# Scaffold-Based Analytics: Enabling Hit-to-Lead Decisions by Visualizing Chemical Series Linked Across Large Datasets

Deepak Bandyopadhyay,[*,†,¶] Constantine Kreatsoulas,[†] Pat G. Brady,[†] Joseph Boyer,[†] Zangdong He,[†] Genaro Scavello Jr.,[†] Tyler Peryea,[‡] Ajit Jadhav,[‡] Dac-Trung Nguyen,[*,‡] and Rajarshi Guha[*,‡,§]

†*GlaxoSmithKline, 1250 S. Collegeville Rd, Collegeville, PA 19426*

‡*National Center for Advancing Translational Science, 9800 Medical Center Drive, Rockville, MD 20850*

¶*Current address: Janssen Pharmaceutical Companies of Johnson and Johnson, McKean and Welsh Roads, Spring House, PA 19477*

§*Current address: Vertex Pharmaceuticals, 50 Northern Avenue, Boston MA 02210*

E-mail: debug22@gmail.com; nguyenda@mail.nih.gov; rajarshi_guha@vrtx.com

**Abstract**

We present a method for visualizing and navigating large screening datasets, taking into account their activities and properties. Our approach is to annotate the data with all possible scaffolds contained within each molecule. We have developed a Spotfire visualization, coupled to a fuzzy clustering approach based on the scaffold decomposition of the screening deck, that is used to drive the hit triage process. Progression decisions can be made using aggregate scaffold parameters and data from multiple datasets merged at the scaffold level. This visualization reveals overlaps that help prioritize

hits, highlight tractable series and posit ways to combine aspects of multiple hits. The SAR of a large and complex hit is automatically mapped onto all constituent scaffolds making it possible to navigate, via any shared scaffold, to all related hits. This scaffold "walking" helps address bias toward a handful of potent and ligand-efficient molecules at the expense of coverage of chemical space. We consider two scaffold generation methods and explored their similarities and differences both qualitatively and quantitatively. The workflow of a Spotfire visualization used in combination with fuzzy clustering and structure annotation provides a intuitive view of large and diverse screening datasets. This allows teams to effortlessly navigate between structurally related molecules and enriches the population of leads considered and progressed in a manner complementary to established approaches.

# Introduction

The advantage and disadvantage of high throughput screening (HTS) campaigns is the large amount of data that is generated. While the value of large scale HTS has been debated,[1] the massive structure-activity datasets generated create a challenge in identifying truly active compounds and their analogs and weeding out false positives. The process of reducing HTS datasets from hundreds of thousands of compounds to a few thousand or hundred active series is termed triaging. Over the last twenty years many approaches to HTS triaging have been described which include activity based thresholds,[2] similarity to known actives,[3] enrichment based approaches,[4,5] ranges of physicochemical properties,[6] crowdsourcing[7] and removal of promiscuous or otherwise undesirable chemotypes.[8] See Shun et al.[9] and Langer et al.[10] for a review of HTS triage approaches.

A key challenge in the triage step is to identify structure-activity series - sets of compounds with similar or analogous structures that exhibit a spectrum of activity (including lack of activity). Identifying such subsets allows one to have some confidence in the presence of a structure-activity relationship amongst the active compounds which enables a more

efficient exploration of the chemical space around the selected hits. A good review of computational methods to extract SAR from screening datasets can be found in Wawer et al.[11].

Given that a SAR series is defined in terms of a core structure along with various decorations, the first step in the triage process is to identify these core structures, or scaffolds. This starts by decomposing the structures in the screening collection, either exhaustively or else using one of the many methods to fragment molecules such as Bemis-Murcko[12,13] or RECAP.[14] These methods lead to a large number of fragments ranging from trivial ones such as a benzene ring to very complex multiring structures. Thus, a key step involves identifying the relevant set of scaffolds and the associated exemplars. This can be challenging since a given compound can contain multiple scaffolds and scaffold relevance can be a subjective decision.[15]

In this work we present a HTS triage workflow based on navigating the scaffold-activity landscape of a screening collection using a fuzzy clustering method to group compounds based on scaffold membership. The workflow includes methods to visualize the activity landscape as well as methods to explore different regions of chemical space via shared scaffolds.

## Related Work

While there are many ways to generate a set of scaffolds from a compound collection, a key step is to identify a relevant subset or else aggregate them in a way that leads to a *useful* clustering of active and inactive compounds. While the term "useful" is rather subjective, it is easy to identify cases that are not actionable by chemistry teams. For example, 5- or 6-member undecorated rings are likely not useful since they will occur in the majority of compounds in a screening collection. At the other extreme, large, extended scaffolds associated with very few compounds are also likely not useful.

As a result, many approaches to scaffold aggregation have been described. A natural approach is to consider a hierarchical aggregation. The Scaffold Tree[16] and Scaffold Network[17] define a hierarchical decomposition from more specialized larger scaffolds to more inclusive

smaller scaffolds. While the Scaffold Tree splits each larger scaffold in exactly one way into two scaffolds with fewer rings, the Scaffold Network performs an exhaustive decomposition into all possible smaller scaffolds with fewer rings. Since some subscaffolds are shared with neighboring scaffolds, this produces a network or graph rather than a tree. Harper et al.[18] use exhaustive enumeration to find all Bemis-Murcko like frameworks in each molecule, and then recursively retain those with highest aggregate activity, removing molecules that contain them until a threshold is met, yielding a set of disjoint frameworks. Other methods have used multiple common substructure (MCS), first proposed for finding protein structural similarity,[19] for example Quintus et al.[20] and the ChemAxon product LibraryMCS. Recent work has leveraged the Graph Edit Distance to define the MCS,[21] and calculated it on reduced graphs(,[22 18]) as a more general way to define similarity and retrieve analogs.

Multiple scaffolds if present in a dataset can be inferred from the scaffold tree decomposition.[23] However in practice, the thresholds used by Clark and Labute[23] miss common scaffolds in HTS-like diverse chemical compound sets. Bandyopadhyay[24] used a common fragment decomposition plug-in for SAReport in order to find and export scaffolds for diverse datasets. This approach assigns at most one scaffold for each molecule, based on the order they are selected by the user from a prioritized list. The user interaction in this analysis introduces subjectivity and reduces repeatability.

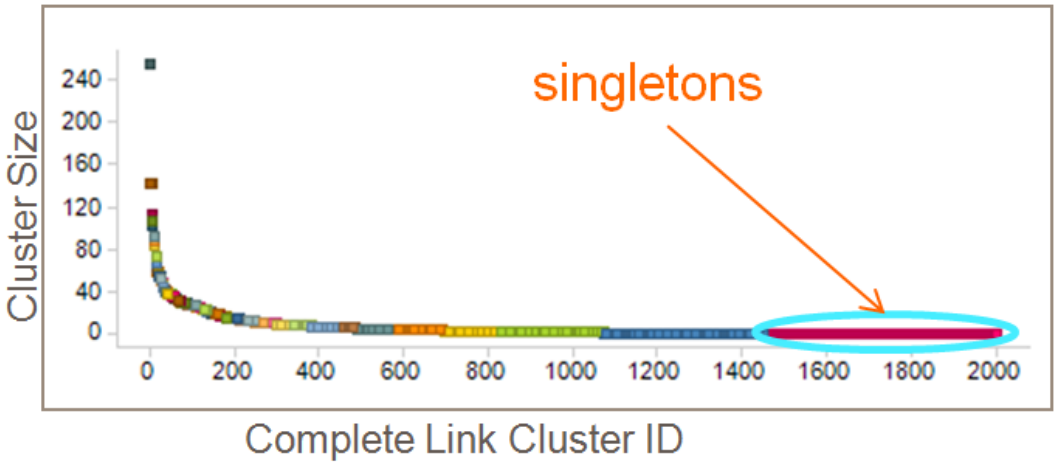



Figure 1: Singletons in a complete-linkage clustering of the TCAMS dataset.

Assigning each molecule to a single cluster partitions the dataset into disparate, non-overlapping groups, also termed a hard clustering; agglomerative fingerprint-based clustering methods such as sphere exclusion, single and complete linkage clustering[25] work this way. These methods tend to place structurally analogous molecules in different clusters where they would not be discovered as analogs of each other - see an example in Supplementary Material Section S8.

On the other hand, a molecule could be assigned simultaneously to multiple clusters, depending on the structural features present. Traditional hard clustering methods will not allow this and will assign such molecules to their own cluster, thus identifying them as singletons. Such a result can be observed in real chemical datasets as well – for example, the complete-linkage clustering of the TCAMS dataset[26,27] has nearly 25% of the 2000 clusters identified with just one compound, as shown in Figure 1.

In contrast to hard clustering, the methods described here rely on fuzzy clustering, where a molecule may be assigned to multiple clusters. Fuzzy clusters have rarely been used in cheminformatics (for example[28,29]), perhaps because they are hard to visualize and navigate. In this work we provide an intuitive visualization framework for a fuzzy clustering with multiple overlapping scaffolds per molecule. We believe this framework can help end users easily navigate a diverse chemical dataset described by a fuzzy clustering.

## Datasets

The datasets used to illustrate and visualize our methods were picked to represent screening datasets we expect the method to be used on in practice. For example, the TCAMS dataset[26] consists of 13,500 diverse hits from an antimalarial screen at GSK, along with pIC50 against a susceptible strain of the malarial parasite (3D7), percentage inhibition against a resistant strain (DD2), Hep G2 hepatotoxicity, a few physical chemical properties (e.g., molecular weight, aromatic ring count, cLogP), and Inhibition Frequency Index (IFI, a measure of promiscuity defined as the percentage of screens in which a molecule inhibits over 50%,[30]).

The dataset is available in ChEMBL's Neglected Tropical Disease section.[31]

The in-house GSK kinase dataset shown ("Kinase X"), comprises 9259 compounds from four screening datasets for a recent kinase program. Three of these have associated activity data from several on-target and off-target assays:

- An *FBDD*[32] (fragment-based drug design) screen run in 2011 with **288** hits.
- A *full HTS* run in 2012 with **4564** pIC50s measured after screening 2 million compounds at a single concentration (10 uM).
- A *top-up HTS* run in 2014 with **3613** pIC50s from screening 350,000 compounds.

A fourth dataset comprises **824** so-called virtual compounds whose activity was inferred from *ELT*, a DNA Encoded Library screen[33] of 130 combinatorial libraries comprising billions of potential molecules. Triaged ELT compounds are synthesized off-DNA to check if they are indeed active.

