Exploring Topological Pharmacophore Graphs for Scaffold-Hopping

Hiroshi Nakano[1], Tomoyuki Miyao,[1,2] Kimito Funatsu,[1,2,3*]

[1]Graduate School of Science and Technology, Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma, Nara, 630-0192, Japan.

[2]Data Science Center, Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma, Nara, 630-0192, Japan.

[3]Department of Chemical System Engineering, School of Engineering, The University of Tokyo, 7-3-1 Hongo. Bunkyo-ku, Tokyo 113-8656, Japan.

*To whom correspondence should be addressed:

Tel: +81-3-5841-7751, Fax: +81-3-5841-7771, E-mail: funatsu@chemsys.t.u-tokyo.ac.jp

Abstract

Primary goal of ligand-based virtual screening is to identify active compounds consisting of a core scaffold that is not found in the current active compound pool. Scaffold-hopping is the term used for this purpose. In the present study, topological representations of pharmacophore features on chemical graphs were investigated for scaffold-hopping. Pharmacophore Graphs (PhGs), which consist of pharmacophore features as nodes and their topological distances as edges, were used as a representation of important information of compounds being active. We investigated ranking methods for prioritizing PhGs for scaffold hopping. The proposed method: *NScaffold*, which ranks PhGs based on the number of scaffolds covered by the PhGs, outperforms other conventional methods. As a demonstrative case, using a thrombin inhibitor data set, we interpreted the highest ranked PhGs by *NScaffold* from the protein-ligand interaction point of view. It resulted that the *NScaffold* method successfully retrieved three known important interactions, showing potential for identifying scaffold hopped compounds with interpretable PhGs.

Introduction

Ligand-based drug design (LBDD), which does not use target macromolecule information, is important especially in the early stage of drug discovery. In the hit-to-lead or lead optimization phases, LBDD often tries to expand the chemical space of active compounds by synthesizing or purchasing compounds based on the knowledge of known actives. The term of "scaffold-hopping" – to identify or generate isofunctional molecules by replacing a core scaffold of an active compound with another is the embodiment of the goal.[1-4] Successful scaffold-hopping increases diversity of the active compound data set, giving opportunities of improving bioactivities and different pathways of synthesis, satisfying other requirements such as absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties[1,2] in addition to avoiding patent conflicts.

Several techniques for generating or identifying scaffold-hopped compounds from known actives have been proposed.[5-17] Some of them utilized three-dimensional molecular representations and matching algorithms.[5-8] Others developed topological (two-dimensional) molecular representations including fingerprints,[9] and combining experimental information.[10] When it comes to generating novel structure, substructure-based fragment replacement in an active molecule is also a direct approach.[11,12]

In the present study, topological molecular representations were investigated.[3,13-17] Topological representations are equivalent to chemical graph-based molecular representations. Broadly speaking, researches on scaffold-hopping based on this type of representation share the concept of pharmacophore on the chemical graphs, termed as Pharmacophore Graphs (PhGs) here. PhGs are

graph representations of pharmacophoric features (PFs), which are defined using atomic-centered rules, such as hydrogen bonding acceptor and donor points. In PhGs, nodes are PFs and their topological distances as edge distances. Schneider et al. introduced this graph-based approach.[3,13] They developed a fixed-length descriptor set, called Chemically Advanced Template Search (CATS), which were produced from frequencies of node-edge-node pairs on PhGs. Using CATS as a set of descriptors enabled fast virtual screening (VS) of databases.[3,13] They also successfully produced several *de-novo* active compounds for various targets by utilizing CATS in their *de-novo* molecular generator.[14,15] Barker et al. developed four levels of abstraction of molecular structures called "reduced graphs", which were converted to pairwise similarity or fingerprint.[16] Stiefl et al. proposed different types of abstraction algorithm of molecular structures called "extended reduced graph".[17] Unfortunately, since these methods are required to convert PhGs to fixed length vectors[3,13,16,17] or pairwise similarity[16], interpretability can be impaired to a certain extent.

PhGs or pharmacophore queries can be prioritized based on the number of compounds supporting the queries.[18,19] The elucidation of queries is particularly important for 3D-based pharmacophores due to the conformation uncertainty and molecular alignment issues.[20] On the other hand, for PhGs, which are usually treated as 2D pharmacophore fingerprints (fixed length vectors), common bits among a set of active compounds can be extracted to form the bit vector of a virtual query.[21] This method, however, may not handle combinations of many PFs with topological distances due to combinatorial explosion. Combinations of 2-3 PFs as bits in the vector were only reported.[21]

Recently, Metivier et al. took data-mining approaches to kinase inhibitors to identify important PhGs by counting the frequency of shared PhGs, although they did not focus on scaffold-hopping.[22] They discussed hierarchical relationships of PhGs and gave intuitive interpretation of the PhGs. In their research, PhGs were ranked based on the number of active compounds covered by the PhGs. Inactive compounds information could also be utilized by penalizing PhGs found among inactive compounds. These methods are explained in the Materials and Methods section.

In this study, our goal was to determine a method for identifying the PhGs that are responsible for compounds being active against the target macromolecule, leading to scaffold-hopping. For prioritizing PhGs that do not depend on molecular scaffolds, a scaffold-based scoring method was newly proposed.

We evaluated the performance for scaffold-hopping by employing PhGs as a set of queries and virtually screening new database compounds. These database compounds consisted of the scaffolds that were different from those in the training data set. We found that the choice of PhGs made remarkable differences in scaffold-hopping performances in VS. This trend was emphasized when training compound diversity is limited in terms of the number of scaffolds. As a demonstrative example, interpretation of a PhG was provided for thrombin inhibitors. We confirmed that the PhG prioritized by the proposed scoring-method contained experimentally validated three binding interactions between the macromolecule and the inhibitors.

