JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

Activity-Aware Clustering of High Throughput Screening Data and Elucidation of Orthogonal Structure−Activity Relationships

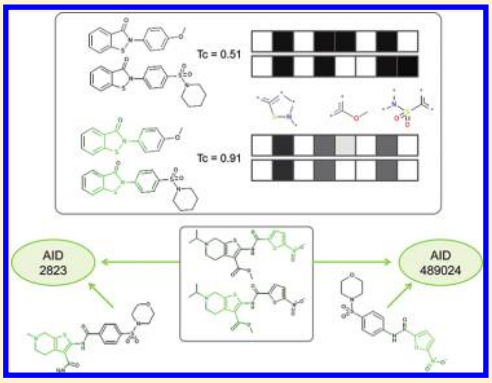
Eugen Lounkine,*[,†] Florian Nigsch,[‡] Jeremy L. Jenkins,[†] and Meir Glick[†]

[†]Novartis Institutes for Biomedical Research, 250 Massachusetts Ave., Cambridge, Massachusetts 02139, United States

[‡]Novartis Institutes for Biomedical Research, Novartis Campus, Forum 1, CH-4056 Basel, Switzerland

Supporting Information

ABSTRACT: From a medicinal chemistry point of view, one of the primary goals of high throughput screening (HTS) hit list assessment is the identification of chemotypes with an informative structure−activity relationship (SAR). Such chemotypes may enable optimization of the primary potency, as well as selectivity and phamacokinetic properties. A common way to prioritize them is molecular clustering of the hits. Typical clustering techniques, however, rely on a general notion of chemical similarity or standard rules of scaffold decomposition and are thus insensitive to molecular features that are enriched in biologically active compounds. This hinders SAR analysis, because compounds sharing the same pharmacophore might not end up in the same cluster and thus are not directly compared to each other by the medicinal chemist. Similarly, common chemotypes that are not related to activity may contaminate clusters, distracting from important chemical motifs. We combined molecular similarity and Bayesian models and introduce (I) a robust, activity-aware clustering approach and (II) a feature mapping method for the elucidation of distinct SAR determinants in polypharmacologic compounds. We evaluated the method on 462 dose−response assays from the Pubchem Bioassay repository. Activity-aware clustering grouped compounds sharing molecular cores that were specific for the target or pathway at hand, rather than grouping inactive scaffolds commonly found in compound series. Many of these core structures we also found in literature that discussed SARs of the respective targets. A numerical comparison of cores allowed for identification of the structural prerequisites for polypharmacology, i.e., distinct bioactive regions within a single compound, and pointed toward selectivity-conferring medchem strategies. The method presented here is generally applicable to any type of activity data and may help bridge the gap between hit list assessment and designing a medchem strategy.



INTRODUCTION

Hit lists, namely sets of biologically active compounds resulting from high throughput screening (HTS) experiments, are often chemically diverse, which makes it difficult to extract meaningful structure−activity relationship (SAR) information.[1,2] This can become a bottleneck in devising a medchem strategy for exploratory chemistry and lead optimization, which rely on understanding the SAR and pharmacophores in the hit list. In addition to the SAR related to the primary target, other biological activities and respective SARs can be equally important, as these may define a desired polypharmacology or off-target effects that should be mitigated.[3−5]

Common approaches to deriving SAR information from diverse sets of molecules consist of two general steps. First, active compounds are grouped on the basis of their chemical structure. In a second step, activity data are analyzed for each group.[6] If compounds within a group share a common core structure, it can be used as a starting point for further chemical diversification, and such privileged scaffolds have been described for several target classes.[7] Optimization rarely is one-dimensional, and compounds may have additional, often unwanted, biological activity that needs to be balanced with the primary activity.[8] The goal then becomes finding structural changes outside the active core structure that retain or improve desired activity, while alleviating unwanted off-target effects.

One widely used approach to identifying common chemotypes through grouping relies on clustering based on pairwise molecular similarity, where entire compounds are compared to each other.[9,10] Binary molecular fingerprints, probably the most common type of molecular representation, are often used in combination with the Tanimoto coefficient (Tc) as the similarity metric to compare molecules.[9,11] Individual bits in molecular fingerprints code for predefined substructures, or for features that are derived from the molecular graph, e.g., in extended connectivity fingerprints (ECFP).[12] Although so derived clusters of known hits provide some information about common chemotypes, they can be easily biased by the screening library. For example, if a large proportion of the screened compounds in the library are sulfonamides, they also will be grouped together among the actives, even if other parts of the compounds are conferring activity and the sulfonamide group is common to both