The GSK Kinase dataset also contains physico-chemical properties of interest for developability, notably the Property Forecast Index (PFI[34]). We chose this dataset to illustrate the power of Scaffold Analytics in joining and merging datasets from multiple screens, combining their SAR to design hybrid molecules, and making inferences about unknown activity in one screen based on known activity in another screen.

Next, we describe several methods we have assembled as part of our workflow that enable multiple scaffolds to be assigned to each molecule and easy navigation between molecules related by these scaffolds.

## Dataset Preprocessing

The typical dataset under consideration is available as a comma-separated text file (CSV), whereas most of the scaffold decomposition methods described expect MDL SD-files. To convert CSV to SDF while preserving non-molecule fields as SDF properties, we have created a simple workflow in Pipeline Pilot,[35] but this can also be done via standard cheminformatics

toolkits such as JChem[36] or RDKit.[37] Prior to SDF conversion, activity or property columns that are not to be aggregated at the scaffold level should be deleted from the CSV file, to accelerate analysis and aggregation. Further quirks specific to one of the methods, the NCATS R-group tool, will be described in the Supplementary Material in section S3.

## Partitioning Method: Complete Linkage Clustering

For comparison purposes, we include the default method used at GSK to visualize groups of molecules in our visualizations. This method produces an output file in which the unique cluster ID (CLink) and number of other molecules in the same cluster (N_Clink) are added as additional fields to the original dataset. Complete linkage chemical clustering is described further in several references such as Downs and Barnard[25], Jain[38].

## Fragmentation Method: NCATS R-Group Tool

The NCATS R-group analysis tool (`https://tripod.nih.gov/?p=46`,[39]) was developed to automatically and exhaustively generate R-group tables from a dataset using all scaffolds, defined as chemical substructures shared by two or more molecules. The scaffolds are defined as molecular fragments generated by exhaustive enumeration of all possible combinations of the Smallest Set of Smallest Rings (SSSR), as described at `https://tripod.nih.gov/?p=160`. For a molecule with $k$ SSSR, the maximum possible number of such scaffolds is $2^k - 1$; however the actual number is usually much lower due to symmetry and additional constraints (e.g., reactivity, synthetic accessibility). Briefly, the fragments are generated based on the following rules:

1. Generate Bemis-Murcko framework by iteratively pruning all pendant atoms except for carbonyl (or any terminal double bond).

2. Exhaustively enumerate all possible combinations of ring system. This is achieved by iteratively breaking non-aromatic bonds and keeping unique set of ring systems.

3. Each fragment generated should be a valid substructure of the parent molecule. This constraint is approximated by ensuring the fingerprint for the fragment is a subset of the parent molecule.

As an example, in Figure 2(b) we see the five scaffolds that were generated from the molecule shown in Figure 2(a) from the TCAMS dataset.

The original NCATS R-group tool focused on scaffold decomposition and the subsequent generation of R-group tables for each scaffold. Since then it has been extended to support scaffold hopping as well as providing contextual data (e.g., literature references, activity summaries) and network visualizations of scaffold relationships. All results from this study were generated in version 8 of the R-group tool, `http://tripod.nih.gov/ws/rgroupbeta/rgrouptool8.jar`). The latest version available at time of writing is `http://tripod.nih.gov/ws/rgroupbeta/rgrouptool11.jar`), which has been tested and produces substantially similar output.

When running from the command-line, ensuring 16G of memory (via `-Xmx16G`) enables datasets of upto 40k compounds with a handful of numeric activity columns to be analyzed without running out of memory. The scaffolds along with R-group tables for each scaffold can be exported in a set of TSV files or a single JSON file. In the current work we employed the TSV format which is described in detail in Supplementary Tables S1 & S2.

## Fragmentation Method: Frameworks

The R-group Tool described above uses the NCATS implementation of Bemis-Murcko Frameworks to generate scaffolds. We compare it to the Harper et al.[18] implementation of Frameworks. The method generates many framework types, from which we retain two that most resemble NCATS scaffolds for analysis: Bemis-Murcko-like[13] (with atom and bond orders retained) and RECAP.[14]

The frameworks found within the same molecule from TCAMS are shown in Figure 2(c). The reader will observe several differences from the R-group tool: there are more scaffolds

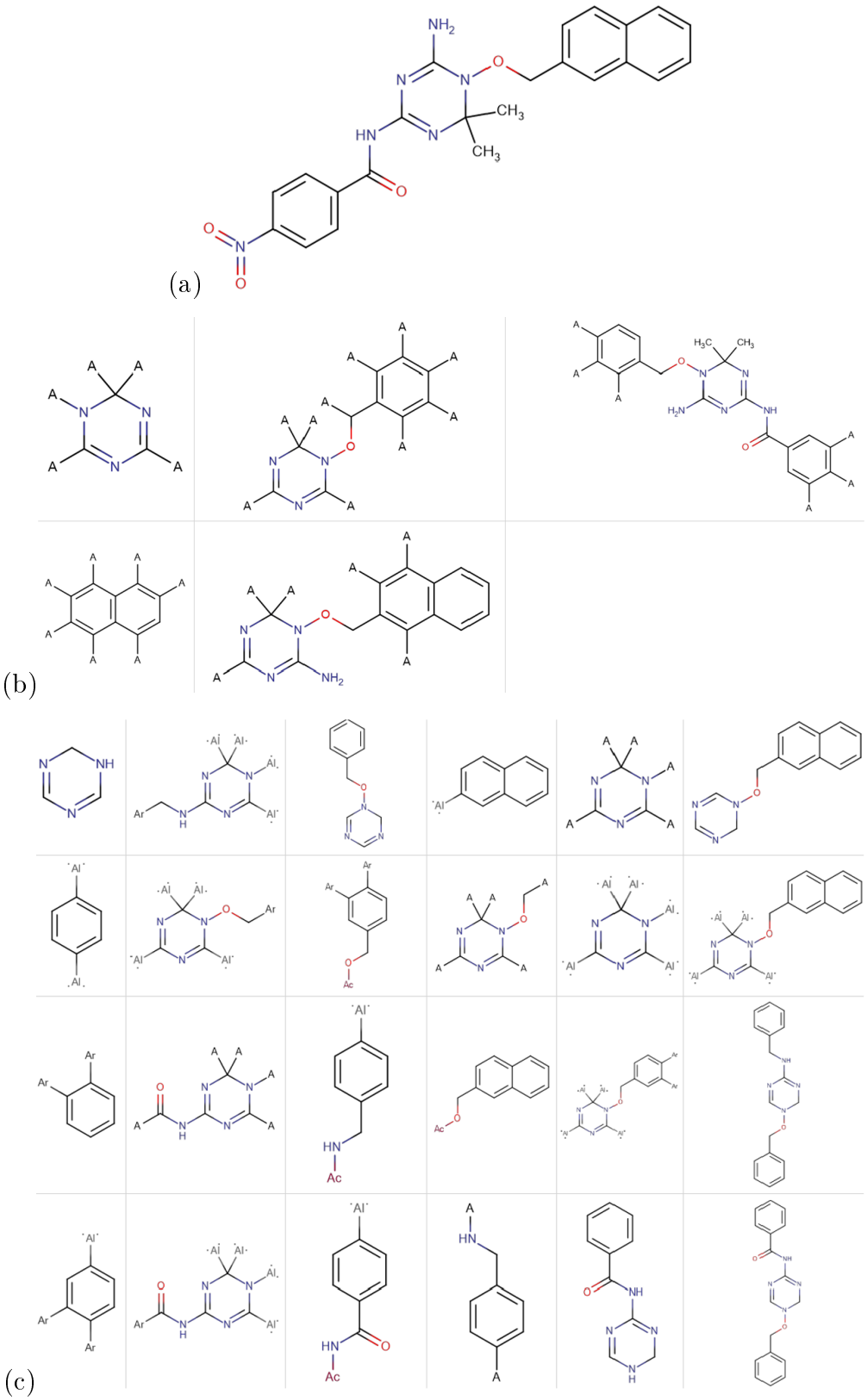



Figure 2: Scaffold Decompositions for (a) Molecule 1 from the TCAMS dataset with Compound ID 536182 (PubChem CID 44522854). (b) 5 scaffolds from NCATS R-Group Tool; (c) 24 Bemis-Murcko Like frameworks.

found, some clipped in the middle of a linker rather than at a ring, and some redundancy between multiple scaffolds. The current implementation does not convert or unify tautomers among scaffolds, again leading to larger numbers of scaffolds.

Further details on how we applied the frameworks code are provided in the Supplementary Material Section S4.

# Methods: Data Integration and Visualization in Spotfire

Next, we describe how tabular scaffold output generated using the NCATS R-group tool and other comparable methods is integrated into Spotfire, a visualization tool of choice at many companies.

## Data Table Generation and Linking

Figure 3 shows how the data tables output by the scaffold generation methods considered here are layered onto the primary data table in Spotfire. This primary data table is usually a direct database import of tabular molecule and activity data or available from public datasets. What gets added follows the schema "Molecule –> Scaffold (including Annotation) –> Related Molecules" shown in the figure. For each Molecule in the dataset, we connect it to every Scaffold/Framework/Cluster it contains, and then to every other molecule containing any of these Scaffolds/Frameworks/Clusters. Slight differences for each individual method are detailed in the Supplementary Material, Section S5.

