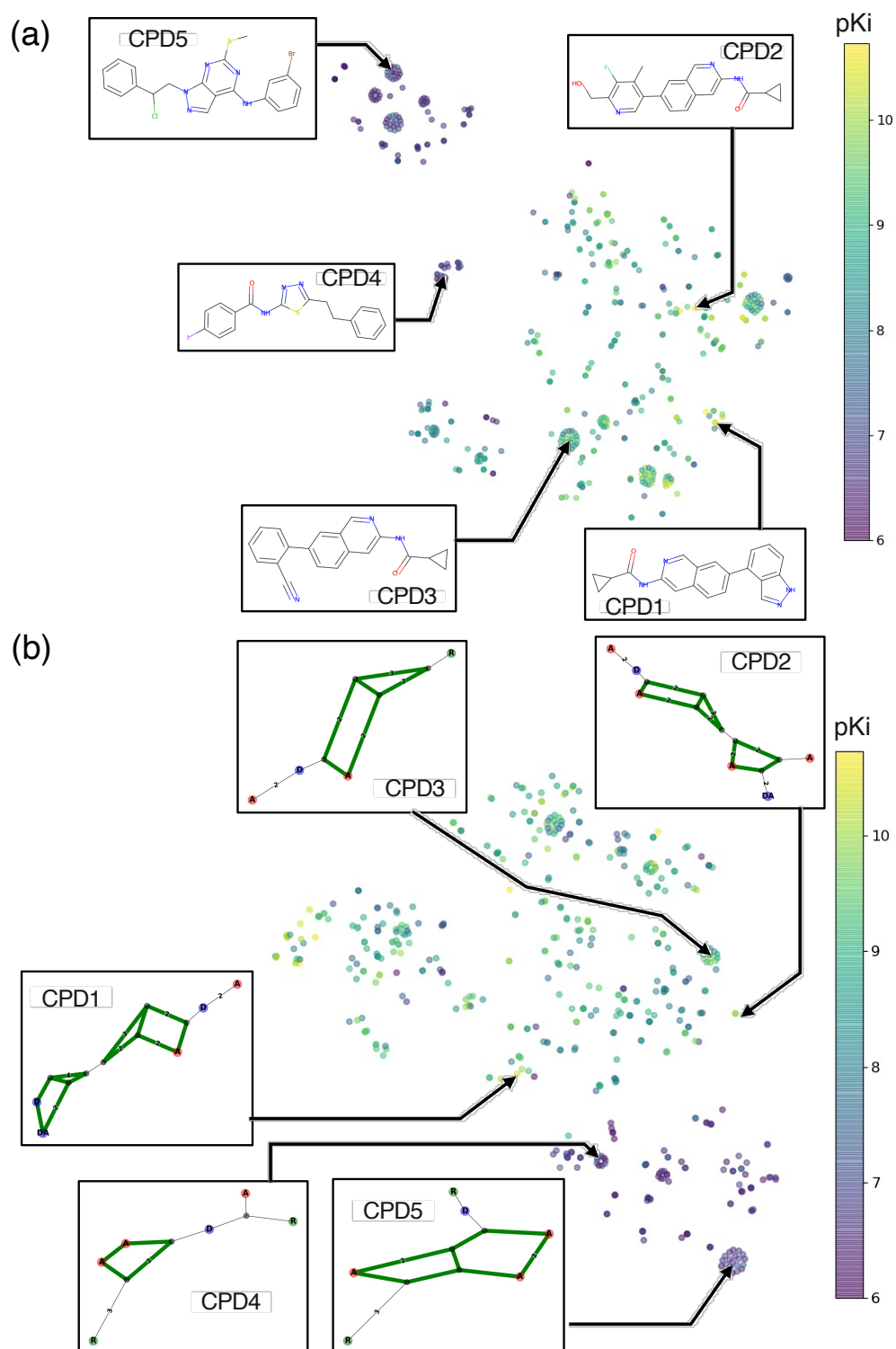


図 4-4 ABL1 の PhFP マップと Mol-SPhG マップ

(a)各点は、それぞれ PhFP を用いてクラスタリングした化合物を表し、点の色は pKi を表す。また代表的な化合物をマップ上に示した。(b) Mol-SPhG を GED による類似度によりクラスタリングした。(a)と同様に点の色は pKi を表す。

4.3.3 SPhG マップ

図 4-3 と図 4-4 に関する議論により Mol-SPhG マップは、scaffold 依存から脱却して、本質的な特徴を捉えていることを示した。以下では、さらにその Mol-SPhG から共通する特徴を抽出した SPhG を用いた SPhG マップについて、各ターゲットに関する高活性化合物のデータセットを用いて議論する。

Thrombin (Thr.)

図 4-5 に Thr. の SPhG マップを示す。

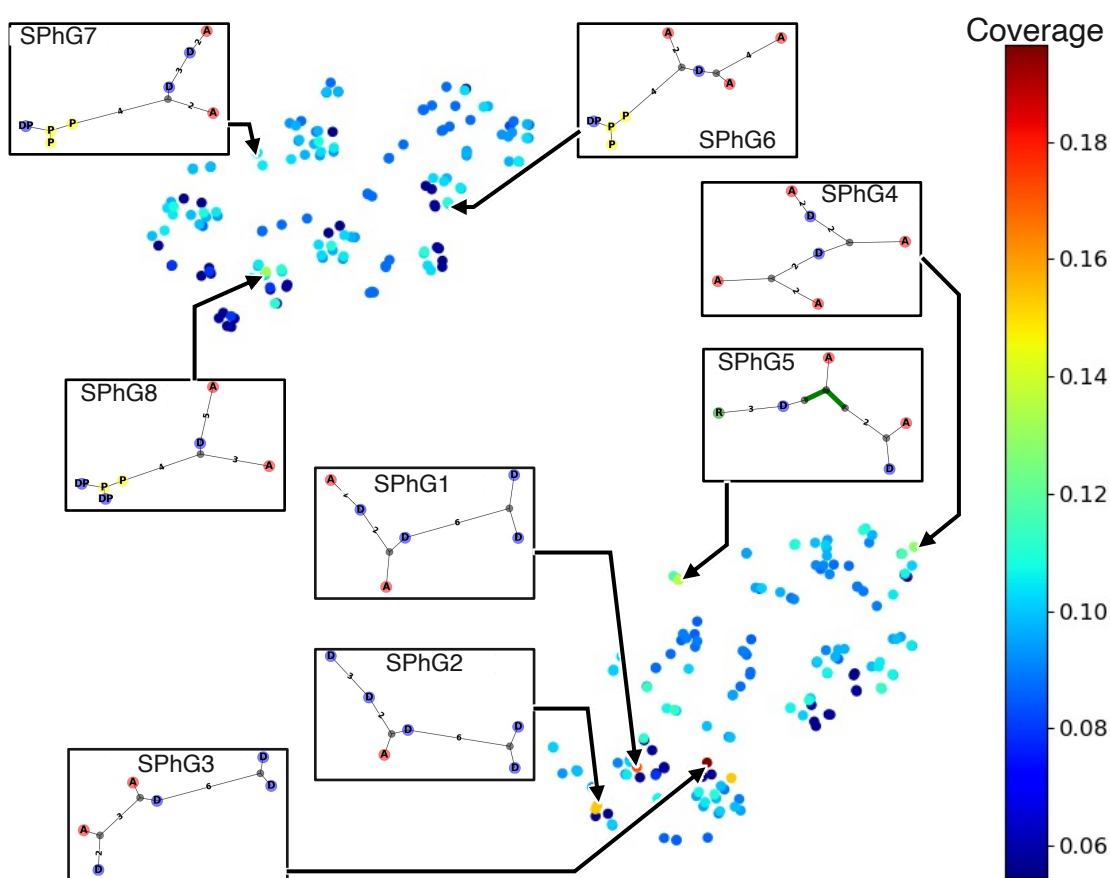


図 4-5 Thr. の SPhG マップ

各点は、GED で類似度評価してクラスタリングした SPhG を表す。点の色は Coverage, また代表的な SPhG をマップ上に示した。

このマップは、大きく 2 つのクラスターに分かれている。SPhG1-3 は 2 つの HBD(D) を持ち、距離 1 の場所にジャンクションノードを共有している。また、右下のクラスター内の、反対側にある SPhG4 と SPhG5 はこの構造を持っていない。左上のクラスター (SPhG6, SPhG7, SPhG8) は、グアニジウム構

