“Molecular Anatomy”: A New Multi-Dimensional Hierarchical Scaffold Analysis tool

Candida Manelfi
Dompé farmaceutici SpA: Dompe farmaceutici SpA https://orcid.org/0000-0002-9175-243X

Marica Gemei
Dompé farmaceutici SpA: Dompe farmaceutici SpA https://orcid.org/0000-0002-9474-0231

Carmine Talarico
Dompé farmaceutici SpA: Dompe farmaceutici SpA https://orcid.org/0000-0003-4789-0955

Carmen Cerchia
Department of Pharmacy, University of Naples Federico II https://orcid.org/0000-0002-6631-5000

Andrea R. Beccari (✉ andrea.beccari@dompe.com)
Dompé farmaceutici SpA: Dompe farmaceutici SpA https://orcid.org/0000-0001-6830-2695

Research article

Keywords: scaffold analysis, HTS data analysis, library design, network visualization, clustering

DOI: https://doi.org/10.21203/rs.3.rs-139903/v1

Abstract

The scaffold representation is widely employed to classify bioactive compounds on the basis of common core structures or correlate compound classes with specific biological activities.

In this paper, we present a novel approach called “Molecular Anatomy” as a flexible and unbiased molecular scaffold-based metrics to cluster large set of compounds. We introduce a set of nine molecular representations at different abstraction levels, combined with fragmentation rules, to define a multi-dimensional network of hierarchically interconnected molecular frameworks. We demonstrate that the introduction of a flexible scaffold definition and multiple pruning rules is an effective method to identify relevant chemical moieties. This approach allows to cluster together different molecular species with similar biological activity, capturing most of the structure activity information, in particular when libraries containing a huge number of singletons are analyzed. We also propose a procedure to derive a network visualization that allows a full graphical representation of compounds dataset, permitting an efficient navigation in the scaffold’s space and significantly contributing to perform high quality SAR analysis.

Introduction

High-throughput screening (HTS) of small-molecule libraries is routinely used in drug discovery process to identify novel leads against clinically relevant targets. Successful HTS require high quality, validated screening assays, but also an effective strategy for chemical structures selection is fundamental for the following hit-to-lead phase. HTS libraries, indeed, comprise some hits of interest, but also many compounds resulting in false positives or promiscuous hits, as well as number of compounds with no relevant biological activities at micromolar concentrations in several biological or biochemical assays [1]. The first fundamental step, affecting the probability of success of the entire lead discovery process, is represented by an incisive preliminary structure activity relationships (SAR) analysis. One of the crucial tasks in the design of large diverse libraries is the chemical space mapping. Selection of a representative subset of the desired chemical space is generally addressed by the identification of three elements: a set of meaningful descriptors [2], a similarity metric allowing to compare molecular structure pairwise [3], and a clustering algorithm for grouping structures according to the calculated pairwise similarity values [4, 5]. Many clustering algorithms exist [6], and many clustering techniques are able to address this task for groups of 10^5 to 10^7 compounds; however, the identification of relevant chemical series within the generated clusters is much more difficult. The generation of clusters organized as “series” in medicinal chemistry is an important asset of the scaffold based techniques.

A chemical scaffold, also referred to as ‘chemotype’ or ‘Markush structure’, can be defined as the common structure characterizing a group of molecules. Compounds sharing the same scaffold are likely to have a similar synthetic pathway as resulting from the concept of molecular template in combinatorial chemistry [7]. Once a scaffold is defined, SAR can be developed analyzing the effects of the substitution patterns [8]. The scaffold approach shows several advantages, in particular its outcomes are both simple to interpret and medicinal chemistry-oriented; additionally, some of the most significant features of the graph-based approaches [9] are combined with molecular fingerprint characteristics and maximum common substructure methods. Bemis et al. [10] introduced a systematic analysis of drugs according to their scaffold/framework representation which it is now a well-established method alongside molecular descriptors, molecular fingerprints and graphs. In the last twenty years, different scaffold definitions have been introduced and numerous scaffold-based computational approaches have been developed for structural classification and biological activities prediction [11]. The introduction in 2005, by Wilkens et al. [12], of hierarchies based on several kinds of scaffold deconstructions, represented a milestone in the development of scaffold-based approaches. In 2007, Schuffenhauer et al. [13] demonstrated the potentiality of combining the scaffold-based approach with *ad hoc* graphical representations through the "Scaffold Tree" algorithm and visualization tool; Schuffenhauer also introduced a rule-based ring disassembly. Since then, other decomposition and visualization tools have been developed, such as Scaffold Hunter in 2009 [14], recently revised and extended, and Scaffold Explorer in 2010 [15]. In 2008 Gianti and Sartori [16] proposed an alternative procedure to address scaffold decoration, pruning and fragmentation as a workflow for the identification of "privileged fragments". Agrafiotis et al. [15], in 2010, demonstrated that the inclusion of relevant side chains and functional groups in the scaffold representation could greatly enhance the derivation of robust SAR, thus indicating that the explicit consideration of the most significant molecular features overcomes the limits associated with "*a priori*" definition of specific pruning rules. In 2011, Varin et al. [17] proposed an extended version of the previously developed Scaffold Tree and demonstrated that rule-based approaches in fragmentation are less useful and flexible than the unbiased ones. Lipkus [18] introduced hierarchies between different abstraction levels and, finally, different hierarchical scaffold decomposition and abstraction approaches were proposed by many authors [19, 20].