## Visualization of Molecules, Scaffolds and Related Molecules

We next describe a minimalistic user interface for exploring the network of molecules, the scaffolds they contain and related molecules that contain the same scaffolds. This interface consists of the following elements:

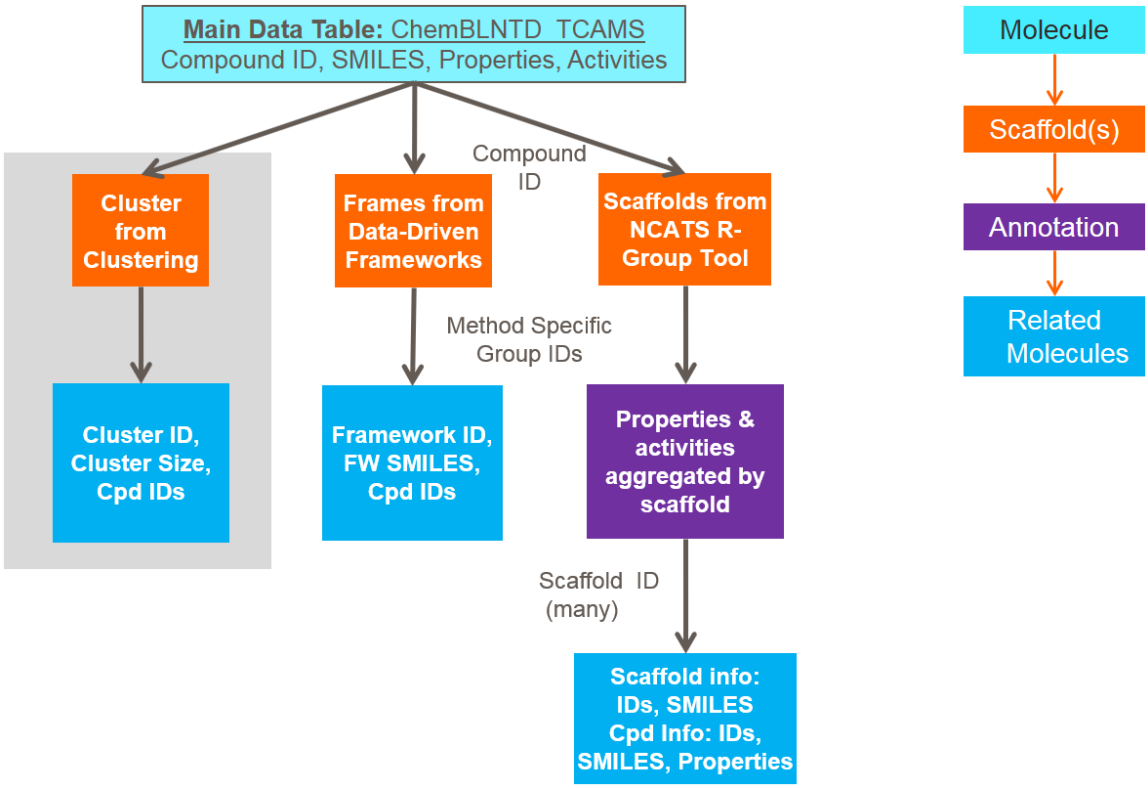



Figure 3: Detailed schematic on how the output from clustering and fragmentation methods are set up as data tables and linked together with the main dataset in Spofire. Right inset: schematic color-coded view of the scaffold-walking navigation that is loosely followed in this diagram.

- **Main Window**: Views defined that allow one to explore the Primary Data Table (with no scaffold information) in the most interpretable way for each dataset. In the canonical example, key activity, selectivity, ligand efficiency and molecular properties may be highlighted on the X, Y, shape, size and color axes on a scatter plot. This view is depicted in compressed form in the top half of Figure 4(a).

- **Related Molecules Tab**: The purpose of this tab is to explore the set of Related Molecules to molecule(s) of interest, and thus implement the Scaffold Walking navigation briefly later in Section . The setup is described for the NCATS R-group Tool decomposition, though this tab applies to and can be set up analogously for any other decomposition. The tab consists of two visualizations, illustrated for the TCAMS dataset in Figure 4(a):

  Miniature version of the **Main window**, allowing the user to select (in Spotfire, mark) molecules of interest without flipping over to the Main tab. Doing so drives one of the following two Details Visualizations on the RGdecomp table, showing only molecules from the scaffolds contained in the marked molecule, i.e., Related Molecules.

  **Scaffold Trellis**: This scatter plot is trellised by Scaffold ID and ideally displays the same properties on the axes as the Main visualization above it. An example is shown in Figure 7. The trellis allows us to break up the SAR for each constituent scaffold individually, identifying promising scaffolds and substituting unproductive ones as we will describe in the Discussion. The trellis visualization suffers from redundancy: the same molecule occurs in multiple trellis panels and the only way to link them is by X and Y coordinates, by observing groups of points that are laid out similarly across multiple trellis panels. Though some of our users still prefer this approach, we now describe a newer solution that better leverages Spotfire's capabilities.

  **Scaffold Pies**: Instead of using a trellis, the Marker shape is changed to Pies, with Colors (which map to pie sectors) by Scaffold ID, and sectors sized by the Count

of molecules containing each scaffold. The result is illustrated in Figure 4(a). This plot shows only one point per related molecule but one sector for each scaffold it shares with the parent molecule. As described in Section , this lets the user quickly and visually home in on key substructures that are important or unimportant for activity.

- **Scaffolds and R-groups Tab**: This tab, currently specific to the NCATS R-group tool method for generating scaffolds, contains two visualizations, as illustrated for the TCAMS dataset in Figure 4(b)–(c):

  The first is a scatter plot display of the Scaffolds table, displaying scaffold Complexity and Count and aggregate activity of each scaffold either on the axes or using the Size and Shape dimensions. Here scaffolds of lesser interest (for example with low complexity or count) can be identified and tagged to remove them from the analysis. Conversely, scaffolds of high interest, for example with many active members or high aggregate ligand efficiency, may be tagged into separate categories.

  The second plot is an R-group table, i.e., a Table view of the RGdecomp table limited to data records that have been marked, i.e., molecules that lie in scaffolds currently marked. The table is sorted first by scaffold and then by primary activity, and molecular fields such as Scaffold SMILES, Molecule SMILES and R-groups $R\_1..R\_n$ are rendered using an appropriate depiction package - at GSK this is JChem.[36] This table may be exported to Excel as an on-the-fly R-group table of the scaffolds of interest.

Alternative visualizations are possible and we describe some of them in the Supplementary Material, Section S6. In addition, we discuss several Spotfire tricks that are instrumental in making our visualization useful to the chemist or biologist user.

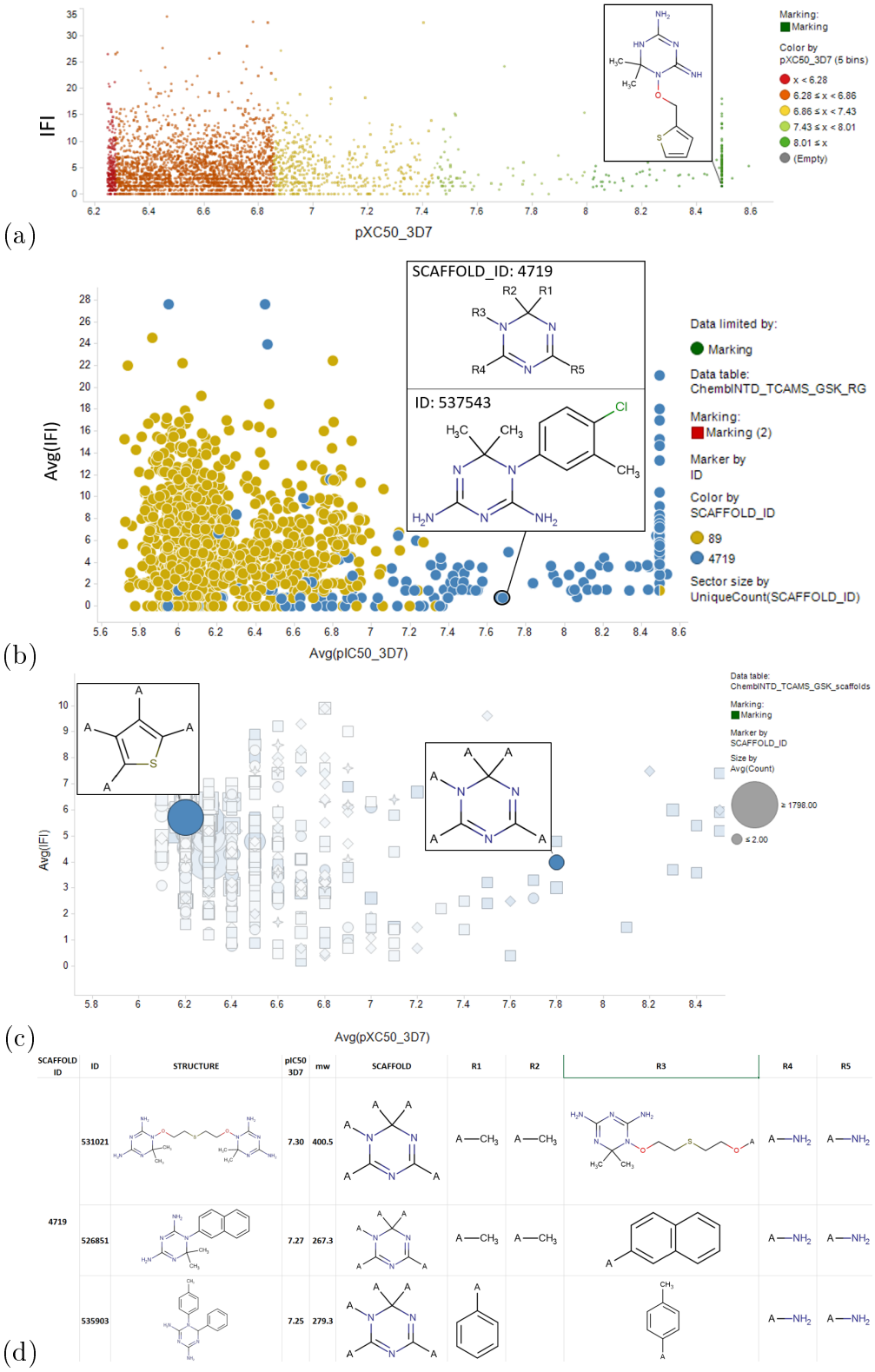



Figure 4: Elements of the minimal Spotfire interface we developed for Scaffold-Based Analytics: Related Molecules page, with (a) molecule selection and (b) related molecules viewed as one pie sector per scaffold shared; (c) Aggregate plot of scaffold statistics; (d) R-group table on selected scaffolds, sorted by scaffold ID and decreasing pIC50 or other desired properties.

## Statistical Comparison of Scaffold Generation Methods

In this section we define the statistical methods used to compare the scaffold generation methods in the context of annotating molecules in a screening dataset with multiple, possibly overlapping scaffolds – a generalization of partitioning the set of molecules into non-overlapping categories, a.k.a. clustering. The code to implement these methods is released as part of the Supplementary Material in Section S10.

The similarities and differences between non-overlapping clusterings have been analyzed in the literature, and one example is the method of Torres et al.[40], who provide a similarity measure between outputs of two partitioning clustering methods. They use Jaccard's similarity coefficient $S_{ij}$, defined as the ratio of intersection size to union size of two groups $i$ and $j$ in the clustering. This coefficient, computed pairwise and summed, yields a simple similarity measure between two clusterings $X$ and $Y$ with $m$ and $n$ clusters, respectively:

$$Sim(X, Y) = \Sigma_{i \leq m, j \leq n} S_{ij} / max(m, n) \tag{1}$$

We adapt this method for SAR analysis using overlapping scaffolds, where the question on the minds of chemists is twofold: what are the fragments in active/desirable molecules, and which other molecules share these fragments? Given two fragmentation schemes, $A$ and $B$, we can evaluate for any molecule how many other molecules share fragments using $A$ or $B$, and thus can be found using fragmentations $A$ alone, $B$ alone, $A$ and $B$, or $A$ excluding $B$. Using these methods we can evaluate the overall similarity of the two fragmentation schemes, and also independently score the usefulness of $A$ and $B$ to connect compound(s) of interest to related molecules.

