Deciphering a Pharmacophore Network: A Case Study Using BCR-ABL Data

Damien Geslin, Alban Lepailleur, Jean-Luc Manguin, Nhat-Vinh Vo, Jean-Luc Lamotte, Bertrand Cuissart, Ronan Bureau

DECIPHERING A PHARMACOPHORE NETWORK: A CASE STUDY USING BCR-ABL DATA

Damien Geslin[†,‡], Alban Lepailleur[†], Jean-Luc Manguin[‡], Nhat-Vinh Vo[‡], Jean-Luc Lamotte[†,⁑], Bertrand Cuissart[‡], and Ronan Bureau[†]

[†]Centre d'Etudes et de Recherche sur le Médicament de Normandie, Normandie Univ, UNICAEN, CERMN, 14000 Caen, France

‡Groupe de Recherche en Informatique, Image, Automatique et Instrumentation de Caen, Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, France

[⁑]Sorbonne Université, UFR 919, 4 place Jussieu F-75252 Paris cedex 05

ABSTRACT. This paper introduces a general method that can be used to create groups of pharmacophores to support their further in-depth analysis. A BCR-ABL molecular dataset was used to calculate graph edit distances between pharmacophores and led to their organization into a novel pharmacophore network. The application of a graph layout algorithm allowed us to discriminate between the pharmacophores associated with active compounds and those associated with inactive compounds. A clustering approach was used to refine the partitioning by grouping the pharmacophores based on their structure, activities, and binding modes. Analysis of a newly-spatialized pharmacophore network provided us with critical insight into structure-activity relationships, most notably those that revealed distinctions between activity

classes and chemical families. As shown, this method permits us to identify families of structurally homogeneous pharmacophores.

INTRODUCTION

We recently described[1] a new approach for the automated detection of pharmacophores based on findings collected in a large chemical dataset.[2] As a case study, the BCR-ABL tyrosine kinase was chosen mainly for two reasons. First, numerous drug discovery programs[3] have contributed to one of the largest collections of records in ChEMBL[4] for a protein kinase target. Second, the active site of the BCR-ABL tyrosine kinase is highly flexible and can adjust to accommodate a variety of inhibitors with distinct and well-defined binding modes. Using the same dataset, our goal here was to analyze the capacity of the computed pharmacophores to differentiate between active and inactive compounds. Classical approaches to this type of structure-activity relationships (SAR) study typically begin with the definition of physicochemical, topological, or topographical descriptors associated with the chemical dataset of interest. This type of analysis has been performed using a BCR-ABL dataset, as described by Zin *et al.*[5] in their study of imatinib derivatives and Kale *et al.*[6] who reported on interactions with 2-phenazinamine derivatives. The descriptors can also be encoded into molecular fingerprints. Pharmacophore fingerprints are a particularly important example of this approach.

The approach used in this study involves the extraction of pharmacophores from a database and the generation of a set of pharmacophoric descriptors that are similar to the aforementioned pharmacophore fingerprints. One of the key features identified by this method is the extent of a pharmacophore which is a value that refers to the number of compounds that are associated. The quality of a pharmacophore is then assessed by using a measure such as the growth rate (GR) which quantifies the capacity to distinguish active from inactive molecules. Our methodology summarizes findings from a large set of pharmacophores by focusing on a representative subset defined by the Maximal Marginal Relevance Feature Selection (MMRFS) algorithm.[7] This algorithm selects a subset of pharmacophores with elements that are

distinctive, discriminative, and representative of a single group. The selected pharmacophores were then organized using a graph edit distance (GED)[8,9] method. The resulting structure can be viewed as a network in which each node is a selected pharmacophore. The edges between the selected pharmacophores are labeled by the GED that separate them. We then analyzed this structure with a graph layout algorithm that provides an initial partitioning and divides the selected pharmacophores into active and inactive subsets. A clustering approach is then used to characterize the diversity of the pharmacophores and to suggest analogies between the binding mode of active compounds supported by pharmacophores within the same cluster.

The pharmacophore network computed by this process is analogous to the concept of chemical space.[10–13] In our earlier publication,[1] the pharmacophore network was used only as a SAR visualization tool. In this study, we aim to use this network approach to group the pharmacophores into subsets in which all elements share both structure and activity.

MATERIALS AND METHODS

Dataset. The BCR-ABL dataset used in this work includes the 1492 compounds described in our earlier publication.[1] In brief, this dataset was collected from ChEMBL[4,14] with the following restrictions: (i) only K_i and IC_{50} values expressed in nM units from biochemical assays reported in CHEMBL_24 (CHEMBL1862 : Target CHEMBL ID) were accepted as bioactivity data; (ii) measurements containing symbols such as ">" or "<" were not included unless they agreed with the threshold value (e.g., "<10 nM" would be retained as a means to identify an active molecule in the case of a 100 nM activity threshold); and (iii) if more than one bioactivity measurement was provided for the same molecule, we included the lowest K_i or the lowest IC_{50} value if no K_i was available. Duplicates were filtered and additional adjustments were performed (e.g., removal of salts, standardization of chemical functions, addition of hydrogens at the heteroatoms, and conversion to a two-dimensional [2D] structure data file [SDF] format) using Pipeline Pilot (BIOVIA, San Diego, CA, USA) components. The

molecules exhibiting K_i or IC_{50} values less than or equal to 100 nM were considered to be active compounds (n = 774); molecules with K_i or IC_{50} values greater than or equal to 1000 nM were considered to be inactive (n = 718). We created this substantial gap between active *versus* inactive molecules to maintain clear differentiation between the two groups.

Pharmacophores. The method presented in this study can be used to analyze molecules with known biological activities and structures that are expressed in a 2D SDF format. As a first step, every molecule is transformed into its respective reduced pharmacophore graph.[15] In this type of graph, the nodes represent a specific pharmacophoric feature, and the edges represent the fewest possible bonds between two nodes. The pharmacophoric features correspond to generalized functionalities that are involved in favorable interactions between ligands and targets, including hydrogen-bond acceptors (A) and donors (D), negatively (N) and positively (P) charged ionizable groups, hydrophobic regions (H), and aromatic rings (R). **Figure 1** depicts the transformation of a molecule (M1) from its 2D molecular structure into a reduced pharmacophore graph. In the second step, the quality of a pharmacophore is addressed. A pharmacophore is of lesser interest if it is detected infrequently. The quality of a given pharmacophore may reflect the quality of the set of molecules in which it appears. In the following trials, this set is defined as the extent of a pharmacophore. The cardinality of this extent provides support for a given pharmacophore. As we would like to retain only those pharmacophores that appear sufficiently frequently in the dataset, our method includes a user-defined threshold; a pharmacophore is considered only if its support exceeds a specific threshold. In the third step, for each pharmacophore detected at the appropriate frequency, we compute its capacity to discriminate between active and inactive molecules. In the implementation of our algorithm, the notion of an emerging pattern (EP) quantifies this imbalance using a measure of growth rate (GR). Based on the partitioning of this dataset into active and inactive molecules, the GR of a given pharmacophore corresponds to the ratio

between the frequencies with which it fits into each of the two subgroups. Extraction of the pharmacophores is based on two parameters: (i) the minimal size of its support (i.e., the minimal number of molecules associated with a given pharmacophore), and the order that results (i.e., the number of features per pharmacophore). In this study, we set n = 10 as the minimal threshold for support. Within these parameters, we have identified all the pharmacophores with three to seven nodes that have emerged based on an evaluation of the transition from active to inactive, and *vice versa*. The order is denoted O_{num}, where *num* denotes the number of nodes. A selection of representative pharmacophores with orders O_3 to O_7 was evaluated using MMRFS,[1,7] which is a feature-selection algorithm borrowed from the Maximal Marginal Relevance heuristic used in information retrieval.[16] Until a stop criterion is reached, MMRFS iteratively selects a pharmacophore in which GR is maximized with the largest number of new molecules added to its support.

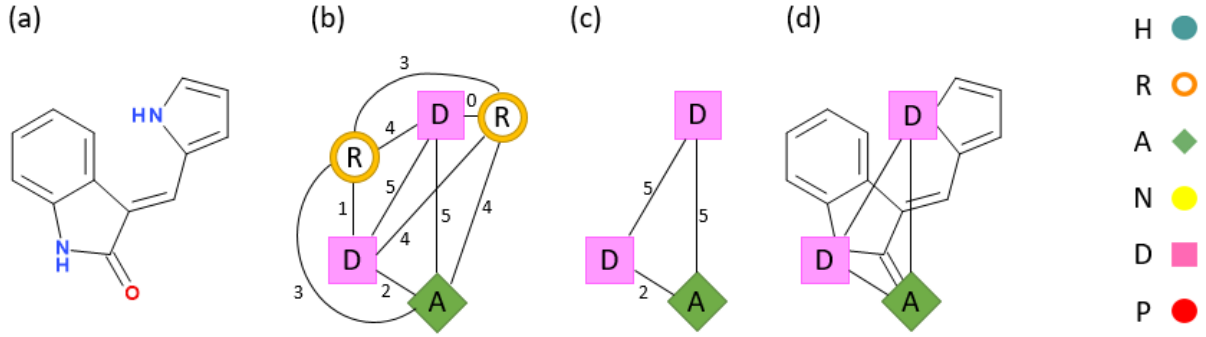


Figure 1. Transformation of molecule M1 (a) into its pharmacophore graph (b). An example of a three-point pharmacophore extracted from the graph in (b) if this pharmacophore satisfies the user-defined support (c). The corresponding fit of the three-point pharmacophore in the molecule M1 (d). The representative symbol for each pharmacophoric feature is indicated on the right: hydrogen-bond acceptors (A) and donors (D), negatively (N) and positively (P) charged ionizable groups, hydrophobic regions (H), and aromatic rings (R).

