ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

Visualization of Topological Pharmacophore Space with Graph Edit Distance

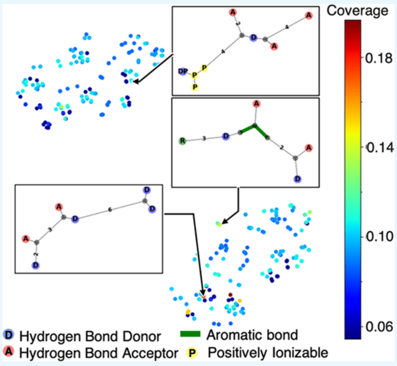
Hiroshi Nakano and Tomoyuki Miyao*

Cite This: *ACS Omega* 2022, 7, 14057−14068

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: A topological pharmacophore (TP) is a chemical graph-based pharmacophore representation, where nodes are pharmacophoric features (PF) and edges are topological distances between PFs. Previously proposed sparse pharmacophore graphs (SPhGs) for TPs were shown to be effective in identifying structurally different active compounds while maintaining the interpretability of the graphs. However, one limitation of using SPhGs as queries is that many structurally similar SPhGs can be identified from a set of active compounds, requiring the classification and visualization of SPhGs, followed by an understanding of the pharmacophore hypotheses. In this study, we propose a scheme for SPhG analysis based on dimensionality reduction techniques with the graph edit distance (GED) metric. This metric enables measuring similarities among SPhGs in a quantitative manner. The visualization of SPhGs, which themselves are the graphs shared by active compounds, can help us understand the pharmacophore hypotheses as well as the data set. As a proof-of-concept study, we generated two-dimensional SPhG-maps using three dimensionality reduction techniques for six biological targets. A comparison with other pharmacophore representations was also conducted. We demonstrated knowledge extraction (interpretation of the data set) from the generated maps. Our findings include a suitable mapping algorithm as well as a pharmacophore hypothesis analysis procedure using an SPhG-map.



INTRODUCTION

A ligand-based pharmacophore is the geometrical arrangement of chemical features in a ligand molecule responsible for molecular interactions against the target macromolecule.[1] Pharmacophoric features (PFs) consist of single atoms or sets of atoms based on interaction types, such as hydrogen bonding and lipophilic interactions. Because a pharmacophore can be regarded as an interaction hypothesis, it can be used as a query for screening chemical libraries.[2−7]

Topological pharmacophore models, coined by Schneider et al., employ chemical graph paths as the distance function among PFs.[8,9] On a chemical graph, nodes and edges correspond to atoms and covalent bonds, respectively.[8,9] The distance between PFs is simply the number of covalent bonds on the path, ignoring bond lengths and types. The distances were proposed as "separation" by Smith et al.[10] In contrast to geometry-based pharmacophores, for which plausible conformations are necessary,[11−13] topological pharmacophores are rigorous and at the same time provide less information on three-dimensional molecular coordinates. They are also less computationally demanding and can be applied to large-scale data sets. Sophisticated multidimensional molecular descriptors embodying this representation have been successfully utilized for identifying novel hit compounds in prospective virtual screening campaigns.[9,14]

Graph representations of the topological pharmacophores, termed pharmacophore graphs (PhGs),[8,9] have been demonstrated to classify a large-scale data set of BCR-ABL tyrosine kinase.[15] A PhG is a complete graph with PFs as nodes and their topological distances on edges. Sharing of PhGs with a number of active compounds becomes the interaction hypothesis against the target macromolecule. In a series of retrospective validation studies, we found that the PhGs extracted from a compound set containing many unique scaffolds became useful queries for identifying structurally different active compounds from the training compounds.[16] In other words, the PhGs shared by diverse active compounds can be useful for scaffold hopping (SH).[8,9,17−19] SH requires a method to capture the functional similarity with a focus on the interaction and freedom from scaffold-based or *structural* similarity.[8,9]

One drawback of PhGs is the lack of interpretability. Complete graphs are hard to interpret because all of the nodes (PFs) are connected to the rest of the nodes. Translating PhGs to the corresponding chemical graphs is not straightforward. In this respect, reduced-graph forms with a small number of edges are preferable. In our previous study, a sparse form of PhGs, termed sparse PhGs (SPhGs), was proposed. We also showed

Received: January 30, 2022
Accepted: March 25, 2022
Published: April 12, 2022

that SPhGs had much fewer edges (close to tree structure) than PhGs, while they were slightly inferior to the PhGs in terms of screening performances.[20] One major difference between SPhGs and other reduced graphs[21−23] in addition to extraction algorithms is that SPhGs are shared graphs found in multiple active compounds, meaning each SPhG manifests a pharmacophore hypothesis. However, one limitation of a set of SPhGs as pharmacophore hypotheses is that there exist a large number of similar SPhGs for a set of active compounds. The classification and visualization of SPhGs are necessary for understating the data set.

A set of PhGs can be visualized as a network where nodes are PhGs and edges are the parent−child relation of PhGs.[15] A child PhG is created by adding a PF to the parent PhG. This type of connection is effective for visualizing a course of PhG development. However, no connection is detected between PhGs with the same number of PFs and slightly different edge distances. Like compound visualization in the field of chemography,[24] a proper similarity measurement (metric) is necessary for understanding a set of PhGs by visual inspection.

In this study, we propose to visualize SPhGs using the graph edit distance (GED) to understand topological pharmacophore relations in a set of active compounds by clustering analysis. GED was previously employed to quantitatively compare reduced graphs for similarity searching.[25] In the previous study, reduced graphs were generated to compare the corresponding compounds: each reduced graph corresponds to a single compound. However, we focus on the visualization of SPhGs, which themselves are common features among active compounds, and visualizing them leads to understanding the relations of the pharmacophore hypotheses. Active compounds against six target macromolecules were analyzed with the proposed methods and the extracted features, and interpretability is discussed.

Python scripts for creating SPhGs from a set of compounds and clustering, including outputs for the six targets in this study, are available in the open access repository: github.com/n-hiroshi/sphg2.

■ MATERIALS AND METHODS

Compound Data Sets. Active compounds for six biological targets were extracted from the ChEMBL database (version 24).[26] The number of highly potent compounds for each data set is listed in Table 1, along with the abbreviation and the CHEMBL ID. The selected targets were Thrombin (Thr.), Tyrosine kinase ABL1 (ABL1), κ-opioid receptor (Kop.), PI3-kinase p100-α subunit (PI3), G protein-coupled receptor 44 (GPCR44), and transmembrane protease serine 6 (TPS6) on the basis of protein family types. Active compounds with more than or equal to 6.0 in terms of pK_i were regarded as highly potent, except for TPS6, whose potency threshold was lowered to 5.0 due to the limited number of eligible compounds (only three if 6.0 was employed). ChEMBL records with a confidence score of 9 were only processed. When multiple pK_i values were available for a single compound, the arithmetic mean was calculated to yield its final potency value as long as all of the values fell into the same order of magnitude; otherwise, the compound was discarded. Compounds with molecular weights between 200 and 600 were used for subsequent analyses to reduce computational burden and to remove compounds with extreme properties. All of the highly potent compounds for the six targets are provided as SMILES with curated pK_i values in the open access repository github.com/n-hiroshi/sphg2. For TPS6, highly potent active compounds with pK_i values are visualized in Figure 1.