For any compound $C$ and fragmentation $A$, define the structure group of $C$ under $A$, $C_A$ as the set of compounds that share fragments from $A$ with compound $C$. Similarly define the structure group of $C$ under $B$, $C_B$ as the set of compounds that share fragments from $B$. The Common Proportion ($CP$) for compound $C$ is then defined as the ratio of the number

of compounds common to $C_A$ and $C_B$ to the total number of unique compounds contained in $C_A$ and $C_B$:

$$CP_{A,B}(C) = \|C_A \cap C_B\| / \|C_A \cup C_B\| \tag{2}$$

Similar statistics can be defined to rank the usefulness of an individual fragmentation given others. The Proportion of Information $PI_A$ calculates the proportion of compounds reachable from $A$ and $B$ that would be reachable from $A$ alone:

$$PI_A(C) = \|C_A\| / \|C_A \cup C_B\| \tag{3}$$

In constrast, the Proportion of Information Unique to $A, PIU_A$ uses the set difference between $C_A$ and $C_B$ to get at the question: if I have $B$, do I still need $A$?

$$PIU_A(C) = \|C_A \setminus C_B\| / \|C_A \cup C_B\| = 1 - PI_B(C) \tag{4}$$

When comparing one fragmentation method against another, we often see that one method utilizes a larger number of shared fragments than the other in order to connect compound $C$ to a very similarly sized structure group. To capture this tendency and reward methods that connect molecules to related ones efficiently rather than exhaustively, we define a Fragment Efficiency measure as follows. Let $frag_A(C)$ be the set of fragments of compound $C$ in fragmentation $A$ that connects $C$ to its structure group $C_A$. Similarly define $frag_B(C)$ for fragmentation $B$. Then:

$$FragEff_A(C) = \|C_A\| / \|frag_A(C)\| \tag{5}$$

The distribution of these statistical measures for our dataset is used to compare the overlapping clustering methods (NCATS RG-Decomposition and Molecular Frameworks) and a partitioning clustering method (Complete linkage) in Section .

# Results

Here we illustrate some of our key findings and use cases on a few datasets.

## Use Case: Scaffold Progression and Prioritization

Aggregate statistics such as maximum, minimum, mean and standard deviation, computed at a per scaffold level, may be useful in prioritizing scaffolds. For example, Figure 5 shows the six scaffolds contained within a tricyclic molecule (Molecule 2, TCAMS Compound ID: 541564, PubChem CID: 44531725) ranked by average activity and IFI. This ordering may be used to determine which substructures are most important for the molecule's activity, and use this information to design or test further compounds.

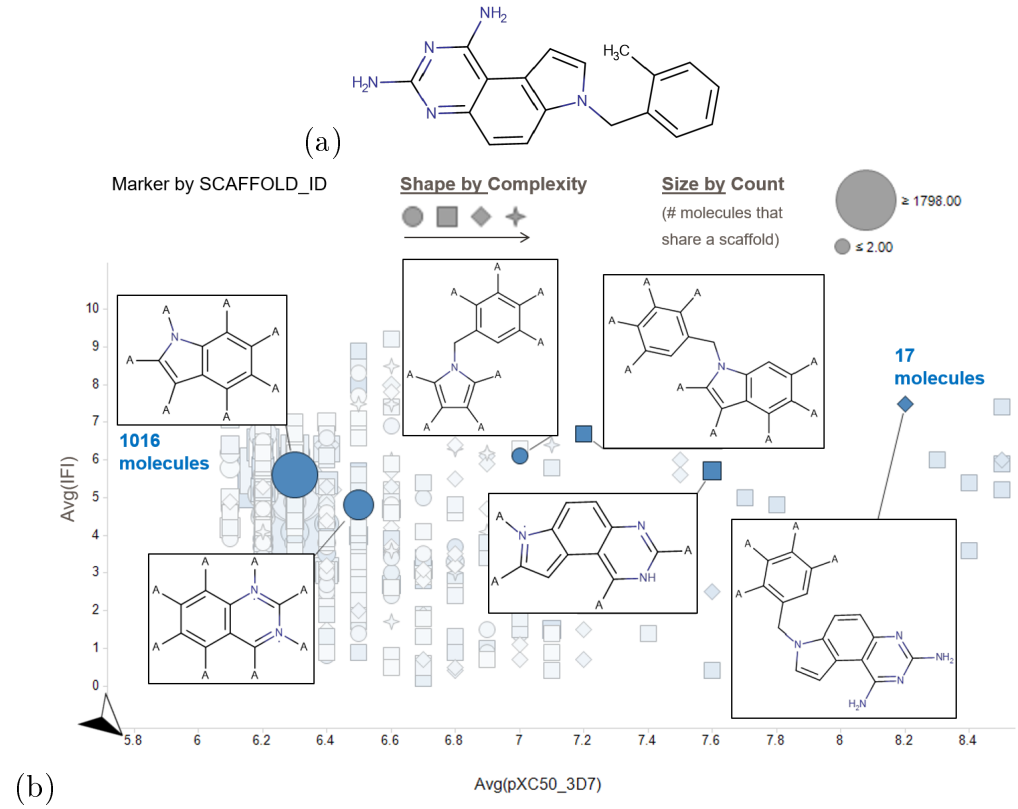



Figure 5: (a) Molecule 2 (TCAMS Compound ID: 541564; PubChem CID: 44531725). (b) All $\tilde{5}000$ scaffolds from the TCAMS dataset ranked by average pIC50 in the P. falciparum 3D7 strain and Inhibition Frequency Index. The six scaffolds contained in Molecule 2 are shown in increasing order of average pIC50.

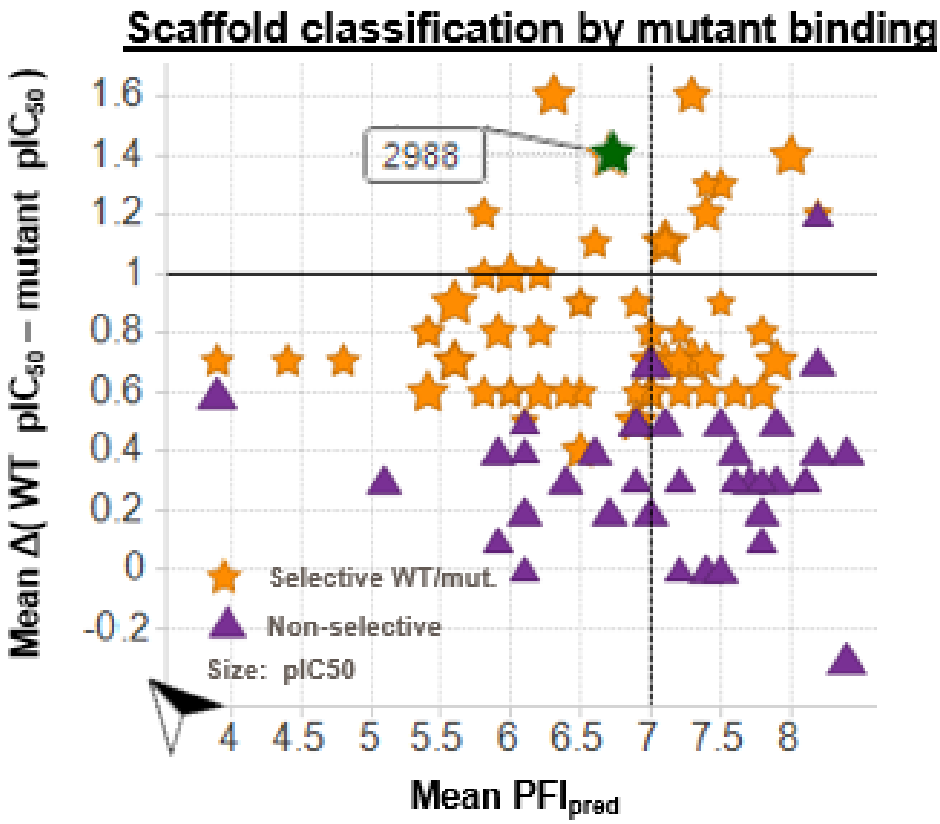
For the in-house kinase dataset, the program team were looking for hit series with activity in a range of orthogonal on-target assays, selectivity against a mutant counter-screen, and good properties to predict developability. We aggregated the activity in each of the primary assays, selectivity, and Property Forecast Index (PFI[34]), using the mean values as filters to identify promising scaffolds. A plot of aggregate selectivity vs. PFI is shown in Figure 6(a).

Once a promising scaffold is identified, one can drill down with a single click into detailed activities and properties of the molecules containing the scaffold to decide whether to pursue the series further. For example, the the scaffold with ID 2988, marked in Figure 6(a) has on the aggregate good selectivity (1.4 log units) and also low PFI (6.7). The activity profiles for 8 molecules sharing this scaffold are shown in Figure 6(b). Clear differentiation is observed between an orthogonal suite of native protein assays and the mutant assay, marking the scaffold as a good one to explore making additional molecules.

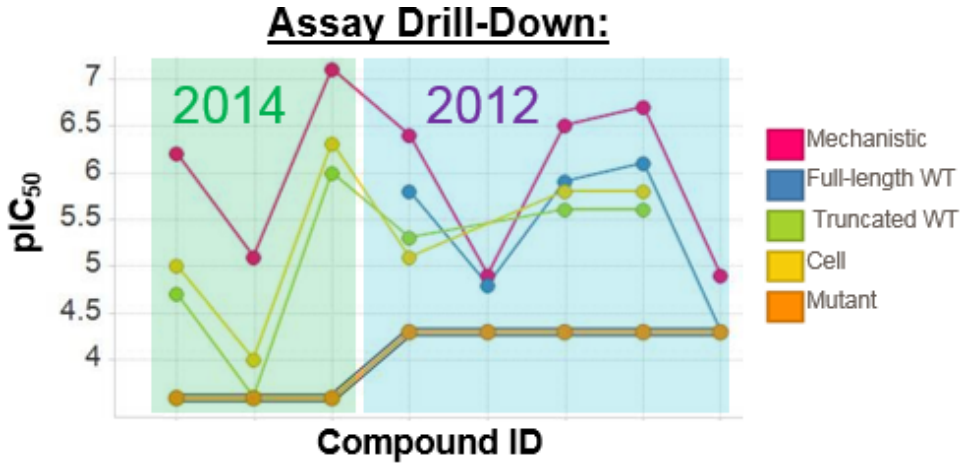
In summary, scaffold-based analytics allows us to make decisions based on the aggregate (average, SD, median, min, max, ...) properties of molecules in scaffolds. An entire scaffold may be rejected or prioritized at a time instead of keeping track of individual molecules. It is important to note that rejecting a scaffold does not reject all molecules containing that scaffold - otherwise rejecting the benzene ring would remove a large percentage of valid leads.