All the above described methods share two major limitations; first a single ring system is represented, decorated with chains of various length, therefore, when pruned, all the molecules collapse into a degenerated cluster. Additionally, there is no relationship between scaffolds when the difference is represented by one or more ring systems. These issues are particularly limiting for the analysis and selection of vendor libraries to build diverse compound collections and, afterward, for HTS campaigns analysis in order to obtain preliminary SAR.

Very recently, Bandyopadhyay et al. [21], in order to overcome the limits related to hard clustering methods, which assign each molecule to a single cluster and so tend to place structurally analogous molecules into different and not related clusters, described a method based on fuzzy clustering that may assign a molecule to different clusters. In this method, for each molecule an exhaustive enumeration of Bemis-Murcko scaffolds, corresponding to all possible combinations of ring systems, was applied and data were annotated and aggregate at scaffold level, allowing to relate molecules on the basis of shared scaffolds. Another recent approach to perform scaffold analysis is based on retrosynthesis rules, which allow to easily identify analog series [22, 23]. Such analog series-based scaffolds can also be associated with activity information to develop possible target hypotheses for other compounds containing the same scaffold [24]. The organization of compounds in analog series leads to the formation of "constellations" of molecules, in chemical space, which can be visualized as a network of all possible molecule–core relationship [25].

However, the main limitation of the network connecting molecules and scaffolds generated with these implementations is that they are based on a unique scaffold representation, not sufficient to effectively map the chemical space of a heterogeneous ensemble of molecules, for example multi-scaffold libraries, and to capture relationships with the relative biological activity. The critical point is that it is not possible to define a priori the best representation of a molecule, because it mainly depends on the biological context and on the nearest-neighbors of the screened library.

Here, we present a novel approach, called “Molecular Anatomy”, for the generation and analysis of correlated molecular frameworks aimed at overcoming the limitations of scaffold analysis based on a single predefined set of rules. In our experience, the here identified molecular frameworks and related fragments are able to capture most of the structure activity information from HTS campaigns, and are also useful for other applications, such as library design and analysis. In particular, the combination of fragments, correlated in frameworks and wireframes, identifies relevant chemical moieties in an efficient manner, clustering together many scaffolds with similar shapes despite, for example, different dispositions of heteroatoms or small differences in bond order. To the best of our knowledge, this is a distinctive feature of our approach, compared to other known methodologies, such as the widely accepted Maximum Common Substructure (MCS) [26].

In the Methods section, the molecular scaffold representations proposed and the fragmentations rules used to generate the related fragments are defined. A COX-2 inhibitors dataset has been chosen to illustrate our approach. Then, we introduce an innovative network representation as a more convenient tool for SAR evaluation and visualization. We first apply this visualization to the molecular frameworks proposed, and then we extend the network visualization also to the underlying fragments, to show the full graphical representation. We also summarize the main advantages of our method compared to the other approaches proposed so far. Finally, we show the general applicability of our approach by performing the SAR analysis of 26092 commercial compounds tested in an HTS campaign aimed at identifying potential inhibitors of the enzyme Histone deacetylase 7 (HDAC7).

Methods

Dataset definition

COX-2

A dataset containing COX-2 inhibitors was prepared and used to evaluate the performances of the “Molecular Anatomy” approach. To this aim, the Integrity™ database was interrogated to search for COX-2 inhibitors, providing 2599 molecules in total. Of these, 819 were in preclinical development or in a higher phase. This subset was used in the following analysis steps to compare different scaffold representations. A Pipeline Pilot protocol [27] was used to standardize the molecular structures, to classify them according to molecular mechanism and highest phase, to perform substructure searches, to generate molecular frameworks according to our definition rules and, finally, to analyze the results in order to compare the different scaffold definitions.

HDAC7

A dataset of 26092 commercial compounds, tested as potential HDAC7 inhibitors during an HTS campaign performed internally within Dompé, was used as a more complex case study. The compounds were stratified in different activity classes according to the value of percent inhibition of the HDAC7 activity obtained at 10 μM concentration (Table S1).

Identification of common scaffold representations and evaluation of their performance

In theory, it is possible to define an arbitrary number of scaffold's representations based on different levels of abstraction and pruning rules. Figure 1 shows an example applied to the COX-2 inhibitor Polmacoxib (1e) [28].

Figure 1a shows the most abstracted representation, obtained removing both bond and atom type, 1b and 1c retain respectively only one of these, whereas the 1d representation corresponds to the Bemis-Murcko scaffold, containing all the rings and chains connecting them of the original molecule.

By using the most abstracted representation 1a of the Polmacoxib scaffold, corresponding to the most common moiety of COX-2 inhibitors, a subset of 224 COX-2 inhibitors was identified. Figure 2 reports the MDL substructure query and the corresponding SMARTS string used to retrieve the molecules containing this substructure.