In this study, pharmacophores were evaluated for their capacity to distinguish between active and inactive compounds. When the GR denotes sufficient association with either an active or inactive subset, the pharmacophore is in turn identified as related to this class. Hence, when a molecule includes a given pharmacophore, one then predicts that it belongs to the subset or class to which this pharmacophore has been assigned. Of note, a molecule may contain several different pharmacophores, including those associated with each of the two different classes; these molecules are classified as "active/inactive." Other molecules may contain no characterized pharmacophores; these molecules are classified as "unpredicted." However, the rate at which we can assign molecules to each of these specific classes is not a sufficient measure of the quality of a given set of pharmacophores, as the categories by their nature require a trade-off between active and inactive properties. Therefore, we evaluated the performance of this method using measures of sensitivity and specificity.[17] In our study, we consider the assignment to a class as a positive event. Thus, the degree of sensitivity reflects the correct assignment of positive events while specificity is defined by the correct assignment of the negative events.

Graph edit distance (GED). As pharmacophores are depicted here as graphs, graph edit distance (GED) can be used to evaluate the structural commonalities shared by two pharmacophores.[18] GED, which is a general and flexible method used to measure similarities between labeled graphs, has been successfully applied in various domains,[19] including chemoinformatics.[8,20,21] GED quantifies an edit path between two input graphs via a sequence of edit operations in which one of the input graphs undergoes a stepwise transformation into the other. An edit operation corresponds to a small change in the graph, for example, insertion, deletion, or substitution of a node or an edge. In practical terms, we used this method to compare pharmacophores of the same order only. Moreover, as pharmacophores are complete graphs, i.e., every two nodes are connected by an edge that is labeled with the distance that

separates their respective features, the edit paths require no insertions or deletions. Thus, we are only required to set the cost of node and/or edge substitution. The cost of a node substitution was set at 10. The cost of an edge substitution was designated at the absolute value of the change over the given distance. While these parameters could be optimized further, we found that the currently assigned costs led to acceptable results in the networks under study.

GED is used to spatialize the pharmacophore network. The input data for this process includes a complete weighted graph in which the nodes are the MMRFS-defined pharmacophores and edges are determined by GED. For the visualization of the pharmacophore network, we used the Gephi software[22] with ForceAtlas2 (default parameters with no overlap option) as the layout algorithm.[23] For this algorithm, the nodes are perceived as separated by repulsive forces (similar to that observed between charged particles) with the edges functioning as virtual springs between nodes. The forces between the nodes lead to movements that converge in a balanced state.[23] Starting from the hypothesis that additional repulsive forces will be required for a complete graph, we focused on the nearest neighbors of each MMRFS-defined pharmacophore that were selected based on GED. The neighbors of each MMRFS-defined pharmacophore were ranked in descending order based on GED values. Using this method, the nearest two, five, and 10 neighbors of each pharmacophore were maintained. Of note, these values corresponded to the minimum number of neighbors because several of the edges within a given network can exhibit identical values for GED.

GED as the basis for clustering. Clustering analyses were carried out separately on five sets of pharmacophores in the following order: O_3, O_4, O_5, O_6, and O_7. Before this analysis, the datasets were analyzed using DBSCAN[24] (Density-Based Spatial Clustering of Applications with Noise) to detect outliers.[25] For this study, an outlier was defined as a pharmacophore represented by a graph that differed from the others. This computation was justified because the inclusion of outliers may substantially change the result of a clustering computation. The

DBSCAN analysis results in clustering that relies strongly on the concept of a neighborhood. In DBSCAN, two pharmacophores are identified as neighbors if they are separated by a distance that does not exceed a threshold known as *Eps.* Then, to compute the clusters, one applies the "core point condition", which states that a pharmacophore that belongs to the same cluster as its central neighbors. A pharmacophore is identified as *central* when its neighborhood contains at least the *MinPts* other pharmacophores. At the end of the partitioning, any pharmacophore that lies alone (i.e., in its own cluster) is considered to be an outlier. To set values for the two parameters of DBSCAN (i.e., *Eps,* the radius that defines a given neighborhood, and *MinPts*, the minimum number of neighbors required to be considered as a central pharmacophore), we implemented an optimization process to identify the highest silhouette index.[26]

The clustering computation[27] was based on the implementation of KMEANS, AGNES, SPECC in R.[28] Six configurations were analyzed using these three clustering methods both with and without preliminary detection of outliers using DBSCAN. For AGNES, the average linkage was used with the function cutree to repartition the pharmacophores into clusters. The three methods were selected for use as each represents a different category: KMEANS is a partitioning algorithm,[29] AGNES is an agglomerative algorithm,[30] and SPECC generates spectral clusterings.[31] Moreover, a recent study has demonstrated that they perform well in practice.[32] Three quality indices were calculated, including Gap[33], Silhouette[26], and Pareto frontier. As the performance of each of the three indices was similar, we chose to focus on the Pareto frontier for this study.

To perform this analysis, we defined $C = \{C_1, \ldots, C_k\}$ as the clustering of a pharmacophore dataset *D* into *k* subsets. In the equations to follow, *n* represents the cardinality of *D*, n_i is the cardinality of C_i ($n = \sum n_i$), and MIN and MAX are the minimum and maximum distances between any two pharmacophores within dataset *D*, respectively. The Pareto frontier was

generated based on two parameters: the intra-cluster similarity [*Intra(C)*; Equation 1] and the inter-cluster dissimilarity [*Inter(C)*; Equation 2]. The number of clusters was chosen to be between three and 30 each including ~100 MMRFS-defined pharmacophores.

$$Intra(C) = \frac{1}{k} \times \left(\sum_{i=1}^{i=k} \left(\frac{2}{n_i \times (n_i - 1)} \right) \times \sum_{x,y \in C_i, x \neq y} 1 - \frac{d(x,y) - MIN}{MAX - MIN} \right) \quad \text{Equation 1}$$

$$Inter(C) = \frac{1}{k} \times \left(\sum_{i=1}^{i=k} \left(\frac{1}{n_i \times (n - n_i)} \right) \times \sum_{x \in C_i, y \notin C_i} d(x,y) \right) \quad \text{Equation 2}$$

An external quality measure was used to evaluate the extent to which the computed clustering matches the initial classifications, *i.e.*, assignment of the pharmacophore to either the active class or inactive class. The quality of this distinction was quantified using the Normalized Mutual Information (NMI) measure.[34] The best clustering models should be capable of separating the pharmacophores with active compounds (i.e., the active pharmacophores) from those that include inactive compounds (i.e., inactive pharmacophores).

Starting from a clustering $C = \{C_1, \ldots, C_k\}$ and a classification $V = \{V_1, V_2\}$(for active *versus* inactive) for *n* pharmacophores, the NMI values are calculated using Equation 3 and the notations defined in the paragraphs above. In Equation 3, $n_{i,j}$ denotes the number of pharmacophores that belong to cluster C_i and class V_j. The terms H(C) and H(V) correspond to the entropies associated with clustering (C) and classification (V). The mean function H (C), H (V) corresponds to the average of these two parameters.

$$NMI(C,V) = \frac{MI(C,V)}{mean(H(C), H(V))} \quad \text{Equation 3}$$

where

$$MI(C,V) = \sum_{i=1}^{k}\sum_{j=1}^{2} P(i,j) \times \log\left(\frac{P(i,j)}{P(i)P'(j)}\right)$$

$$P(i,j) = \frac{n_{i,j}}{n} \qquad P(i) = \frac{n_i}{n} \qquad P'(j) = \frac{|V_j|}{n}$$

$$H(C) = -\sum_{i=1}^{k} P(i) \times \log\left(P(i)\right) \qquad H(V) = -\sum_{j=1}^{2} P'(j) \times \log\left(P'(j)\right)$$

RESULTS AND DISCUSSION

Sensitivity and specificity. The active pharmacophores correspond to those that emerged from the active set (774 compounds) to the inactive set (718 compounds) when using a minimum GR threshold value of 3. Inversely, the inactive pharmacophores correspond to those that emerged from the inactive set to the active set when using the same minimum GR threshold. The total number of pharmacophores generated (called *all* pharmacophores in **Table 1**) varies from 1,327 for O_3 to 22,730 for O_7. The results of sensitivity and specificity testing are described in **Table 1**. For example, the application of the classification rule based on the results from 1327 active O_3 pharmacophores derived from the total 1492 compounds (i.e., full dataset) yielded scores of 0.98 and 0.15 for sensitivity and specificity, respectively. This result indicates that the active compounds were correctly identified (i.e., high sensitivity); by contrast, inactive compounds were defined active with these active O_3 pharmacophores leading to a low specificity. We observed an overall tendency toward increasing specificity values with increasing order of the pharmacophores. We have observed a clear improvement in specificity when considering pharmacophores with five features compared to those with only three or four features. We note that our computational methods also result in the development of a new type of fingerprinting method with pharmacophores that are potentially larger than the three or four-points pharmacophores used classically.[35,36] This point is important for SAR studies based on the identification of pharmacophores. The opposite trend was observed in our evaluation of

sensitivity. We found that the constraints for fitting a molecule associated with a pharmacophore with a larger number of features (e.g., O_7), are more difficult to fulfill. Our consideration of the support required (i.e., the minimum number of compounds associated with a given pharmacophore) plays a critical role in determining the number of pharmacophores extracted and consequently the coverage of the chemical structures associated with a given dataset. In our dataset, the decrease of sensitivity observed when progressing from O_3 to O_7 was less apparent for active compared to inactive compounds. In summary, in contrast to measurements of specificity, the sensitivity decreased in parallel with increases in the order of the pharmacophores (from O_3 to O_7).

Table 1. Statistical measures of the performance of functional classification within increasing pharmacophore order.

	All pharmacophores						MMRFS-defined pharmacophores					
	Active pharmacophores[a]			Inactive pharmacophores[a]			Active pharmacophores[a]			Inactive pharmacophores[a]		
	N[b]	Sn[c]	Sp[d]	N	Sn	Sp	N	Sn	Sp	N	Sn	Sp
O_3	1,327	0.98	0.15	1,632	0.98	0.33	103	0.97	0.45	148	0.97	0.63
O_4	6,396	0.98	0.20	5,716	0.98	0.39	114	0.93	0.81	146	0.91	0.81
O_5	12,935	0.94	0.51	9,588	0.85	0.63	100	0.90	0.92	113	0.75	0.85
O_6	19,067	0.92	0.75	14,322	0.58	0.85	102	0.88	0.95	74	0.54	0.93
O_7	22,730	0.86	0.86	19,564	0.46	0.93	91	0.82	0.96	52	0.42	0.96

[a]growth-rate (GR) = 3; [b]N, number of pharmacophores; [c]Sn, sensitivity; [d]Sp, specificity.