Materials and Methods

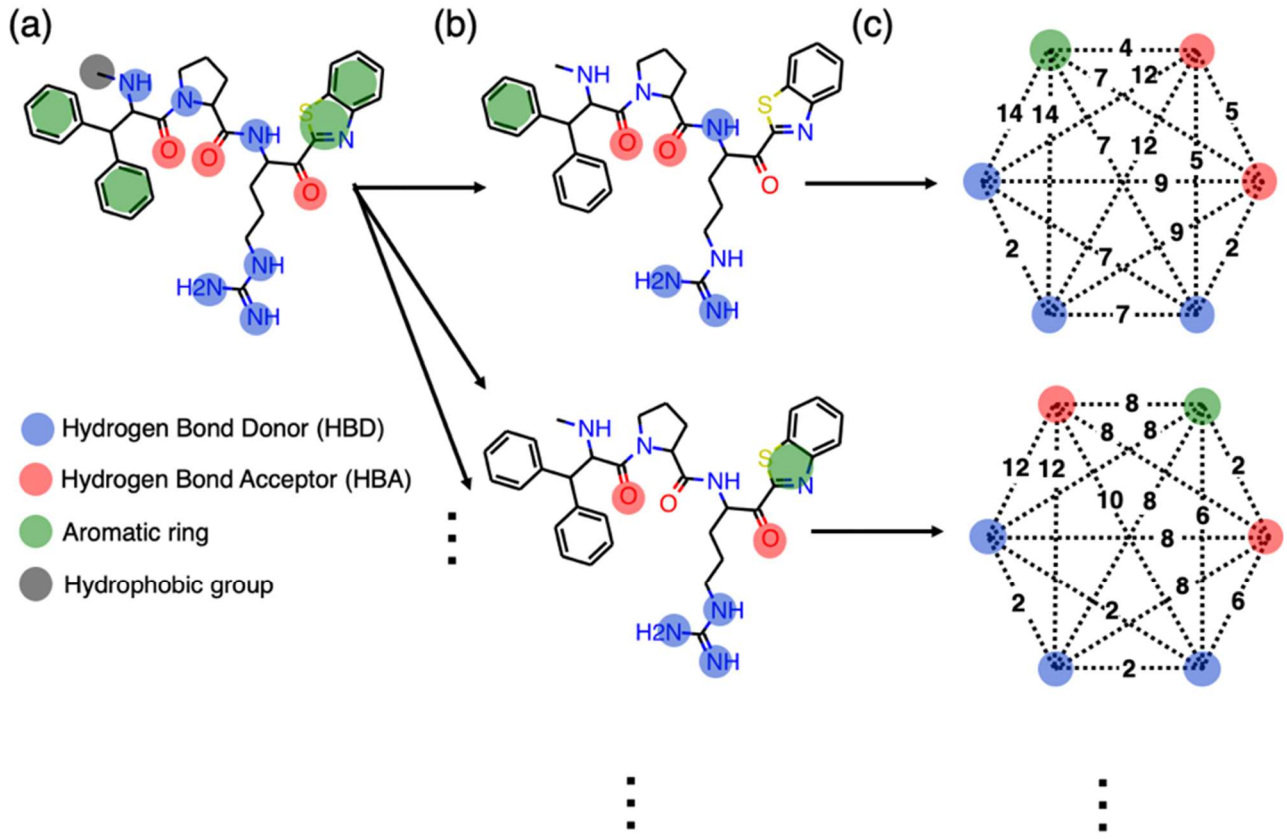
Pharmacophore graph

A pharmacophore graph is a graph-based representation of a chemical structure. It consists of pharmacophoric features (PFs)[3] as nodes and their topological distances as edges. PFs are chemical features in ligand molecules that characterize ligand-receptor interactions, such as hydrogen bonding and lipophilic interaction. They are entirely determined by the atomic environments of a focused site. Thus, linear notations like SMARTS[23] and SLN[24] are usually used for identifying the sites of potential pharmacophore points (PPPs) in molecules. In the present study, employed PFs were hydrogen bond donor (HBD), hydrogen bond acceptor (HBA), aromatic ring, positively ionizable, negatively ionizable, zinc binder, and hydrophobic. As a definition of PFs, we employed RDKit implementation described under the file name of "BaseFeatures.fdef"[25,26] except for hydrophobic features. For the hydrophobic features, we followed the definition in Métivier's study.[22] Topological distances are defined by the numbers of bonds on the shortest path between one PF and another PF. In the case that one PF contains more than one atom, distance (bond paths) is measured from the atom that is the closest to its counterpart PF.

Construction of PhGs based on an active compound is illustrated in **Figure 1**. First, PFs were assigned on sites in the compound to identify PPPs of the compound (**Figure 1a**). All the combinations of a specific number of PPPs were individually extracted and converted to PhGs (**Figure 1b and 1c)**. Only

unique PhGs were selected based on a hash representation of PhGs. In the present work, the number of PPPs in a PhG was set to six, considering computational cost.

Figure 1. Extraction of pharmacophore graphs (PhGs).



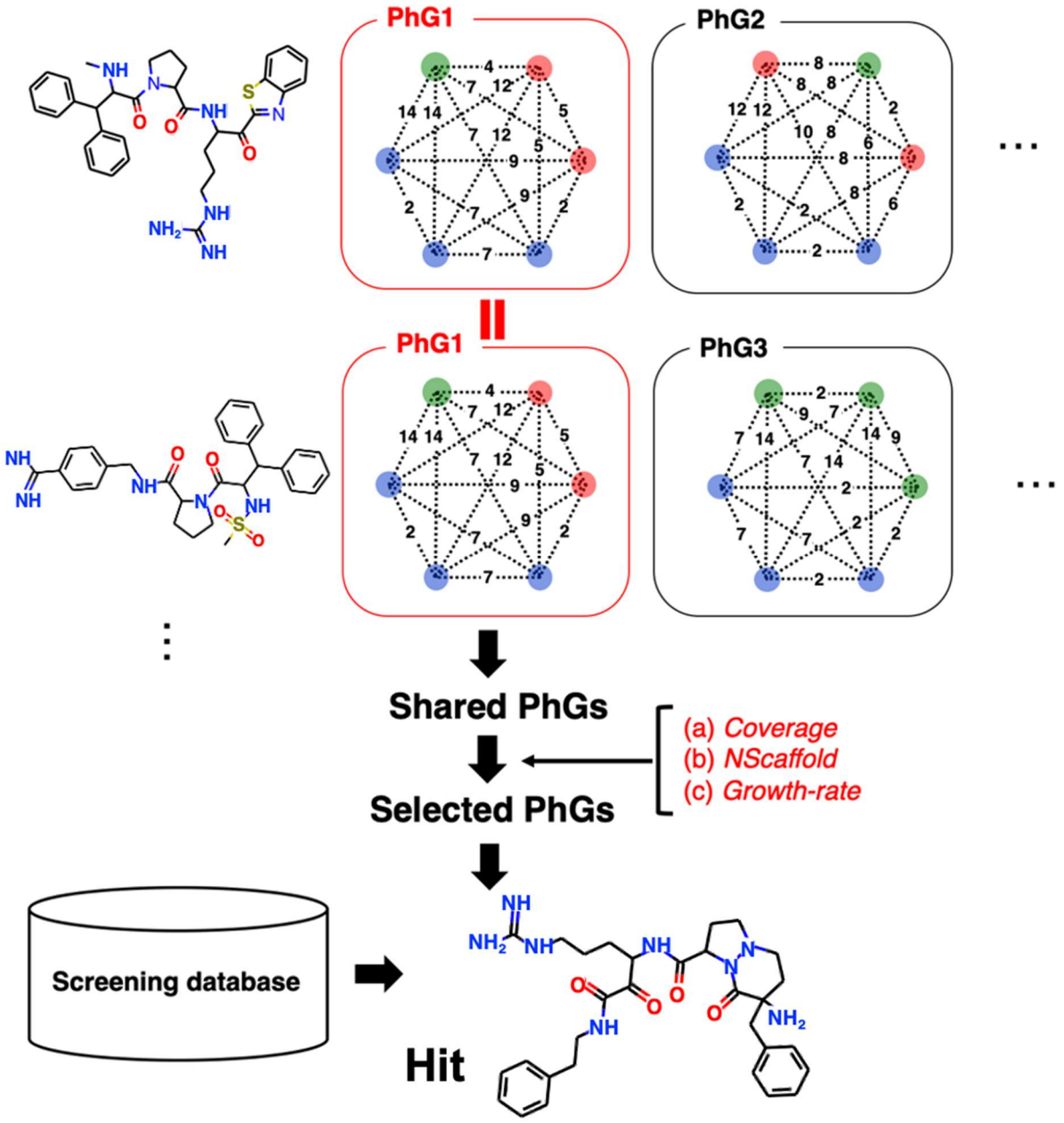
Pharmacophoric features (PFs) are assigned on components of a chemical structure (a). All the combinations of the predefined number of PFs are tested (the number of PFs is 6 in (b)). Each combination forms a pharmacophore graph (c). In pharmacophore graphs, PFs correspond to nodes, and topological distances between two PFs are regarded as the length between nodes.