Received: October 19, 2011
Published: November 19, 2011

active and inactive compounds. In addition to screening library bias, common chemical groups used to diversify compounds, such as halogen substituents, may define clusters, even though they are not specific for activity. Such features can contaminate otherwise informative clusters, compromising the identification of privileged scaffolds[7] or other features associated with activity.[13−15]

A variety of alternative data set analysis and visualization techniques have been developed to account for these shortcomings.[6] Scaffold decomposition according to well-defined rules[16] allowed identification of chemically intuitive and activity-privileged scaffolds.[17,18] While interpretable from a chemistry point of view, scaffolds focus on ring systems, and activity is projected onto calculated scaffolds, rather than influencing the fragmentation procedure itself. Thus, key functional groups are neglected, such as hydroxyl in phenol or the amine in aniline that can participate in specific H-bond interactions with the protein. Such constraints imposed by a ring-centric scaffold definition[16,19] have led to the development of tools that allow for a knowledge-based, less restrictive definition of scaffolds by medicinal chemists.[20] This definition of core structures, however, is still imposed by experts, rather than being data-driven.

We therefore asked the question: how can we influence the chemical clustering procedure itself to group compounds with common activity-characteristic chemotypes? One way to incorporate activity data into chemical space is via naïve Bayesian modeling. Naïve Bayesian models identify individual chemical features that are enriched in active versus inactive compounds[21,22] and have proven tolerant to noisy activity data.[23] Typically, Bayesian scores are calculated for each molecule individually and are thus not directly applicable to clustering compounds. However, model-derived feature weights, essentially a byproduct of Bayesian modeling, have served as a means to project observed *in vitro* bioactivity,[24] as well as more complex phenotypes,[25] into a common chemical reference space.[26]

We combined the two distinct concepts of molecular similarity and naïve Bayesian modeling to enable activity-aware clustering of HTS data sets in different biological contexts. Furthermore, we identified polypharmacologic compounds for which distinct parts of the structure conferred the distinct activities, pointing toward medchem strategies that optimize selectivity.

METHODS

Data Sets. We assembled 462 confirmatory dose−response assays from Pubchem[27] containing at least 100 tested compounds, and 20−1000 active compounds, where we only considered compounds with an outcome of "active" or "inactive" (Supporting Information, Table 1). Compounds were washed using Pipeline Pilot (version 8): counterions and solvents were removed, and stereochemistry was discarded. InChi representations of compounds were calculated, and unique compound structures were represented using standard InChiKeys.[28]

Weighted Fingerprint Comparison. In order to focus the comparison of compounds on activity-conferring substructures, rather than entire molecules, we combined Bayesian modeling and Tanimoto-based molecular similarity. A multiclass naïve Bayesian model across all selected Pubchem bioassays was trained using extended connectivity (ECFP_4) fingerprints in Pipeline Pilot (version 8)[32] to distinguish active compounds from a large set of diverse inactive compounds. For each assay, the Bayesian model determined a Laplace-corrected, log-odds weight for each ECFP_4 feature based on its enrichment in the active compounds (Figure 1A). Thus, "good" features that preferentially occurred in active, rather than inactive, compounds received positive weights, while features characteristic of inactive compounds were penalized by negative weights.[22]

Fingerprints of active molecules were then calculated and modified using the Bayesian weights by weighting each feature:

$$\text{feature weight} = 5 \times (\text{Bayesian weight} + 1) \text{ if Bayesian weight} > 0$$
$$\text{feature weight} = 1 \text{ if Bayesian weight} \leq 0$$

We used the general form of the Tanimoto coefficient[9] to compare the weighted fingerprints. Thus, in our approach, molecules sharing activity-characteristic features were more similar to each other than molecules sharing the same number of low-scoring features (Figure 1A). The coefficient of 5 was found by comparing six different weighting schemes and the corresponding Tc value distributions (Supporting Information). The selected scoring scheme prioritized "good" features over other features of the molecule but did not discard "bad" features entirely. This allowed robust comparison of active molecules with or without many high-scoring features, while focusing on activity-characteristic parts of the molecular structures.

Molecule Clustering. Active molecules in each bioassay were clustered using complete-linkage hierarchical agglomerative clustering[29] in R (version 2.13).[30] This clustering approach starts with each molecule assigned to a distinct cluster. Most similar clusters are then iteratively joined until the predefined number of clusters is reached. Cluster similarity is calculated as the average similarity between all members of the two clusters. We have chosen the number of clusters based on the total number of active compounds for each bioassay. For n compounds, the number of clusters was set to the square root of $n/2$.[31] This allowed direct comparison of unweighted, i.e., conventional, and weighted clustering.