GED and pharmacophore networks. Figure 2 documents the different pharmacophore networks based on graph edit distances (GEDs) for O_3 to O_7 as a function of the number of nearest neighbors (NNs). From O_3 to O_6, the separation between the active and inactive pharmacophores improves when the number of NNs increases. By contrast, the separation between the groups in O_7 is apparent regardless of the number of NNs, however, visualization of this graph is most straightforward in the case of two NNs as there are fewer edges that need to be displayed.

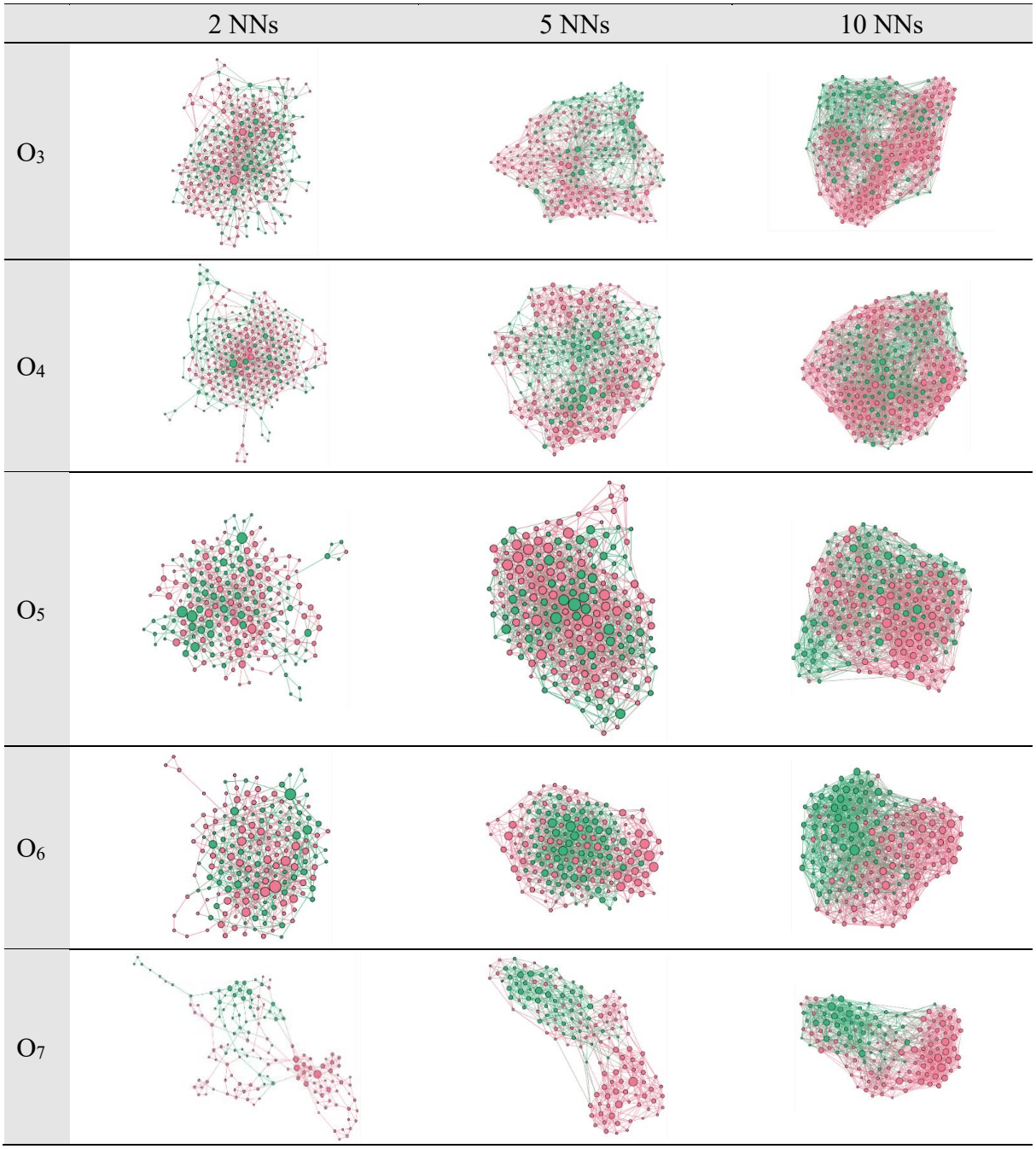


Figure 2. Pharmacophore networks (O_3 – O_7) were identified by considering two, five, or 10 nearest neighbors (NNs). Red nodes represent active pharmacophores, while green nodes represent inactive pharmacophores. The individual nodes are scaled in size based on the overall number of their respective NNs.

As shown in **Figure 3**, the pharmacophore network is helpful to analyze local SARs. This network highlights groups of active pharmacophores included in closely-related molecules, for

example, the series of substituted 6,6-fused nitrogenous heterocyclic compounds (indicated in A), or the diverse series of chemical compounds that share specific anchoring features (as shown in B). This presentation of the pharmacophore network also facilitates the identification of discontinuous local SARs by highlighting pharmacophores with opposite activity labels that are nonetheless connected to one another. A subset of pyrazolopyrimidine derivatives (C), whose potency is achieved by targeting a crucial glutamic acid residue (Glu286),[37] provides an example of how one might detect discontinuity in the activity data. Similar results can be obtained by detection of classical activity cliffs[38,39]; however, using our methodology, the information is extracted at the level of the pharmacophores, not of the molecules. As a comparison, we computed a similarity matrix (Tanimoto coefficient) from the BCR-ABL dataset using a 2D pharmacophore fingerprint implemented in Pipeline Pilot (PHFP_3, in which pharmacophores consist of triplets of molecular features with corresponding bond distances). We then examined several highly similar molecules within a neighborhood of two (Tanimoto coefficient [Tc] > 0.75) that belong to different activity classes. We identified 156 pairs of structurally similar compounds that correspond to activity cliffs (data not shown). Among these compounds, 25 are related to the aforementioned pyrazolopyrimidine derivatives. These 25 activity cliffs are summarized by only one triplet of pharmacophores within our network (C). We conclude that these two approaches are likely to be complementary to one another; our method focuses on frequent pharmacophoric substitutions that strongly influence biological activity and does not permit us to capture specific and sometimes rare chemical modifications.

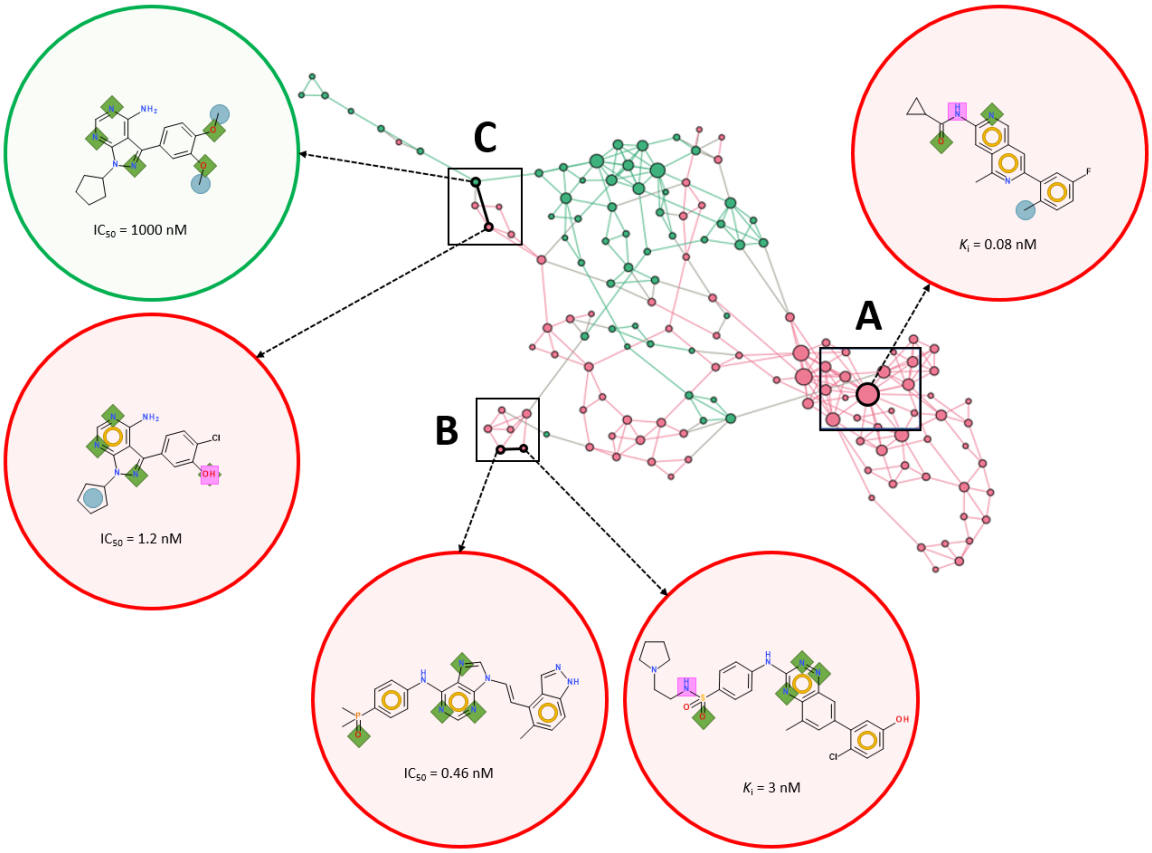


Figure 3. Detailed views of three locations of interest in the O_7 pharmacophore network (generated with 2 NNs). Example molecules with properties that are predicted by the corresponding pharmacophores are shown in the red (active) and green (inactive) circles.

Clustering and GED. Given the six different clustering methods, five different descriptions (O_3 to O_7), and 28 distinct clusters (3 to 30 clusters), we can consider a total of 840 configurations. Among these configurations, 333 partitions were defined by Pareto frontiers (**Figure 4**). We ranked these configurations based on their NMI values (**Table 2**, **Figure 4**). The most reliable partitions are those associated with O_7 (while those associated with O_4 are the least effective) with an average rank of 40 (nb: O_7 has 62 partitions with an average optimum rank value of 31). The optimum partition was obtained with AGNES/DBSCAN by considering eight clusters and six outliers (e.g., NMI of 0.31). The results for each order (from O_3 to O_7) are summarized in **Table 2**.