Table 1. Compound Data Sets

ChEMBL ID	target	code	#Highly potent CPDs[a]
CHEMBL204	thrombin	Thr.	514
CHEMBL1862	tyrosine kinase ABL1	ABL1	515
CHEMBL237	κ-opioid receptor	Kop.	1425
CHEMBL2498	PI3-kinase p110-α subunit	PI3	812
CHEMBL5701	G protein-coupled receptor 44	GPCR44	686
CHEMBL1795139	transmembrane protease serine 6	TPS6	21

[a] Highly potent CPDs: compounds exhibiting pK_i values greater than or equal to 6.0 except for TPS6 (5.0).

TP Representations. Three representations for topological pharmacophores were tested: conventional pharmacophore fingerprints (PhFP),[27] molecular sparse pharmacophore graphs (Mol-SPhGs), and sparse pharmacophore graphs (SPhGs).[20] PhFP is a bit vector, and the other representations are graphs. The three representations are illustrated in Figure 2.

PFs. A PF is a chemical feature of a ligand molecule that characterizes an interaction between the ligand and the target macromolecule. Such interactions include hydrogen bonding and electrostatic interactions. Consequently, hydrogen bond donor (HBD), hydrogen bond acceptor (HBA), aromatic rings, and positive/negative ionizable groups are frequently used as PFs. In this study, we employed the RDKit implementation described under the file name of “BaseFeatures.fdef” to identify atoms or groups of atoms matching PFs.[27,28]

PhFPs. A PhFP is a set of combinations of PFs with the topological distances among them.[27] Each combination represents the pharmacophore pattern containing a fixed number of PFs and the distances among them, forming a bit in the fingerprint vector (Figure 2b). In a bin, a range of distance instead of an exact distance takes bond length ambiguity into account. For avoiding combinatorial explosion and too sparse bit vectors, the number of PFs is usually limited to 3 and the maximum distance to 8.[27] Similar atom-pair-based fingerprints are proposed by Capecchi et al.[29] In this study, the RDKit function of topological pharmacophore with the default parameter values was used.[27]

Mol-SPhGs. A Mol-SPhG is a reduced graph of a chemical graph. Nodes of Mol-SPhG are PF-assigned atoms (termed PF nodes) or junction atoms.[20] The junction atoms are nodes without PFs, which are introduced to keep the original distances among PF nodes. Because Mol-SPhG holds the topological distance between every pair of PF nodes, no information is lost in terms of TPs (Figure 2c). Details of the construction algorithm of Mol-SPhG from a chemical graph have been reported by our group.[20]

SPhGs. An SPhG is a sparse representation of a TP in terms of the number of edges and nodes (Figure 2d).[20] This form of pharmacophore has a good balance of trade-off between intuitive understanding of the TP and keeping topological distances among PF nodes. The previous study using an active compound data set for thrombin showed that more than 90%

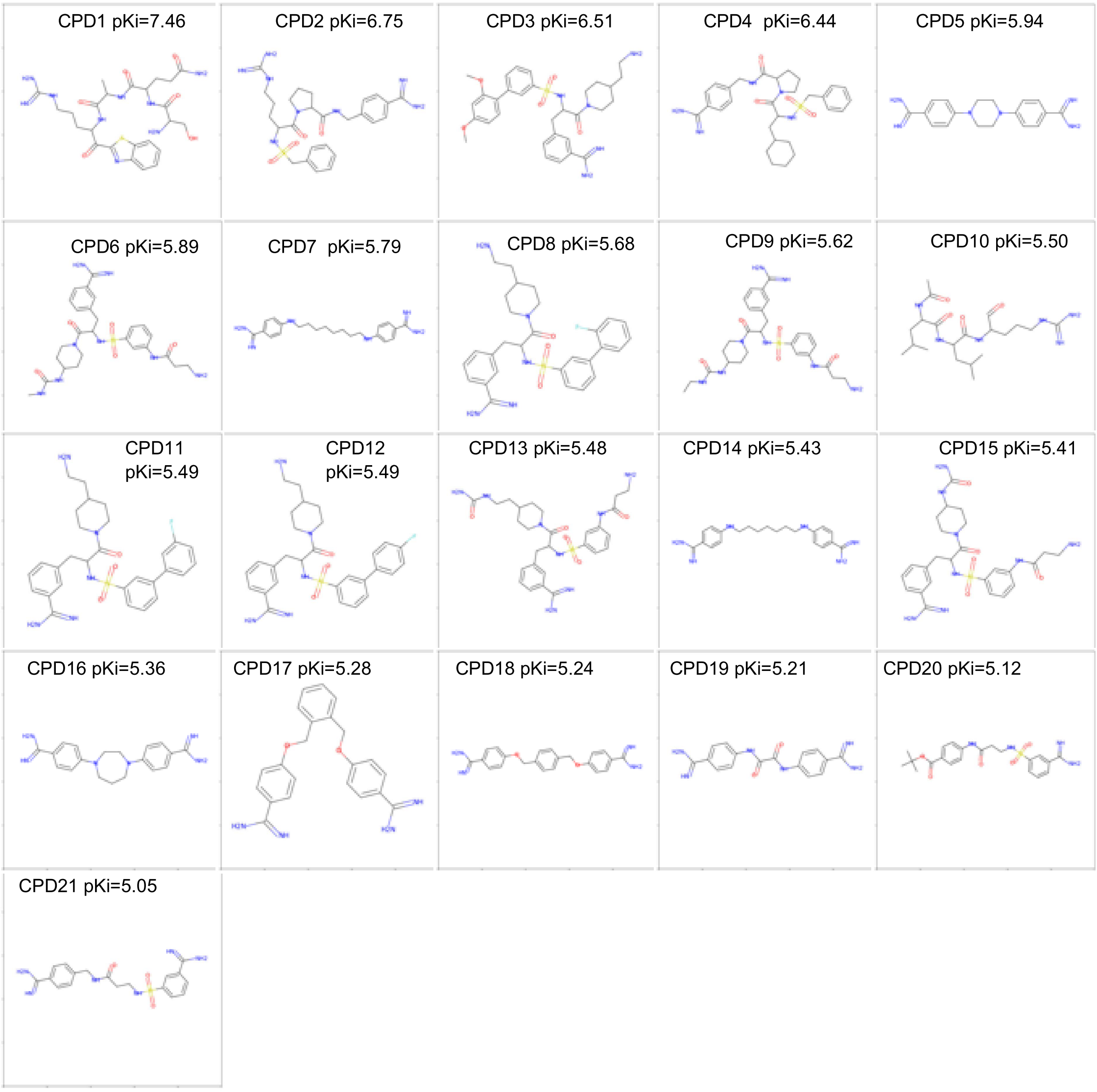


Figure 1. Highly potent compounds for TPS6.

of SPhGs kept the topological distances, while a sparse index of 1.02 was achieved on average. The sparse index is defined as

$$\text{sparse index(SPhG)} = \frac{N_E}{N_N - 1}$$

where N_E is the number of edges and N_N is the number of nodes. The average value of 1.02 implied that most SPhGs are tree structures.

For obtaining SPhGs, candidate graphs for SPhGs (candidate SPhGs) are generated from Mol-SPhG by selecting a predefined number of PF nodes and applying a node reduction algorithm. Our proposed algorithm is to remove unnecessary nodes and convert aromatic ring features to aromatic bonds. The candidate SPhGs are further filtered based on the number of active compounds or scaffolds containing the candidate SPhGs. The SPhGs passing the filter represent shared TPs among active compounds.