These 224 molecules correspond to 84 different scaffolds if the less abstracted 1d representation is used (Table 1), thus resulting impossible to associate them each other as belonging to the same substructure. Furthermore, we found that 53 out of 84 Bemis-Murko scaffolds (63.1%), have one or more additional rings, corresponding to 82 of the 224 molecules (36.6%) and clustered in 34 groups according to the 1a representation, whereas the remaining 142 molecules, with exactly 3 rings, corresponding to 31 scaffolds based on the 1d representation, collapse to only one cluster if the 1a representation is used.

Table 1. Clusterization of the 819 COX2 inhibitors in preclinical development or in a higher phase, and of the subset of those matching the MDL substructure reported in Figure 2. Clusters were obtained for each representation type, distinguishing the number of those containing only molecules with exactly 3 or more than 3 rings.

	All	Matching the MDL substructure	With only 3 rings	With more than 3 rings
N. of molecules	819	224	142	82
Representation type	**N. of clusters**			
1a	280	35	1	34
1b	369	59	11	48
1c	415	78	27	51
1d	437	84	31	53

Identification of nine correlated scaffold representations

In “Molecular Anatomy” we used, as starting point, the widely accepted scaffold abstraction representation (here called *Basic Scaffold*), which is generated by removing all side chains and terminal atoms. Then, we defined a set of nine molecular frameworks (MF) at different abstraction levels to match different side chain definitions, as showed in Figure 3 for the COX-2 inhibitor Polmacoxib. We used two sets of pruning rules able to determine a multidimensional hierarchy.

The first set of rules is based on an increased level of structural information with respect to the basic scaffold. In a first step, terminal atoms with bond order greater than one are maintained (*Decorated Scaffold*); in a second step, the longest atom chain, considering also substitutions, is retained but all terminal non-carbon atoms, belonging to side chains, are iteratively pruned (*Augmented Scaffold*). In case that no terminal atoms remain removing all terminal non-carbon atom with a bond order equal to 1, decorated and augmented scaffold coincide. Some examples reported in Figure 4 explain these rules, applied to different possible cases.

The second set of rules, conversely, increases chemical abstraction by removing the atom type label and then the bond order, generating, respectively, a *Framework* and a *Wireframe* for each level of the scaffold (basic, decorated and augmented), thus finally producing nine molecular representations with a hierarchical correlation.

Fragmentation rules definition

To further overcome one general limitation of the scaffold based techniques [12, 14, 29] that, by definition, molecules sharing the same scaffold only partially belong to distinct clusters, in “Molecular Anatomy” approach we have implemented an unbiased fragmentation scheme that can be applied in parallel to all nine scaffold representations described above. These rules are explained in Figure 5, applied to specific molecules chosen on representative purpose. The first rule (Figure 5a) depicts an example of fragmentation based on an exhaustive and progressive elimination of all the internal chains from the scaffold. As second rule, unbiased ring disassembly was also implemented; the methodology for ring decomposition involves the removal of all fused rings, allowing their opening into fragments (Figure 5b). For sake of consistency, we also introduced a third rule to remove internal rings (Figure 5c).

The here reported fragmentation and deconstruction introduce other hierarchies, meaning that each fragment of the original scaffold is related with all the other representations in a multi-dimensional space. As a result of this multi-dimensional hierarchical scaffold analysis, the entire set of generated molecular frameworks are highly interconnected, and it is possible to move from one to another following SAR. To clarify this concept, Figure 6 reports an example showing how the combination of fragments and molecular frameworks at different abstraction levels allows to cluster molecules with different scaffolds.

Network representation of “Molecular Anatomy”'s frameworks

The software Cytoscape was used for creating and visualizing an MF-based network, which was also integrated with activity data for the SAR analysis. This network provides a full graphical representation of the dataset composition, allowing to easily navigate through the molecular frameworks and their hierarchical correlation. A Pipeline Pilot protocol was implemented to prepare the data matrix needed for the visualization, in the format required for the import process.

Each molecule from the dataset of 819 COX-2 inhibitors was described according to the nine molecular representations implemented in the “Molecular Anatomy” approach (Table S2); then, a unique list of frameworks was obtained (Table S3), keeping the less abstracted one in case of duplicate structures (when a same scaffold structure was obtained with different representations), corresponding to the nodes of the network. All possible parent-child relationships between the nine molecular frameworks of each molecule were generated, as reported in Table S4, according to the hierarchical relationships between the representation types shown in Figure 3, corresponding to the edges of the network.

To fully exploit this graphical representation, the network data matrix can be integrated with the enrichment factors (EF, see supplementary information) calculated for each molecular framework according to the activity data of the corresponding molecules, keeping the highest EF value in case of duplicate structures when the unique list of frameworks was generated. Applying filters based on the EF associated to each node allows to focus the network visualization on the most relevant dataset information, as described in the HDAC7 case study (see Results and Discussion)

A fully connected network representation by means of “Molecular Anatomy”'s frameworks and fragments.

“Molecular Anatomy” allows, as already described, to derive trees in multiple dimensions such as *wireframe > framework > scaffold*, or *augmented > decorated > basic* or *wireframe > ring disassembly > fragments* and in all the other possible directions maximizing the SAR information of the dataset. The network visualization can be extended also to the fragments to obtain a fully connected network, considering that the smallest fragments (e.g. benzene ring) are shared by a huge number of the original molecules. In this implementation, the network nodes can be molecular frameworks, fragments or entire small molecules, and the direction of the edges, defined by the fragmentation rule, starts from the originator fragment and end up into the corresponding fragments.