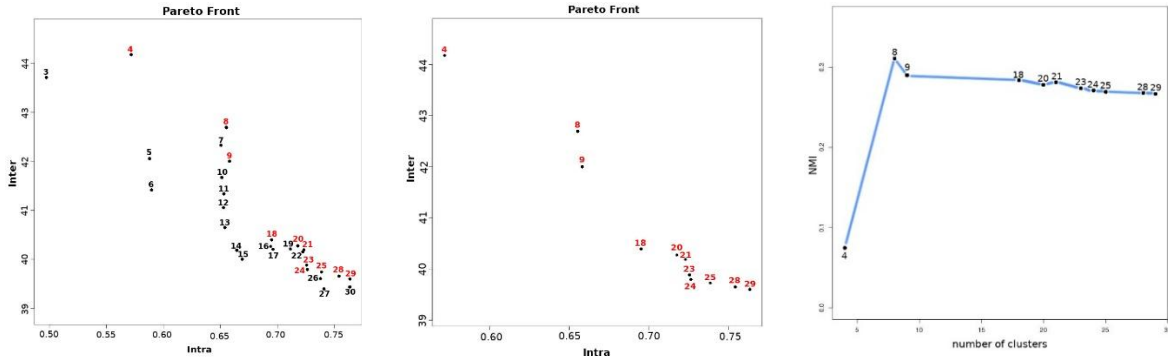


Figure 4. Representations of intra-cluster similarity *versus* inter-cluster dissimilarity (left), Pareto frontier (center), and NMI values with respect to the function of k clusters (right) generated by AGNES/DBSCAN by considering the O_7 description.

Table 2. Qualities of the partitions and best-clustering methods according to NMI values.

	O_3	O_4	O_5	O_6	O_7
Number of partitions	68	57	72	74	62
Average rank of partitions	227	263	171	138	40
Best clustering methods	KMEANS / DBSCAN	KMEANS	AGNES / DBSCAN	AGNES	AGNES / DBSCAN
Outliers	18	-	1	-	6
Number of clusters	30	27	6	30	8
NMI	0.11	0.08	0.14	0.20	0.31

A cluster was defined as active when the ratio between the active and the inactive pharmacophores included within the cluster was >3. This value is in agreement with the GR values associated with the extraction of the pharmacophores. A cluster was defined as inactive when the ratio was <0.33. The composition of the clusters and the corresponding dendrograms are shown in **Figure 5**. Note that 30 pharmacophores with a ratio between active and inactive pharmacophores of 2.33 favoring the active class were not classified and grouped in Cluster 5.

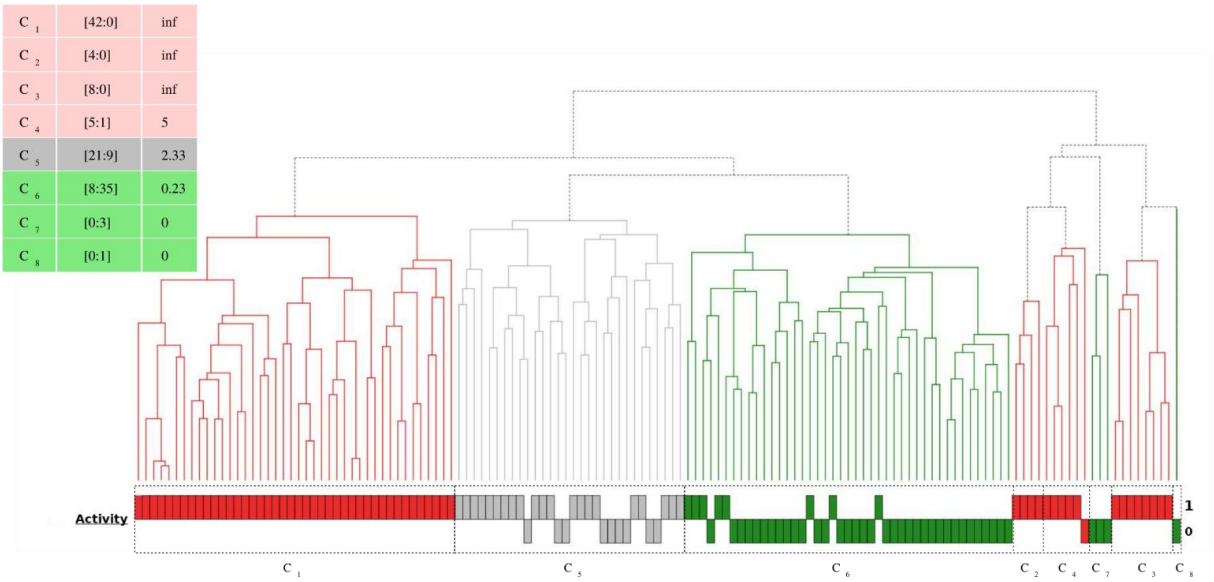


Figure 5. Dendrogram depicting the contents of the eight clusters associated with O_7. The pharmacophores included in the active ($C_1 – C_4$) and inactive clusters ($C_6 – C_8$) are represented in red and green, respectively. The pharmacophores that were not classified (C_5) are shown in grey. The activity histogram in the lower part of the figure shows the initial classification of the pharmacophores as active (1) or inactive (0).

The results shown in **Figure 6** document clustering on the initial layout of O_7 that was generated using two or 10 NNs. In this figure, the shape of the nodes indicates the activity of the pharmacophores (triangle for active and circle for inactive). The six outliers are depicted in gold and described in **Figure 7**. Four of these outliers are adjacent pharmacophores (**Figure 6**, 2 NNs, and 10 NNs) with only hydrogen bond acceptor features (except for the one labeled active that exhibits a positive ionizable site). The other two outliers correspond to the only two pharmacophores with no hydrogen bond acceptor features (of the 143 MMRFS-defined pharmacophores for O_7). As shown in **Figure 6**, the 42 active pharmacophores in cluster C_1 are on the right side of the network. By contrast, the four active pharmacophores in cluster C_2 are at the bottom center, and the eight active pharmacophores of C_3 are on the left. The five active and one inactive pharmacophore of C_4 are localized in the middle of the graph. Our results revealed that 512 of the 774 active compounds (66%) were associated with active

clusters; 246 out of the 718 inactive compounds (34%) were associated with inactive clusters. The sensitivity values observed for the MMRFS-defined pharmacophores document these results with respect to O_7 (**Table 1**).

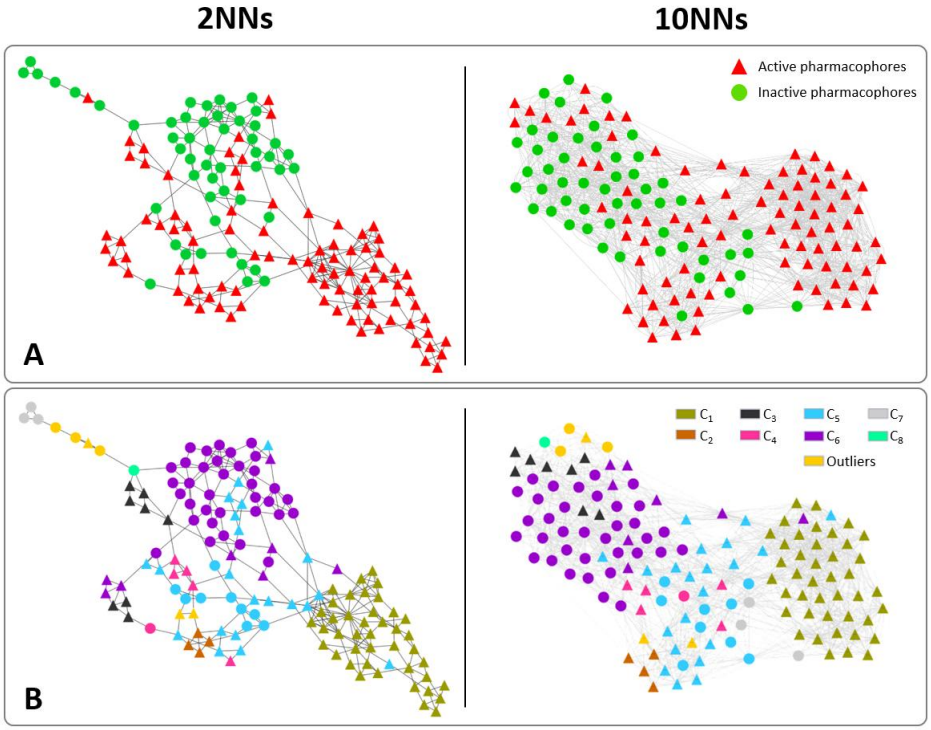


Figure 6. An initial representation of the O_7 pharmacophore network generated with two or 10 NNs (A). The same layout with consideration of clustering (B).

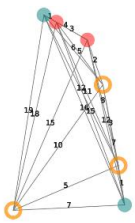
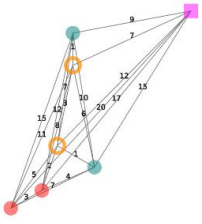
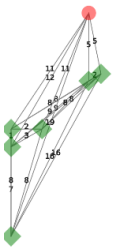
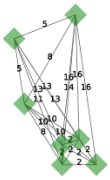
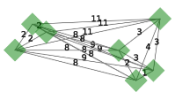
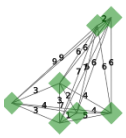
Active pharmacophores			
Support	n = 25	n = 25	n = 13
Inactive pharmacophores			
Support	n = 10	n = 12	n = 13

Figure 7. Descriptions of six pharmacophores that are considered to be outliers from O_7.