造に対応する 4 つの PI(Positively Ionizable) という特徴を持つ部分構造が共通している。また、HBD と HBA(A)を持ち、2-3 個のジャンクションノードを持っていて、共通の特徴を示している。

また、図 4-5 内に例示した SPhG の中で互いに近い SPhG を比べると、ターゲットの結合についてより深い洞察が得られる。例えば、SPhG1-3 は 2 つのドナーを共有しており、距離 1 のところでそれらの間にジャンクションノードがある。ジャンクションノードから 6 つ離れたところに別のドナーがあり、2 つのドナーの反対側に距離 2 のドナー-アクセプターのペアがある。SPhG1 は第 3 章の図 3-13a で取り上げた SPhG と同じである。この SPhG の持つ PF は、すべて PDB から得られたターゲットと化合物の結合状態の構造から、実際に結合に寄与することが確認されている [2,14]。

Tyrosine kinase ABL1 (ABL1)

図 4-6 に、ABL1 の PhFP マップを示す。

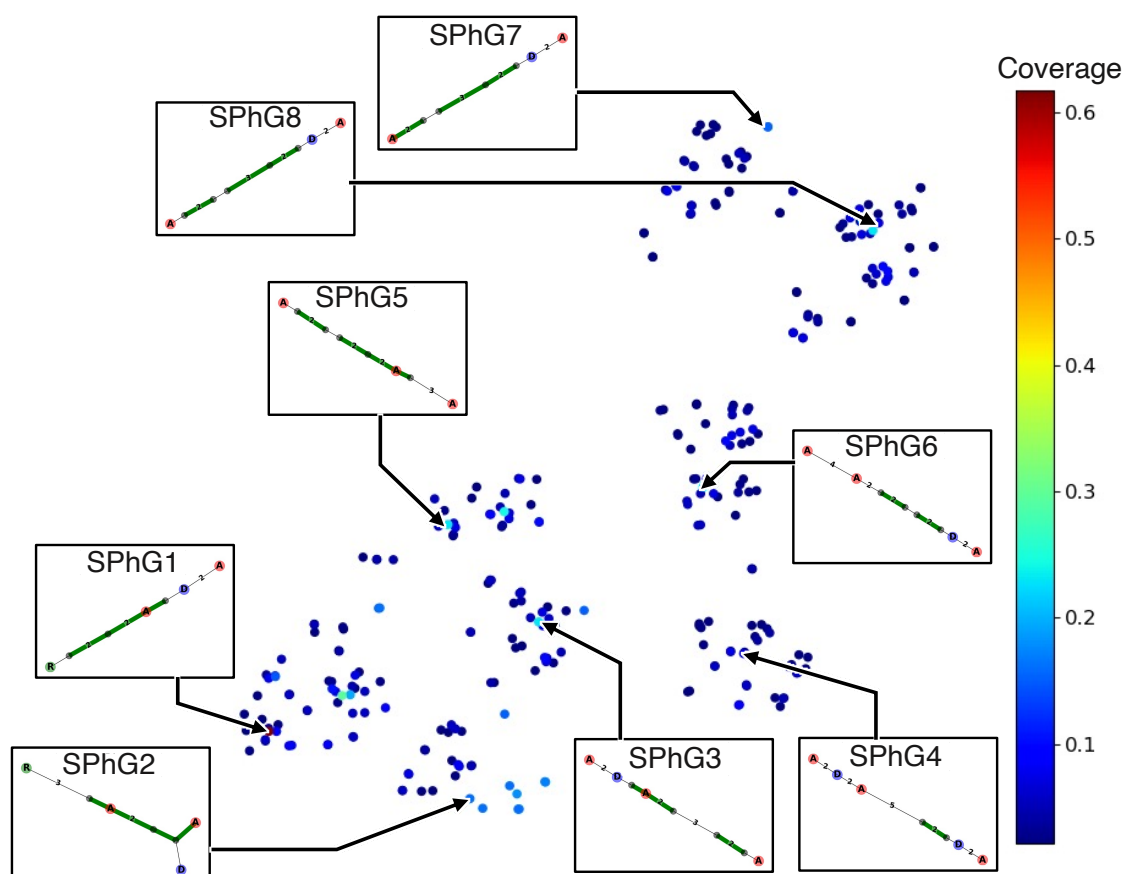


図 4-6 ABL1 の SPhG マップ

マップ中の SPhG1 は 0.6 付近で最も高い Coverage を持つ。この SPhG は、縮合芳香環を共有する活性 CPD が 2 つの環から構成され、少なくとも 1 つの環に HBA があることを示している。この概念は、SPhG2 や SPhG5 にも共通しており、図 4-6 の CPD1、CPD2、CPD3 に含まれるイソキノリン類も含まれる。また、SPhG1 は、この縮環構造から、トポロジカル距離で 1 つ離れたところにもう一つの芳香環を有し、さらにその反対側に D と A のペアを持つことがわかる。それ以外にも、SPhG1 の一部を変更した SPhG が各クラスタを形成している。例えば、SPhG5 は、SPhG1 における、縮環構造から距離 1 にある別の芳香環にさらに HBA 構造が付加していることを表す。このように SPhG マップは各クラスタ内の SPhG を調べることで、データセット内のカバレッジ情報付きで分子設計指針を得ることができる。

κ-opioid receptor (Kop.)

図 4-7 は、Kop. の PhFP マップである。

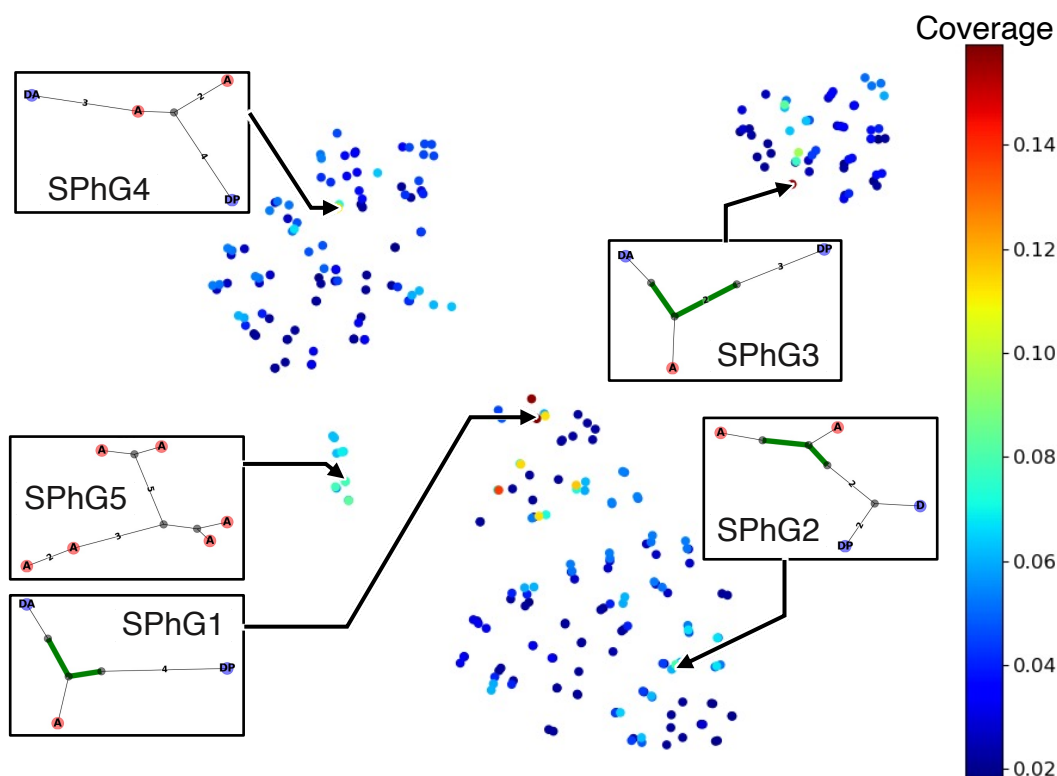


図 4-7 Kop. の SPhG マップ

SPhG1 と SPhG2 は、長さが 2 の芳香族結合の PF を持ち、結合の途中で HBA への分岐を持つ。右上のクラスターに属する SPhG3 は、2 ではなく 3 の長さを持つ芳香族結合を持っている。SPhG1 と SPhG3 は似たようなグラフに見えて