In this study, six PF nodes were selected to form candidate SPhGs. For each target, the top 300 SPhGs were selected in terms of the number of the Bemis−Murcko scaffolds of the compounds matching the candidate SPhGs (the *NScaffolds* criterion), identical to the conditions in the previous studies.[15,16,20,30]

Distance Metrics for TP Representations. A similarity of pharmacophore graphs is quantitatively measured by the GED. Jaccard distances measured how (dis)similar a pair of PhFP bit vectors is.

GED. The GED of graphs A and B is the minimum cost of converting graph A to graph B by editing nodes and edges of graph A.[31] In other words, the graph similarity is measured by how easily graph A is transformed to graph B. For calculating the GED, editing operations and associated cost definition are necessary. We used six edit operations: node substitution, node insertion, node deletion, edge substitution, edge insertion, and edge deletion. Based on the work by Garcia-Hernandez et al.,[25] costs of all node and edge operations were newly defined, which are reported in Tables 2 and 3, respectively. According to these tables, the cost of node insertion and deletion is 1, and the cost of changing a node from one PF type to another is 2, which equals the sum of the node deletion of the old PF and the insertion of the new PF. Also, the cost of removing one PF (e.g., removing only D) from a node with two PFs (e.g., DP) is the same as the general node deletion cost of 1. The cost of

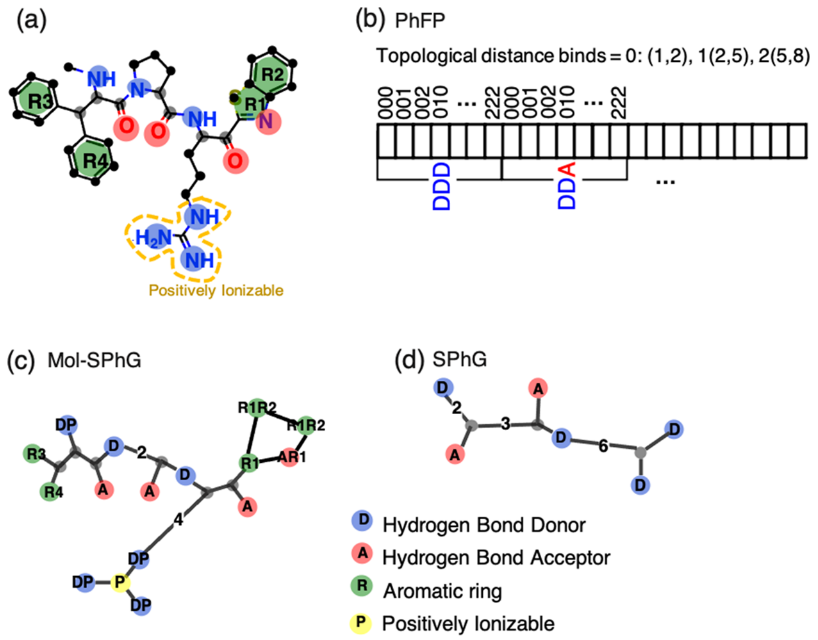


Figure 2. Overview of the three TP representations. (a) Sample molecule with PFs. Blue circles represent HBDs, red circles HBAs, and green circles ARs. (b) Example of PhFP. Each box has a value of 0 or 1. The value of 1 in a box means that the corresponding pharmacophoric pattern(s) exist. Distances between PFs are binned into three categories as mentioned in the parenthesis (lower, upper) in (b). (c) Mol-SPhG converted from the molecule depicted in (a). The letters representing each PF are written on the nodes (D: hydrogen bond donor, A: hydrogen bond acceptor, R: aromatic ring, P: positively ionizable). A node with multiple PFs has the corresponding multiple letters. (d) SPhG generated from Mol-SPhG (c).

Table 2. Node Edit Costs in GED Calculation

	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

[a] D: hydrogen bond donor. [b] A: hydrogen bond acceptor. [c] P: positively ionizable. [d] N: negatively ionizable. [e] R: aromatic ring. [f] J: junction. [g] Double and triple symbols mean the node to which two or three PFs are assigned.

changing a node with two PFs to a new PF(s) is set to 2. In addition to the original definition, our modification of the cost tables for SPhGs or Mol-SPhGs includes three major points. First, the definition for junction nodes, represented as J, is added to the node operation table (Table 2). Every cost of this node modification is defined as 2 because type J is equally (dis)similar to other node types. Second, the cost of edge distance operations is newly defined as shown in Table 3. The cost of the replacement of two nonaromatic edges with lengths n and m ($n > m$) is defined as

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

This monotonical decreasing cost with the edge length matches our intuition about molecules. For example, changing an edge with a length of one to an edge with a length of two has a higher impact than changing an edge with six to seven. The cost of substitutions of a nonaromatic edge for an aromatic edge with the same length is defined as 3 times their edge length based on ref 25. In a similar way, the substitution of two aromatic edges with different lengths costs 10 times more than the corresponding nonaromatic edges, as a 10 times cost was given for the insertion and deletion of a single, double, or triple bond in ref 25. The robustness of GEDs on edge cost functions was confirmed by testing other forms of functions. High distance correlation coefficients of GEDs were observed

Table 3. Edge Edit Costs in GED Calculation

	nonaromatic edge with a length of n^a	aromatic edge with a length of n^a
nonaromatic edge with a length of m^a	$\sum_{k=m+1}^{n} \frac{1}{k}$	$3n \quad (n = m)$ $3n + \sum_{k=m+1}^{n} \frac{1}{k} \quad (n>m)$
aromatic edge with a length of m^a	$3m \quad (n = m)$ $3m + \sum_{k=m+1}^{n} \frac{1}{k} \quad (n>m)$	$\sum_{k=m+1}^{n} \frac{10}{k}$
insertion	0.1	1.0
deletion	0.1	1.0

[a] Without the loss of generalizability, the inequality $n \geq m$ can be assumed.

when using a square root of k or a square k function instead of k in eq 1 (Table S1).

While calculating GEDs of Mol-SPhGs, a time limitation was introduced, which was implemented in the *networkx* library.[32] Mol-SPhGs have more PF nodes than SPhGs, which sometimes results in too much time taken for GED calculation.[18] The time-limitation option causes the minimum graph edit path search to be terminated after a predefined time and the current minimum distance to be given as output. In this study, a value of ten seconds was set, resulting in the consumption of 283 h of CPU time for calculating GEDs for the Kop. data set, which contained 1425 compounds (1 016 025 comparisons). For SPhGs, the calculation time was reduced to around 1.5 h of CPU time for 300 SPhGs (45 300 comparisons) due to the sparseness of SPhGs. It