Pharmacophore graphs as queries for virtual screening

A direct way of making use of PhGs in virtual screening (VS) is to employ them as queries for screening a database (**Figure 2**). In the present study, database compounds having at least one PhG in the query-set were regarded as hit compounds. Because there are a number of PhGs generated from a set of active compounds, selection of PhGs for a set of queries is important. Two conventional scoring methods to rank PhGs were tested in addition to our newly proposed one, which focuses on the identification of new chemotypes (scaffold-hopped compounds).

Execution time of generating PhGs for around 1,600 molecules and selecting PhG with one of the three criteria was around 3 hours using a single core in a desktop PC with the CPU of Intel Core-i9-7960X 2.80GHZ.

Figure 2. Pharmacophore graph (PhG) -based virtual screening.



The procedure of pharmacophore graph (PhG)-based virtual screening is described (**a**). Training active and inactive compounds are transformed into PhGs according to the procedure explained in **Figure 1**. Specific number of PhGs are chosen from a set of shared PhGs based on ranking. In the present study, we have tested three scoring methods: (a) *Coverage*, (b) *NScaffold*, and (c) *Growth-rate*. These methods are illustrated in the main text as well as in **Figure 3**.

Scoring pharmacophore graphs

Conventional scoring methods for prioritizing PhGs are *Coverage* and *Growth-rate*.[22] The *Coverage* score

$$Coverage(PhG) = N_{PhG}$$

, where N_{PhG} is the number of active compounds covered by the PhG. The *Coverage* method simply ranks PhGs based on the number of active compounds in the training set covered by the PhGs.

Growth-rate is the ratio of the *Coverage* score to the number of the inactive compounds covered by the PhG defined as:

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

, where $N_{\mathrm{inactive,PhG}}$ is the number of inactive compounds covered by the PhG. In this study, compounds were assumed to be inactive when their pK_i values were less than 6.0.

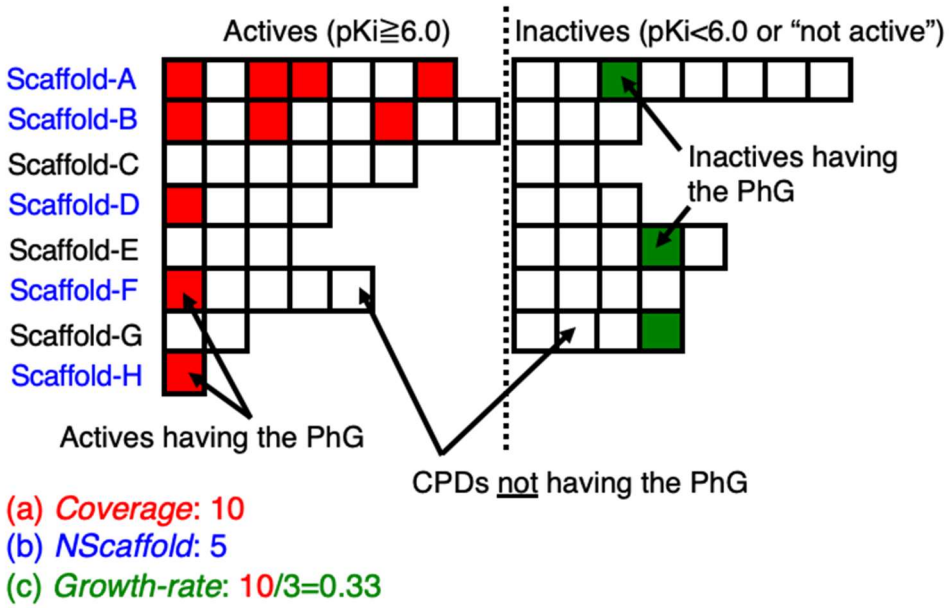
Our proposed scoring method: *NScaffold* is the number of scaffolds covered by a PhG, instead of the number of active compounds, i.e. *Coverage*. The definition of *NScaffold* is

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

Identifying PhGs that are shared by different scaffolds is expected to be equivalent to the extraction of fundamental hypotheses for activity. Therefore, selected PhGs based on *NScaffold* could identify new chemotypes, which are composed of scaffolds different from ones in the training active compounds. Calculation of three scores can be explained using **Figure 3**. In the figure, each square corresponds to one compound and squares on a row share the same scaffold. Active compounds are placed on the left

and inactive ones are on the right. Filled cells are compounds containing the PhG. For *Coverage* the score is 10, for *NScaffold* 5 by counting the number of scaffolds which include at least one active compound containing the PhG, and for *Growth-rate* 10/3 because three inactive compounds contain the PhG. All PhGs extracted from training active compounds are ranked by the scores, and the top-*N* PhGs are used as a set of queries. In this study, the numbers of *N* were set to 1, 3, 10, 30, 50, and 100.

Figure 3. Scoring methods for query Pharmacophore Graph (PhG) selection.



Three methods for selecting PhGs are illustrated. In *Coverage*, PhGs are ranked based on the number compounds covered by the graphs. The *Coverage* score for the exemplified PhG is ten (**a**). In *NScaffold*, the numbers of scaffolds covered by a PhG is the score. Here, five scaffolds are covered by the PhG (**b**). Growth-rate is the ratio of the *Coverage* to the number of the inactive compounds covered by the PhG (**c**).

Compound data sets

As shown in **Table 1**, six macromolecular targets were selected from the ChEMBL database[27] version 24 based on the numbers of compounds listed for the targets, and macromolecular types: two kinases, two proteases, and two G-protein coupled receptors. Bioactive compounds with high confidence score were extracted and annotated with equilibrium constant (K_i) values. When multiple K_i values were available for a compound, their geometric mean was calculated to yield its final potency value as long as all the values fell into the same order of magnitude. Inactive compounds were extracted from the ChEMBL database as well, those which were reported as 'not active' or 'inactive' for each target. In this study, compounds whose pK_i values were less than 6.0 were also regarded as inactive according to the previous research.[22] Compounds with the molecular weight of less than 100 and more than 800 were discarded. During the calculation of *NScaffold* scores, we tested two types of scaffolds: Bemis-Murcko (BM) scaffolds[28] and scaffolds based on compound-core relationships (CCR).[29] For deriving CCR-based scaffolds, covalent bonds in molecules were dissected when they are single (CCR Single) or defined by retrosynthesis rules (CCR RECAP).[29,30] Data set profiles are reported in **Table 1**. For the six targets, the number of scaffolds ranged from 209 to 1380 in the BM scaffolds definition, 101 to 1041 in CCR Single and 262 to 1737 in CCR RECAP. As negative compounds, 250,000 compounds were randomly selected from ZINC version 15.[31]

Table 1. Compound data set profiles

CHEMBL ID [4]	Target	Code	#CPDs [a] (Highly potent CPDs [b])	#Scaffolds		
				Bemis Murko	CCR Single	CCR RECAP
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: all compounds for a target in the present study including active and inactive ones.