も、カバーする化合物の範囲が異なる．図 4-8 に 2 つの高活性化合物を例にとって説明する．

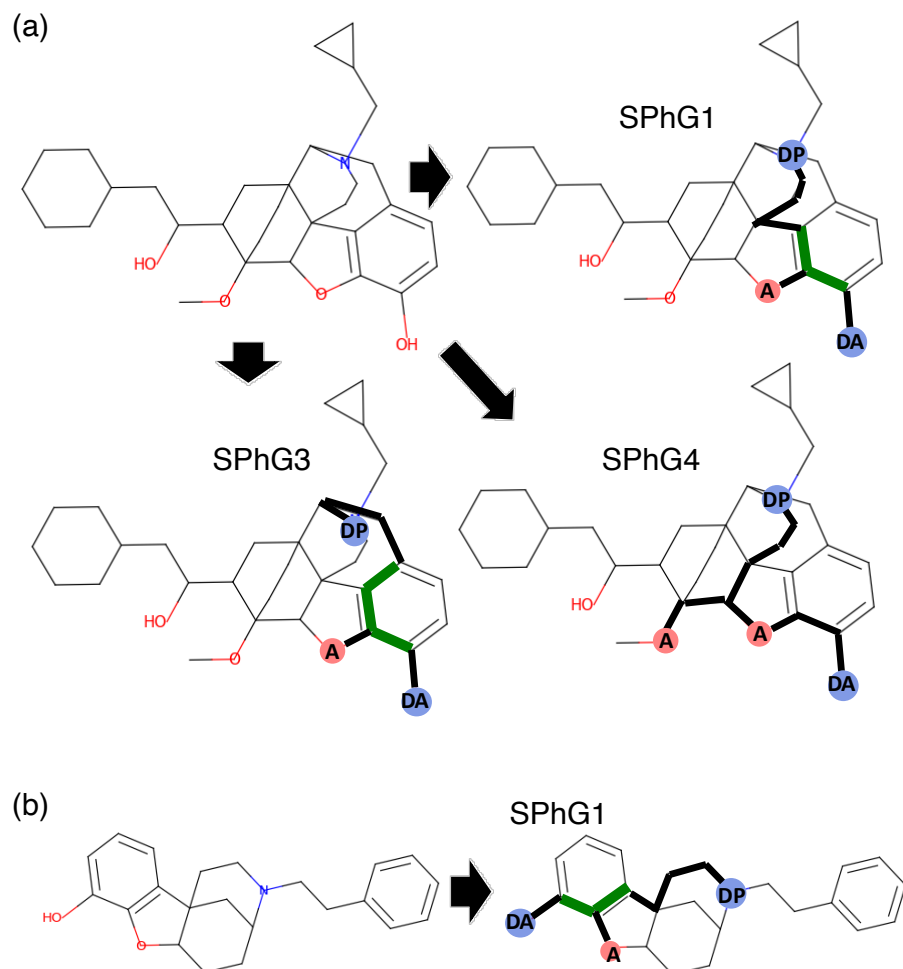


図 4-8 Kop.の活性化合物と SPhG の関係

(a) SPhG1, SPhG3, SPhG4(図 4-7)の 3 つの SPhG を含むモルフィネに類似の化合物. OH 基は HBD と HBA の両方の PF が割当られ, DA と表現されている. また, 三級アミンは HBD と Positively Ionizable (PI)の特徴が付与されている. (b) 別のペンタゾシンの類似の分子構造と, SPhG1 との対応. この分子構造は SPhG3 と SPhG4 は含まない.

図 4-8a に示したのは、モルフィネと類似の高度に縮環した多環構造を持つ化合物である。この化合物は、図 4-7 内の SPhG1, SPhG3, および SPhG4 に該当する。この 3 つの SPhG は、縮環構造の異なるパスを通ることで CPD1 にあてはまる。一方で、モルフィネと異なる Kop. に作用する化合物として知られているペンタジソンの類似構造を図 4-8b に示す。この化合物は、SPhG1 のみにあてはまり、SPhG3 と SPhG4 にはマッチしない。つまり、SPhG1 は異なる scaffold 間で共有される特徴の表現をうまく導き出している。

また、図 4-7 の左側にある SPhG5 を含む小さなクラスターは、CPD5 に対応する異なる種類の化合物に対応する。SPhG マップにより、2 つの異なるタイプの活性化合物がコンセプトレベルで検出され、クラスター化できている。

Transmembrane Proatase Serine 6 (TPS6)

TPS6 の高活性化合物のデータセットは、表 4-S1 に示した通り 21 個しかなく、データセットのサイズが小さい。このとき、SPhG マップと、CPD ベースのマップ (Mol-SPhG マップと PhFP マップ) との間に顕著な違いが見られる。図 4-S6 に示すように、PhFP マップはデータセット中の化合物の数と同じ数のポイントがクラスターにならずに分散している状態しか表現できない。

例えば、右上のクラスターは、グアニジウム部分と他の水素結合の特徴を反対側に持つデザインコンセプトに対応している。また、左上のクラスターは、水素結合点の配置に重点を置いた分子設計指針を提供している。SPhG1 と SPhG2 は 2 つの HBD を表し、6 つ以上のトポロジカル距離を持つ他の PF から 1 つのトポロジカル距離を持つジャンクションノードを共有している。どちらも Thr. に似たコンセプトである。特に、図 4-5 に示す Thr. の SPhG1 と図 4-9 に示す TPS6 の SPhG3 は同一の SPhG である。これは、この共通の SPhG を含む化合物は、Thr. と TPS6 の両方に活性を持つことを示唆している。実際に、文献[15]に示された化合物が、この SPhG を含み、かつ Thr. と TPS6 の両方に活性を持つことを確認できた。³

一方、左下のクラスターは、CPD5 と CPD16 に相当する、全く異なる設計指針の分子構造を表している。このようにごく少ない活性化合物のデータからも SPhG を抽出し、クラスターリングすることで、異なる複数の分子設計指針を導出できる。

³ なお、今回クラスターリング解析を行った 6 つのターゲットの全てのペアについて共通する SPhG が存在するか確認したところ、Thr. と TPS6 以外には共通する SPhG は存在しなかった。Thr. と TPS6 は両方 Serine Protease であり、類似の化合物に活性を持つことは妥当である。また、他のターゲットについては、類似のターゲットがなく、SPhG も共通するものがなかったと考えられる。