Results And Discussion

Comparison between common scaffold representations and “Molecular Anatomy” to perform SAR analysis

As shown in methods section for to the COX-2 inhibitors dataset, scaffold representations with high level of abstraction, showed in Fig. 1a-b for Polmacoxib, perform generally better than the others in the identification of relevant chemotypes. Table 1 summarizes the results obtained for each representation in terms of number of clusters generated, starting, on one hand, from all the 819 COX-2 inhibitors in preclinical development or in a higher phase, and, on the other hand, from the subset of the COX-2 inhibitors matching the MDL substructure reported in Fig. 2, the most common COX-2 inhibitor moiety. In particular, the number of clusters containing the molecules matching the common substructure with exactly or more than 3 rings was specified.

Representation 1a clusters together most of the well-known marketed drugs, such as valdecoxib and celecoxib, as well as many others leads and experimental drugs, and collapses all the 142 active molecules with exactly 3 rings to a single cluster. This cluster likely includes also several inactive molecules. Interestingly, we can note that, even though this representation is used, still almost the 40% of the structural scaffolds information, corresponding to the molecules with additional rings, would be lost in unrelated clusters, impairing the identification of the most relevant additional structural information.

Using the less abstracted representation 1d, we can retrieve and distinguish the most diverse COX-2 inhibitor scaffolds, even if this information is distributed in 84 clusters considering both those with 3 or more rings. Furthermore, an intermediate representation as 1b, where only the atom type information is removed, could allow a more effective clustering of the relevant structural information, identifying only 11 different frameworks containing molecules with exactly 3 rings, instead of 31; but, almost the same number of clusters containing molecules with more than 3 rings is generated with the two representations (48 instead of 53).

This example on COX-2 inhibitors clearly shows how this kind of analysis strongly depends on the nature of the dataset; each scaffold abstraction of Fig. 1 provides some important structural information but none of them is sufficient, alone, to capture the complexity of the heterogeneous ensemble of molecules. Only the integration of the information captured from the different scaffold abstractions, in a Multi-Dimensional Hierarchical Scaffold Analysis, allows to effectively map the entire chemical space of multi scaffold libraries. Furthermore, the combination of the “Molecular Anatomy” approach, the fragmentation rules and the network representation allows to immediately focus the attention on the most interesting and useful structural information, easily navigating among several structural clusters, moving from a molecular framework to another on the basis of their hierarchy and according to the SAR.

Attempts to identify more relevant chemical moieties have been presented in the past, for example the rule-based decompositions proposed by Schuffenhauer et al [29], schematized in Fig. 7 for three COX-2 inhibitor scaffolds. However, a clear limitation resides in the difficulty to define a priori a set of rules able to maintain a general consistency with SAR information.

The method that we propose, involving the combination of correlated molecular frameworks and fragments, is able to efficiently identify relevant chemical moieties, and to cluster together different molecular species showing similar biological activity (also in the nanomolar range) within HTS campaigns, capturing most of the SAR information.

To fully exploit the hierarchical correlation among the molecular frameworks and to generate a full graphical representation of the analyzed dataset, we also propose a network visualization. Actually, the combination of the MF approach with a network representation provides a more convenient tool for SAR evaluation and visualization [30–33], usefully guiding the user from a molecular framework to another, on the basis of their hierarchy in the direction of increasing or decreasing level of abstraction and according to the SAR.

Figure 8 shows the complete network obtained for the dataset of 819 COX-2 inhibitors. As reported in the list of statistical parameters (Fig. 8b), 280 connected components were generated, corresponding to the clusters obtained using the most abstracted (basic wireframe) representation. It is possible to clearly note the biggest cluster at the top of Fig. 8a corresponding to the 142 molecules with exactly 3 rings (Table 1), all sharing the basic wireframe 1a. Figure 8c reports the hierarchical visualization of a smaller cluster, to further show how this graphical representation of the data matrix consists in an oriented network, where nodes are in general molecular frameworks, and the direction of the edges is defined by the direction of increasing abstraction level of the molecular representations.

Furthermore, it is possible to retrieve the relationships among the diverse representations within this cluster and, focusing on the most interconnected frameworks, to identify the structural characteristic representative of the active molecules, as shown in Fig. 9. On the other hand, the network visualization clearly shows the high number of singletons that would be dispersed considering only the representation 1a. Here, thanks to the use of the fragmentation, these singletons can be related each other if containing the same fragments, allowing to easily verify if they contain characteristics in common with relevant clusters of actives.

Focusing on the fragments related to the basic wireframe representation, all the clusters identified in Fig. 8a can be connected each other in a unique network, as can be visualized in Figure S1.

Furthermore, Figure S2 shows the two fragments, cyclohexane and cyclopentane, with the highest indegree value, which means the highest number of fragments connected within the network in Figure S1.