[b] Highly potent CPDs: compounds exhibiting pK_i values greater than or equal to 6.0.

Bemis Murko: Bemis Murko scaffold using RDKit.[25,28]

CCR Single: Compound core-relationship (CCR)-based scaffolds.[29]

CCR RECAP: CCR-based scaffolds, but bonds were dissected based on the RECAP rules.[30]

Performance measures for scaffold-hopping

An effectiveness of VS can be measured by the ability of identifying new chemotypes (scaffold-hopped compounds). Therefore, for each target, training and test data sets were compiled so that there were no shared scaffolds. Scaffold combinations for training/test split were randomly chosen 10 times for each of the three scaffold definitions: BM, CCR Single, or CCR RECAP. [28-30] A total of thirty trials were averaged in VS performances. The tested ratios between training and test scaffolds were 1:1 and 1:4. In the latter case, a limited number of scaffolds were included in a training data set on the assumption of the initial stages of drug design. Screening performances were evaluated using the values of correctly identified active compounds (TP), of wrongly predicted active ones (FN) and of wrongly predicted inactive compounds in the ZINC database (FP). Precision (TP/(TP+FP)) and recall (TP/(TP+FN)) were used metrics. In general, the two metrics show a trade-off relationship and the number of PhGs in a query set is the parameter affecting balances of the trade-off. Using only the top-1 PhG is expected to show a high precision value at the expense of recall. On the other hand, using top-100 PhGs could identify more active compounds at the risk of increasing FPs. Because hit compounds were defined as the ones containing at least one of the PhGs, metrics for standard VS performance evaluation, such as AUC-ROC, were not employed in this study. Therefore, visualization of relationships between precision and recall is a way to evaluate screening performances.

Results and Discussion

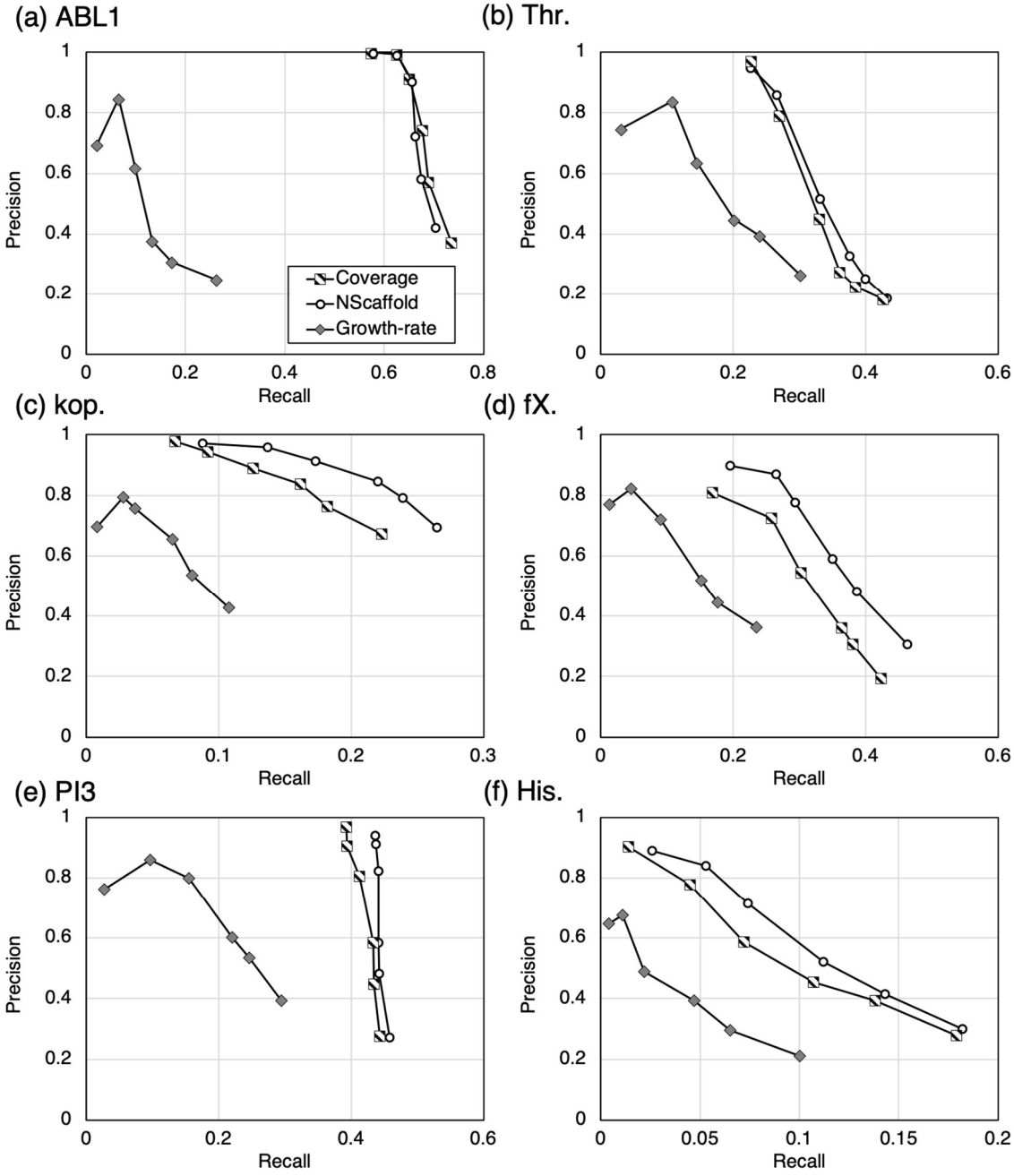
Study design

Topological representations of PFs have been studied for extracting potential characteristic of protein-ligand interaction in ligand-based approaches.[3,13,16,17,22] However, no prior researches have focused on ranking methods for prioritizing PhGs as a set of queries. PhGs should be of activity relevance because there could be a number of PFs generated from a molecule, leading to many false-positive PhGs, which are not activity relevant. Using true-positive PhGs is important for identifying active compounds in different chemotypes. In the present study, three scoring methods were compared with respect to the ability of identifying active compounds with different scaffolds from those in the training data set. Ranking PhGs based on the number of scaffolds specified by the PhGs instead of active compounds, i.e. *NScaffold,* was newly proposed aiming at identifying (designing) scaffold-hopped compounds. Three definitions of scaffolds were tested and two ratios of training and test data sets were tested: 1:1 and 1:4 for the six various target macromolecules. The number of PhGs as a query set varied from 1 to 100. For each combination of target macromolecules, training-test ratios, targets, scoring methods and the numbers of queries, 10 randomly split data sets for each scaffold-type[28-30] were tested for ensuring statistical validness.