Feature Mapping and Visualization. In order to identify and visualize the parts of the molecule that were characteristic of activity in each bioassay, we mapped ECFP_4 features with a positive Bayesian weight onto the molecules by monitoring the ECFP_4#A property in Pipeline Pilot. Atom scores were calculated for each non-hydrogen atom as a weighted average of all features that mapped that atom:[32] for each atom in a molecule, the sum of Bayesian weights of features mapping that atom was divided by the total sum of Bayesian weights of good features mapping that molecule. Atom scores were then scaled to the range [0, 1] by dividing by the maximal atom score. For visualization purposes, atom scores were linearly mapped to shades of green in molecular graph depictions. The atom with the highest score in a molecule was colored green (RGB value "0, 255, 0"), and all other scores were linearly mapped to the range [0, 255] for the green component.

Feature mapping and the resulting atom score vectors enabled comparison of activity-characteristic molecule regions across different bioassays (Figure 1B). For each molecule that was active in at least two bioassays, we used the general Tc in order to assess the overlap of mapped molecular regions. Low Tc values in this case indicated that distinct parts of the molecule were characteristic of different activities.

RESULTS

Molecular Similarity Using Weighted Fingerprints. We incorporated activity information into molecular similarity using

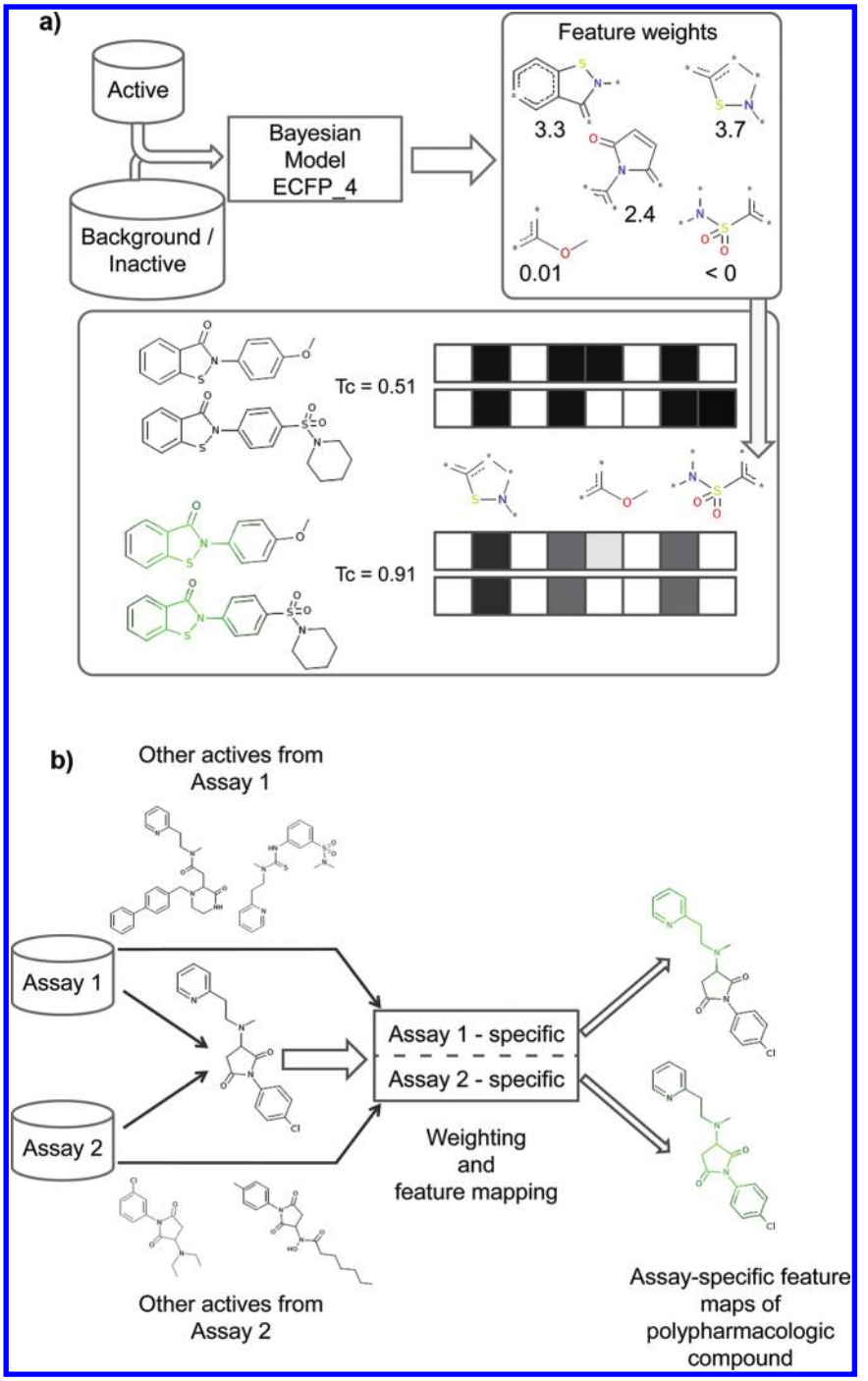


Figure 1. Weighted fingerprints and feature mapping. (a) Fingerprint weighting using Bayesian weights. Given a set of active compounds and a background set that includes inactives, a Bayesian model is trained, yielding positive weights for features enriched in active compounds. Features not characteristic of biological activity in the assay of interest receive low or negative weights. ECFP_4 fingerprints of active compounds are then modified using the Bayesian weights. This yields higher Tanimoto coefficient (Tc) values for compounds sharing many activity-characteristic features. In this example, the compound pair is not very similar when considering all fingerprint bits (Tc = 0.51), but weighting ignores many features that are distinct between the two molecules and yields a much higher Tc value of 0.91. The compounds shown are glucose-6-phosphate dehydrogenase inhibitors (AID 504765). (b) Assessment of orthogonal SAR determinants in polypharmacologic compounds. For polypharmacologic compounds active in multiple assays, feature weighting and mapping can be carried out for each assay individually. The weights and resulting feature maps are influenced by other actives in each bioassay. Distinct active cores point to different parts of the molecule characteristic of the distinct activities.

Bayesian feature weights to preferentially compare parts of molecules that were characteristic of activity. Changes in similarity values between pairs of active compounds included both an increase and a decrease in similarity (Figure 2). Similarity values were higher when two molecules shared a common activity-characteristic substructure that was mapped by many "good" ECFP_4 features (Figure 1A). Many molecule pairs, however, shared substructures that were not characteristic of active