Some qualitative considerations about the obtained networks can be done. As a first point, it is reasonable that highly connected singletons tend to be small fragments shared by a large number of molecules included in the library (as shown in Figure S2). On the contrary, low molecular weight singletons involved in a small number of connections represent potential interesting decorations of a specific group of the original molecules. If this group is enriched in a specific activity of interest, the corresponding singleton fragments connecting all the molecules included in the group, could represent a pharmacophore. As a second point, high molecular weight singleton fragments, connecting cluster of molecules with enriched activity, could represent chemical scaffolds or the “minimal chemical entity” that confers the selected activity to the cluster. As a third point, it is comprehensible that the meaning of the singleton constituting the networks may change according to the fragmentation rules used. While the approach suggested herein consists in a purely informatics fragmentation procedure, an alternative method is possible, where singletons consist in reaction intermediates derived applying retrosynthetic rules to the original molecules. In other words, in this case the network would contains, as “fragments” the precursors used to synthesize larger molecules, and as pathways connecting couple of singletons, possible synthetic strategies to attach a specific interesting low molecular weight singleton to another one representing, for example, a scaffold.

In our experience, the “Molecular Anatomy”' approach allows deciphering more easily the connections between chemotypes. In particular, filtering by EF and ranking by number of connections for each cluster allow to focus the analysis on the highly connected singletons. These frameworks have high relevance, considering that they connect different chemotypes without overlapping fragments and, then, could, suggest the most significant parts of active molecules, the fragments that could be exchanged, and the bond order and the atom type relevant for SAR derivation. This approach allows to include in SAR analysis also molecules usually underestimated because singletons, or compounds with small ligand efficacy, but here connected to relevant clusters corresponding to specific series of compounds. In this way, a valuable information could be added in the SAR of this major hit series, connecting them to additional latent ones [34]. This method could be considered an extension of the already proposed compound set enrichment [17, 35, 36], based on an implementation of an higher level of abstraction, potentially able to identify new hit series connected with the conventional one.

Case study: SAR analysis of an HTS campaign on HDAC7.

In order to better illustrate the molecular scaffold representations and the fragmentations rules that we introduced and with the intent to clarify the advantages to use the network visualization proposed for SAR evaluation, we present, as case study, the SAR analysis of the HTS campaign on HDAC7 performed for 26092 compounds.

First, the set of nine molecular frameworks at different abstraction levels were generated for the entire dataset. For each of the nine frameworks, the EF was calculated (according to the formula provided in SI), based on the inhibition data of the corresponding molecules; molecules were considered as active if belonging to the activity classes moderate, strong and very strong (Table S1).

Figure 10 shows the complete network obtained with Cytoscape, as described in Methods section, for this dataset, that clearly appears a more complex case study compared to the previous one, thus chosen to show the potentiality of our approach. 3061 connected components were generated, corresponding to the clusters obtained using the most abstracted (basic wireframe) representation.

The most interesting basic wireframe in terms of SAR evaluations are selected (Figure S3), filtered by the highest values of EF and number of connected active molecules, to focus the analysis on the abstracted scaffolds accounting for more actives.

Figure 11a reports the network corresponding to one of these selected clusters, using a hierarchical layout for a better visualization. The complexity of this specific network is due to the high number of nodes corresponding to all the molecules (on top, in light blue) and relative molecular frameworks (all other nodes) matching the basic wireframe reported in Fig. 11c. This complex network may however be considerably simplified removing nodes with EF value equal to 0, that is, removing all the nodes connected to inactive molecules. Applying this filter, a more clear and useful oriented network can be obtained (Fig. 11b), with exactly the most relevant dataset information. In this way, it is possible to easily extract only the interesting pathways in terms of SAR analysis, starting from a huge number of connections that ensure a complete evaluation of the structural information.

In more detail, starting from the basic wireframe selected (Fig. 11c), thanks to the network visualization, two more interesting sub-clusters can be identified corresponding to the decorated wireframe (definition in Fig. 3) reported in Fig. 12. The EF values of this decorated wireframe are higher than that of the basic wireframe in common, meaning that such approach allows focusing on specific characteristics of the active compounds. Furthermore, it is also possible to move to the less abstracted representation within the network, the decorated frameworks also reported in Fig. 12, that provide information about the bond order characteristics common to the active compounds. And so on, moving back through the network toward the lowest abstraction level is it possible to visualize the original molecules.

A first interesting consideration about these results concerns the introduction of decorations in our scaffold representations: defining a description level in which protruding bonds are added to the basic scaffold allows to better identify and distinguish the requirement essential for the activity. This point is clearly showed in Figs. 11 and 12, where moving from the basic to the decorated wireframes with higher values of EF and number of connections, it is possible to retrieve all the clusters containing the active molecules. On the other hand, 12 decorated wireframes and 37 decorated frameworks are identified in common with inactive molecules, another useful information to rationalize which scaffold decorations are responsible of decrease or even loss of activity.

Finally, we want to show how the most useful SAR information can be obtained extending the analysis and the network to the fragments. When the fragmentation rules are applied to the dataset, the network visualization of the fragmented library allows to interconnect all the molecular frameworks containing the same fragment and the EF can be recalculated for each fragment according to the activity data of all the molecules connected via the corresponding molecular frameworks.

In particular, focusing the attention on the interesting structures above identified, Fig. 13 reports the same scheme of Fig. 12, with the EF values recalculated considering all the clusters identified by molecular frameworks corresponding to superscaffolds of the scaffolds visualized (superframeworks).