Comparing screening methods

Figure 4 reports the precision recall curves for the three scoring methods by changing the number of PhGs as a set of queries (*N* = from 1 to 100) when training and test data set sizes were identical (1:1). In **Figure 4**, the marks (squares, circles, and diamonds) of the lowest recall correspond to using a single query *N*=1, and recall values increase as *N* increases. Six marks on each line correspond to *N* = 1, 3, 10, 30, 50, and 100, respectively. The VS performances demonstrated trade-off relationships between recall and precision. For four out of six targets: κ-opioid receptor, factor Xa inhibitors, PI3-kinase p110-alpha subunit inhibitors, and Histamine H3 receptor ligands, the *NScaffold* method consistently outperformed the other scoring methods. This indicates that the *NScaffold* method can extract more scaffold-independent features than the other two methods. In other words, PhGs selected with the *Coverage* and *Growth-rate* method might possess scaffold-dependent characteristics. However, the performances of *Coverage* in Tyrosine kinase ABL1 outperformed *NScaffold* and the performances by the *Coverage* and *NScaffold* methods for Thrombin inhibitors had no significant difference.

Figure 4. Target-wise comparison of three types of queries in screening performance (training: test = 1:1)



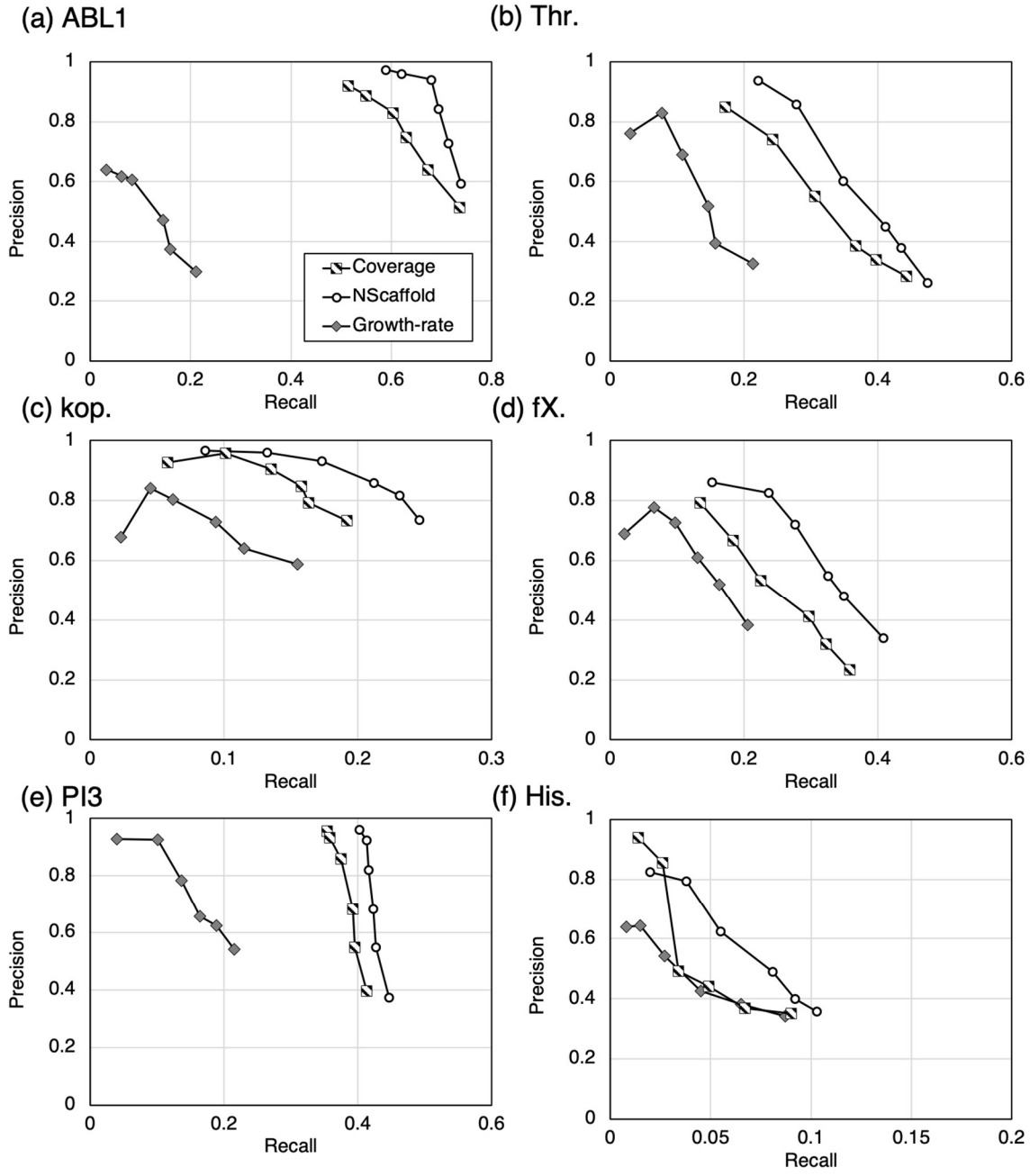
Reported are target-basis screening performances in terms of averaged precision and recall by changing the number of PhGs as a query set (from 1 to 100). The procedure was repeated 30 times. Training and test data sets were randomly split to with the ratio of 1:1 on the basis of BM, CCR Single,

or CCR RECAP scaffolds. Ten trials were based on BM scaffold, another ten times on CCR Single, and the other ten times on CCR RECAP.

Figure 5 reports the precision recall curves when training and test data set sizes were 1:4. In this condition, the number of scaffolds in a training data set was reduced to 40% compared to **Figure 4**. **Figure 5** clearly showed that *NScaffold* was superior to the other ranking criteria: *Coverage* and *Growth-rate*. For factor Xa inhibitors and for PI3-kinase p110-alpha subunit inhibitors, the gaps in the performances between *NScaffold* and the other methods were widened compared to equally split data sets (**Figure 4**). Moreover, for the thrombin inhibitors, *NScaffold* could identify most active compounds with different scaffolds even using a single PhG. The only exception was the *Coverage* method for Tyrosine kinase ABL1 inhibitors with N=100. The performance of *Coverage* was comparable with that of *NScaffold*.

These results indicated that PhGs identified in high rank by the *NScaffold* method could become useful queries for finding scaffold-hopped compounds compared to the other conventional methods, in particular, when using a limited number and diversity of training compounds. It should be noted that among the six targets, for histamine H3 and and κ-opioid receptor, the three methods gave the lowest performances (**Figure 4** and **Figure 5)** in terms of recall. For these two targets, the numbers of scaffolds were greater than those for the other targets. For histamine H3, the number of BM (CCR Single) scaffolds was 1268 (667), for κ-opioid receptor 1380 (1041). On the other hand, thrombin, for which the number of BM (CCR Single) scaffolds was the third-greatest: 884 (659), exhibited relatively high recall. This target is characterized by its sold binding pocket, which will be discussed later.