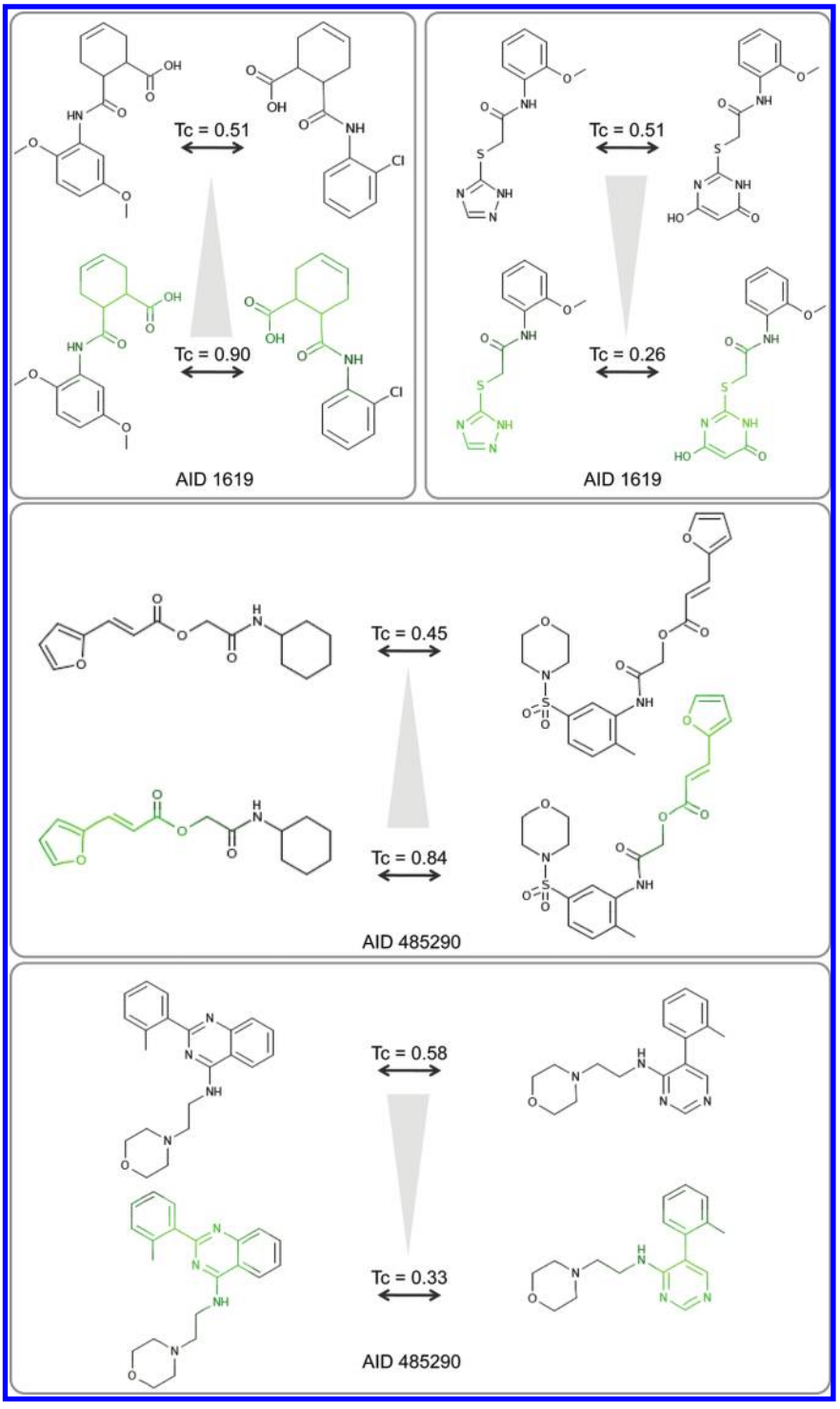


Figure 2. Weighted fingerprints influence similarity. Examples of similarity value changes. For two bioassays, examples of increased and decreased similarity values due to fingerprint weighting are shown. Activity-characteristic core structures are mapped in green. Tc: Tanimoto coefficient. AID: Pubchem bioassay ID.

compounds but rather reflected chemical series among the tested (i.e., both active and inactive) compounds. This led to decreased similarity between weighted compound pairs, because their activity-characteristic moieties differed substantially (Figure 2). The median change of Tc values in individual assays ranged from −0.07 to +0.16, with 446 (97%) of all assays having a negative median change. Despite these overall moderate changes in similarity values, individual compound pairs changed their similarity substantially when compared using weighted fingerprints (Figure 2): the maximal increase in the Tc value observed was 0.48, and the most pronounced decrease was −0.39.

Feature mapping facilitated rationalization of changes in similarity. Since ECFP_4 has been designed to be sensitive to substitution patterns,[12] even minor differences in activity-characteristic