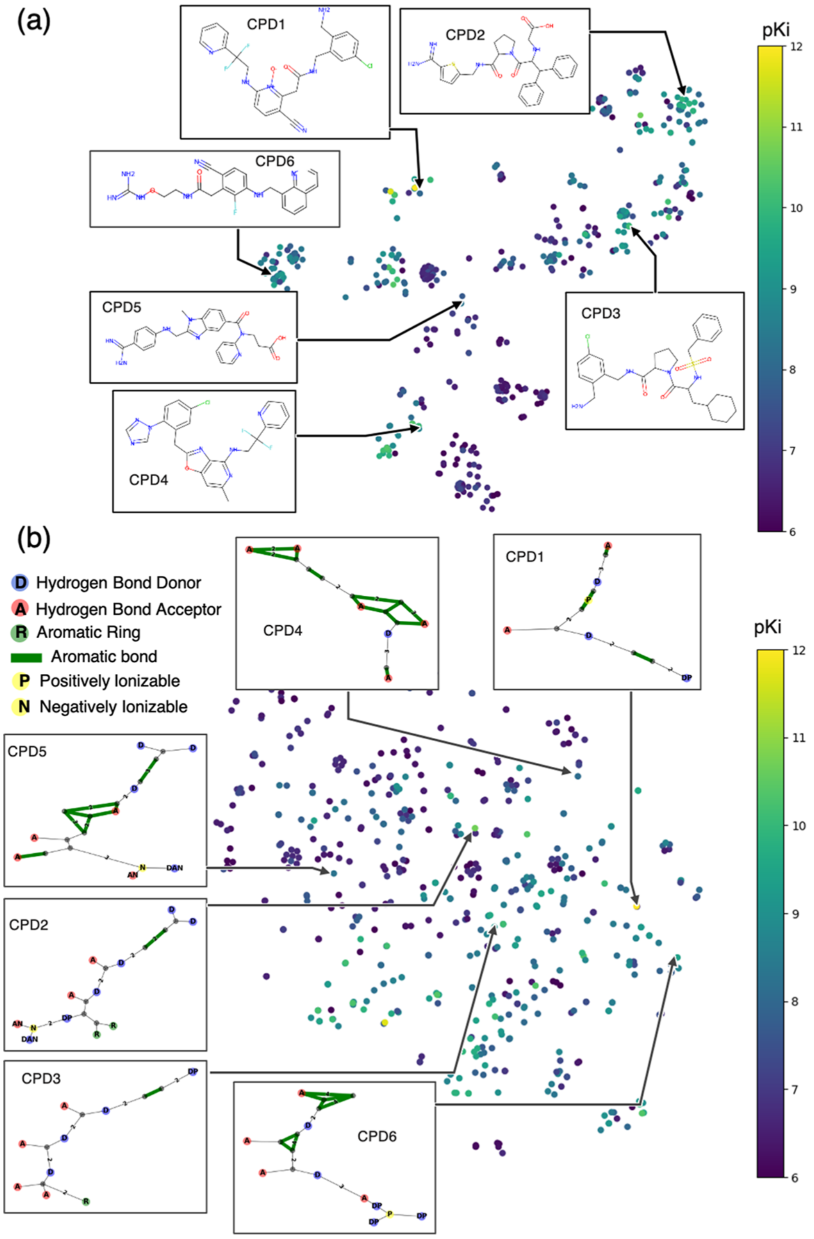


Figure 3. Maps for Thrombin (Thr.). (a) PhFP-map of Thr. Typical active compounds (CPD) are shown on the clustering map. Each point represents each CPD, and its color is defined by its pK_i. Six exemplified Mol-SPhGs selected by the k-means method are displayed. (b) Mol-SPhG-map of Thr. Each dot corresponds to a compound. The Mol-SPhGs of compounds CPD1−6 are displayed along with their locations.

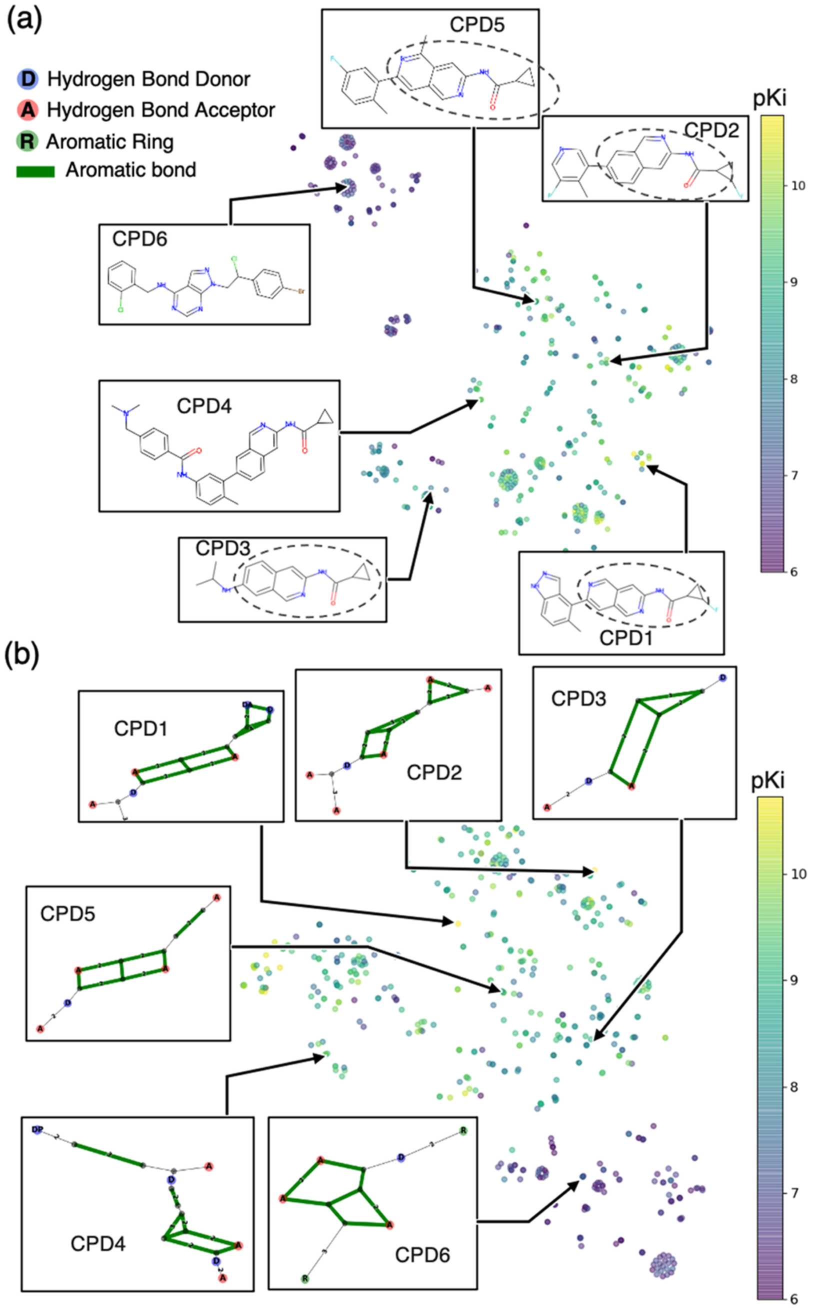


Figure 4. Maps for tyrosine kinase ABL1 (ABL1). (a) PhFP-map of ABL1 typical active compounds (CPD) are shown on the clustering map. Each dot represents each CPD, and its color is defined by its pK_i. Six exemplified Mol-SPhGs selected by the k-means method are displayed. (b) Mol-SPhG-map for ABL1. Each dot corresponds to a compound. The Mol-SPhGs of compounds CPD1−6 are displayed along with their locations.

should be noted that a pairwise comparison can be parallel, leading to the further reduction of computation time. We used the approximated GED implemented in the *networkx* library (version 2.5).[32]

Visualization of the TP Space. Distance metrics enable visualizing data sets in terms of pharmacophore representations by means of unsupervised learning techniques. Three-dimensionality reduction methods were tested: t-distributed stochastic neighbor embedding (t-SNE),[33] isomap,[34] and multidimensional scaling (MDS),[35] all of which are implemented in the scikit-learn library (version 0.23.2).[36]

These three visualization methods were employed with the three pharmacophore representations: PhFPs, Mol-SPhGs, and SPhGs, generating nine maps for a single biological target. We call these maps PhFP-map, Mol-SPhG-map, and SPhG-map, respectively, while ignoring the mapping algorithms.