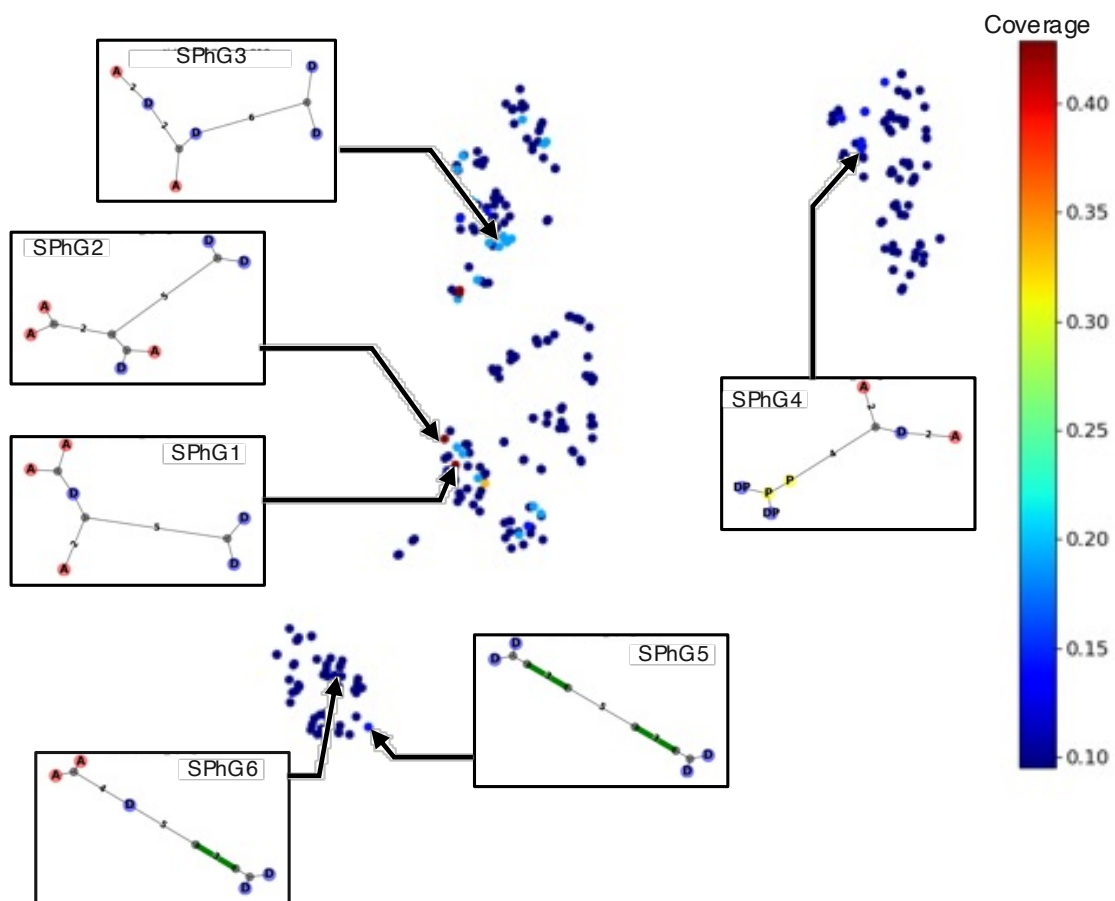


図 4-9 TPS6 の SPhG マップ

4.4 結論

本章では、NScaffold 基準で得られた 300 個の SPhG を用いて、ターゲットとの相互作用に関与する特徴を捉え、候補化合物の設計指針を得ることができるクラスタリング手法を提案した。

SPhG の類似性を評価する手法として GED を導入し、高活性化合物のデータセットから得られた SPhG のマッピングに t-SNE を用いた。今回は、Thr., ABL1, Kop., PI3, GPCR, および TPS6 の 6 つのターゲットに可視化手法の比較をした。まず、トポロジカル・ファーマコフォアの表現として PhFP マップと、SPhG の生成時に作成される Mol-SPhG マップを比較したところ、Mol-SPhG マップは、PhFP マップに比較して scaffold に依存しない特徴を抽出できることがわかった。次に、化合物ごとに作成された Mol-SPhG から共通項を抽出して作成された SPhG を 300 個抽出し、GED を用いた類似性評価を用いてクラスタリング解析を行った。その結果、SPhG マップは、ターゲットと化合物の相互作用に必要な本質的な特徴を抽出でき、クラスターに分類された分子設計指針を提示できることがわかった。今回、提案した SPhG マップを用いることで、scaffold に依存しない化合物探索が可能となり、今後の LBDD に貢献することができると考えている。

4.5 参考文献

1. Nakano, H.; Miyao, T.; Funatsu, K. Exploring Topological Pharmacophore Graphs for Scaffold Hopping. *J. Chem. Inf. Model.* **2020**, *60*, 2073–2081.
2. Nakano, H.; Miyao, T.; Swarit, J.; Funatsu, K. Sparse Topological Pharmacophore Graphs for Interpretable Scaffold Hopping. *J. Chem. Inf. Model.* **2021**, *61*, 3348–3360.
3. Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M.P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. 'The ChEMBL Database in 2017.' *Nucleic Acids Res.* **2017**, *45*, D945–D954.
4. Landrum, G., RDKit: Open-Source Cheminformatics Software <http://www.rdkit.org> (accessed Dec 28, 2019)
5. <https://github.com/rdkit/rdkit/blob/master/Data/BaseFeatures.fdef> (accessed Dec 28, 2019)
6. Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.
7. Abu-Aisheh, Z.; Raveaux, R.; Ramel, J. Y.; Martineau, P. An Exact Graph Edit Distance Algorithm for Solving Pattern Recognition Problems. *ICPRAM 2015 - 4th Int. Conf. Pattern Recognit. Appl. Methods, Proc.* **2015**, *1*, 271–278.
8. Garcia-Hernandez, C.; Fernández, A.; Serratos, F. Ligand-Based Virtual Screening Using Graph Edit Distance as Molecular Similarity Measure. *J. Chem. Inf. Model.* **2019**, *59* (4), 1410–1421.
9. Hagberg A, Swart P, S Chult D. Exploring network structure, dynamics, and function using NetworkX. **2008**.
10. Maaten, L. van der; Hinton, G. Multiobjective Evolutionary Algorithms to Identify Highly Autocorrelated Areas: The Case of Spatial Distribution in Financially Compromised Farms. *J. Mach. Learn. Researsch* **2008**, *9*, 2579–2605.
11. Tenenbaum, J. B.; Silva, V. de; Langford, J. C. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science* **2000**, *290*, 2319 LP – 2323.
12. Torgerson, W.S. Multidimensional scaling I: Theory and method. *Psychometrika*, **1952**, *17*:401–419
13. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python, *J. Mach. Learn. Researsch* **2011**, *12*, 2825–2830.

14. Berman, H. M.; Westbrook, J. D.; Feng, Z.; Gilliland, G. L.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, 28, 235–242.
15. Béliveau, F.; Tarkar, A.; Dion, S. P.; Désilets, A.; Ghinet, M. G.; Boudreault, P. L.; St-Georges, C.; Marsault, É.; Paone, D.; Collins, J.; et al. Discovery and Development of TMPRSS6 Inhibitors Modulating Heparin Levels in Human Hepatocytes. *Cell Chem. Biol.* **2019**, 26, 1559–1572.e9.

4.6 補足

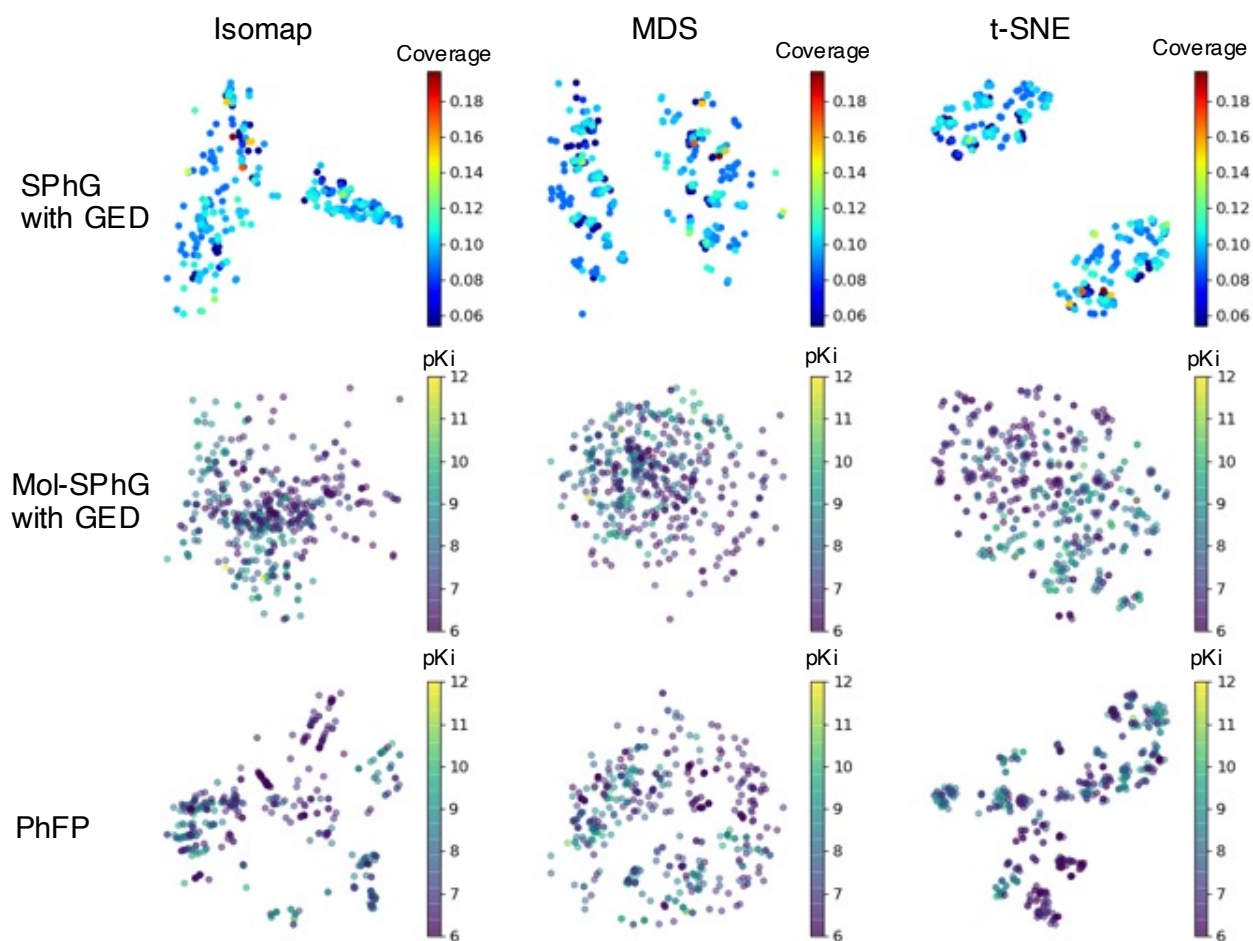


図 4-S1 Thr. のデータセットを用いた全マップ

最初の行は、3つのマッピングアルゴリズム Isomap, MDS, t-SNE による GED に基づいた SPhG マップを表す。各点の色は、そのノードの coverage を示している。2 番目の行は、1 番目の行と同じ GED に基づいて作成した Mol-SPhG マップを表す。下段は RDKit で実装された PhFP マップを表す。図 4-S2～図 4-S6 も同様。