## Use Case: Dataset Fusion and Hybridization

In the previous section and Figure 6(b), we noted that the 8 molecules sharing scaffold 2988 came from two datasets, viz. HTS hits from 2012 and 2014. This is significant since the 5 molecules found during the 2012 campaign were not prioritized since they did not jump to the top using visual inspection, clustering or other methods used to triage the hits – this Scaffold-Based Analytics method was not available. However, when the hits from both screens were combined and analyzed with this method in 2014, the pattern of activity, selectivity and properties that made these desirable hits to pursue was automatically found and readily flagged.




(a)




(b)

Figure 6: (a) Aggregate scaffold statistics for molecules from the Kinase X dataset, with average selectivity against a mutant assay in log units on the Y-axis and average PFI (Property Forecast Index) on the X-axis. Scaffolds are tagged as selective or non-selective after exploring their detailed properties. (b) Detailed properties of 8 molecules that share scaffold 2988 (structure not shown), split over two HTS datasets from 2014 and 2012. Lines are color-coded by assay type, and the mutant assay is shown in orange across the bottom of the graph.

In this way, Scaffold-Based analytics may be used to combine multiple datasets using the scaffold as the common unit of comparison and merging, with dataset labels identifying molecules in the resulting merged dataset. If the datasets have comparable activities, as the Kinase X HTS datasets did, aggregate statistics are easy to compute. If they do not, e.g., if one dataset has $pIC_{50}$ and others primary screening activity or a different measure of desirability, normalized activities may be computed to enable aggregation at the scaffold level.

Sometimes one of the datasets does not contain activities at all, e.g., in the case of the "virtual" hits from DNA Encoded Library Technology[33] (ELT) screening that are yet to be synthesized off-DNA, or for molecules from a vendor catalog that are yet to be ordered and assayed. In Figure 7 we show an example where two computed properties, Molecular Weight and PFI were used to identify promising untested molecules to make (ELT dataset) or test (FBDD dataset[32]).

## Use Case: Scaffold Walking Navigation

Scaffold Walking is our term for navigating from molecule(s) through scaffolds (implicitly) to Related Molecules. This contrasts with Scaffold Hopping, which is usually defined as a complete replacement of a scaffold by another 2D-dissimilar but 3D-similar or bioisosteric scaffold. Scaffold Walking is meant to be a gradual change to the molecule, at each step retaining at least one element of its maximal Murcko scaffold (i.e., at least one among the multiple scaffolds it shares with other molecules in the dataset). The concept is similar to using Graph Edit Distance.[22] In the process the SAR gets deconvoluted in terms of these scaffolds, allowing us to determine visually both the most essential scaffolds in a molecule and the best Related Molecules containing them.

As an example, consider Molecule 1 (TCAMS Compound ID: 536182, PubChem CID 44522854) in Figure 8, as a hit that we want to explore SAR of and optimize. This molecule contains 5 scaffolds as determined by the NCATS R-group tool. Using the Scaffold Pie visu-

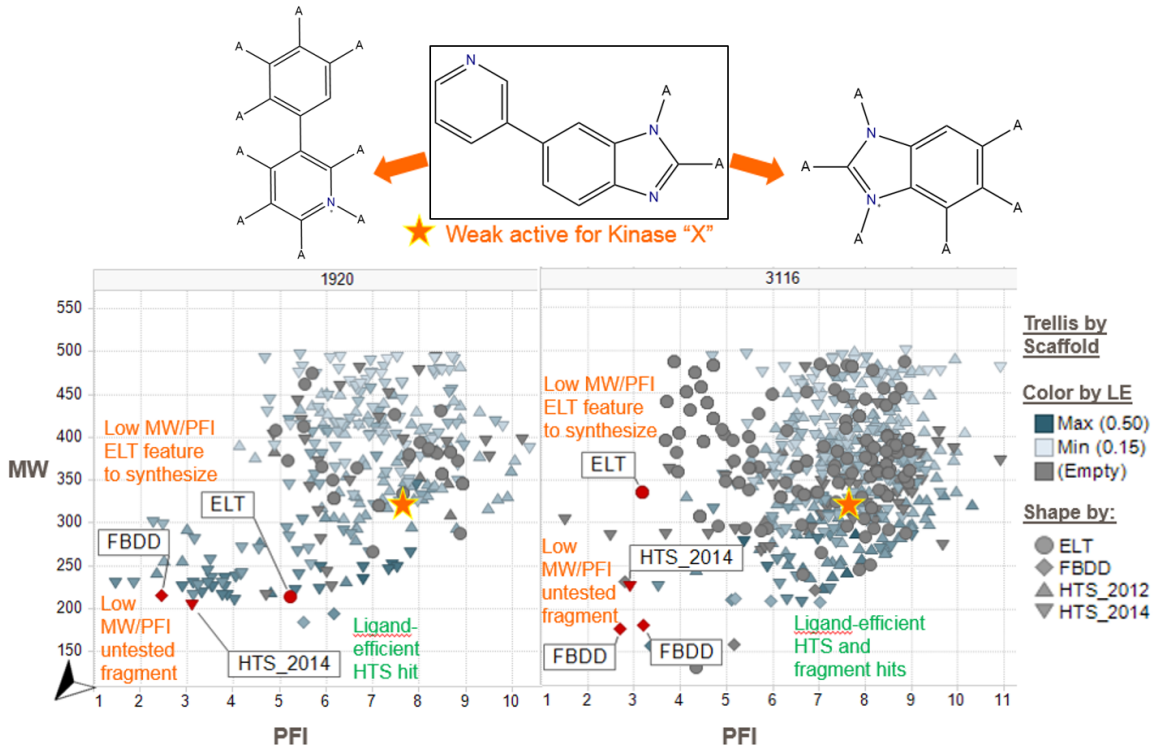



Figure 7: Two scaffolds with ID 1920 (pyridophenyl) and 3116 (benzimidazole) shown in a trellised plot of molecules related to a weak active in the Kinase X 2014 HTS screen. The axes of the plots are computed properties, MW and PFI, to allow molecules without measured activity or virtual molecules not yet synthesized to be included in the plot. The compass arrow device points in the direction of improved properties – lower MW and PFI. Interesting related molecules are labeled on the plot, including untested fragment-like molecules from the FBDD dataset, virtual molecules from the ELT dataset, and another ligand-efficient hit from the HTS2014 dataset, all with low MW/PFI. The unmeasured and unmade molecules are good candidates for future synthesis and testing.

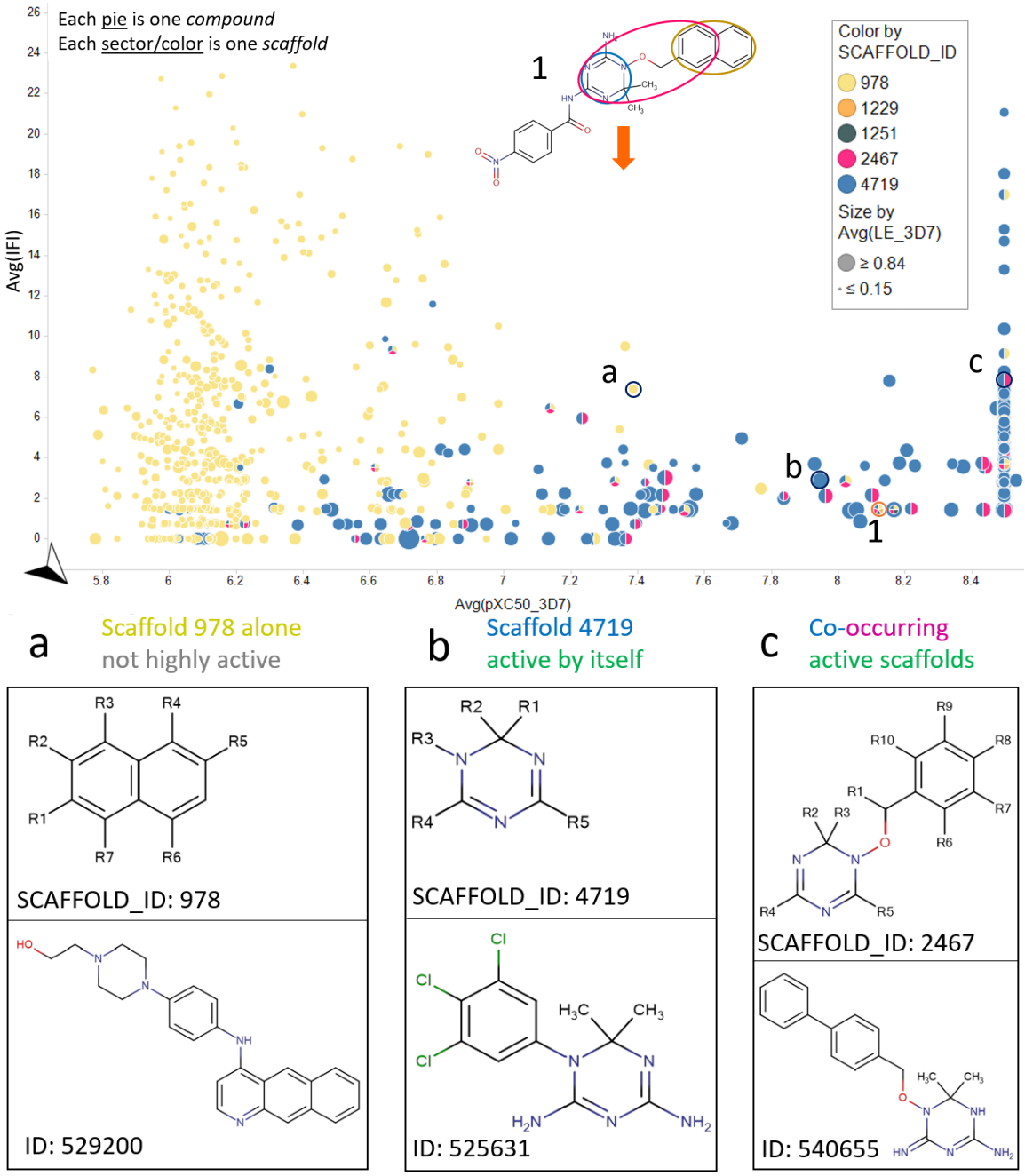



Figure 8: Related Molecules scaffold pies visualization for Molecule 1 (TCAMS Compound ID: 536182, PubChem CID 44522854). Each pie here is one related molecule, and each pie sector and color is a scaffold that it shares with the parent molecule. The star symbol is added to show the location of the parent molecule in this plot, and the compass device at the origin shows the direction of favorable properties (in this case towards the +X and -Y axes). Insights derived from the plot are highlighted in the bottom part of the figure (showing structures) and also discussed in the text.

alization, we observe that the naphthyl scaffold (#978, yellow) is by itself only moderately active. The dihydrotriazine (#4719, blue) scaffold is observed to always occur where the dihydrotriazine-phenethyl-ether (#2467, pink) one does, implying the substructure relationship between them visually even if one did not know it beforehand. Scaffold #4719 also exists and is active without #2467, implying that the phenethyl ether can be substituted and the dihydrotriazine may be sufficient for activity by itself.