On a map using PhFP or Mol-SPhG representation (PhFP-map or Mol-SPhG-map), each dot matches one Mol-SPhG or one vector of PhFP, which also corresponds to each CPD for which the representation is generated. Furthermore, dots are colored according to pK_i values. On an SPhG-map, dots correspond only to pharmacophore graphs. The dots are colored according to the coverage, which is defined as the ratio of the compounds covered by the SPhG over the total number of compounds in the data set.

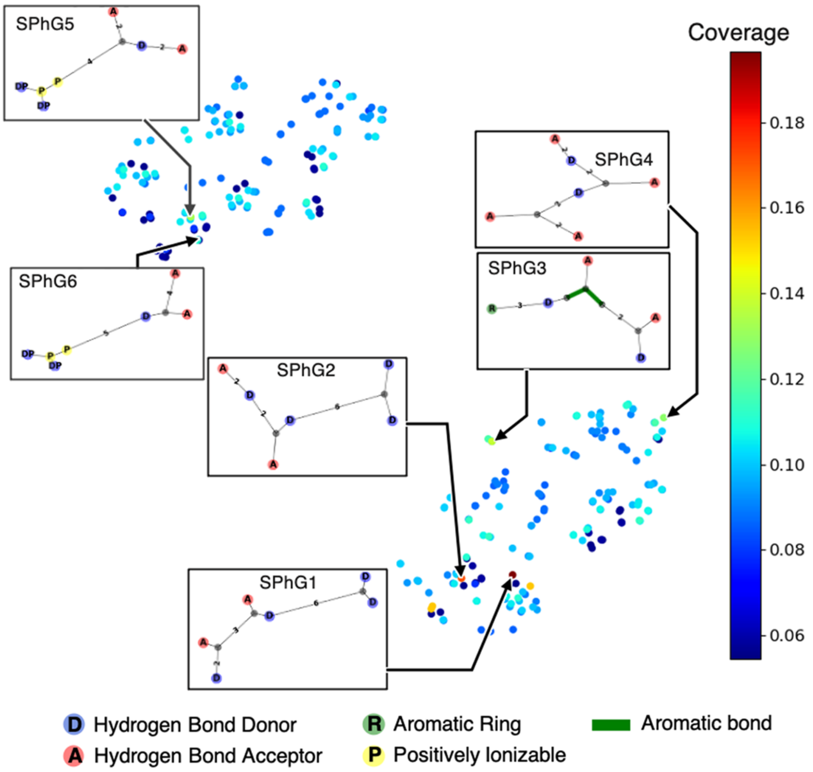


Figure 5. SPhG-map for Thr. Each point represents SPhG, and its color is defined by its coverage. Selected SPhGs are shown on the map. The SPhGs with the first and second-highest coverages in each of the classes categorized by the k-means method are displayed.

■ RESULTS AND DISCUSSION

TP Maps. Visualization of SPhGs gives us an intuitive understanding of the SPhG relation, which cannot be achieved by inspecting the chemical space spanned by molecular descriptors including TP fingerprints. The main difference between these two maps is that SPhGs are shared features among active compounds, not compounds themselves.

We employed the three-dimensionality reduction algorithms Isomap, MDS, and t-SNE to make two-dimensional maps for a set of highly potent compounds represented by PhFP, Mol-SPhG, and SPhG, resulting in nine maps for each target. All of the maps for all of the six biological targets are reported in Figures S1−S6 in the Supporting Information. t-SNE was selected because it formed clusters and was suitable for the later discussion of pharmacophore space. On most maps created by MDS, clustered regions were not created at all, and dots (compounds or SPhGs) were overlapped one another on some maps by Isomap, although this algorithm could make clustered regions. Furthermore, t-SNE mapping showed the best ability to preserve distances in GEDs between SPhGs, in particular for similar SPhGs. For six out of the seven targets, when measuring the shortest 1% distances, t-SNE showed the highest correlation coefficients between GEDs and Euclidian distances on the maps ranging from 0.708 to 0.857. The distance correlation coefficients using the thresholds of 1, 3, and 10% are reported in Table S2. Thus, we decided to further discuss using the maps by the t-SNE algorithm. In the following section, first, the difference between Mol-SPhGs and PhFP as a molecular representation is clarified. Then, using SPhG-maps, the TP information, which could be extracted for the highly potent CPDs, is discussed.

Comparison of Mol-SPhGs with PhFP. Two maps using PhFP and Mol-SPhG representations for Thr. are reported in Figure 3a,b, respectively. Example compounds in Figure 3a were selected based on the *k*-means clustering. The number of clusters was determined so that the sum of the squared errors inside the clusters reached a 90% reduction for the first time as the number of clusters increased. In each cluster, one compound with the highest pK_i value is displayed. Figure 3b shows the Mol-SPhG-map, where each point corresponds to a compound as in PhFP. The CPD1 to 6 in Figure 3a were represented as Mol-SPhGs on the map.

On the PhFP-map, there were more clusters than on the Mol-SPhG-map in Figure 3. For example, the cluster to which CPD3 belonged consisted of CPDs with the same scaffold in terms of the Bemis−Murcko scaffolds. However, these clusters did not exist on the Mol-SPhG-map. On the Mol-SPhG-map, CPD3 belonged to a single cluster with molecules containing different scaffolds. Molecular scaffold-based clustering could miss the actual (topological) pharmacophore. Actually, several compounds belonging to different scaffolds were found to interact with thrombin on the same binding site, supported by X-ray co-crystallization complexes.[37−41] Three of the example CPDs, CPD2, CPD5, and CPD6, contained amidine substructures. However, CPD1 and CPD3 had no substructures similar to amidines.

The chemical structures of CPD2 and CPD5 shared no common scaffolds. However, their Mol-SPhGs-based scaffolds were relatively similar to each other (Figure 3b). These Mol-SPhGs contained the two Ds (HBDs) connected with a two-length bond and the aromatic bonds next to the junction node between the two Ds. Furthermore, negatively ionizable features, carboxy groups, were located on the opposite side of the graphs to the two Ds. This indicated that the mapping using Mol-SPhGs with GED clustered CPDs in a less structurally dependent manner.

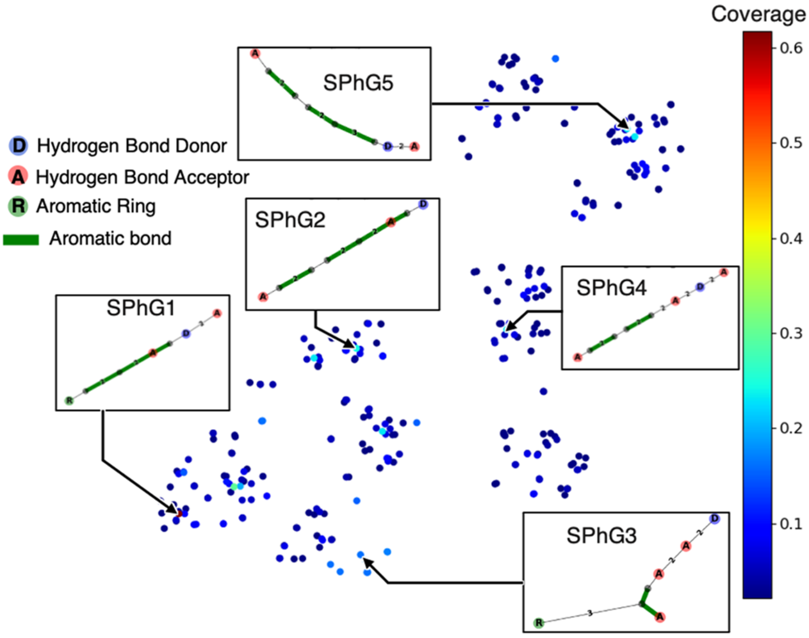


Figure 6. SPhG-maps for tyrosine kinase ABL1 (ABL1). SPhG-map of ABL1. Each point represents SPhG, and its color is defined by its coverage. Selected SPhGs are shown on the map. The SPhGs with the highest coverages in each of the clusters categorized by the *k*-means method are displayed.