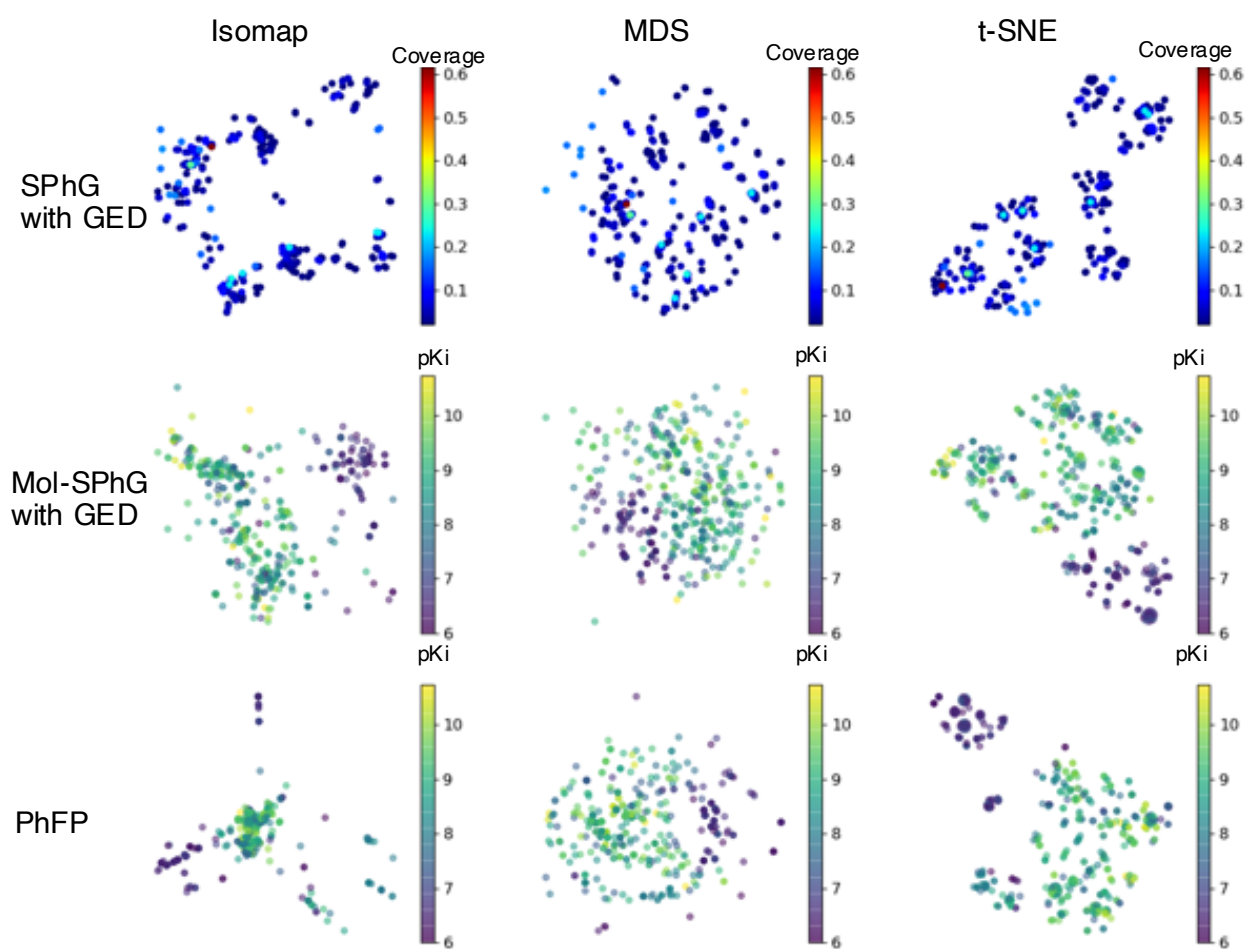


図 4-S2 ABL1 のデータセットを用いた全マップ
各行・列の説明は図 4-S1 を参照.

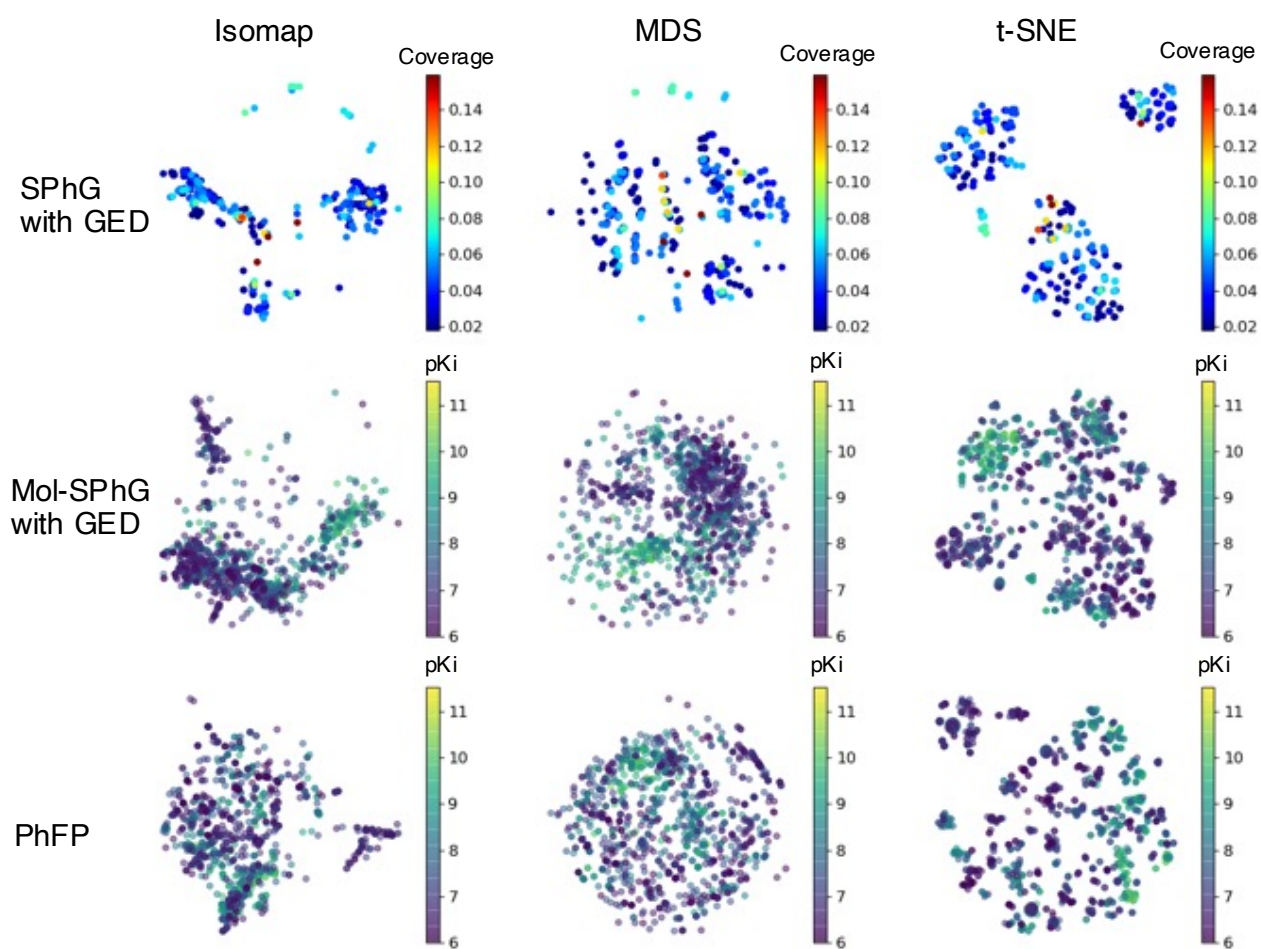


図 4-S3 Kop.のデータセットを用いた全マップ
各行・列の説明は図 4-S1 を参照.