Figure 5. Target-wise comparison of three types of queries in screening performance (training: test = 1:4)



Reported are target-basis screening performances in terms of averaged precision and recall by changing the number of PhGs as a query set (from 1 to 100). The procedure was repeated 30 times according to those in **Figure 4**.

Comparison with Conventional Similarity Search

PhGs extracted from a training data set can be utilized as queries for VS as well as for interpretation of the common PPPs among active compounds. As a control calculation, conventional similarity search (1-NN) was carried out using all active compounds in the training data set. Morgan fingerprints as a 2048-bit vector form were used as molecular representations[25] and Tanimoto similarity was as a metric for similarity calculation. For this benchmark calculation, one pair of training and test active compound data sets was randomly compiled so that the scaffold ratio of the training data set to the test was 1:4 (CCR Single). As a source of inactive compounds for VS, either the ZINC data set or the inactive compound data set was prepared for each of the six targets.

Table 2 reports the precision and recall of the active compounds when using the top PhG ranked by *NScaffold* for each target. For similarity searching, the precision value at the same recall value as using PhG is reported. Using the best PhG selected by the *NScasffold* method showed precision values as high as using conventional similarity searching (**Table 2a**). However, for the data sets consisting of inactive and active compounds, using the best PhG as a query showed higher precision values than those using similarity searching (**Table 2b**).

Table 2. Precision for active compounds using the best PhG as a query and the corresponding similarity search performance. (a) Database consisted of ZINC 250,000 negative compounds and active compounds. (b) Database consisted of inactive and active compounds.

(a)

Code	Recall	Precision	
		PhG	Similarity Search
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)

Code	Recall	Precision	
		PhG	Similarity Search
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

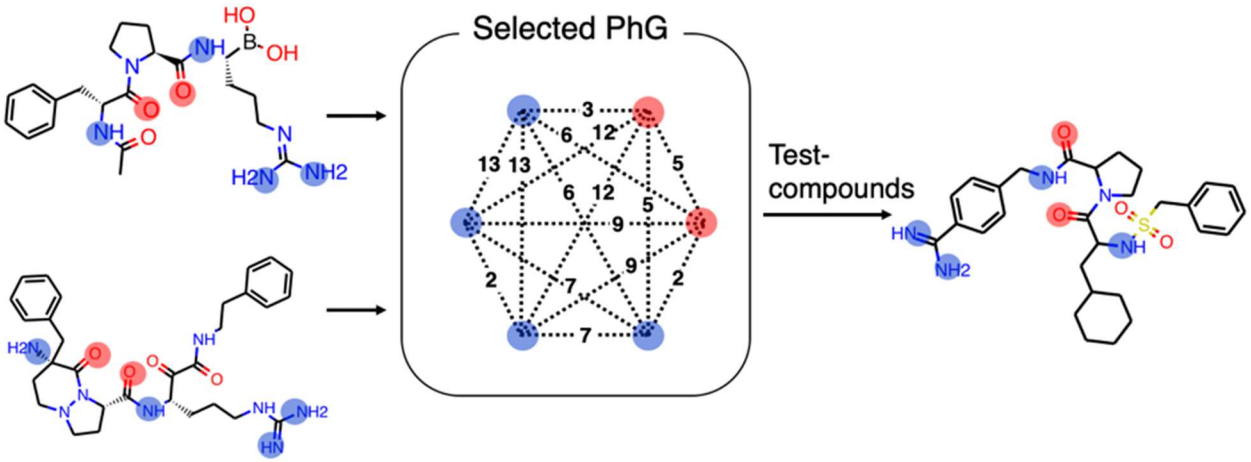
Identified active compounds

Figure 6 reports the top PhGs identified by the *NScaffold* (**Figure 6a**) and *Coverage* (**Figure 6b**) methods for thrombin inhibitors using 1:4 splitting of the data set based on CCR Single. In the figure, exemplary training (**left**) and a test (**right**) compounds are also reported. Using the *NScaffold* method, four hydrogen bond donors and two acceptors were selected. All of the six PPPs were consistent with the previously investigated binding mode of thrombin inhibitors. According to structure-based interpretation of the PPPs, two hydrogen bond donors with distance two corresponded to an arginine residue in a thrombin inhibitor, which interacts with the carbonyl oxygens of Asp189.[32-36] A pair of a hydrogen bond donor and an acceptor with distance two (amide group), which are seven paths apart from the NH_2 in the arginine residue was reported to direct toward the residue of Ser214,[32] and the other carbonyl parts of the inhibitors were reported to interact with Gly216.[33] These features were proved to be shared in many highly active thrombin inhibitors. This explanation was also confirmed with the bound conformations measured by X-ray crystallography and registered in Protein Data Bank,[34] such as 3UTU,[33] 1C4V,[35] and 1LHC.[36] On the other hand, the *Coverage* method seemed to fail to retrieve common features for protein-inhibitor interaction, sticking to a great number of active compounds with limited scaffolds, i.e. analogous compounds. In **Figure 6b**, the top PhG contained one aromatic ring, one lipophilic part, two hydrogen bond donors and acceptors. The PhG was shared by the great number of training active compounds. Since many analog active compounds were found in the training data set, it was unlikely to identify active test compounds with different scaffolds using

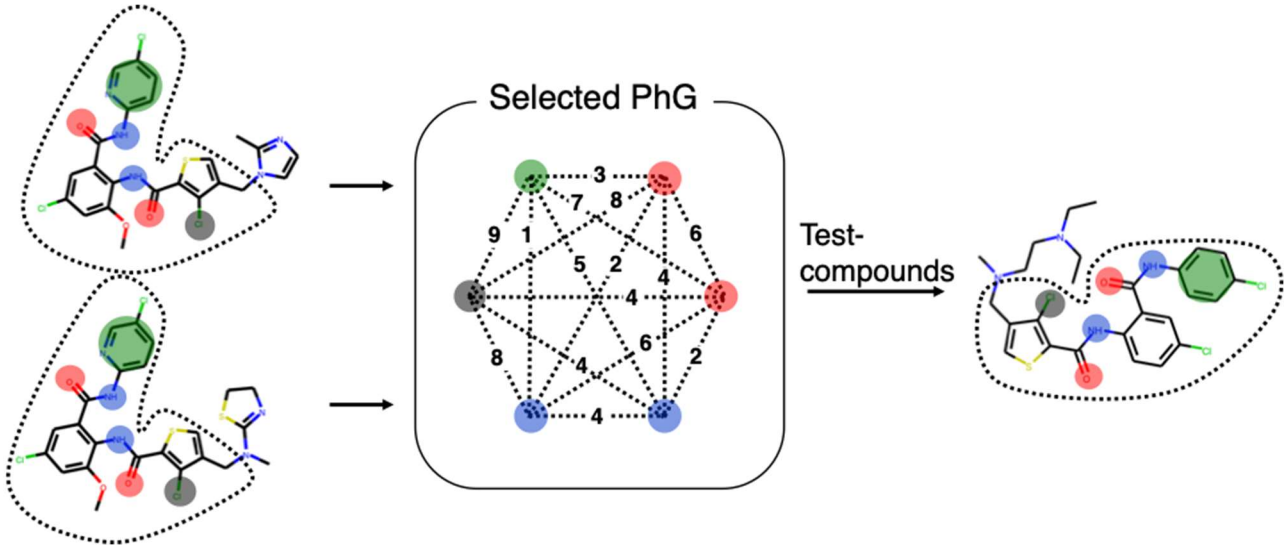
the PhG, which was supported by the fact that the test and training compounds shared a substructure (**Figure 6b**). In the PhG, two PFs were not hydrogen-bonding related. One was aromatic ring which would usually make a weak interaction, and one hydrophobic feature which might bring about dispersive interactions in addition to entropic effect.[37,38]