Similar characteristics were observed for the other targets. For example, for ABL1 inhibitors, the substructure of dashed circles on CPD1, CPD2, CPD3, and CPD5 in Figure 4a became a core of these compounds, and they were relatively dispersed. On the other hand, in the form of Mol-SPhG (Figure 4b), these CPDs were located closer to each other. Furthermore, Mol-SPhGs made interpretation easier because they reflected how easy (difficult) one graph can be modified to another. For example, CPD1 and CPD5 were distinct on the PhFP-map but not on the Mol-SPhG-map. The Mol-SPhGs of these two CPDs had the same heterocyclic structure consisting of two fused pyridines, and the difference in substituents was measured by GED, resulted in the relatively short distance between these two CPDs on the map in Figure 4b.

SPhG-maps. Figures 3 and 4 show that the capturing TP information by Mol-SPhG-maps was less dependent on the structural scaffolds. In the following, we further discuss SPhGs-maps. It should be noted again that SPhGs are the extracted common subgraphs of Mol-SPhGs. The visualization of SPhGs is conceptually different from visualizing CPDs on Mol-SPhG-maps.

For each target, the number of Bemis−Mucko scaffolds found in the compounds containing the selected 300 SPhGs was counted. The average number was 35.9 for Thr., 12.9 for ABL1, 29.2 for Kop, 68.7 for PI3, 29.7 for GPCR44, and 2.2 for TPS6. Although the number of scaffolds for TPS6 was small due to a small data set size, selected SPhGs were indeed common features of active molecules not dependent on molecular scaffolds. For the SPhG examples found in the following SPhGs-maps, the number of scaffolds is reported in Table S3.

The number of clusters on the SPhG-maps was determined using the same criteria used in Figures 3a and 4a. The SPhGs examples shown in the figures exhibited the highest and the second-highest coverages, as indicated in the figure captions.

Thrombin (Thr.). SPhGs were clustered into two distinct regions in Figure 5. On the top left cluster, SPhG5 and SPhG6 contained the same subgraph with four positively ionizable features (Ps) following by a long chain without any PFs. The positively ionizable feature corresponded to the guanidium substructure. On the other cluster on the right bottom, there were no Ps in the SPhGs forming the cluster. SPhG1 and SPhG2 had two hydrogen bond donors (Ds in Figure 5), which commonly had a junction node with a distant one. SPhG3 and SPhG4 did not have this subgraph.

The SPhG-map displayed graphs with six PFs and a few additional junction nodes, commonly identified among active CPDs. This led to pharmacophore hypotheses of the ligand−target interaction. For example, in Figure 5, SPhG1-2 had a common substructure of two HBDs (Ds) and a junction between them at a distance of 1. Another donor was found at a distance of 6 from the junction, and a pair of D and HBA (A) at a distance of 2 on the opposite side of the two HBDs was typical. These features here were also consistent with those explained by the X-ray cocrystallized structures listed in the Protein Data Bank (PDB).[37]

Tyrosine Kinase ABL1 (ABL1). The SPhG-map for ABL1 along with the selected SPhGs colored based on the coverage for ABL1 are shown in Figure 6. The SPhG with the highest coverage was SPhG1 (61.7%), meaning that over 60% of the active compounds contained SPhG1. Overall, the SPhGs on the maps resembled one another. SPhG1 contained one fused aromatic ring consisting of two rings, with an HBA on one of the rings. This substructure was commonly detected in SPhG2 and SPhG3, which were also included in a substructure of isoquinoline in CPD1, CPD2, and CPD3 in Figure 4. Furthermore, a pattern of HBA and HBD separated by a distance of two followed by an aromatic bond was found in SPhG1, 4, and 5. While these features might be detected from the Mol-SPhG-map in Figure 4 by careful inspection, the SPhG-map represented the relations. The design concept of the ABL1 inhibitors could be interpreted with the help of the SPhG-map.

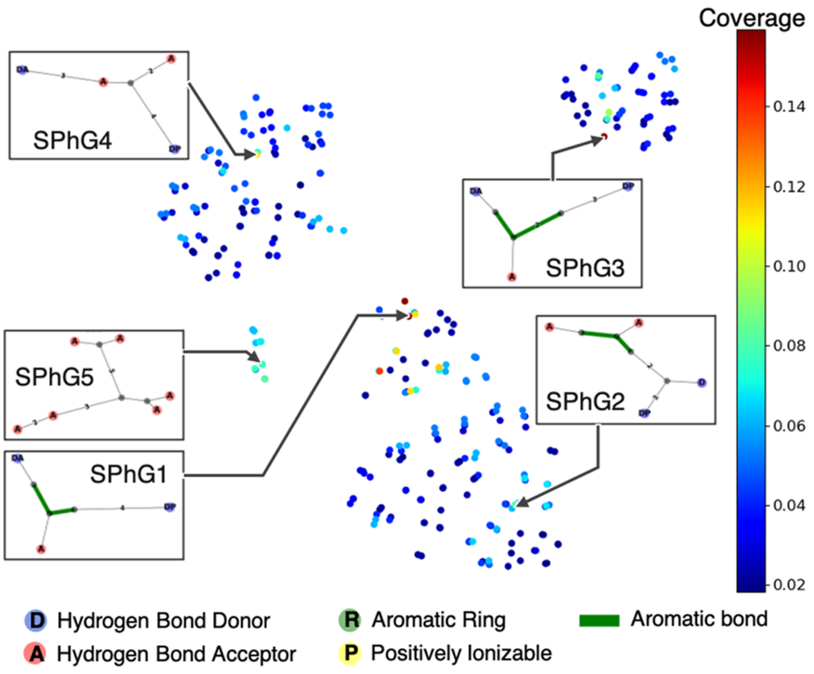


Figure 7. SPhG-map for κ-opioid receptors (Kop.). SPhG-map of Kop. Each point represents SPhG, and its color is defined by its coverage. Selected SPhGs are shown on the map. The SPhGs with the highest coverages in each cluster categorized by the k-means method are displayed.

κ-Opioid Receptors (Kop.). On the SPhG-map for Kop., as shown in Figure 7, SPhG1 and SPhG2 belonged to the same cluster, in which SPhGs contained an aromatic bond feature with a length of two and a branch to an HBA starting at the middle of the bond. SPhG3 in the cluster at the upper right corner had an aromatic bond feature with a length of three (not two). Although SPhG1 and SPhG3 seemed similar and the GED distance between SPhG1 and SPhG3 was 3.25 (7.5 percentile of the whole pairwise distances for all of the SPhG pairs in Figure 7), the encoded features (PFs with bonds) were different (Figure 8a). SPhG1 and SPhG3 matched different paths to the same PFs on the same compound. Out of the three SPhGs, only SPhG1 was detected in the active compounds with different scaffolds, similar to that of pentazocine as shown in Figure 8b. SPhG4 also matched the compound in Figure 8a without introducing aromatic bonds, focusing only on the hydrogen bonds. SPhGs in the small cluster including SPhG5 on the left side of the map only matched the different chemotypes represented by the compounds shown in Figure S7. Note, as shown in Figure 8, an SPhG could contain a node with two PFs (e.g., DA). This meant that a substructure matching both PFs, such as a hydroxyl group, was necessary. If only one of them had been required for activity, the mined SPhG would have contained a node with only the PF.