dx.doi.org/10.1021/ci2004994 |J. Chem. Inf. Model. 2011, 51, 3158–3168

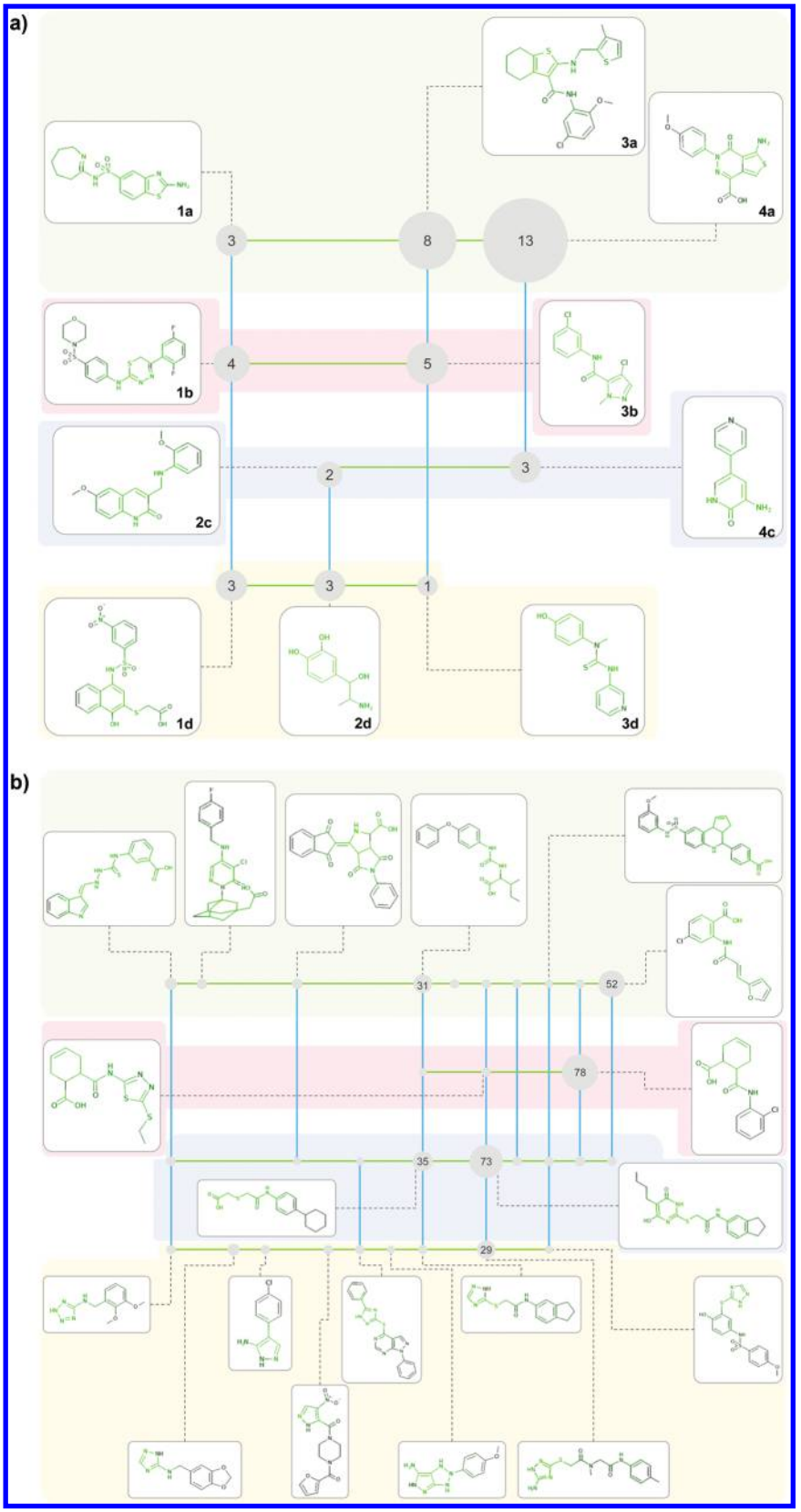


Figure 3. Clustering comparison. (a) For tau fibril formation inhibitors (AID 1558), the distribution of compounds among weighted clusters (horizontal green lines, a−d) and unweighted clusters (vertical blue lines, 1−4) is shown. Each circle is one intersection of a weighted and an unweighted cluster. Circle size and numbers in circles signify the number of compounds. One representative compound is shown for each intersection. Each background color highlights a different weighted cluster. (b) Four of in total 20 weighted clusters are shown for the inhibitors of *Plasmodium falciparum* M17-family leucine aminopeptidase (AID 1619).

dx.doi.org/10.1021/ci2004994 |*J. Chem. Inf. Model.* 2011, 51, 3158–3168

molecular moieties tended to reduce the overall similarity between otherwise similar molecules (Figure 2, bottom panel). Nevertheless, our method successfully ignored major differences in functional groups that were not characteristic of activity (Figure 2, mid panel).

Activity-Aware Clustering of Active Compounds. Next, we assessed the influence of the altered similarity scores on clustering. We monitored the number of molecule pairs found in unweighted clusters that were separated using our method and *vice versa*. Across all assays, the median fraction of molecule pairs originally clustered together, but separated by weighted clustering, was 57% (interquartile range: 39−70%). Similarly, the median fraction of compound pairs found in weighted clusters that were separated in unweighted clustering was 56% (39−69%). For a minority of 16 assays (3%), the clusters did not change.