To better explain this step, we report in Fig. 14, as example, one decorated wireframe of Fig. 13 and the corresponding five decorated wireframes retrieved in the fragmented library containing it as a fragment. For each of these decorated wireframes, the EF value is reported and that of the central wireframe, here treated as a fragment, is recalculated, adding the contribution of the other five ones. Comparing Figs. 12 and 13, it is possible to identify the molecular frameworks, the EF of which increases when they are considered as fragments, thus containing relevant structural characteristic of active molecules.

We can observe that, among the nine molecular frameworks, in this particular case study, the decorated wireframe turned out to be the most useful representation to obtain SAR information. Thus, in general we can conclude that the integration of all molecular frameworks and fragments in the network visualization is crucial for capturing the most relevant information in compound libraries analysis.

Conclusions

We propose “Molecular Anatomy” as a fast and flexible method for the analysis of the chemical space, library design and SAR studies.

This set of tools could be useful in the management of large compounds collections, for example in the analysis of HTS campaign results, as well as in focused libraries design. On the other side, this kind of data organization allows to efficiently analyze scaffold-activity relationship, identify relevant clusters and easily connect different chemotypes with biological activity. The limitation in the underestimation of the side chain effect can be easily circumvented combining the “Molecular Anatomy” approach with other techniques, such as matched molecular pairs (MMP) [37–39]. In this case, the identified molecular frameworks can make MMP equally or even more efficient and consistent than other methods [40, 41]. Furthermore, using MA with a higher level of abstraction, it is possible to compare effectively SAR behaviors on multiple scaffolds, or support scaffold hopping strategies.

Another interesting application, still in an early phase of evaluation, is the possibility to annotate a library according to therapeutic areas information of classified drugs, in order to accelerate the identification of target-based or disease-based libraries, using for example annotated database such as MDDR, WOMBAT [42], or public databases [43].

Furthermore, molecular frameworks can be profitably used for compounds clustering and database indexing. One of the most critical tasks in the design of large diverse libraries is the comprehensive mapping of chemical space. The generation of groups that can be considered as “series” in a medicinal chemist’s perception represents an important asset of the scaffold-based techniques. However, the clusters generated by currently available approaches generally tend to contain an elevated number of scaffolds, hampering the selection of chemical series for follow-up activities. Some improvements have been introduced to overcome this problem, for example by means of Maximum Overlapping Set (MOS) [44]. The main advantage of scaffold-based clustering techniques is that they do not require the calculation of similarity indices, nor pairwise similarities: indeed, the scaffold structure itself represents the aggregation rule, so that each molecule is assigned to a cluster regardless of the nature of the neighbors. In this sense, the approach can be defined as a “Natural Clustering” (NC) method. Another important feature is that no bias, like the average number of molecules per group or the expected number of scaffolds, has to be introduced. This is very useful when some over-represented scaffolds may drive the analysis. Therefore, MA enables the hierarchical clustering, considerably extending the potentialities of scaffold-based clustering.

Finally, NC makes relational databases ideal for chemical graph-based compound clustering applications. The composition of a specific cluster is independent from the chemical structure of the other clusters, and, once the scaffold abstraction is defined, isn’t needed to re-cluster the whole library.

Organizing molecules within a database by means of clustered SQL indices (based on the MA), can dramatically reduce the time required for substructure searches, as reported by Wilkens et al. [45] and Masciocchi et al. [46]. In our implementation, due to the higher abstraction of the frameworks and wireframes, it is also possible to further speed up substructure searches using these representations as a wild character-like query, such as “any atom” or “any bond”. Interestingly, MA are, per se, searchable molecular representations, and this allows to define local similarities in substructure searches space. For example, the scaffold of a target molecule could be searched with either a lower or higher similarity to the reference template. Besides that, it is also possible to constrain the local diversity of the scaffold by requiring, for example, the presence of a specific hydrogen bond acceptor at a given position on the scaffold, or even specifying a LogP range.

All these prospective applications make the MA approach a valuable cheminformatics tool that can considerably improve any kind of structural data analysis.

Abbreviations

HTS: High-throughput screening; SAR: structure activity relationships; MCS: Maximum Common Substructure; HDAC7: Histone deacetylase 7; MF: molecular frameworks; EF: enrichment factor; MOS: Maximum Overlapping Set; NC: Natural Clustering; MMP: matched molecular pairs; MOS: Maximum Overlapping Set.

Declarations

Acknowledgements

Author’s contribution

All authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

Datasets generated and analysed during this study are included in this published article and its supplementary information files.

The protocols employed for the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