Figure 8 displays some representative compounds associated with each active cluster (cluster center/ECFP4 (1024 bit)/Tanimoto metric).[40] The number of representative compounds was a function of the distinct areas within each active cluster as shown in **Figure 6**. Several observations can be made from these findings. The pharmacophores included in Cluster C_1 are found in 379 compounds. The pharmacophore **P1** alone includes 271 of these compounds. This group originates from a series of substituted 6,6-fused nitrogenous heterocyclic compounds and was previously discussed in our earlier publication.[1] The pharmacophores included in C_1 are scaffolds that occupy the adenine binding site and the hydrophobic back pocket adjacent to the ATP binding cleft. These interactions are specific for type I inhibitors that bind to the active conformation of the kinase and compete with ATP. No crystallographic data are available on the PDB for any of the compounds included in C_1, although one that was closely related (Compound **8**) was reported in the form of a complex with GSK3β.[41] In this crystal structure, characteristics that are typical of type I binding mode can be observed (**Figure 9A**), including (i) the activation loop is in the DFG-in conformation, (ii) the pyridine and its cyclopropylamide substituent forms two critical hydrogen bonds with the Hinge region, and (iii) the isoquinoline moiety fills the hydrophobic back pocket. Similarly, 72 compounds that have been clearly designed to bind to the kinase in its inactive conformation are included in C_2. All of these compounds exhibit a piperazine group, for example, compound **2** (**Figure 8**). This corresponds to the positive ionizable pharmacophore function that fills the allosteric pocket created when the activation loop moves to the DFG-out position. These ligands are referred to as type II inhibitors. The tyrosine kinase inhibitor, ponatinib (**9**, **Figure 9B**) provides a classic example of this binding mode.[42] The 62 compounds represented by cluster C_3 include pharmacophores that can be divided into two subgroups based on the pharmacophore network and 2 NNs (**Figure 6**). Representative pharmacophores of these two subgroups are shown in **Figure 8** (**P3**, **P4**). Co-crystallographic data are available for

compounds belonging to both subgroups. **Figure 9C** reveals the crystal structure of ABL1 complexed with PD166326 (Compound **10**), which is a pyridopyrimidine derivative that functions as a type I inhibitor.[43] Our consideration of the pyrazolopyrimidine series includes a multi-targeted kinase inhibitor (PP102, Compound **11**) found in a complex with c-Src (nb: no structural data have been reported for complexes with ABL). The pyrazolopyrimidine moiety corresponds to the position of the adenine of ATP and projects an aryl substituent into the hydrophobic back pocket.[37] It should also be noted that all the compounds associated with cluster C_3 appear to hold their activation loops in the DFG-in conformation. The C_4 cluster is even more scattered across the network than C_3; this is because the pharmacophores included in C_4 correspond to three different areas (**Figure 6**). These findings are summarized by the representative pharmacophores **P5**, **P6**, and **P7** (**Figure 8**). In contrast to other active clusters, C_4 includes one pharmacophore associated with the inactive set of molecules (**P7**). There are no 3D data for this cluster, although Compound **6** displays a methyl-imidazole moiety that is similar to that of nilotinib (**12**, **Figure 9D**), which is a potent BCR-ABL inhibitor that binds to the inactive conformation of the kinase.[44] According to the co-crystallography data and the size of the active C_4 pharmacophores, we can state confidently that the related active compounds all bind to ABL in the DFG-out conformation and thus fill the allosteric pocket. All the chemical structures and the associated clusters are described as supplementary information (OSF[45] for the deposit [see Notes]).

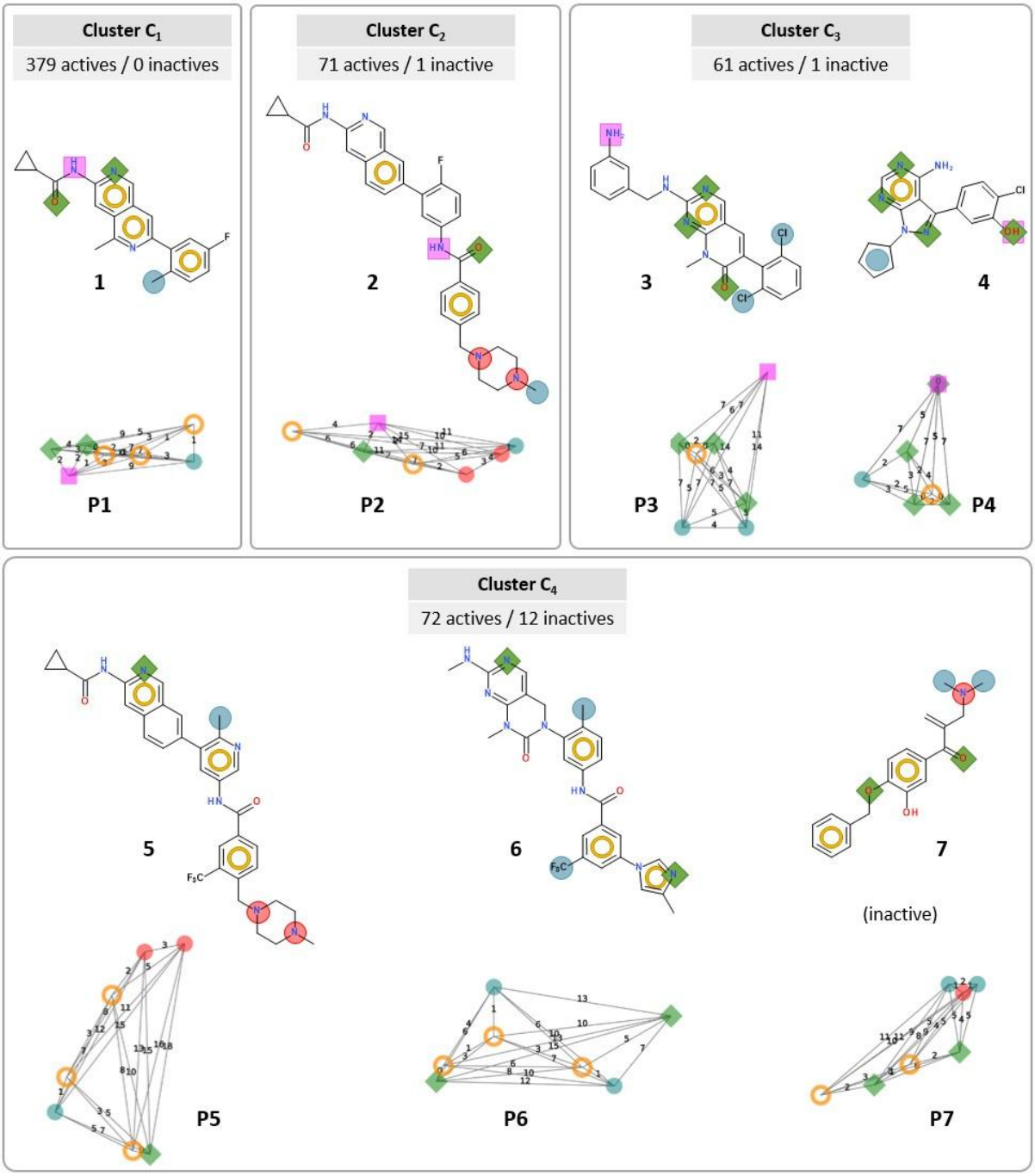


Figure 8. Representative compounds from each active cluster and associated pharmacophores.

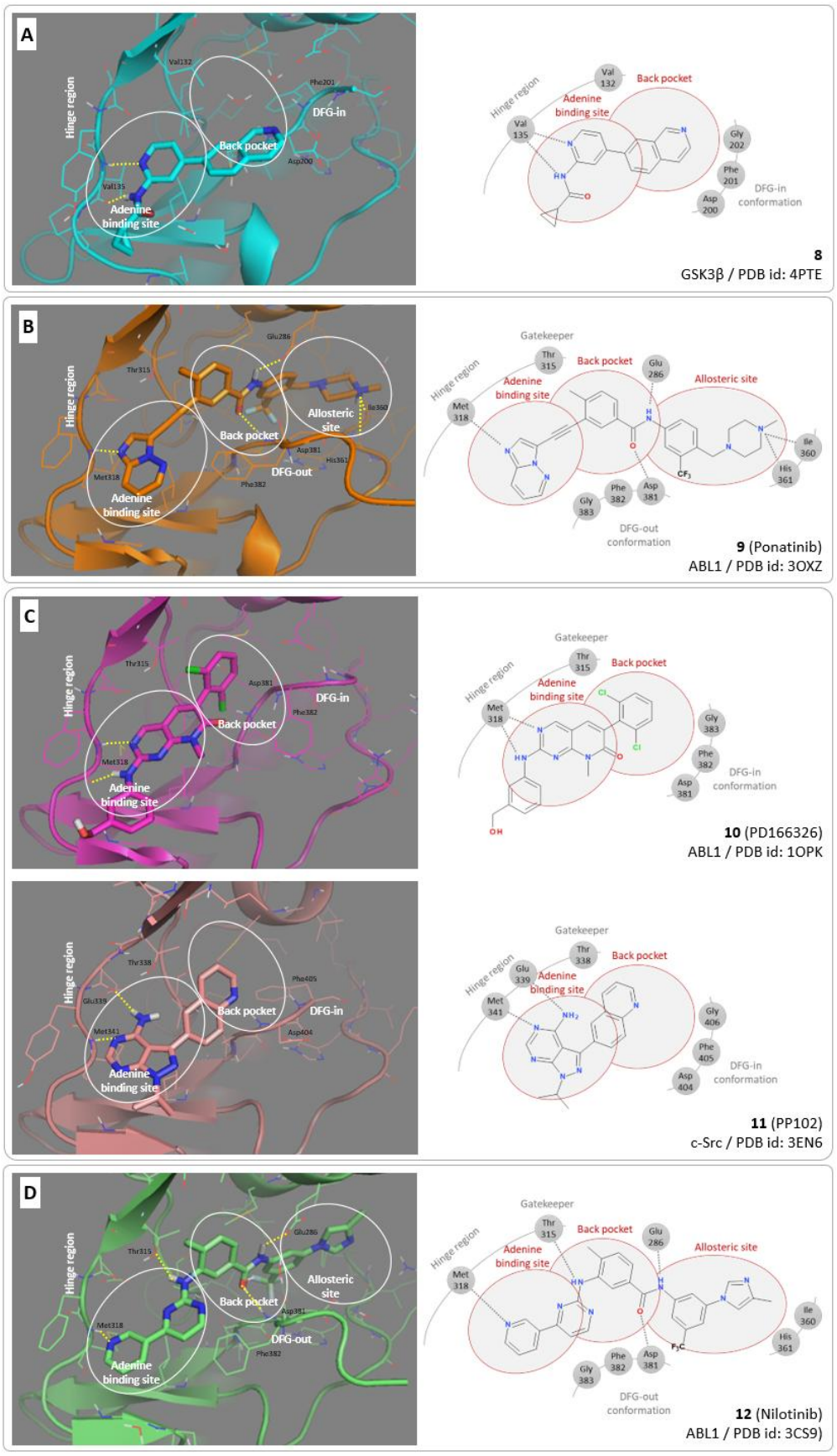


Figure 9. Co-crystallographic data that supports the links between clusters C_1 –C_4 and diverse binding modes (e.g., DFG conformation).