We then assessed how weighted clusters facilitated the identification of activity-characteristic core structures that were hard to identify using the conventional, unweighted clustering approach. Plotting weighted versus unweighted clusters (Figures 3A,B) revealed clusters with different degrees of overlap. For example, tau inhibitors containing a thiophene or thiazole core were grouped together by weighted clustering (cluster "a" in Figure 3A, green), and 13 of 24 of these compounds also were grouped by unweighted clustering, constituting the greatest overlap between the two methods observed for that assay ("4a" in Figure 3A). Three compounds characterized by a 2-pyridone core originally also belonged to that cluster but were grouped with other 2-pyridones using our approach (cluster "c" in Figure 3A, blue). Unweighted clustering grouped together sulfonamides, but because this substructure was not characteristic of tau inhibitory activity, they were separated into phenols (cluster "d" yellow), diazines ("b", red), and thiazoles ("a", green) by weighted clustering (Figure 3A). For larger hit lists, this visualization allowed us to identify compounds sharing a relatively small active core structure that were grouped together by weighted clustering but were separated by the conventional, unweighted clustering approach. For example, azoles were grouped together among *Plasmodium falciparum* aminopeptidase inhibitors using weighted clustering (yellow cluster in Figure 3B). Similarly, compounds defined by a 3-cyclohexene carboxylic acid core were grouped together by weighted clustering (red cluster). Another big weighted cluster (green) contained more general carboxylic acids that were largely dispersed among unweighted clusters.