Further examples of the value of Scaffold Walking are in the Supplementary Material, Section S7.

## Statistical Comparison of Scaffold-Generation Methods and Clustering

Recall the concepts of structure group and common proportion defined in Section , and illustrated with an example in Supplementary Material, Section S9. Let us now apply these to analyze and compare the scaffold generation methods Frameworks (denoted $A$ below) and NCATS R-group tool ($B$), and compare $B$ with clustering method Complete-Linkage clustering ($D$).

Comparing these methods for the TCAMS dataset, we show the aggregate statistics over the entire dataset in Figure 9.

We compare the Common Proportion, Proportion of Information Unique (PIU) and Fragment Efficiency (FragEff) statistics for the NCATS R-group tool (method "B") with both the Frameworks (method "A") and the Complete Linkage Clusters (method "D") for all the 13.5k compounds in TCAMS in Figure 10. This figure has the PIU of these methods on the axes, and is sized by the ratio of their Fragment Efficiencies for all 13.5k molecules.

We can make a few observations from this data and these plots:

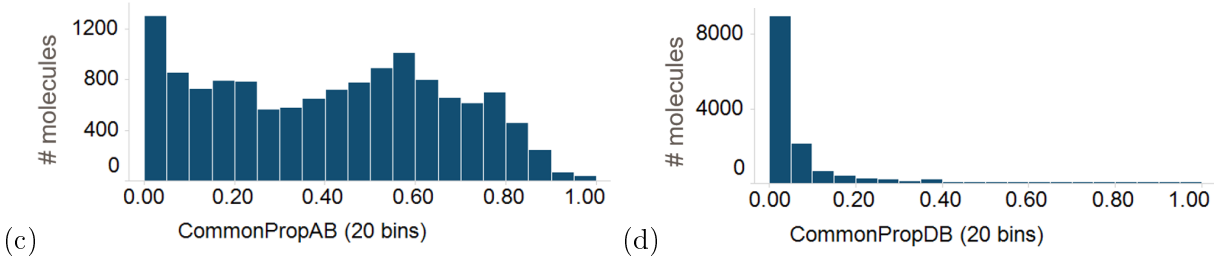
1. **Clustering (method $D$) is able to access significantly fewer molecules** as structurally related analogs for most molecules in TCAMS, compared to the overlapping

(a)

| Column | P10 | Median | P90 | Avg |
|---|---|---|---|---|
| $C_A$ | 420.00 | 1849.00 | 3203.50 | 1845.54 |
| $C_B$ | 83.50 | 888.00 | 2151.00 | 1047.93 |
| $CP$ | 0.05 | 0.43 | 0.76 | 0.41 |
| $PI_A$ | 0.76 | 0.96 | 1.00 | 0.91 |
| $PI_B$ | 0.09 | 0.53 | 0.87 | 0.50 |
| $PIU_A$ | 0.13 | 0.47 | 0.91 | 0.50 |
| $PIU_B$ | 0.00 | 0.04 | 0.24 | 0.09 |
| $Frag_A$ | 10.00 | 32.00 | 74.00 | 38.96 |
| $Frag_B$ | 2.00 | 5.00 | 11.00 | 5.93 |
| $FragEff_A$ | 15.60 | 52.12 | 140.23 | 69.72 |
| $FragEff_B$ | 21.25 | 150.80 | 454.00 | 211.80 |

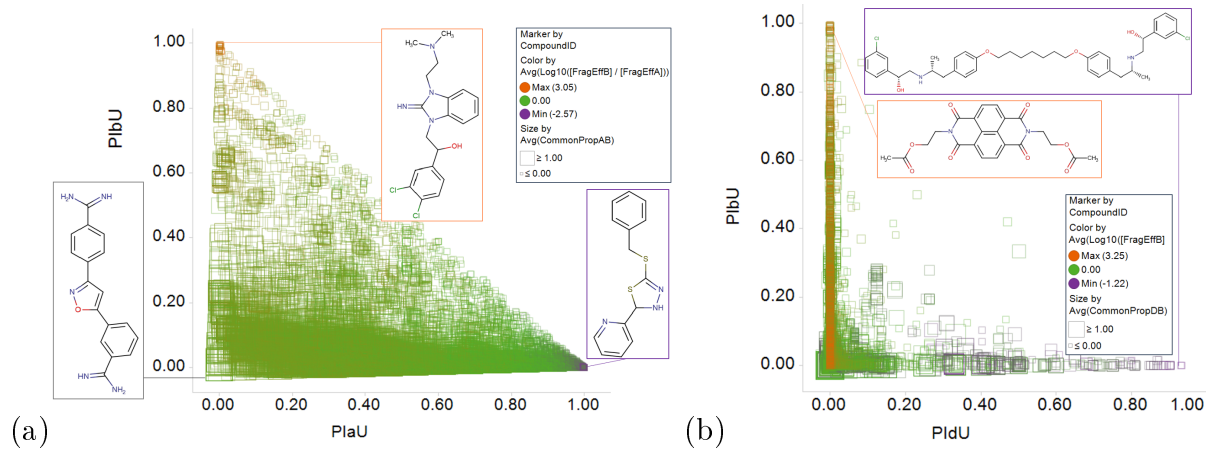
(b)

| Column | P10 | Median | P90 | Avg |
|---|---|---|---|---|
| $C_D$ | 3.00 | 17.00 | 88.00 | 33.11 |
| $C_B$ | 83.50 | 888.00 | 2151.00 | 1047.93 |
| $CP$ | 0.00 | 0.02 | 0.17 | 0.07 |
| $PI_D$ | 0.00 | 0.02 | 0.2 | 0.09 |
| $PI_B$ | 0.98 | 1.00 | 1.00 | 0.98 |
| $PIU_D$ | 0.00 | 0.00 | 0.02 | 0.02 |
| $PIU_B$ | 0.79 | 0.98 | 1.00 | 0.91 |
| $Frag_D$ | 1.00 | 1.00 | 1.00 | 1.00 |
| $Frag_B$ | 2.00 | 5.00 | 11.00 | 5.93 |
| $FragEff_D$ | 3.00 | 17.00 | 88.00 | 33.11 |
| $FragEff_B$ | 21.25 | 150.80 | 454.00 | 211.80 |




(c) (d)

Figure 9: (a) Statistics computed for 10th/90th percentile, median and average value for structure group size ($Ca/Cb$), common proportion, proportional information unique to $A$ and $B$, number of fragments in and fragment efficiency of method $A$ (frameworks) and $B$ (NCATS R-group tool). (b) Same comparison for $B$ (NCATS R-group tool) and $D$ (Complete-Linkage Clustering). (c)-(d) Histograms of Common Proportion for the datasets tabulated above in (a)-(b).




(a) (b)

Figure 10: Comparison of the Common Proportion, PIU and FragEff statistics (described in the text) between the NCATS R-group Tool ("**B**"), PIU on Y axis, and (a) Frameworks (method "**A**") or (b) Complete Linkage Clustering (method "**D**"). PIU for the methods being compared are on X and Y axes, and points are colored by the log ratio of Fragment Efficiencies between the two methods being compared and sized by their Common Proportion.

scaffold-based methods $A$ and $B$ (which were comparable). Most molecules have a low Common Proportion between clustering and the scaffold-based method, shown in Figure 10(d). Almost all the information (PIU) is contributed by the scaffold method ($B$) for most molecules, not surprising when we recall the high singleton rate from Figure 1 where the clustering contributes no information. Even allowing for fewer (one) fragments per molecule, the partitioning cluster method still had a lower Fragment Efficiency than any scaffold-based method.

2. The two scaffold-based methods ($A$: Frameworks and $B$: NCATS R-Group Tool) allow one to access different sets of molecules starting from any molecule in TCAMS - the average overlap in their coverage ($CP$) is 40%, seen in Row 3 of Figure 9(a). The distribution of $CP$ seen in Figure 9(c) is bimodal with peaks at 0 and 55%. The few molecules with almost complete overlap in coverage have high Common Proportion, low $PIU$ for both methods $A$ and $B$, and hence are represented by the large squares towards the bottom left of Figure 10(a).

3. On the average, frameworks connected to a compound add more unique information than NCATS scaffolds connected to the same compound - this is seen in the higher density near the X-axis in Figure 10(a), and higher numbers for $PIU_A$ than $PIU_B$ in Figure 9.

4. On the average, one can link to about twice as many molecules with the frameworks, as seen by comparing $C_A$ to $C_B$ in Figure 9(a); however, this is because on the average there are 6 times more frameworks ($Frag_A$) than NCATS scaffolds ($Frag_B$).

5. The outliers in Figure 10(a) are interesting. At one end, compounds in a rare tautomer are unified with the dominant one by the NCATS tool (high fragment efficiency), but left as singletons by the frameworks (low fragment efficiency). And compounds whose only link with other molecules would be a benzene ring or similar low complexity scaffold remain singletons with the R-group tool (lower fragment efficiency).

6. Singleton clusters are outliers in Figure 10(b), connected to other molecules via scaffolds but not clusters. And a minority of molecules are connected to others via clusters only, mostly those with low complexity linear scaffolds whose benzene rings were excluded from scaffold but not cluster analysis.