Transmembrane Protease Serine 6 (TPS6). The number of active compounds for TPS6 was 21 (Figure 1). For this small-sized data set, the SPhG-map could categorize a number of SPhGs into different clusters (Figure 9). Because each SPhG represented a TP hypothesis, extracting common SPhGs followed by the clustering analysis gave insights into the hypotheses, as opposed to PhFP and Mol-SPhG-maps in Figure S6. The top right cluster on the map in Figure 9 might correspond to the hypothesis of the guanidium moiety and other hydrogen bonding features on the opposite side as exemplified in SPhG2. On the other hand, SPhG1, SPhG4, and SPhG5 in other clusters corresponded to the arrangement of hydrogen bonding. These three SPhGs had two HBDs (Ds in Figure 9) and a junction node between them. From the junction node, another HBD is placed at a distance of six, followed by two HBAs (As in Figure 9) with a distance of three. These SPhGs were similar to ones for Thr., and SPhG5 in Figure 9 was identical to SPhG2 in Figure 5. An experimental study showed that the compounds containing SPhG2 in Figure 5, which is identical to SPhG5 in Figure 9,

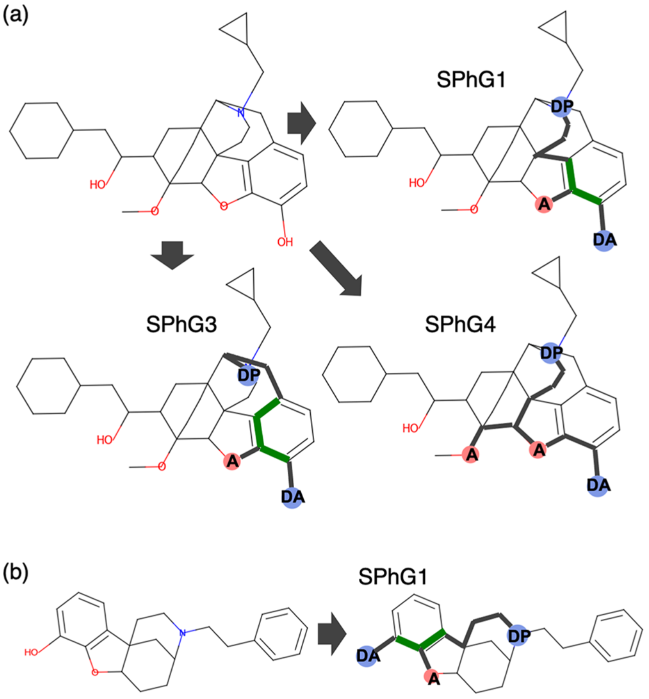


Figure 8. Three different SPhGs derived an active compound for Kop. (a) Active compound analogous to morphine containing three different SPhGs. (b) Active compound containing a different scaffold but SPhG1 as a subgraph.

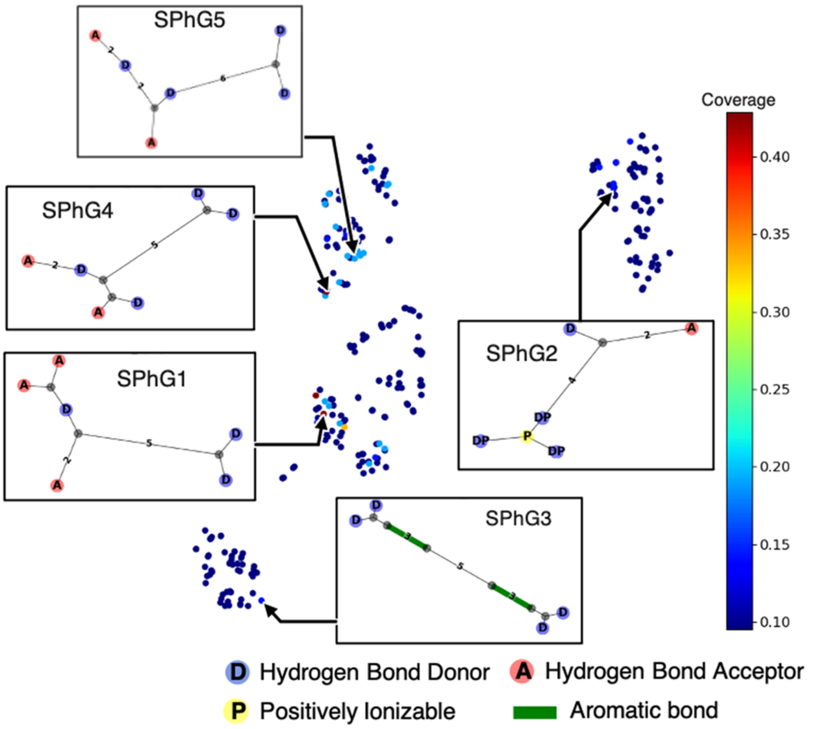


Figure 9. SPhG-map for transmembrane protease serine 6 (TPS6). SPhG-map of TPS6. Each point represents SPhG, and its color is defined by its coverage. Selected SPhGs are shown on the map. All active CPDs with pK_i > 5.0 for TPS6 in our data set are displayed. The SPhG with the highest coverage in each cluster categorized by the k-means method is shown.

were active for both Thr. and TPS6.[42] The common SPhG was successfully identified in this study. The SPhGs on the bottom left cluster where SPhG3 were representatives were completely different hypotheses and matched CPD5 and CPD16 in Figure 1. These types of SPhGs were not found in Thr. This implied that the CPDs, which included SPhG3 and did not include SPhG5 in Figure 9, were expected to be active for TPS6 and not for Thr.

■ CONCLUSIONS

The visualization of topological pharmacophores (TPs) is important for understanding the ligand−target binding hypotheses. In this study, GED was introduced as a metric to evaluate the similarity among SPhGs, which were sparse representations of pharmacophore graphs. Among the three tested dimensionality reduction algorithms, t-SNE was the best based on the visual inspection and local-distance preservation of GEDs. For evaluating the maps and demonstrating the use case of the maps, we generated SPhG-maps using active compounds against the six biological targets: Thr., ABL1, Kop., PI3, GPCR44, and TPS6. First, we compared the two TP representations using the maps PhFP and Mol-SPhG and found that Mol-SPhG was less structurally dependent than PhFP. Then, for each target, the top 300 SPhGs identified from a set of active compounds were visualized on an SPhGs-map with the GED metric. The classification of SPhGs and TP knowledge extraction were demonstrated using the maps.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.2c00173.