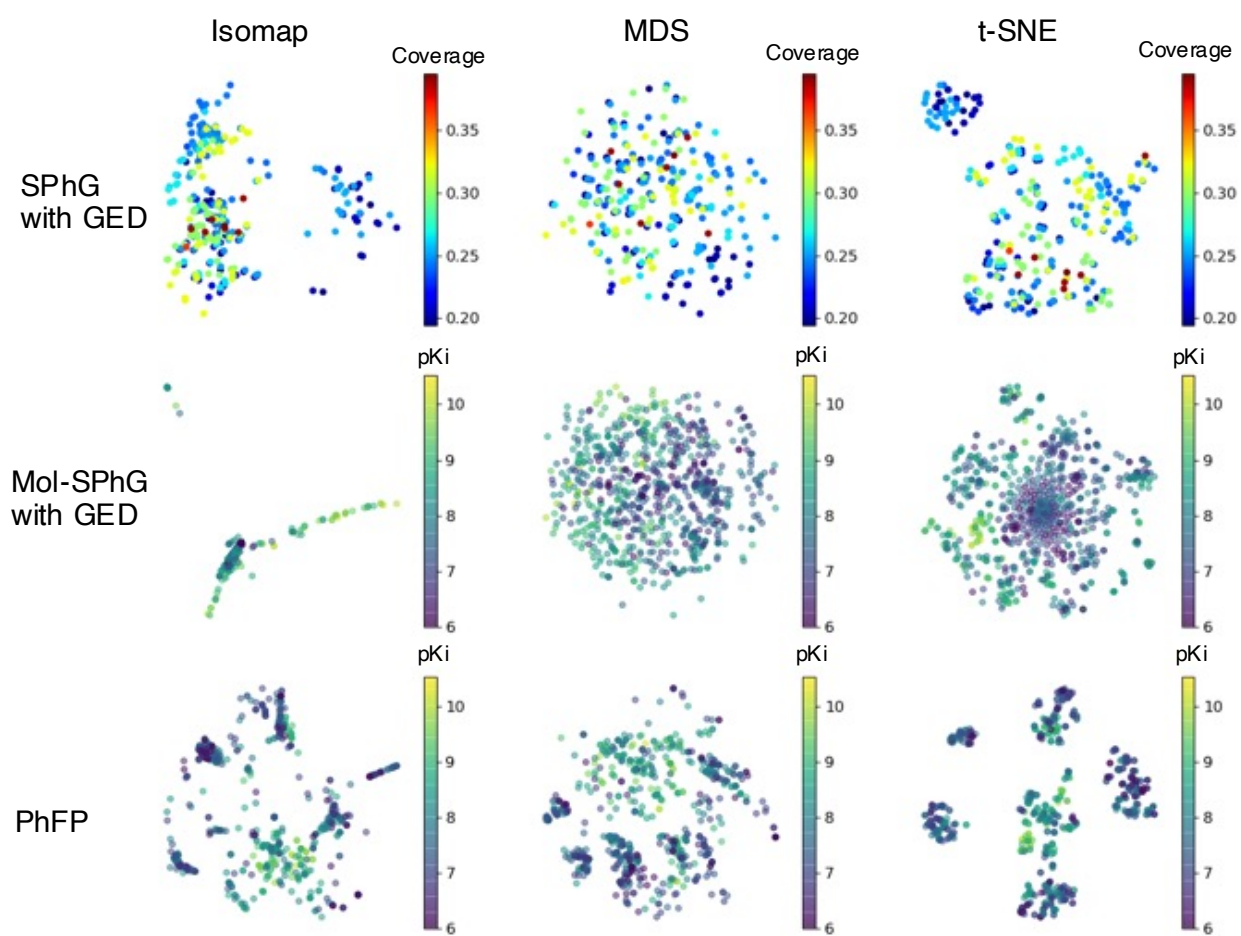


図 4-S4 PI3 のデータセットを用いた全マップ
各行・列の説明は図 4-S1 を参照.

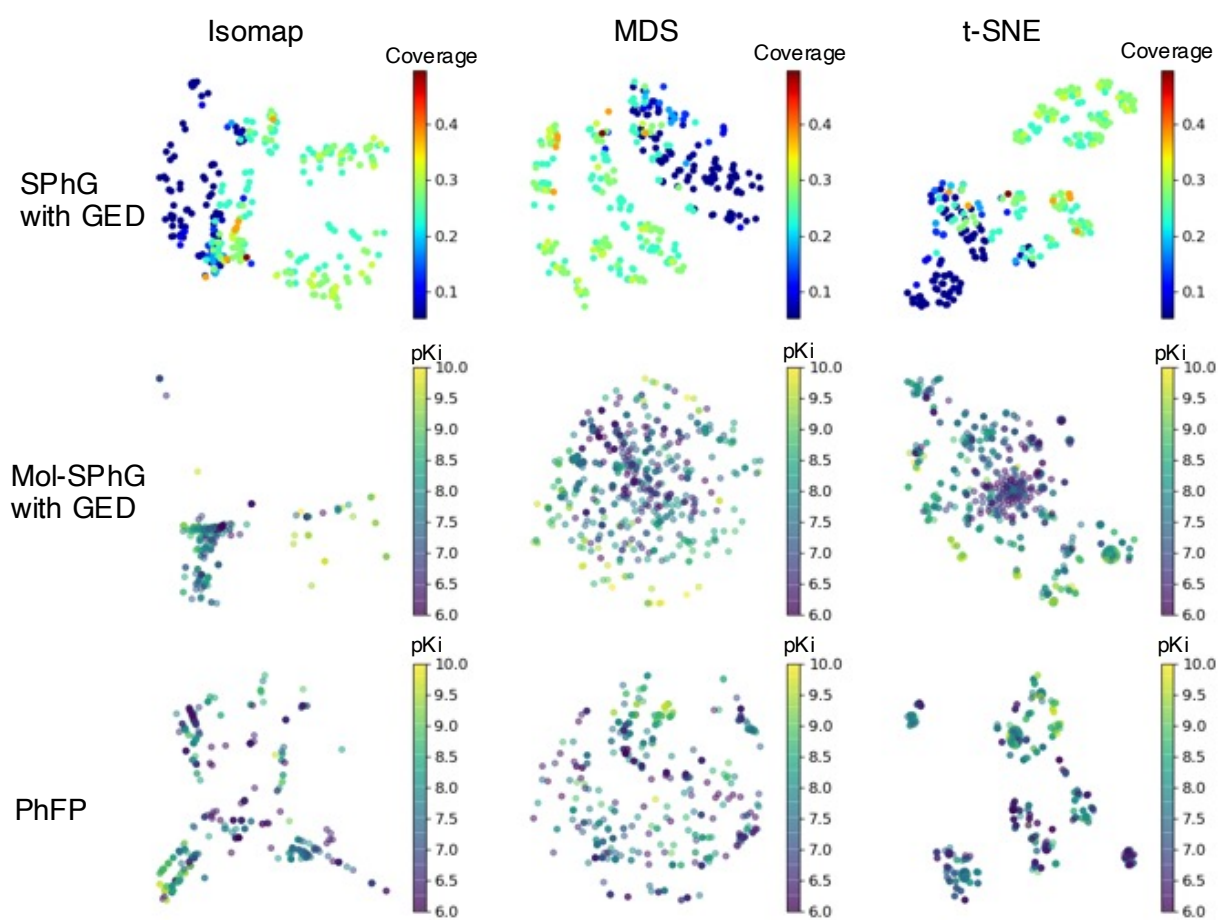


図 4-S5 GPCR44 のデータセットを用いた全マップ
各行・列の説明は図 4-S1 を参照.