To gather small but informative examples of such regrouping using weighted activity-aware clustering, we selected groups of nine molecules that clustered differently for weighted and unweighted fingerprints (Figure 4A−E, Supporting Information Figure S2A−J). Each time, the nine molecules were taken from three weighted and three unweighted clusters. However, any three molecules that belonged to the same unweighted cluster belonged to different weighted clusters, and *vice versa*. These sets of nine compounds reflected the typical changes in similarity values observed previously for individual compound pairs and focused on activity-characteristic core structures. For example, among luciferase inhibitors (AID 1379), aminoprop-2-en-1-ones were grouped together by weighted clustering (Figure 4A, cluster "c"). This substructure has been previously identified as a core structure by manual luciferase SAR analysis.[33] Conversely, conventional clustering assigned these three compounds to separate clusters (clusters "1", "2", and "3" in Figure 4A). In our approach, any substructure, even if it is not a "classical" ring-centric scaffold, as in this case, will influence clustering if it is characteristic of active compounds. In Figure 4B, the conventional, unweighted cluster "2" grouped together compounds that shared a sulfonamide group. However, it was not characteristic of activity in the particular assay (AID 493003 − TIM10-1 inhibitors). Conversely, a thiourea group in combination with an adjacent imino group was characteristic of active compounds and led to the formation of cluster "c". Compounds with this structural motif are known to inhibit protein translocation via TIM10-1.[34]

Although the sulfonamide group was not characteristic of TIM10-1 inhibitors, for another bioassay (AID 1631, activators of pyruvate kinase, Figure 4C), this moiety was characteristic of actives and thus contributed to the formation of a cluster containing compounds otherwise dispersed among three unweighted clusters (Figure 4C, cluster "c"). The SAR of sulfonamides has been explored for the activation of pyruvate kinase to alter the metabolism of cancer cells.[35] The unweighted clusters in this example appeared to be influenced by other, non-activity-relevant groups such as nitroxy (cluster 3) and toluoyl (cluster 2) substituents.

As already seen in changes of similarity values, activity-characteristic substructures were well-distinguished if they had distinct substitution patterns. For example, for TRPC6 calcium channel inhibitors (AID 2696, Figure 4D), an ethanolyl substitution at the piperazine ring was highly characteristic of active compounds (Figure 4D, cluster a), while the piperazine ring alone was not characteristic of actives in this assay (Figure 4D, molecule 3a). Core structures used to describe compound series (like the aminoprop-2-en-1-ones in Figure 4A) were successfully mapped in several cases; e.g., pyridazinone derivatives were grouped among GSK-3 inhibitors (AID 463203) by weighted clustering (Figure 4E, cluster "a") and previously have been introduced as GSK-3 inhibitors.[36] Conversely, conventional clustering separated them in the bioassay (Figure 3E, compounds "2a" and "3a").

Overall, compounds in weighted clusters shared activity-characteristic features, as visualized by feature mapping and could be highly variable in other regions (Figure 4A−E, Supporting Information Figure S2A−J). While these active cores often included ring systems, they were not limited to any predefined notion of a scaffold but emerged from mapping of activity-characteristic fingerprint features.

Elucidation of Orthogonal SAR. A key observation from weighted clustering was that one and the same substructure could be characteristic of one biological activity but behave as an unspecific functional group for another set of bioactive compounds (Figures 4B,C). Going beyond the primary activity, we compared distinct feature maps for the same compound that were calculated on the basis of distinct assays. Our goal was to find compounds with multiple biological activities for which each activity could be attributed to a distinct characteristic substructure of the molecule (Figures 5A−C, Supporting Information Figure S3A−G). Using an approach analogous to fingerprint comparison, we were able to find feature maps that had minimal overlap for selected compounds.