Visualization results (maps) for the six biological targets (Figures S1−S6); minor chemotype in Kop. and its corresponding SPhG (Figure S7); GED correlation coefficients using different cost functions from Table 3 (Table S1); correlation coefficients between GEDs and Euclidian distances on maps (Table S2); and number of Bemis−Murcko scaffolds found in the SPhGs explained in the text (Table S3) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Tomoyuki Miyao − *Graduate School of Science and Technology, Nara Institute of Science and Technology, Ikoma, Nara 630-0192, Japan; Data Science Center, Nara Institute of Science and Technology, Ikoma, Nara 630-0192, Japan;* orcid.org/0000-0002-8769-2702; Phone: +81-743-72-6065; Email: miyao@dsc.naist.jp; Fax: +81-743-72-6037

Author

Hiroshi Nakano − *Graduate School of Science and Technology, Nara Institute of Science and Technology, Ikoma, Nara 630-0192, Japan;* orcid.org/0000-0003-2745-9340

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.2c00173

Notes

The authors declare no competing financial interest.
All the scripts used in this study including the generation of SPhGs from active compound data sets, clustering a group of SPhGs with GED, and visualizing SPhGs as simple colored graphs are found in the GitHub repository under the name SPhG2 (https://github.com/n-hiroshi/sphg2). All the highly

potent compounds for the six targets are also provided as SMILES with curated pK_i values in the repository.

■ ACKNOWLEDGMENTS

This work was supported by the Grant-in-Aid for Transformative Research Areas (A) 21A204 Digitalization-driven Transformative Organic Synthesis (Digi-TOS) from the Ministry of Education, Culture, Sports, Science & Technology, Japan. This work was also supported by the JSPS KAKENHI Grant number JP20K19922.

■ REFERENCES

(1) Khedkar, S.; Alpeshkumar, M. K.; Evans, C. C.; Sudha, S. Pharmacophore Modeling in Drug Discovery and Development: An Overview. *Med. Chem.* **2007**, *3*, 187−197.

(2) Kirchmair, J.; Distinto, S.; Markt, P.; Schuster, D.; Spitzer, G. M.; Liedl, K. R.; Wolber, G. How to Optimize Shape-Based Virtual Screening: Choosing the Right Query and Including Chemical Information. *J. Chem. Inf. Model.* **2009**, *49*, 678−692.

(3) Pascual, R.; Almansa, C.; Plata-Salamán, C.; Vela, J. M. A New Pharmacophore Model for the Design of Sigma-1 Ligands Validated on a Large Experimental Dataset. *Front. Pharmacol.* **2019**, *10*, No. 519.

(4) Glaab, E.; Manoharan, G. B.; Abankwa, D. Pharmacophore Model for SARS-CoV-2 3CLpro Small-Molecule Inhibitors and in Vitro Experimental Validation of Computationally Screened Inhibitors. *J. Chem. Inf. Model.* **2021**, *61*, 4082−4096.

(5) Opo, F. A. D. M.; Rahman, M. M.; Ahammad, F.; Ahmed, I.; Bhuiyan, M. A.; Asiri, A. M. Structure Based Pharmacophore Modeling, Virtual Screening, Molecular Docking and ADMET Approaches for Identification of Natural Anti-Cancer Agents Targeting XIAP Protein. *Sci. Rep.* **2021**, *11*, No. 4049.

(6) Ehmki, E. S. R.; Rarey, M. Exploring Structure−Activity Relationships with Three-Dimensional Matched Molecular Pairs—A Review. *ChemMedChem* **2018**, *13*, 482−489.

(7) Wolber, G.; Langer, T. LigandScout: 3-D Pharmacophores Derived from Protein-Bound Ligands and Their Use as Virtual Screening Filters. *J. Chem. Inf. Model.* **2005**, *45*, 160−169.

(8) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. "Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chem. Int. Ed.* **1999**, *38*, 2894−2896.

(9) Schneider, G.; Schneider, P.; Renner, S. Scaffold-Hopping: How Far Can You Jump? *QSAR Comb. Sci.* **2006**, *25*, 1162−1171.

(10) Smith, D. H.; Carhart, R. E.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.

(11) Renner, S.; Schneider, G. Scaffold-Hopping Potential of Ligand-Based Similarity Concepts. *ChemMedChem* **2006**, *1*, 181−185.

(12) 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*, No. 44.

(13) Grisoni, F.; Merk, D.; Byrne, R.; Schneider, G. Scaffold-Hopping from Synthetic Drugs by Holistic Molecular Representation. *Sci. Rep.* **2018**, *8*, No. 16469.

(14) 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.

(15) Métivier, J.; 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) Nakano, H.; Miyao, T.; Funatsu, K. Exploring Topological Pharmacophore Graphs for Scaffold Hopping. *J. Chem. Inf. Model.* **2020**, *60*, 2073−2081.

(17) Böhm, H.; Flohr, A.; Stahl, M. Scaffold Hopping. *Drug Discovery Today Technol.* **2004**, *1*, 217−224.

(18) Hu, Y.; Stumpfe, D.; Bajorath, J. Recent Advances in Scaffold Hopping. *J. Med. Chem.* **2017**, *60*, 1238−1246.

(19) 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*, No. 54.

(20) Nakano, H.; Miyao, T.; Swarit, J.; Funatsu, K. Sparse Topological Pharmacophore Graphs for Interpretable Scaffold Hopping. *J. Chem. Inf. Model.* **2021**, *61*, 3348−3360.

(21) 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.

(22) 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.

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

(24) Oprea, T. I.; Gottfries, J. Chemography: The Art of Navigating in Chemical Space. *J. Comb. Chem.* **2001**, *3*, 157−166.

(25) Garcia-Hernandez, C.; Fernández, A.; Serratosa, F. Ligand-Based Virtual Screening Using Graph Edit Distance as Molecular Similarity Measure. *J. Chem. Inf. Model.* **2019**, *59*, 1410−1421.

(26) 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.

(27) Landrum, G. RDKit: Open-Source Cheminformatics Software., http://www.rdkit.org. (accessed Dec 28, 2019).

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

(29) Capecchi, A.; Probst, D.; Reymond, J. L. One Molecular Fingerprint to Rule Them All: Drugs, Biomolecules, and the Metabolome. *J. Cheminform.* **2020**, *12*, No. 43.

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

(31) Abu-Aisheh, Z.; Raveaux, R.; Ramel, J. Y.; Martineau, P. *An Exact Graph Edit Distance Algorithm for Solving Pattern Recognition Problems*, ICPRAM 2015—4th International Conference Pattern Recognition Applications Methods, Proc., 2015; pp 271−278.

(32) Hagberg, A.; Swart, P.; Schult, D. *Exploring Network Structure, Dynamics, and Function Using NetworkX*; Los Alamos National Lab. (LANL): Los Alamos, NM (United States), 2008.

(33) van der Maaten, L.; Hinton, G. Multiobjective Evolutionary Algorithms to Identify Highly Autocorrelated Areas: The Case of Spatial Distribution in Financially Compromised Farms. *J. Mach. Learn. Res.* **2008**, *9*, 2579−2605.

(34) Tenenbaum, J. B.; Silva, V. de.; Langford, J. C. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science* **2000**, *290*, 2319−2323.

(35) Torgerson, W. S. Multidimensional scaling I: Theory and method. *Psychometrika* **1952**, *17*, 401−419.

(36) 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. Res.* **2011**, *12*, 2825−2830.

(37) 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.

(38) Friedrich, R.; Steinmetzer, T.; Huber, R.; Stürzebecher, J.; Bode, W. The Methyl Group of N$^{\alpha}$(Me)Arg-Containing Peptides Disturbs the Active-Site Geometry of Thrombin, Impairing Efficient Cleavage. *J. Mol. Biol.* **2002**, *316*, 869−874.

(39) 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.

(40) 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.

(41) 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.

(42) 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 Hepcidin Levels in Human Hepatocytes. *Cell Chem. Biol.* **2019**, *26*, 1559−1572.