## Statistical Basis of Structure-Activity Relationships (SAR)

We have also explored the Common Proportion measure to get at the overlap between structure and activity, i.e., to investigate the statistical basis of SAR based on overlapping scaffolds. To do this, we define an Activity Group of compound $C$, $C_{act}$ as a set of compounds with activity bordering that of $C$ (in this case pIC50 in the 3D7 antimalarial assay). The activity group of each compound $C$ is defined to be the same size as the structure group, denoted $C_A$ or $C_B$ above, and generalized here as $C_{str}$. This is done arbitrarily for ease of statistical calculations, being cognizant that it may sometimes lead to issues such as breaking a large list of compounds with the same activity arbitrarily, especially when picking a small list of activity neighbors for compounds with a small structure group.

The structure-activity common proportion, $CP_{str,act}(C)$ then measures the overlap between the structure neighbors of $C$, i.e., compounds sharing any scaffold with $C$, and its activity neighbors, i.e., compounds in a similar activity range.

$$CP_{str,act}(C) = \|C_{str} \cap C_{act}\| / \|C_{str} \cup C_{act}\| \tag{6}$$

For comparing groups of different sizes we use the normalized Common Proportion, dividing by the expected overlap of the same activity group with a structure group determined purely by chance. With the structure group being the same size as the activity group, this is the probability that if we pick $\|C_{act} - 1\|$ compounds at random they will lie in $C_{act}$, i.e., $(\|C_{act} - 1\|)/(N - 1)$. The subtraction by 1 accounts for the fact that the compound itself always lies in its own structure and activity neighborhoods. So then:

$$NormCP_{str,act}(C) = CP_{str,act}(C)/(ExpectedCPforRandomStr) \quad (7)$$

We computed this NormCP measure for all compounds, which fit a normal distribution centered at 1, implying that for all compounds considered as an aggregate, structural similarity does not imply activity similarity. However, the top 200 most active compounds, which are activity outliers, are also outliers in this normal distribution of NormCP as shown in Figure 11(a).

Among these compounds there are some that have over 40 times as much overlap between structure and activity than would be expected by chance, shown on the right of Figure 11(b) and in Figure 11(c). This implies a strong confidence in the activity being genuine, and the series being worthy of making or measuring the activity of further analogs.

On the other hand, there are also some compounds, highlighted on the left side of Figure 11(b) and in Figure 11(d–e), that though they are among the top 200 by activity, have NormCP much less than 1, i.e., much less overlap between structure and activity than would be expected by chance. This would imply that structural neighbors of these compounds are largely less active that the original compound, no matter which of the overlapping scaffolds within the molecule we use to define those structural neighbors! This gives us a way to quantify flagpoles or false positive hits, which may not be worth following up.

# Discussion

While threshold-based hit selection is a prevalent approach in the analysis of high throughput screening datasets, it ignores the extra information encoded in chemical structure. Thus, scaffold based analysis of high throughput screening datasets represents a truly data-driven approach to hit triage that attempts to make use of all the data collected from a high throughput screen. Ranking scaffolds is a key step in prioritizing hits in a scaffold-based approach, and while there are many ways to generate a ranking, it is not obvious that there

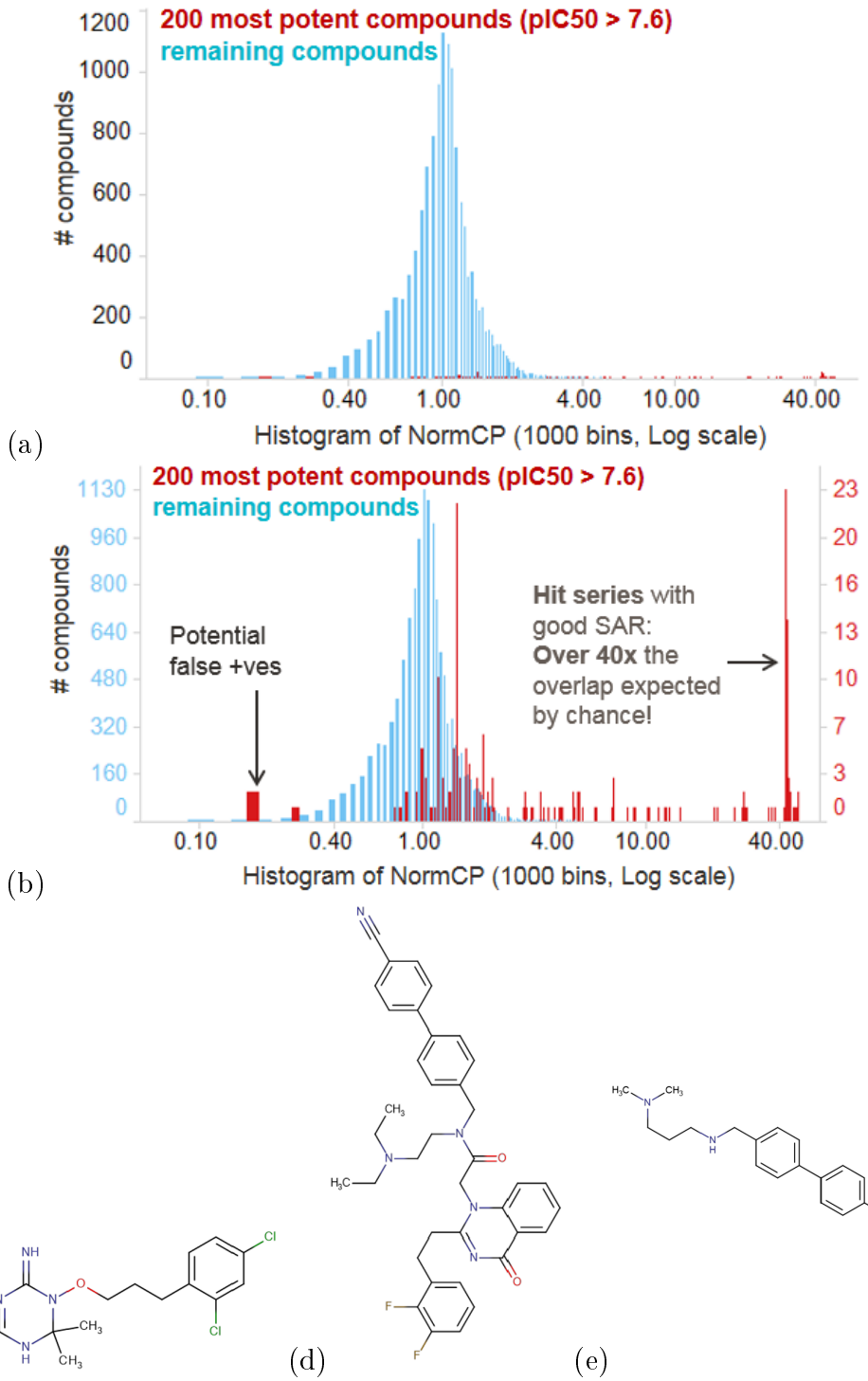



Figure 11: (a) Normalized common proportion scores between structure (data-driven frameworks) and activity ($pIC50_{3D7}$) for all 13.5k molecules in TCAMS. The histogram has 1000 bins and a log scale on the X-axis. The top 200 most active compounds ($pIC50_{3D7} > 7.6$) are marked in red, and occupy many of the outlier bars with high NormCP $> 10$. (b) Making the distribution of the top 200 active compounds clearer by giving them a separate scale, at the RHS. (c) One of the 23 compounds with NormCP $= 42.5$, Compound ID 524739. (d)–(e) two active compounds with low NormCP in the 0.16-0.17 range: (d) Compound ID 541941 and (e) Compound ID 531249. For both compounds, they alone are highly active among a structure group of 96 compounds, raising the probability that these compounds are flagpoles and not worth following up as hits.

is a single, optimal method.

In this study we have presented some alternate approaches to scaffold prioritization that make use of aggregate statistics based on overlapping scaffolds, with the goal of providing a quantitative basis for the comparison of different scaffold-based analysis schemes. Also, the overlapping scaffold approach described here avoids the phenomenon of similar molecules being arbitrarily assigned to exclusive clusters, which affects partitioning-based methods such as complete-linkage clustering.Instead, using overlapping scaffolds ensures that molecules that differ only in decorations off a shared scaffold will be considered within the same group.

Given the different ways to generate scaffolds and to compute overlapping scaffolds, a quantitative approach to characterizing differences in these approaches is necessary. The use of common proportion (CP), fragment efficiency (FragEff) and proportion of information unique to a method (PIU) places such differences within a sound statistical framework, allowing for an objective comparison of fragmentation methods for a given screen. They also extend the applicability of methods to compare the output of different partitioning clustering methods such as Torres et al. [40], allowing them to be used for non-overlapping fuzzy clusters. Furthermore quantifying structure-activity overlap using NormCP is a novel contribution, though similar in spirit and purpose to local hit rate calculations that have been proposed for HTS triage,[41] with a comprehensive structural neighbor metric based on overlapping scaffolds.

We observe that the approach applied to triage the kinase "X" dataset can be a powerful tool to identify promising hits. The Spotfire workflow is simple to implement and allows interactive drill-down from aggregate properties to individual compounds. In essence, the workflow described here enables decisions on individual compounds using the aggregated data as a filter. Another advantage of this workflow is that it supports the inclusion of "virtual datasets" where there is no measured activity, as highlighted in the data fusion use case. Inclusion of such datasets can be useful as they provide an opportunity to directly highlight untested regions of chemical space. When multiple datasets are included in the

data fusion, some with more accurately measured activities, it increases confidence in noisy data by merging data for scaffolds across the datasets.

In summary, the combination of anecdotal and statistical methods to compare scaffold schemes and the resultant analysis of HTS datasets highlights the fact that no single fuzzy clustering method is optimal, and the most appropriate approach should be selected based on the types of analyses described here. Screening scientists have traditionally used "chemical intuition" to select and examine scaffolds, which can lead to biased selections of scaffolds and subsequently of leads. Our Scaffold-Based Analytics approach described in this study combines data and intuitive visualizations to help scientists combat such biases.

## Acknowledgement

The GSK authors thank Subhas Chakravorty, Neysa Nevins, Ami Lakdawala Shah, Eric Manas, Todd Graybill, Stan Martens, Mike Ouellette, Tony Jurewicz, Rob Young, Ken Lind and Jeff Messer for valuable feedback and suggestions while developing the method and visualizations. We dedicate this work to the memory of Christopher Louer, our colleague and cheminformatics wizard at GSK who always encouraged us to innovate in order to help chemists.