References

1. Macarron R (2015) Chemical libraries: How dark is HTS dark matter? Nat Chem Biol 11:904–905. https://doi.org/10.1038/nchembio.1937
2. Bender A, Jenkins JL, Scheiber J, et al (2009) How similar are similarity searching methods? A principal component analysis of molecular descriptor space. J Chem Inf Model 49:108–119. https://doi.org/10.1021/ci800249s
3. Todeschini R, Consonni V, Xiang H, et al (2012) Similarity Coefficients for Binary Chemoinformatics Data: Overview and Extended Comparison Using Simulated and Real Data Sets. J Chem Inf Model. https://doi.org/10.1021/ci300261r
4. Brown RD, Martin YC (1996) Use of Structure−Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. J Chem Inf Comput Sci 36:572–584. https://doi.org/10.1021/ci9501047
5. McGregor MJ, Pallai P V (1997) Clustering of Large Databases of Compounds: Using the MDL "Keys" as Structural Descriptors. J Chem Inf Comput Sci 37:443–448. https://doi.org/10.1021/ci960151e
6. Raymond JW, Blankley CJ, Willett P (2003) Comparison of chemical clustering methods using graph- and fingerprint-based similarity measures. J Mol Graph Model 21:421–433
7. Katritzky AR, Kiely JS, Hebert N, Chassaing C (2000) Definition of Templates within Combinatorial Libraries. J Comb Chem 2:2–5
8. Hu Y, Bajorath J (2011) Target family-directed exploration of scaffolds with different SAR profiles. J Chem Inf Model 51:3138–3148. https://doi.org/10.1021/ci200461w
9. Bonchev D, Rouvray DH (1991) Chemical graph theory : introduction and fundamentals. Abacus, New York ; London

10. Bemis GW, Murcko MA (1996) The properties of known drugs. 1. Molecular frameworks. J Med Chem 39:2887–2893. https://doi.org/10.1021/jm9602928
11. Hu Y, Stumpfe D, Bajorath J (2016) Computational Exploration of Molecular Scaffolds in Medicinal Chemistry. J Med Chem 59:4062–4076. https://doi.org/10.1021/acs.jmedchem.5b01746
12. Wilkens SJ, Janes J, Su AI (2005) HierS: hierarchical scaffold clustering using topological chemical graphs. J Med Chem 48:3182–3193. https://doi.org/10.1021/jm049032d
13. Schuffenhauer A, Ertl P, Roggo S, et al (2007) The scaffold tree–visualization of the scaffold universe by hierarchical scaffold classification. J Chem Inf Model 47:47–58. https://doi.org/10.1021/ci600338x
14. Wetzel S, Klein K, Renner S, et al (2009) Interactive exploration of chemical space with Scaffold Hunter. Nat Chem Biol 5:581–583. https://doi.org/10.1038/nchembio.187
15. Agrafiotis DK, Wiener JJ (2010) Scaffold explorer: an interactive tool for organizing and mining structure-activity data spanning multiple chemotypes. J Med Chem 53:5002–5011. https://doi.org/10.1021/jm1004495
16. Gianti E, Sartori L (2008) Identification and selection of “privileged fragments” suitable for primary screening. J Chem Inf Model 48:2129–2139. https://doi.org/10.1021/ci800219h
17. Varin T, Schuffenhauer A, Ertl P, Renner S (2011) Mining for bioactive scaffolds with scaffold networks: improved compound set enrichment from primary screening data. J Chem Inf Model 51:1528–1538. https://doi.org/10.1021/ci2000924
18. Lipkus AH, Yuan Q, Lucas KA, et al (2008) Structural diversity of organic chemistry. A scaffold analysis of the CAS Registry. J Org Chem 73:4443–4451. https://doi.org/10.1021/jo8001276
19. Vogt M, Huang Y, Bajorath J (2011) From activity cliffs to activity ridges: informative data structures for SAR analysis. J Chem Inf Model 51:1848–1856. https://doi.org/10.1021/ci2002473
20. Hu Y, Stumpfe D, Bajorath J (2011) Lessons learned from molecular scaffold analysis. J Chem Inf Model 51:1742–1753. https://doi.org/10.1021/ci200179y
21. Bandyopadhyay D, Kreatsoulas C, Brady PG, et al (2019) Scaffold-Based Analytics: Enabling Hit-to-Lead Decisions by Visualizing Chemical Series Linked across Large Datasets. J Chem Inf Model 59:4880–4892. https://doi.org/10.1021/acs.jcim.9b00243
22. Stumpfe D, Dimova D, Bajorath J (2016) Computational Method for the Systematic Identification of Analog Series and Key Compounds Representing Series and Their Biological Activity Profiles. J Med Chem 59:7667–7676. https://doi.org/10.1021/acs.jmedchem.6b00906
23. Dimova D, Stumpfe D, Hu Y, Bajorath J (2016) Analog series-based scaffolds: computational design and exploration of a new type of molecular scaffolds for medicinal chemistry. Futur Sci OA 2:FSO149. https://doi.org/10.4155/fsoa-2016-0058
24. Cerchia C, Dimova D, Lavecchia A, Bajorath J (2017) Exploring Structural Relationships between Bioactive and Commercial Chemical Space and Developing Target Hypotheses for Compound Acquisition. ACS Omega 2:7760–7766. https://doi.org/10.1021/acsomega.7b01338