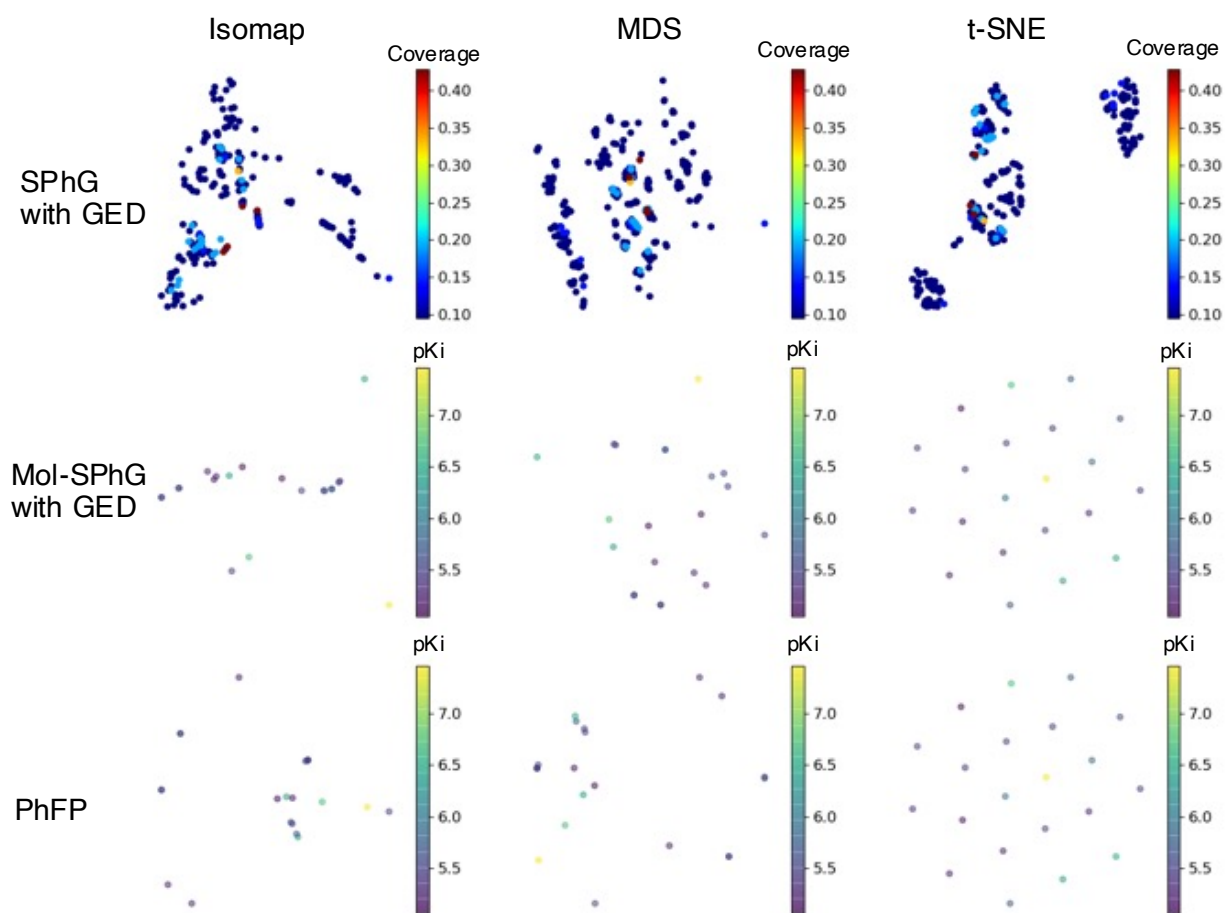


図 4-S6 TPS6 のデータセットを用いた全マップ
各行・列の説明は図 4-S1 を参照.

表 4-S1 TPS6 をターゲットとする活性化合物の SMILES と pKi 値

TPS6 においては「活性化合物」の定義を pKi>5.0 とした.

CPD	ChEMBL ID	SMILES	activity
CPD1	CHEMBL3352840	<chem>CC(NC(=O)C(CCC(N)=O)NC(=O)C(N)CO)C(=O)NC(CCCNC(=N)N)C(=O)c1nc2ccccc2s1</chem>	7.46
CPD2	CHEMBL1215083	<chem>N=C(N)NCCCC(NS(=O)(=O)Cc1ccccc1)C(=O)N1CCCC1C(=O)NCc1ccc(C(=N)N)cc1</chem>	6.75
CPD3	CHEMBL3219087	<chem>COc1ccc(-c2cccc(S(=O)(=O)NC(Cc3cccc(C(=N)N)c3)C(=O)N3CCC(CCN)CC3)c2)c(OC)c1</chem>	6.51
CPD4	CHEMBL1215085	<chem>N=C(N)c1ccc(CNC(=O)C2CCCN2C(=O)C(CC2CCCC2)NS(=O)(=O)Cc2ccccc2)cc1</chem>	6.44
CPD5	CHEMBL9126	<chem>N=C(N)c1ccc(N2CCN(c3ccc(C(=N)N)cc3)CC2)cc1</chem>	5.94
CPD6	CHEMBL3219091	<chem>CNC(=O)NC1CCN(C(=O)C(Cc2cccc(C(=N)N)c2)NS(=O)(=O)c2cccc(NC(=O)CCN)c2)CC1</chem>	5.89
CPD7	CHEMBL3594083	<chem>N=C(N)c1ccc(NCCCCCCCNC2ccc(C(=N)N)cc2)cc1</chem>	5.79
CPD8	CHEMBL3219073	<chem>N=C(N)c1cccc(CC(NS(=O)(=O)c2cccc(-c3ccccc3F)c2)C(=O)N2CCC(CCN)CC2)c1</chem>	5.68
CPD9	CHEMBL3219093	<chem>CCNC(=O)NC1CCN(C(=O)C(Cc2cccc(C(=N)N)c2)NS(=O)(=O)c2cccc(NC(=O)CCN)c2)CC1</chem>	5.62
CPD10	CHEMBL191881	<chem>CC(=O)NC(CC(C)C)C(=O)NC(CC(C)C)C(=O)NC(C=O)CCCNC(=N)N</chem>	5.5
CPD11	CHEMBL3219075	<chem>N=C(N)c1cccc(CC(NS(=O)(=O)c2cccc(-c3ccccc3F)c2)C(=O)N2CCC(CCN)CC2)c1</chem>	5.49
CPD12	CHEMBL3219077	<chem>N=C(N)c1cccc(CC(NS(=O)(=O)c2cccc(-c3ccc(F)cc3)c2)C(=O)N2CCC(CCN)CC2)c1</chem>	5.49
CPD13	CHEMBL3219088	<chem>N=C(N)c1cccc(CC(NS(=O)(=O)c2cccc(NC(=O)CCN)c2)C(=O)N2CCC(CCNC(N)=O)CC2)c1</chem>	5.48
CPD14	CHEMBL3594082	<chem>N=C(N)c1ccc(NCCCCCCCNC2ccc(C(=N)N)cc2)cc1</chem>	5.43
CPD15	CHEMBL3219089	<chem>N=C(N)c1cccc(CC(NS(=O)(=O)c2cccc(NC(=O)CCN)c2)C(=O)N2CCC(NC(N)=O)CC2)c1</chem>	5.41
CPD16	CHEMBL269032	<chem>N=C(N)c1ccc(N2CCCN(c3ccc(C(=N)N)cc3)CC2)cc1</chem>	5.36
CPD17	CHEMBL181480	<chem>N=C(N)c1ccc(OCc2ccccc2COc2ccc(C(=N)N)cc2)cc1</chem>	5.28
CPD18	CHEMBL365916	<chem>N=C(N)c1ccc(OCc2ccc(COc3ccc(C(=N)N)cc3)cc2)cc1</chem>	5.24
CPD19	CHEMBL3828410	<chem>N=C(N)c1ccc(NC(=O)C(=O)Nc2ccc(C(=N)N)cc2)cc1</chem>	5.21
CPD20	CHEMBL2159291	<chem>CC(C)(C)OC(=O)c1ccc(NC(=O)CCNS(=O)(=O)c2cccc(C(=N)N)c2)cc1</chem>	5.12
CPD21	CHEMBL2159271	<chem>N=C(N)c1ccc(CNC(=O)CCNS(=O)(=O)c2cccc(C(=N)N)c2)cc1</chem>	5.05

第5章 総括

5.1 結論

第1章では、創薬における CADD の重要性から始まり、ある化合物がターゲットに対して活性を持つための本質的な特徴を表現する重要性と、その適用先として、SH という課題があること述べた。また、その中で、このような本質的な特徴を表すトポロジカルな表現法を探求すること、またその表現に基づく SH 手法を創出することを本研究の目的とすることを述べた。