Figure 6. Top pharmacophore graphs (PhGs) and exemplified test compounds matching the PhGs for thrombin inhibitors.

(a)



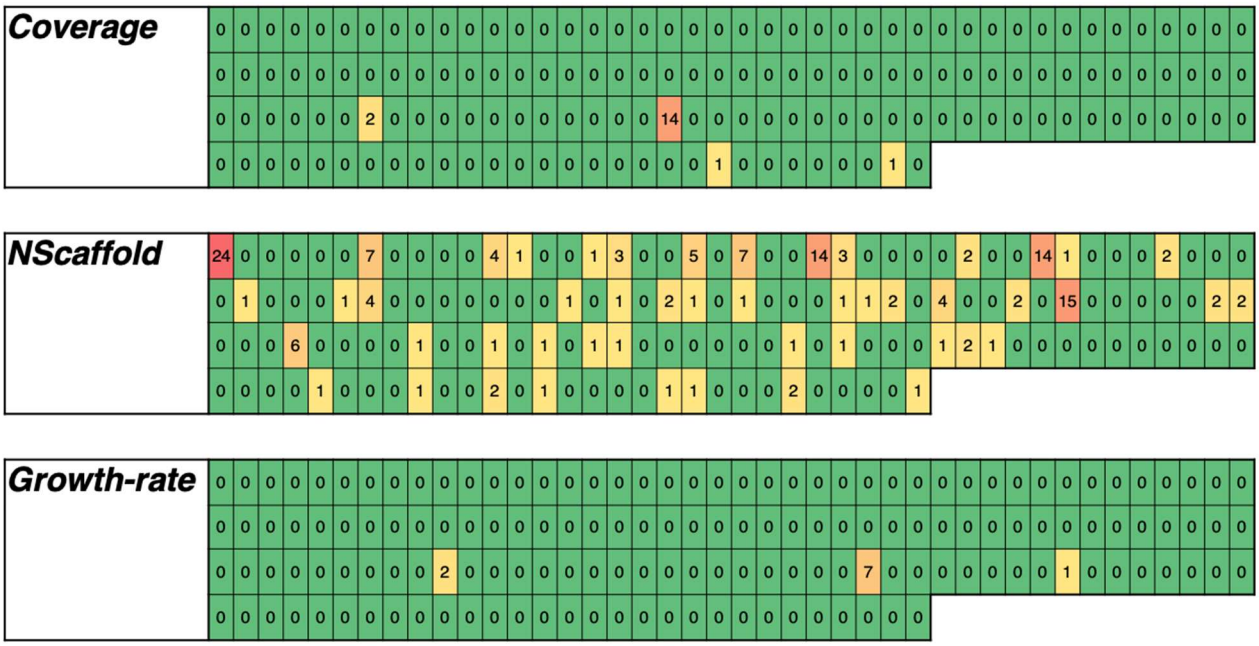
(b)



For thrombin, the top PhGs based on *NScaffold* (a) and *Coverage* (b) selection criteria are reported as selected PhGs with training compounds consisting of the PhGs. On the rightmost part, exemplified hit compounds are illustrated.

Distributions of selected test compounds and scaffolds were visualized in **Figure 7**. The same target, data sets, and the PhG were employed in this figure as for **Figure 6**. For this trial in the test data set, the number of scaffolds and compounds were 155 and 512, respectively. The PhG identified by the *Coverage* method, which covered 31.8 percentage of the training active compounds, could identify only four test scaffolds, matching 18 active compounds. The PhG determined by the *Growth-rate* method covered 25.0 percentage of training active compounds and could identify only three test scaffolds only with ten active compounds. However, the PhG identified by *NScaffold*, which covered 22.7 percentage of the training active compounds, could identify 49 active scaffolds distinct from those in the training data set. In this exemplary case, the PhG ranked as the top by the *NScaffold* method successfully retrieved universal features of protein-ligand interaction. Compared to *NScaffold* method, the other two methods only detected a limited type of scaffolds in the test data set.

Figure 7. Covered compounds and scaffolds.



For thrombin, using three types of queries: *Coverage*, *NScaffold*, and *Growth-rate*, covered scaffolds and number of compounds in a test data set are reported. One cell represents an individual scaffold, and a number on the cell the number of test compounds consisting of the scaffold. Ten PhGs were used as a set of queries for screening.

Conclusions

Herein, we have investigated graph representations of potential pharmacophore points (PPPs) as well as scoring methods, aiming at the identification of compounds consisting of scaffolds different from those in the training data set. The proposed *NScaffold* method selects pharmacophore graphs (PhGs) based on the number of scaffolds instead of based on the number of active compounds. The method fundamentally tries to identify PhGs which do not depend on the molecular scaffolds. As a proof of concept study, PhG identification and virtual screening for six bioactive targets from different target classes were carried out. It resulted that the method displayed preferable precision and recall relationships to the *Coverage* and *Growth-rate* methods. When training data-set size was getting smaller, which corresponds to the initial stage of drug discovery, the *NScaffold* method was far superior to the other methods. As a demonstrative case for thrombin inhibitors, the proposed method successfully prioritized previously investigated hydrogen bonding against the thrombin residues Asp189, Ser214, and Gly216. This result supported that the PhGs selected by the method may provide an interpretable scaffold-independent PPPs. The method is very simple and no additional parameters are needed. We hope that combining PhGs with the scoring method would further provide opportunity for the identification of scaffold-hopped active compounds in drug discovery projects.

Acknowledgements

We thank OpeneEye Scientific Sofrware, Inc., for providing a free academic license of the OpenEye chemistry toolkits. We thank Dr. Swarit Jasial for carefully proofreading the manuscript. We also appreciate Prof. Dr. Jürgen Bajorath for allowing us to use the CCR method developed in his group.