The authors declare the following sources of funding: author DB was funded by GSK while undertaking this work and by Janssen Pharmaceuticals in the later stages of writing it up. Authors CK, PB, JB, ZH and GS were funded by GSK. Authors TP, DTN and AJ were funded by NIH. Author RG was funded by NIH while undertaking this work and by Vertex Pharmaceuticals in the later stages of writing it up.

## Supporting Information Available

Supplementary material is available online for this article.

# References


(1) Macarron, R.; Banks, M. N.; Bojanic, D.; Burns, D. J.; Cirovic, D. A.; Garyantes, T.; Green, D. V. S.; Hertzberg, R. P.; Janzen, W. P.; Paslay, J. W.; Schopfer, U.; Sittampalam, G. S. Impact of high-throughput screening in biomedical research. Nat. Rev. Drug Discov. **2011**, 10, 188–195.

(2) Mulrooney, C. A. et al. An informatic pipeline for managing high-throughput screening experiments and analyzing data from stereochemically diverse libraries. J. Comp. Aided Mol. Des. **2013**, 27, 455–468.

(3) Shanmugasundaram, V.; Maggiora, G. M.; Lajiness, M. S. Hit-directed nearest-neighbor searching. J. Med. Chem. **2005**, 48, 240–248.

(4) Varin, T.; Gubler, H.; Parker, C. N.; Zhang, J. H.; Raman, P.; Ertl, P.; Schuffenhauer, A. Compound set enrichment: a novel approach to analysis of primary HTS data. J Chem Inf Model **2010**, 50, 2067–2078.

(5) Pu, M.; Hayashi, T.; Cottam, H.; Mulvaney, J.; Arkin, M.; Corr, M.; Carson, D.; Messer, K. Analysis of high-throughput screening assays using cluster enrichment. Stat. Med. **2012**, 31, 4175–4189.

(6) Cox, P. B.; Gregg, R. J.; Vasudevan, A. Abbott Physicochemical Tiering (APT)–a unified approach to HTS triage. Bioorg Med Chem **2012**, 20, 4564–73.

(7) Peng, Z.; Gillespie, P.; Weisel, M.; So, S.; So, W.; Kondru, R.; Narayanan, A.; Hermann, J. A Crowd-Based Process and Tool for HTS Hit Triage. Mol. Inf. **2013**, 32, 337–345.

(8) Dahlin, J. L.; Walters, M. A. The essential roles of chemistry in high-throughput screening triage. Future Med. Chem **2014**, 6, 1265–1290.

(9) Shun, T. Y.; Lazo, J. S.; Sharlow, E. R.; Johnston, P. A. Identifying actives from HTS data sets: practical approaches for the selection of an appropriate HTS data-processing method and quality control review. J. Biomol. Screen. **2011**, 16, 1–14.

(10) Langer, T.; Hoffmann, R.; Bryant, S.; Lesur, B. Hit finding: towards 'smarter' approaches. Curr. Opin. Pharmacol. **2009**, 9, 589–593.

(11) Wawer, M.; Lounkine, E.; Wassermann, A. M.; Bajorath, J. Data structures and computational tools for the extraction of SAR information from large compound sets. Drug Discov. Today **2010**, 15, 630–639, [DOI:10.1016/j.drudis.2010.06.004] [PubMed:20547243].

(12) Bemis, G. W.; Murcko, M. A. Properties of known drugs. 2. Side chains. J. Med. Chem. **1999**, 42, 5095–5099.

(13) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. J. Med. Chem. **1996**, 39, 2887–2893.

(14) Lewell, X.; Judd, D.; Watson, S.; Hann, M. RECAP - Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. J. Chem. Inf. Comput. Sci. **1998**, 38, 511–522.

(15) Hu, Y.; Stumpfe, D.; Bajorath, J. Computational Exploration of Molecular Scaffolds in Medicinal Chemistry. J. Med. Chem. **2016**, 59, 4062–4076.

(16) Ertl, P.; Schuffenhauer, A.; Renner, S. The scaffold tree: an efficient navigation in the scaffold universe. Methods Mol. Biol. **2011**, 672, 245–260.

(17) Varin, T.; Schuffenhauer, A.; Ertl, P.; Renner, S. Mining for bioactive scaffolds with scaffold networks: improved compound set enrichment from primary screening data. J Chem Inf Model **2011**, 51, 1528–1538.

(18) Harper, G.; Bravi, G. S.; Pickett, S. D.; Hussain, J.; Green, D. V. The reduced graph descriptor in virtual screening and data-driven clustering of high-throughput screening data. J Chem Inf Comput Sci **2004**, 44, 2145–2156.

(19) Koch, I.; Lengauer, T. Detection of distant structural similarities in a set of proteins using a fast graph-based method. Proc Int Conf Intell Syst Mol Biol **1997**, 5, 167–178.

(20) Quintus, F.; Sperandio, O.; Grynberg, J.; Petitjean, M.; Tuffery, P. Ligand scaffold hopping combining 3D maximal substructure search and molecular similarity. BMC Bioinformatics **2009**, 10, 245.

(21) Sayle, R.; Batista, J.; Grant, A. 244th American Chemistry Society National Meeting, Philadelphia, PA, August 19-23, 2012.

(22) Garcia-Hernandez, C.; FernÃ¡ndez, A.; Serratosa, F. Ligand-Based Virtual Screening Using Graph Edit Distance as Molecular Similarity Measure. Journal of Chemical Information and Modeling **2019**, 59, 1410–1421, PMID: 30920214.

(23) Clark, A. M.; Labute, P. Detection and assignment of common scaffolds in project databases of lead molecules. J. Med. Chem. **2009**, 52, 469–483.

(24) Bandyopadhyay, D. 244th American Chemistry Society National Meeting, Philadelphia, PA, August 19-23, 2012.

(25) Downs, G.; Barnard, J. Clustering methods and their uses in computational chemistry. Reviews in computational chemistry **2003**, 18, 1–40.

(26) Gamo, F.-J.; Sanz, L. M.; Vidal, J.; de Cozar, C.; Alvarez, E.; Lavandera, J.-L.; Vanderwall, D. E.; Green, D. V. S.; Kumar, V.; Hasan, S.; Brown, J. R.; Peishoff, C. E.; Cardon, L. R.; Garcia-Bustos, J. F. Thousands of chemical starting points for antimalarial lead identification. Nature **2010**, 465, 305–310.

(27) Calderón, F.; Barros, D.; Bueno, J. M.; Coterón, J. M.; Fernández, E.; Gamo, F. J.; Lavandera, J. L.; León, M. L.; MacDonald, S. J. F.; Mallo, A.; Manzano, P.; Porras, E.; Fiandor, J. M.; Castro, J. An invitation to open innovation in malaria drug discovery: 47 quality starting points from the TCAMS. ACS Medicinal Chemistry Letters **2011**, 2, 741–746.

(28) Holliday, J. D.; Rodgers, S. L.; Willett, P.; Chen, M.-Y.; Mahfouf, M.; Lawson, K.; Mullier, G. Clustering files of chemical structures using the fuzzy k-means clustering method. Journal of chemical information and computer sciences **2004**, 44, 894–902.

(29) Richmond, N. 6th Joint Sheffield Conference on Chemoinformatics, Sheffield, UK, July 22-24, 2013.

(30) Chakravorty, S. J.; Chan, J. A.; Luengo, J.; Greenwood, N. M.; Popa-Burke, I.; Macarron, R. 245th American Chemistry Society National Meeting, New Orleans, LA, April 7-11, 2013.

(31) ChEMBL Neglected Tropical Disease datasets. `https://chembl.gitbook.io/chembl-ntd`, [Dataset 1 – compound IDs differ from ChEMBL and PubChem IDs].

(32) Erlanson, D. In Fragment-Based Drug Discovery and X-Ray Crystallography; Davies, T., Hyvönen, M., Eds.; Topics in Current Chemistry; Springer: Berlin, 2011; Chapter Introduction to Fragment-Based Drug Discovery.

(33) Clark, M. A. et al. Design, synthesis and selection of DNA-encoded small-molecule libraries. Nat. Chem. Biol. **2009**, 5, 647–654.

(34) Young, R. J.; Green, D. V. S.; Luscombe, C. N.; Hill, A. P. Getting physical in drug discovery II: The impact of chromatographic hydrophobicity measurements and aromaticity. Drug Discovery Today **2011**, 16, 822–830.

(35) Hassan, M.; Brown, R. D.; Varma-O'brien, S.; Rogers, D. Cheminformatics analysis and learning in a data pipelining environment. Mol. Divers. **2006**, 10, 283–299.

(36) Csizmadia, F. JChem: Java applets and modules supporting chemical database handling from web browsers. J Chem Inf Comput Sci **2000**, 40, 323–324.

(37) RDKit: Open-source cheminformatics. `http://www.rdkit.org`, [Online; accessed 11-April-2013].

(38) Jain, A. K. Data clustering: 50 years beyond K-means. Pattern Recognition Letters **2010**, 31, 651–666.

(39) Peryea, T.; Braisted, J.; Jadhav, A.; Guha, R.; Southall, N.; Nguyen, D.-T. 245th American Chemistry Society National Meeting, New Orleans, LA, April 7-11, 2013.

(40) Torres, G. J.; Basnet, R. B.; Sung, A. H.; Mukkamala, S.; M, B. A Similarity Measure for Clustering and its Applications. Int J Electr Comput Syst Eng **2009**, 3, 164–170.

(41) Posner, B. A.; Xi, H.; Mills, J. E. J. Enhanced HTS hit selection via a local hit rate analysis. Journal of Chemical Information and Modeling **2009**, 49, 2202–2210.

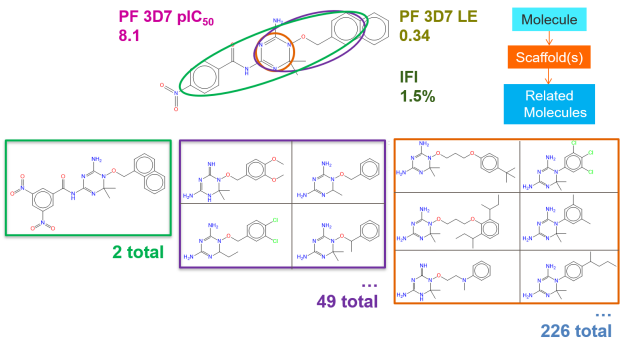



Figure 12: For Table of Contents Only