For comparison purposes, we described the BCR-ABL dataset using 2D pharmacophore fingerprints. As described above, PHFP_3 fingerprints were used. The best partitioning was obtained using the KMEANS method with a consideration of nine clusters (NMI = 0.22). We then compared the content of these clusters to the subsets of molecules identified by our original approach. A quantitative comparison of the contents of each of these clusters is described in **Table 3**. By focusing on the clusters with predominantly active compounds, we found that 611 unique active compounds of the original 774 were assigned to clusters $C_1 – C_5$ (O_7 pharmacophores; 79%). By contrast, only 426 of these active compounds were associated with clusters $C'_1 – C'_4$ using PHFP_3 (55%).

Table 3. Composition of the clusters based on the method used: O_7 MMRFS pharmacophores *versus* PHFP_3 pharmacophore fingerprint.

O_7		PHFP_3	
Cluster	Active/Inactive[a]	Cluster	Active/Inactive
C_1	379/0	C'_1	254/0
C_2	71/1	C'_2	111/0
C_3	61/1	C'_3	61/1
C_4	72/12	C'_4	10/4
C_5	257/67	C'_5	138/193
C_6	109/255	C'_6	17/25
C_7	0/21	C'_7	148/388
C_8	0/10	C'_8	35/93
		C'_9	0/14

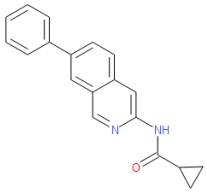
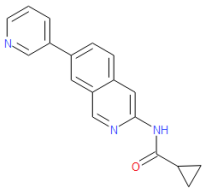
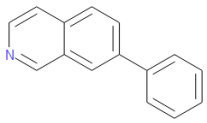
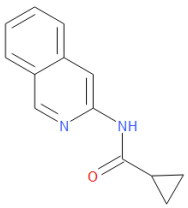
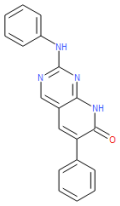
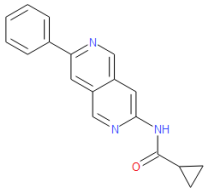
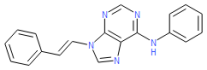
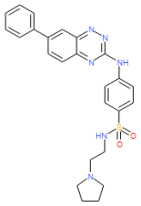
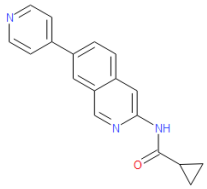
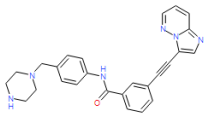
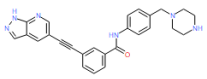
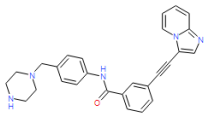
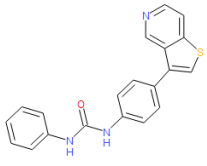
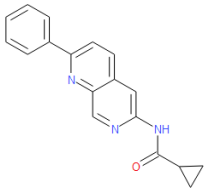
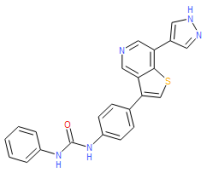
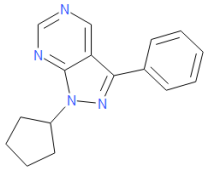
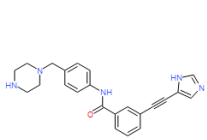
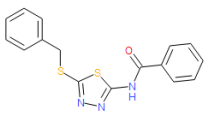
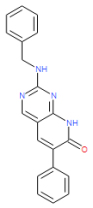
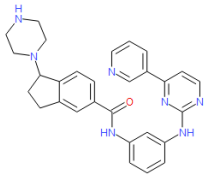
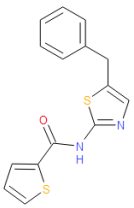
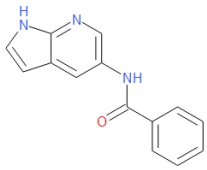
[a]A single molecule can be associated with several clusters using our method.

We evaluated these findings in light of the different families of active molecules discussed in **Figure 8**. We noted that using PHFP_3, cluster C_1 was split into two groups (C'_1 and C'_2); cluster C_2 corresponded to C'_3 and the contents of clusters C_3 and C_4 were scattered among the remaining groups. These results suggest that while our method was better for grouping the active compounds, it might be interesting to analyze clusters $C'_4 – C'_8$ to determine whether other binding modes were captured. This is of importance because the frequency parameter we

used (i.e., a minimum threshold value of 10 for the number of molecules needed to support a pharmacophore) could prevent us from detecting rare variant binding modes.

We also performed a scaffold-based clustering of the BCR-ABL dataset using components implemented in Pipeline Pilot. First, each input molecule was trimmed to create the Bemis-Murcko framework. Next, rings were selected and trimmed following the rules described by Schuffenhauer *et al.*[46] to generate a hierarchical scaffold tree. Finally, the clusters were computed based on the Bemis-Murcko framework level, *i.e* the largest scaffold at the top of the tree for each molecule. From the 1492 compounds in the original dataset, 502 were grouped into 36 scaffold-based clusters with at least six compounds per cluster. Among those remaining, 603 are singletons. As shown in **Table 4**, 318 out of the 770 active compounds were grouped into 22 clusters that contain predominantly active molecules. A cross-analysis of the content of the scaffold-based clusters and the subsets of molecules identified by our approach led to the following observations: (i) scaffolds S_1 – S_4, S_6 – S_9, S_{14}, and S_{22} correspond to cluster C_1 and contain most of the active compounds; (ii) scaffolds S_{10} – S_{12}, S_{17}, and S_{20} correspond to the cluster C_2; and (iii) scaffolds S_5, S_{13}, S_{15}, S_{16}, and S_{19} correspond to the cluster C_3. Concerning the scaffold C_4, only substructures corresponding to the left part of compound **5** can be retrieved (S_2 and S_4 which are rather associated to C_1). We also note that scaffolds S_{18} and S_{21} have no counterparts within the O_7 MMRFS-defined pharmacophores. The very few compounds associated with S_{18} and S_{21} might be an explanation for this observation since the frequency parameter was set to 10 molecules for the detection of our pharmacophores. While a significant strength of our approach is the high level of abstraction of pharmacophores, this comparison study highlights the impact of decreasing the frequency threshold to encompass the full diversity of a dataset.

Table 4. Representation of the Bemis-Murcko frameworks associated with scaffold-based clusters containing predominantly active compounds. The number of active and inactive compounds included in each cluster is as indicated in parentheses (active/inactive).

			
S_1 (60/0)	S_2 (52/0)	S_3 (22/0)	S_4 (20/0)
			
S_5 (20/0)	S_6 (18/0)	S_7 (12/0)	S_8 (10/0)
			
S_9 (9/0)	S_{10} (8/0)	S_{11} (7/0)	S_{12} (7/0)
			
S_{13} (6/0)	S_{14} (6/0)	S_{15} (15/1)	S_{16} (9/1)
			
S_{17} (8/1)	S_{18} (7/1)	S_{19} (6/1)	S_{20} (5/1)
			
S_{21} (5/1)	S_{22} (6/2)		

Problems associated with inactive compounds. The chemical and biological data in ChEMBL were extracted from the scientific literature. Consequently, the data collected are often the result of hit-to-lead optimization programs that originate from a limited number of chemical subfamilies. Beginning with a first hit, medicinal chemists frequently perform pharmacomodulations on a selected molecular scaffold to obtain a lead compound to target. In our study, the difference between the two groups or clusters is noticed once it achieves O_5; O_7 is considered to be the ideal in terms of NMI, sensitivity, and specificity values. Consequently, while small molecular scaffolds or pharmacophores exist for both active and inactive compounds, a clear differentiation between the two groups is achieved only when data from larger pharmacophores are integrated. The most interesting layouts are observed with O_7, notably the identification of an important subfamily of active pharmacophores that were associated with nearly 25% of the compounds in the entire dataset (**Figure 3**). The number of inactive MMRFS-defined pharmacophores is higher than the number of active ones when O_3, O_4, and O_5 are considered. However, in analyses limited to O_6 and O_7, the opposite trend is observed (**Table 1**). This result is related to the support value (set at 10 for this study) because this value becomes a significant constraint only when the order of the pharmacophores is sufficiently high. Under these conditions, lower measures of sensitivity and reduced coverage of the inactive compounds were obtained (**Table 1**). These data may be interpreted as follows: during the drug design process, structural optimization is performed to generate a new compound from an original "hit." This action provides a new set of leads that are similar but not identical to the first hit. Therefore, starting with findings reported in ChEMBL, the probability of collecting close structural neighbors (our study requires at least 10) will be much lower for the inactive compounds compared to those shown to be active. This results in reduced sensitivity values for the inactive pharmacophores (**Table 1**), particularly among those that are O_6 (0.54) and O_7 (0.42). To address this problem, we replaced the BCR-ABL inactive

compounds (718 compounds with a K_i or an IC_{50} value greater than or equal to 1000 nM) with 10,885 decoys generated for these targets and available from the DUD·E repository[47]. We applied the same starting parameters (i.e., 10 supporting molecules and GR ≥ 3), and pharmacophores of order O_7 were generated for both the active compounds and the decoys. MMRFS was then used to select representative O_7 pharmacophores, which led to the retention of 88 MMRFS-defined pharmacophores that fitted the active compounds (coverage of 86%) and 520 MMRFS-defined pharmacophores for the decoys (coverage of 35%). The corresponding pharmacophore network that was generated by this approach is shown in **Figure 10**. A clear distinction between the two groups can be seen. Only a few of the MMRFS-defined active pharmacophores were found among the MMRFS-defined decoy pharmacophores. This observation can be related to another result that suggested that only 24 decoys (of a total of 10885 compounds) aligned with MMRFS-defined active pharmacophores. However, no discontinuities in the local SAR can be considered in this context because while the decoys are presumed to be inactive, this has not been formally tested.