References

1. Schneider, G.; Schneider, P.; Renner, S. Scaffold-Hopping: How Far Can You Jump? *QSAR Comb. Sci.* **2006**, *25*, 1162–1171.
2. Böhm, H. J.; Flohr, A.; Stahl, M. Scaffold Hopping. *Drug Discov. Today Technol.* **2004**, *1*, 217–224.
3. 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.
4. Hu, Y.; Stumpfe, D.; Bajorath, J. Recent Advances in Scaffold Hopping. *J. Med. Chem.* **2017**, *60*, 1238–1246.
5. 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.

6. Renner, S.; Schneider, G. Scaffold-Hopping Potential of Ligand-Based Similarity Concepts. *ChemMedChem* **2006**, *1*, 181–185.

7. 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.

8. Grisoni, F.; Merk, D.; Byrne, R.; Schneider, G. Scaffold-Hopping from Synthetic Drugs by Holistic Molecular Representation. *Sci. Rep*. **2018**, *8*, 16469.

9. Rabal, O.; Amr, F. I.; Oyarzabal, J. Novel Scaffold Fingerprint (SFP): Applications in Scaffold Hopping and Scaffold-Based Selection of Diverse Compounds. *J. Chem. Inf. Model.* **2015**, *55*, 1–18.

10. 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.

11. Vainio, M. J.; Kogej, T.; Raubacher, F.; Sadowski, J. Scaffold Hopping by Fragment Replacement. *J. Chem. Inf. Model.* **2013**, *53*, 1825–1835.

12. Lamberth, C. Agrochemical Lead Optimization by Scaffold Hopping. *Pest Manag. Sci.* **2018**, *74*, 282–292.

13. Reutlinger, M.; Koch, C. P.; Reker, D.; Todoroff, N.; Schneider, P.; Rodrigues, T.; Schneider, G. Chemically Advanced Template Search (CATS) for Scaffold-Hopping and Prospective Target Prediction for 'Orphan' Molecules. *Mol. Inform.* **2013**, *32*, 133–138.

14. Hartenfeller, M.; Zettl, H.; Walter, M.; Rupp, M.; Reisen, F.; Proschak, E.; Weggen, S.; Stark, H.; Schneider, G. Dogs: Reaction-Driven *de Novo* Design of Bioactive Compounds. *PLoS Comput. Biol.* **2012**, *8*, e1002380.

15. Rodrigues T.; Roudnicky F.; Koch C. P.; Kudoh T.; Reker D.; Detmar M; Schneider G. *De novo* design and optimization of Aurora A kinase inhibitors. *Chem. Sci*. **2013**, *4*, 1229–1233.

16. 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.

17. Stiefl, N.; Watson, I. A.; Baumann, K.; Zaliani, A. ErG: 2D Pharmacophore Descriptions for Scaffold Hopping. *J. Chem. Inf. Model.* **2006**, *46*, 208–220.

18. Podolyan, Y.; Karypis, G. Common Pharmacophore Identification Using Frequent Clique Detection Algorithm. *J. Chem. Inf. Model.* **2009**, *49*, 13–21.

19. Schneidman-Duhovny, D.; Dror, O.; Inbar, Y.; Nussinov, R.; Wolfson, H. J. Deterministic Pharmacophore Detection via Multiple Flexible Alignment of Drug-Like Molecules. *J. Comput. Biol.* **2008**, *15*, 737–754.

20. 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.

21. Hoksza, D.; Škoda, P. 2D Pharmacophore Query Generation. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer Verlag, **2014**; Vol. 8492, pp 289–300.

22. 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.

23. Daylight Chemical Information Systems, Inc. Daylight Theory Manual. http://www.daylight.com/dayhtml/doc/theory/index.html (accessed Dec 28, 2019)

24. 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.

25. Landrum, G., RDKit: open-source cheminformatics software http://www.rdkit.org (accessed Dec 28, 2019)

26. https://github.com/rdkit/rdkit/blob/master/Data/BaseFeatures.fdef (accessed Dec 28, 2019)

27. 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.

28. Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.

29. 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.

30. 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.

31. Sterling, T.; Irwin, J. J. ZINC 15 – Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, 55, 2324–2337.

32. 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.

33. 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.

34. 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.

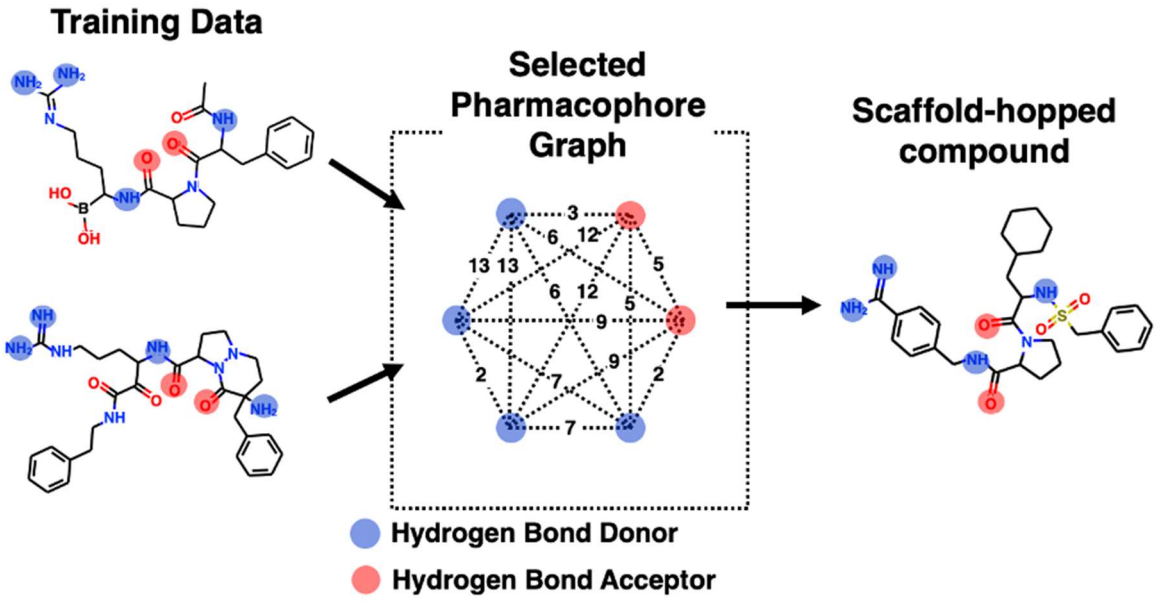
35. Krishnan, R.; Mochalkin, I.; Arni, R.; Tulinsky, A. Structure of Thrombin Complexed with Selective Non-Electrophilic Inhibitors Having Cyclohexyl Moieties at P1. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2000**, *56*, 294–303.

36. 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.

37. Desiraju, G. R. Hydrogen Bridges in Crystal Engineering: Interactions without Borders. *Acc. Chem. Res.* **2002**, *35*, 565–573.

38. Schneider, G.; Baringhaus, K. H. *Molecular Design. Concepts and Applications*; Wiley-VCH: Weinheim, Germany, **2008**; pp. 54–57.

TOC


Training Data
Selected Pharmacophore Graph
Scaffold-hopped compound
Hydrogen Bond Donor
Hydrogen Bond Acceptor