第2章では、PF を頂点に、PF 間のトポロジカル距離を辺の長さとして持つ完全グラフ型の PhG という概念に着目した。骨格に依存しない本質的な特徴を抽出するために、PhG のマイニング手法を考案した。実際に従来の方法よりも、高い SH 性能を得ることかできた。

第3章は、第2章で説明した PhG に、解釈可能性を付与するために、完全グラフであった PhG の代替として、分子構造に近いスパースなグラフ Sparse Pharmacophore Graph (SPhG) を提案した。さらに、複数の活性化合物から共通する特徴を抽出して SPhG として提示するまでのアルゴリズムについて説明した。

第4章では、6つのターゲットに対して、SPhG を用いた解析を行った結果について示した。活性化合物のデータセットから得られる複数の SPhG を、Graph Edit Distance (GED) というグラフの類似度評価手法を用いてクラスタリング解析を行い、本研究で開発した SPhG が、ターゲットと化合物の相互作用の本質的な特徴を抽出できることを示した。

5.2 今後の予定

今後は、本研究で創出したトポロジカル・ファーマコフォアの表現である SPhG をさらに発展させ、また応用範囲を広げていきたい。例えば、SPhG の抽出や SH 化合物のスクリーニングには、完全一致方式を採用していて、ある SPhG の辺の長さや PF の種類が少しでも異なれば、化合物に含まれないと判定していた。それをある程度の類似度までは許容するいわゆるファジー化に取り組むことで、より本質的なファーマコフォア表現に昇華させたいと考えている。図 5-1 にその一例を示す。Thr.の活性化合物である図 5-1a は、図 5-1b の SPhG を含んでいる。また別の Thr.の活性化合物である図 5-1c は、図 5-1d の SPhG を含む。しかし、図 5-1b の SPhG は図 5-1c の化合物に含まれず、図 5-1d は図 5-1a の化合物に含まれない。それは、図 5-1b と d に青線と赤線示し

た、トポロジカル距離 5 と 6 の辺が異なるからである。しかし、実際図 5-1c の距離 5 に相当する部分は、ターゲットに結合するか否かという観点で、距離 6 とほぼ同等である。実際、その距離 5 の区間にチオフエンの C-S 結合があり、C-C 結合よりも物理的距離が長く、また、一重結合の角度や二面角などの自由に変形できる結合によって、そのトポロジカル距離違いを吸収できる。したがって、このトポロジカル距離 1 の差が活性に与える影響は小さい。実際、図 5-1a,c の pKi はそれぞれ 10.4 と 10.7 で差は小さい。本研究で用いた SPhG をクエリとする活性予測では完全一致を用いているため、図 5-1b は図 5-1c に、図 5-1d は図 5-1a に含まれないが、図中に示すとおり、ほとんど含まれるに等しい状態を考慮した SPhG のマイニング手法を開発することで、SH を含む活性・不活性予測において、より高い性能を発揮でき、より本質的なファーマコフォア表現を創出できると考えている。

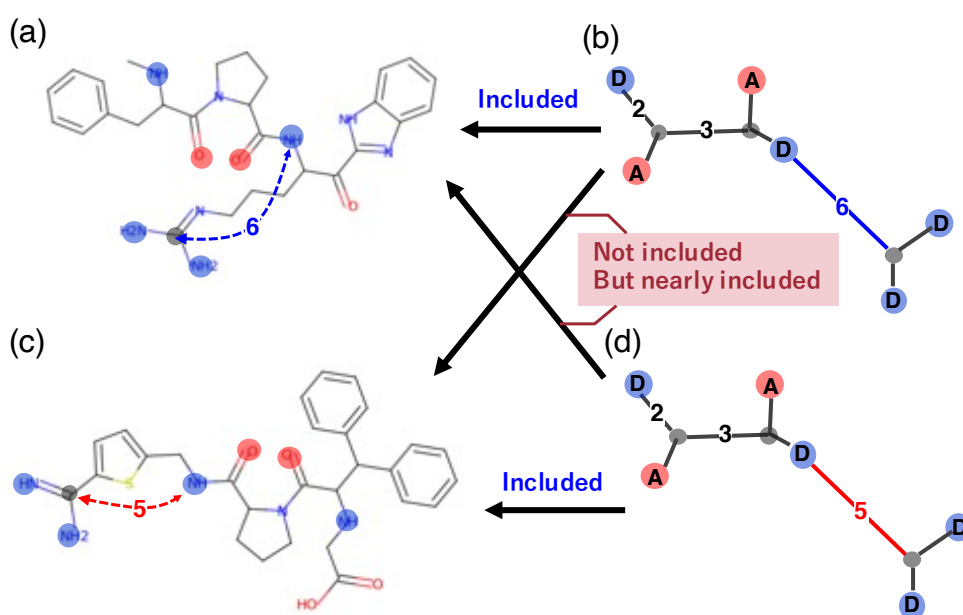


図 5-1 完全一致方式では考慮できていない SPhG

(a) HBD とジャンクションノードのトポロジカル距離 6 の分子 (b) 分子 a に含まれる SPhG (c) トポロジカル距離 5 の分子 (d) 分子 c に含まれる SPhG

またファジー化以外の SPhG の改良として、3 次元情報を取り込めるようにすることも改善案の一つである。また、SPhG の改善に加え、創薬における SH 以外の応用として、類似構造を持つ化合物において、大きく活性が異なる (例えば pKi で 2 以上) 現象として定義される Activity Cliff の解析や理解にも役立つ可能性があると考えている。また、チャレンジングな目標としては、既存

のデータセットのみから、より高い活性を持つ化合物を見つける外挿問題にも取り組みたいと考えている。

さらに、今回は PF を用いているため医薬品が対象となるが、あらゆる機能性有機材料に対して、今回の手法を改良することで適用可能である。例えば、有機エレクトロニクス材料・色素材料・ポリマー設計などにも有効であると考えている。今回は、創薬をターゲットとしていたため、その薬理活性を発現するために、ターゲットに結合する可能性のある部分構造を、PPP として検出し、PhG または SPhG では、頂点として扱った。他の分野に応用するには、PPP に替えて、各目的の機能を発現するために必要な部分構造を改めて定義し、今回のグラフ化手法にあてはめることが、最初の一步である。

このように、今回開発した技術は、創薬をより深化させる方向にも、また創薬以外の分野に広げる方向にも応用できると考えている。今後も、本論文の成果を発展させて、ケモインフォマティクス・マテリアルズインフォマティクスを用いた材料設計により、優れた材料を世に生み出し、社会に貢献していきたい。

謝辞

本研究開始当初から，研究全般に渡り，2020 年度まで主指導教官として丁寧かつ的確な御指導いただきました船津公人特任教授，3 年間に亘り，直議何度も議論をし，ご指導いただいた宮尾知幸准教授には深く御礼申し，心から敬意を表します．また，データ解析を手伝っていただいた Swarit Jasial 助教，2021 年度より主指導教員としてご指導いただいた浦岡行治教授に感謝致します．

また，同研究室立ち上げ当初のメンバーとして，活発な議論をさせていたいた田村俊祐様(現在同研究室博士後期課程二年)・青島慎一郎様(2019 年度博士前期課程修了)に感謝いたします．博士学位審査員であります浦岡行治教授，上久保裕生教授，および中村哲教授におかれましては，研究全般における有用なアドバイスを頂き，深く御礼申し上げます．データ駆動型化学研究室の方々には研究サポート以外にも様々な会話・親交を通して，公私共々大変お世話になりました．この場をお借りして感謝の意を表します．

最後に，平日は社会人として働きながら，その業務外の時間を使って奈良先端科学技術大学院大学で研究を進めた 3 年間，素晴らしい学習の機会を与えてくれた，妻と二人の娘に心より感謝いたします．

研究業績

I. 学位論文の主たる部分を公表した論文

1. “Exploring Topological Pharmacophore Graphs for Scaffold Hopping”
Hiroshi Nakano, Tomoyuki Miyao, and Kimito Funatsu,
J. Chem. Inf. Model. 2020, 60, 2073–2081.
2. “Sparse Topological Pharmacophore Graphs for Interpretable Scaffold Hopping”
Hiroshi Nakano, Tomoyuki Miyao, Swarit Jassial and Kimito Funatsu,
J. Chem. Inf. Model. 2021, 61, 3348–3360.

II. 学会発表

1. Pharmacophore Graph Study for Scaffold Hopping”
Hiroshi Nakano, Tomoyuki Miyao, and Kimito Funatsu,
第6回ケモインフォマティクス秋の学校, Nov. 2019.