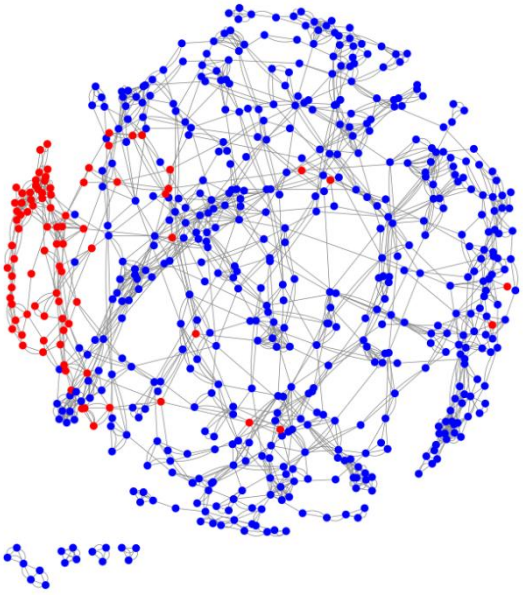


Figure 10. Active *versus* decoy pharmacophore networks for O_7 generating by using a two nearest neighbors (NNs) approach. Red nodes represent active pharmacophores and blue nodes denote decoy pharmacophores.

Application to other datasets. Our previous publication featured a critical analysis of eight datasets[1]. Here, two of these datasets were chosen for additional analysis of the results and to test our methodology on datasets that vary in size and ratios of active/inactive compounds (i.e., balancing). The AChE (acetylcholinesterase) dataset includes 3030 compounds, 861 of which are active. The hERG (human ether-à-go-go-related gene/potassium channel 1) dataset includes 4537 compounds, of which 1539 are active. The selection of representative O_7 pharmacophores using MMRFS resulted in the retention of 61 active (42% coverage) and 86 inactive compounds (32% coverage) from the AChE dataset. Similarly, 91 MMRFS-defined active pharmacophores (42% coverage) and 213 MMRFS-defined inactive pharmacophores (32% coverage) were selected from the hERG dataset. The corresponding pharmacophore networks are shown in **Figure 11**. Localized areas were identified that contained either active or inactive pharmacophores. These areas merit further analysis in future studies.

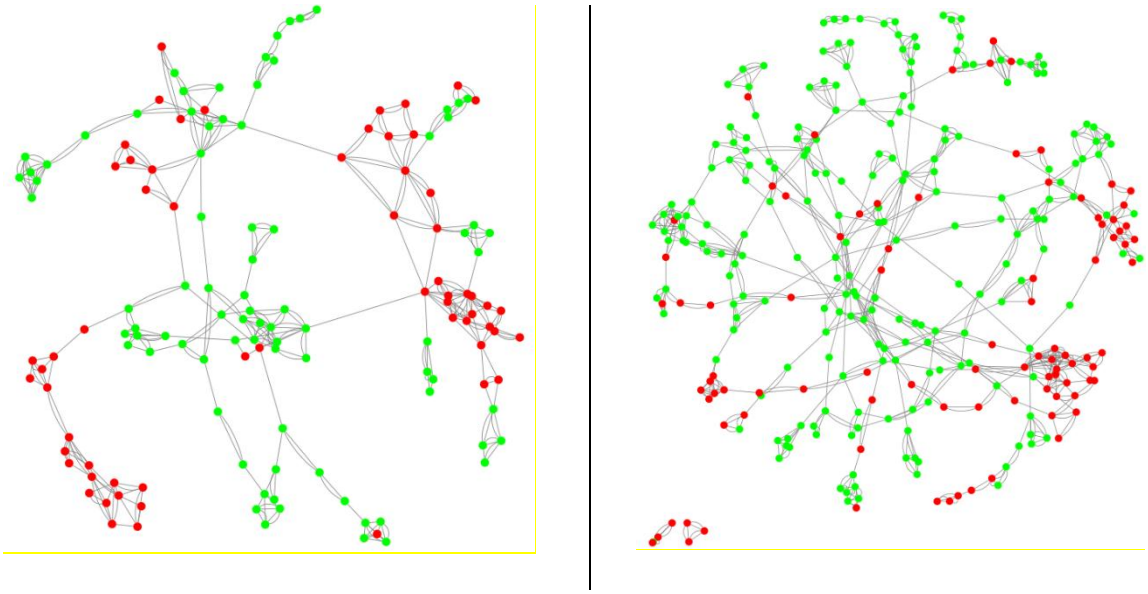


Figure 11. Active *versus* inactive pharmacophore networks for the AChE (left) and the hERG (right) datasets with O_7 pharmacophores and a two NNs. Red nodes indicate active pharmacophores and green nodes denote inactive pharmacophores.

CONCLUSION

In this study, we described the computation of graph edit distances (GEDs) between pharmacophores extracted from a BCR-ABL dataset. The application of a graph layout algorithm provided us with a visual separation between the pharmacophores associated with active compounds from those associated with inactive ones. A clustering approach was then used to characterize groups of pharmacophores that contain molecules with similar structures and biological activities. Optimized results were obtained with the highest order of pharmacophore considered in this study, *i.e.,* ones that included seven pharmacophore features (O_7). In future studies, we plan to analyze the potential of this method to identify and analyze compounds that interact with multiple targets via a cross-analysis of pharmacophores. We will also explore methods that might be used to generate efficient biological predictions from pharmacophore data based on approaches currently in development in our laboratory, including constraint-based and neural networks-based clustering approaches.

DATA AND SOFTWARE AVAILABILITY

Dataset and notebooks (scripts written with R / Python) related to this work are available at: https://osf.io/t9cbn/?view_only=7ad8c89110e245a1b603be7bfbb51fe9.

Image_Dockers for Norns and GED are available by contacting the main author: ronan.bureau@unicaen.fr

Pipeline Pilot (BIOVIA Pipeline Pilot, Release 7.5, San Diego: Dassault Systems) is commercial software.

AUTHOR INFORMATION

Corresponding Author

*Phone: (33)2-31-56-68-20

E-mail: ronan.bureau@unicaen.fr

ORCIDs

Damien Geslin : 0000-0002-5039-2837

Alban Lepailleur : 0000-0003-0202-1588

Jean-Luc Manguin : 0000-0003-4671-0107

Nhat-Vinh Vo : 0000-0002-6512-4958

Jean-Luc Lamotte : 0000-0001-6493-1769

Bertrand Cuissart: 0000-0003-4964-5427

Ronan Bureau : 0000-0001-9404-8117

Notes

The authors declare no competing financial interests.

FUNDING SOURCES

Damien Geslin received funding from Normandy's regional council (RIN 100%). Nhat-Vinh Vo received funding from Normandy's regional council (RIN AGAC). This work was supported by the ANR InvolvD project (ANR-20-CE-23-0023).

ABBREVIATIONS USED

EP, Emerging Pattern; GR, Growth Rate; MMRFS, Maximal Marginal Relevance Feature Selection; SAR, Structure-Activity Relationships.

ASSOCIATED CONTENT

Supporting Information available: Molecular data (SMILES format) for all compounds and molecular structures in each cluster. Pharmacophores with their associated compounds for each order.

REFERENCES

(1) 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. https://doi.org/10.1021/acs.jmedchem.7b01890.

(2) Langer, T.; Hoffmann, R. D. *Pharmacophores and Pharmacophore Searches*; John Wiley & Sons, 2006.

(3) Lambert, G. K.; Duhme-Klair, A.-K.; Morgan, T.; Ramjee, M. K. The Background, Discovery and Clinical Development of BCR-ABL Inhibitors. *Drug Discov. Today* **2013**, *18*, 992–1000. https://doi.org/10.1016/j.drudis.2013.06.001.

(4) 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. https://doi.org/10.1093/nar/gkw1074.

(5) Kyaw Zin, P. P.; Borrel, A.; Fourches, D. Benchmarking 2D/3D/MD-QSAR Models for Imatinib Derivatives: How Far Can We Predict? *J. Chem. Inf. Model.* **2020**, *60*, 3342–3360. https://doi.org/10.1021/acs.jcim.0c00200.

(6) Kale, M.; Sonwane, G.; Choudhari, Y. Searching for Potential Novel BCR-ABL Tyrosine Kinase Inhibitors Through G-QSAR and Docking Studies of Some Novel 2-Phenazinamine Derivatives. *Curr. Comput. Aided Drug Des.* **2020**, *16*, 501–510. https://doi.org/10.2174/1573409914666181022142934.

(7) Cheng, H.; Yan, X.; Han, J.; Hsu, C.-W. Discriminative Frequent Pattern Analysis for Effective Classification. In *2007 IEEE 23rd International Conference on Data Engineering*; 2007; pp 716–725. https://doi.org/10.1109/ICDE.2007.367917.

(8) 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. https://doi.org/10.1021/acs.jcim.8b00820.

(9) Blumenthal, D. B.; Boria, N.; Gamper, J.; Bougleux, S.; Brun, L. Comparing Heuristics for Graph Edit Distance Computation. *VLDB J.* **2020**, *29*, 419–458. https://doi.org/10.1007/s00778-019-00544-1.

(10) Fechner, N.; Papadatos, G.; Evans, D.; Morphy, J. R.; Brewerton, S. C.; Thorner, D.; Bodkin, M. ChEMBLSpace—a Graphical Explorer of the Chemogenomic Space Covered by the ChEMBL Database. *Bioinformatics* **2013**, *29*, 523–524. https://doi.org/10.1093/bioinformatics/bts711.