25. Naveja JJ, Medina-Franco JL (2019) Finding Constellations in Chemical Space Through Core Analysis . Front. Chem.   7:510
26. Hariharan R, Janakiraman A, Nilakantan R, et al (2011) MultiMCS: a fast algorithm for the maximum common substructure problem on multiple molecules. J Chem Inf Model 51:788–806. https://doi.org/10.1021/ci100297y
27. Dassault Systèmes BIOVIA (2016) BIOVIA Pipeline Pilot
28. Penning TD, Talley JJ, Bertenshaw SR, et al (1997) Synthesis and biological evaluation of the 1,5-diarylpyrazole class of cyclooxygenase-2 inhibitors: identification of 4-[5-(4-methylphenyl)-3-(trifluoromethyl)-1H-pyrazol-1-yl]benze nesulfonamide (SC-58635, celecoxib). J Med Chem 40:1347–1365. https://doi.org/10.1021/jm960803q
29. Ertl P, Schuffenhauer A, Renner S (2011) The scaffold tree: an efficient navigation in the scaffold universe. Methods Mol Biol 672:245–260. https://doi.org/10.1007/978-1-60761-839-3_10
30. Xiong B, Liu K, Wu J, et al (2008) DrugViz: a Cytoscape plugin for visualizing and analyzing small molecule drugs in biological networks. Bioinformatics 24:2117–2118. https://doi.org/10.1093/bioinformatics/btn389
31. Shannon P, Markiel A, Ozier O, et al (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–2504. https://doi.org/10.1101/gr.1239303
32. Iyer P, Stumpfe D, Bajorath J (2011) Molecular mechanism-based network-like similarity graphs reveal relationships between different types of receptor ligands and structural changes that determine agonistic, inverse-agonistic, and antagonistic effects. J Chem Inf Model 51:1281–1286. https://doi.org/10.1021/ci2001378
33. Lepp Z, Huang C, Okada T (2009) Finding key members in compound libraries by analyzing networks of molecules assembled by structural similarity. J Chem Inf Model 49:2429–2443. https://doi.org/10.1021/ci9001102
34. Varin T, Didiot MC, Parker CN, Schuffenhauer A (2012) Latent hit series hidden in high-throughput screening data. J Med Chem 55:1161–1170. https://doi.org/10.1021/jm201328e
35. Varin T, Gubler H, Parker CN, et al (2010) Compound set enrichment: a novel approach to analysis of primary HTS data. J Chem Inf Model 50:2067–2078. https://doi.org/10.1021/ci100203e
36. Kruger F, Stiefl N, Landrum GA (2020) rdScaffoldNetwork: The Scaffold Network Implementation in RDKit. J Chem Inf Model 60:3331–3335. https://doi.org/10.1021/acs.jcim.0c00296
37. Griffen E, Leach AG, Robb GR, Warner DJ (2011) Matched molecular pairs as a medicinal chemistry tool. J Med Chem 54:7739–7750. https://doi.org/10.1021/jm200452d
38. Wassermann AM, Bajorath J (2011) Large-scale exploration of bioisosteric replacements on the basis of matched molecular pairs. Future Med Chem 3:425–436. https://doi.org/10.4155/fmc.10.293
39. Leach AG, Jones HD, Cosgrove DA, et al (2006) Matched molecular pairs as a guide in the optimization of pharmaceutical properties; a study of aqueous solubility, plasma protein binding and

oral exposure. J Med Chem 49:6672–6682. https://doi.org/10.1021/jm0605233

40. Hussain J, Rea C (2010) Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. J Chem Inf Model 50:339–348. https://doi.org/10.1021/ci900450m
41. Hu X, Hu Y, Vogt M, et al (2012) MMP-Cliffs: systematic identification of activity cliffs on the basis of matched molecular pairs. J Chem Inf Model 52:1138–1145. https://doi.org/10.1021/ci3001138
42. Keiser MJ, Setola V, Irwin JJ, et al (2009) Predicting new molecular targets for known drugs. Nature 462:175–181. https://doi.org/10.1038/nature08506
43. Zhou Y, Zhou B, Chen K, et al (2007) Large-scale annotation of small-molecule libraries using public databases. J Chem Inf Model 47:1386–1394. https://doi.org/10.1021/ci700092v
44. Stahl M, Mauser H, Tsui M, Taylor NR (2005) A robust clustering method for chemical structures. J Med Chem 48:4358–4366. https://doi.org/10.1021/jm040213p
45. Wilkens SJ (2006) Relational database driven two-dimensional chemical graph analysis. Chem Biol Drug Des 68:135–138. https://doi.org/10.1111/j.1747-0285.2006.00426.x
46. Masciocchi J, Frau G, Fanton M, et al (2009) MMsINC: a large-scale chemoinformatics database. Nucleic Acids Res 37:D284-90. https://doi.org/10.1093/nar/gkn727

Supplementary Information

Additional file 1: Table S1. Dataset of 26092 commercial compounds tested at 10 μM concentration against HDAC7, classified on the basis of enzyme activity percent of inhibition. Figure S1: Cytoscape network visualization of the 819 COX-2 inhibitors subset where nodes includes fragments related to the basic wireframe representation, contributing to create a fully connected unique network.

Figure S2: The two fragments (cyclohexane and cyclopentane, shown in the foreground), extracted from the basic wireframe representation, with the highest number of connections (indegree) in the Cytoscape network visualization reported in Figure S1.

Figure S3: Selection of the most interesting basic wireframe, corresponding to the most abstracted representation in common within each cluster of the network, filtered by the highest values of EF and number of connected active molecules of the corresponding cluster.

Additional file 2: Table S2. Dataset of 819 COX-2 inhibitors in preclinical development or in a higher phase, described according to the nine molecular representations of “Molecular Anatomy”.

Additional file 3: Table S3. List of unique frameworks obtained from the molecules in Table S2.

Additional file 4: Table S4. List of parent-child relationships between the nine molecular frameworks of molecules listed in Table S2.