(11) Orlov, A. A.; Khvatov, E. V.; Koruchekov, A. A.; Nikitina, A. A.; Zolotareva, A. D.; Eletskaya, A. A.; Kozlovskaya, L. I.; Palyulin, V. A.; Horvath, D.; Osolodkin, D. I.; Varnek, A. Getting to Know the Neighbours with GTM: The Case of Antiviral Compounds. *Mol. Inform.* **2019**, *38*, 1800166. https://doi.org/10.1002/minf.201800166.

(12) Kringelum, J.; Kjaerulff, S. K.; Brunak, S.; Lund, O.; Oprea, T. I.; Taboureau, O. ChemProt-3.0: A Global Chemical Biology Diseases Mapping. *Database* **2016**, *2016*, 1–7. https://doi.org/10.1093/database/bav123.

(13) Maggiora, G. M.; Bajorath, J. Chemical Space Networks: A Powerful New Paradigm for the Description of Chemical Space. *J. Comput. Aided Mol. Des.* **2014**, *28*, 795–802. https://doi.org/10.1007/s10822-014-9760-0.
(14) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL Bioactivity Database: An Update. *Nucleic Acids Res.* **2014**, *42*, D1083–D1090. https://doi.org/10.1093/nar/gkt1031.
(15) Gillet, V. J.; Willett, P.; Bradshaw, J. Similarity Searching Using Reduced Graphs. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 338–345. https://doi.org/10.1021/ci025592e.
(16) Carbonell, J.; Goldstein, J. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*; SIGIR '98; Association for Computing Machinery: New York, NY, USA, 1998; pp 335–336. https://doi.org/10.1145/290941.291025.
(17) Japkowicz, N.; Shah, M. Performance Measures I. In *Evaluating Learning Algorithms: A Classification Perspective*; Cambridge University Press, 2011; pp 74–110. https://doi.org/10.1017/CBO9780511921803.004.
(18) Sanfeliu, A.; Fu, K. A Distance Measure Between Attributed Relational Graphs For Pattern-Recognition. *IEEE Trans. Syst. MAN Cybern.* **1983**, *13*, 353–362. https://doi.org/10.1109/TSMC.1983.6313167.
(19) Gao, X.; Xiao, B.; Tao, D.; Li, X. A Survey of Graph Edit Distance. *PATTERN Anal. Appl.* **2010**, *13*, 113–129. https://doi.org/10.1007/s10044-008-0141-y.
(20) Litsa, E. E.; Pena, M. I.; Moll, M.; Giannakopoulos, G.; Bennett, G. N.; Kavraki, L. E. Machine Learning Guided Atom Mapping of Metabolic Reactions. *J. Chem. Inf. Model.* **2019**, *59*, 1121–1135. https://doi.org/10.1021/acs.jcim.8b00434.
(21) Garcia-Hernandez, C.; Fernandez, A.; Serratosa, F. Learning the Edit Costs of Graph Edit Distance Applied to Ligand-Based Virtual Screening. *Curr. Top. Med. Chem.* **2020**, *20*, 1582–1592. https://doi.org/10.2174/1568026620666200603122000.
(22) Bastian, M.; Heymann, S.; Jacomy, M. *Gephi: An Open Source Software for Exploring and Manipulating Networks*; https://gephi.org; 2009.
(23) Jacomy, M.; Venturini, T.; Heymann, S.; Bastian, M. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLOS ONE* **2014**, *9*, e98679. https://doi.org/10.1371/journal.pone.0098679.
(24) Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*; KDD'96; AAAI Press: Portland, Oregon, 1996; pp 226–231.
(25) Wang, H.; Bah, M. J.; Hammad, M. Progress in Outlier Detection Techniques: A Survey. *IEEE ACCESS* **2019**, *7*, 107964–108000. https://doi.org/10.1109/ACCESS.2019.2932769.
(26) Rousseeuw, P. J. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7.
(27) Saxena, A.; Prasad, M.; Gupta, A.; Bharill, N.; Patel, O. P.; Tiwari, A.; Er, M. J.; Ding, W.; Lin, C.-T. A Review of Clustering Techniques and Developments. *Neurocomputing* **2017**, *267*, 664–681. https://doi.org/10.1016/j.neucom.2017.06.053.
(28) R: The R Project for Statistical Computing https://www.r-project.org/ (accessed 2020 -09 -04).
(29) Hartigan, J. A.; Wong, M. A. Algorithm AS 136: A K-Means Clustering Algorithm. *J. R. Stat. Soc. Ser. C Appl. Stat.* **1979**, *28*, 100–108. https://doi.org/10.2307/2346830.

(30) Kaufman, L.; Rousseeuw, P. J. *Finding Groups in Data: An Introduction to Cluster Analysis.*; https://doi.org/10.1002/9780470316801; John Wiley, 1990.
(31) Ng, A.; Jordan, M.; Weiss, Y. On Spectral Clustering: Analysis and an Algorithm. In *Advances in Neural Information Processing Systems*; Dietterich, T., Becker, S., Ghahramani, Z., Eds.; MIT Press, 2002; Vol. 14.
(32) Rodriguez, M. Z.; Comin, C. H.; Casanova, D.; Bruno, O. M.; Amancio, D. R.; Costa, L. da F.; Rodrigues, F. A. Clustering Algorithms: A Comparative Approach. *PLOS ONE* **2019**, *14*, 1–34. https://doi.org/10.1371/journal.pone.0210236.
(33) Tibshirani, R.; Walther, G.; Hastie, T. Estimating the Number of Clusters in a Data Set via the Gap Statistic. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2001**, *63*, 411–423. https://doi.org/10.1111/1467-9868.00293.
(34) scikit-learn: machine learning in Python — scikit-learn 0.23.2 documentation https://scikit-learn.org/stable/ (accessed 2020 -11 -04).
(35) Qing, X.; Lee, X.; De Raeymaecker, J.; Tame, J.; Zhang, K.; De Maeyer, M.; Voet, A. Pharmacophore Modeling: Advances, Limitations, and Current Utility in Drug Discovery. *J. Recept. Ligand Channel Res.* **2014**, *7*, 81–92. https://doi.org/10.2147/JRLCR.S46843.
(36) McGregor, M. J.; Muskal, S. M. Pharmacophore Fingerprinting. 1. Application to QSAR and Focused Library Design. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 569–574. https://doi.org/10.1021/ci980159j.
(37) Apsel, B.; Blair, J. A.; Gonzalez, B.; Nazif, T. M.; Feldman, M. E.; Aizenstein, B.; Hoffman, R.; Williams, R. L.; Shokat, K. M.; Knight, Z. A. Targeted Polypharmacology: Discovery of Dual Inhibitors of Tyrosine and Phosphoinositide Kinases. *Nat. Chem. Biol.* **2008**, *4*, 691–699. https://doi.org/10.1038/nchembio.117.
(38) Dimova, D.; Bajorath, J. Advances in Activity Cliff Research. *Mol. Inform.* **2016**, *35*, 181–191. https://doi.org/10.1002/minf.201600023.
(39) Stumpfe, D.; Hu, Y.; Dimova, D.; Bajorath, J. Recent Progress in Understanding Activity Cliffs and Their Utility in Medicinal Chemistry. *J. Med. Chem.* **2014**, *57*, 18–28. https://doi.org/10.1021/jm401120g.
(40) *BIOVIA Pipeline Pilot, Release 7.5*; San Diego: Dassault Systèmes.
(41) Sivaprakasam, P.; Han, X.; Civiello, R. L.; Jacutin-Porte, S.; Kish, K.; Pokross, M.; Lewis, H. A.; Ahmed, N.; Szapiel, N.; Newitt, J. A.; Baldwin, E. T.; Xiao, H.; Krause, C. M.; Park, H.; Nophsker, M.; Lippy, J. S.; Burton, C. R.; Langley, D. R.; Macor, J. E.; Dubowchik, G. M. Discovery of New Acylaminopyridines as GSK-3 Inhibitors by a Structure Guided in-Depth Exploration of Chemical Space around a Pyrrolopyridinone Core. *Bioorg. Med. Chem. Lett.* **2015**, *25*, 1856–1863. https://doi.org/10.1016/j.bmcl.2015.03.046.
(42) Zhou, T.; Commodore, L.; Huang, W.-S.; Wang, Y.; Thomas, M.; Keats, J.; Xu, Q.; Rivera, V. M.; Shakespeare, W. C.; Clackson, T.; Dalgarno, D. C.; Zhu, X. Structural Mechanism of the Pan-BCR-ABL Inhibitor Ponatinib (AP24534): Lessons for Overcoming Kinase Inhibitor Resistance. *Chem. Biol. Drug Des.* **2011**, *77*, 1–11. https://doi.org/10.1111/j.1747-0285.2010.01054.x.
(43) 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. https://doi.org/10.1016/S0092-8674(03)00194-6.
(44) Weisberg, E.; Manley, P. W.; Breitenstein, W.; Brüggen, J.; Cowan-Jacob, S. W.; Ray, A.; Huntly, B.; Fabbro, D.; Fendrich, G.; Hall-Meyers, E.; Kung, A. L.; Mestan, J.; Daley, G. Q.; Callahan, L.; Catley, L.; Cavazza, C.; Mohammed, A.; Neuberg, D.; Wright, R. D.; Gilliland, D. G.; Griffin, J. D. Characterization of AMN107, a Selective

Inhibitor of Native and Mutant Bcr-Abl. *Cancer Cell* **2005**, *7*, 129–141. https://doi.org/10.1016/j.ccr.2005.01.007.
(45) OSF https://osf.io/ (accessed 2021 -04 -14).
(46) Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. A.; Waldmann, H. The Scaffold Tree − Visualization of the Scaffold Universe by Hierarchical Scaffold Classification. *J. Chem. Inf. Model.* **2007**, *47*, 47–58. https://doi.org/10.1021/ci600338x.
(47) Mysinger, M. M.; Carchia, M.; Irwin, John. J.; Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **2012**, *55*, 6582–6594. https://doi.org/10.1021/jm300687e.

Table of Contents graphics

DECIPHERING A PHARMACOPHORE NETWORK: A CASE STUDY USING BCR-ABL DATA Damien Geslin, Alban Lepailleur, Jean-Luc Manguin, Vinh Vo Nhat, Jean-Luc Lamotte, Bertrand Cuissart, Ronan Bureau*