For example, the compound shown at the top in Figure 5A increased the expression of NFκB (AID 1241) and also inhibited human lipoxygenase 2 (AID 881). The benzodioxole moiety was characteristic of the first activity, as supported by other NFκB expression inducers sharing this substructure, while the thiourea moiety was not characteristic of this activity. In addition to

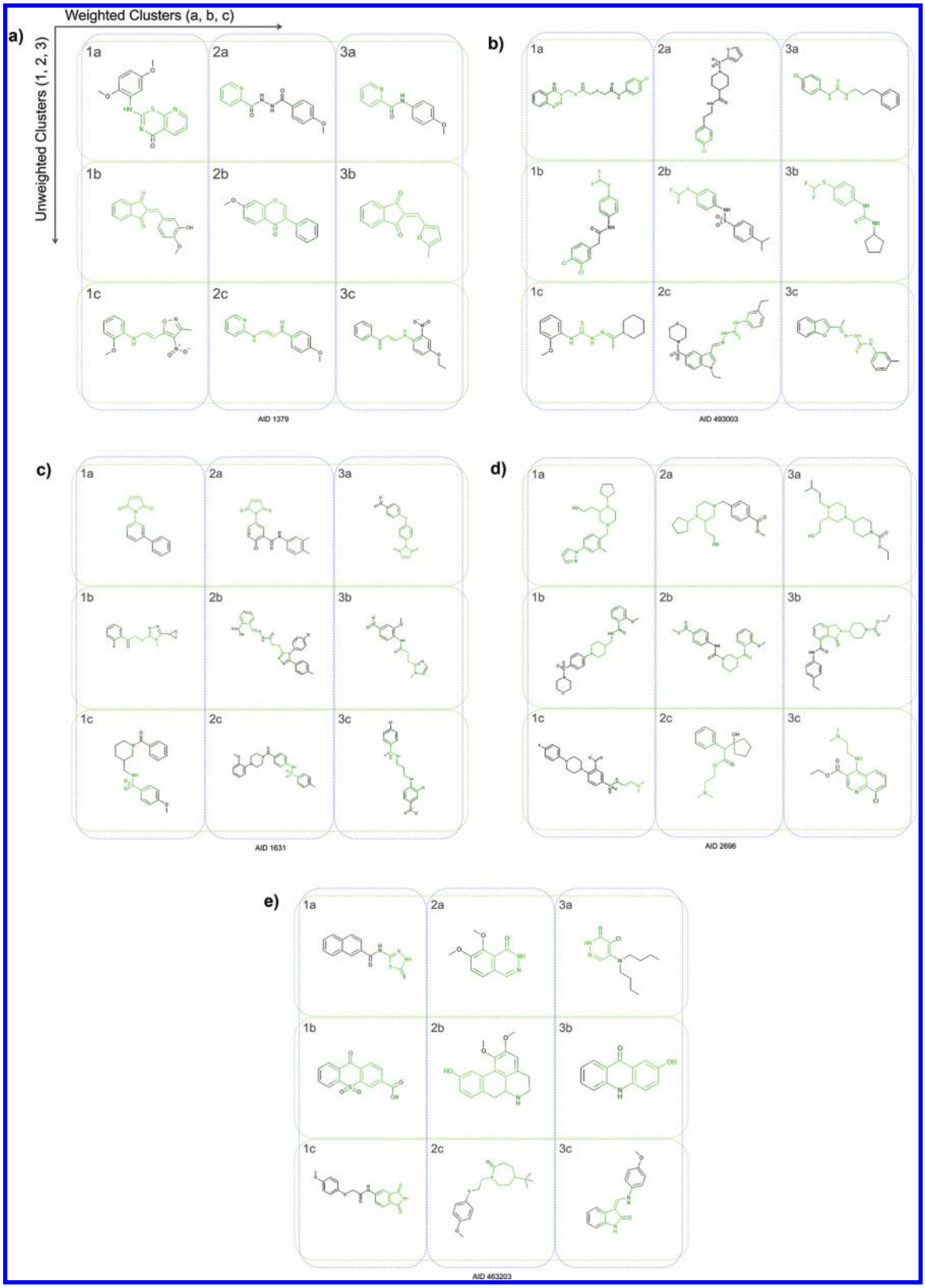


Figure 4. Informative cluster differences. (A−E) For five bioassay sets, representative examples of cluster differences between nonweighted and weighted clustering are shown. Nine molecules are shown that belong to three unweighted (vertical blue boxes, numbers) or three weighted clusters (horizontal green boxes, letters), respectively. Molecules from each nonweighted cluster belong to three distinct weighted clusters, and vice versa. Activity-characteristic substructures are mapped in green. AID: Pubchem Bioassay ID.

benzodioxole, the similar benzodioxan fragment was also mapped in one of the molecule's closest neighbors based on weighted fingerprint similarity (bottom left compound in Figure 5A). The correspondence between these two fragments

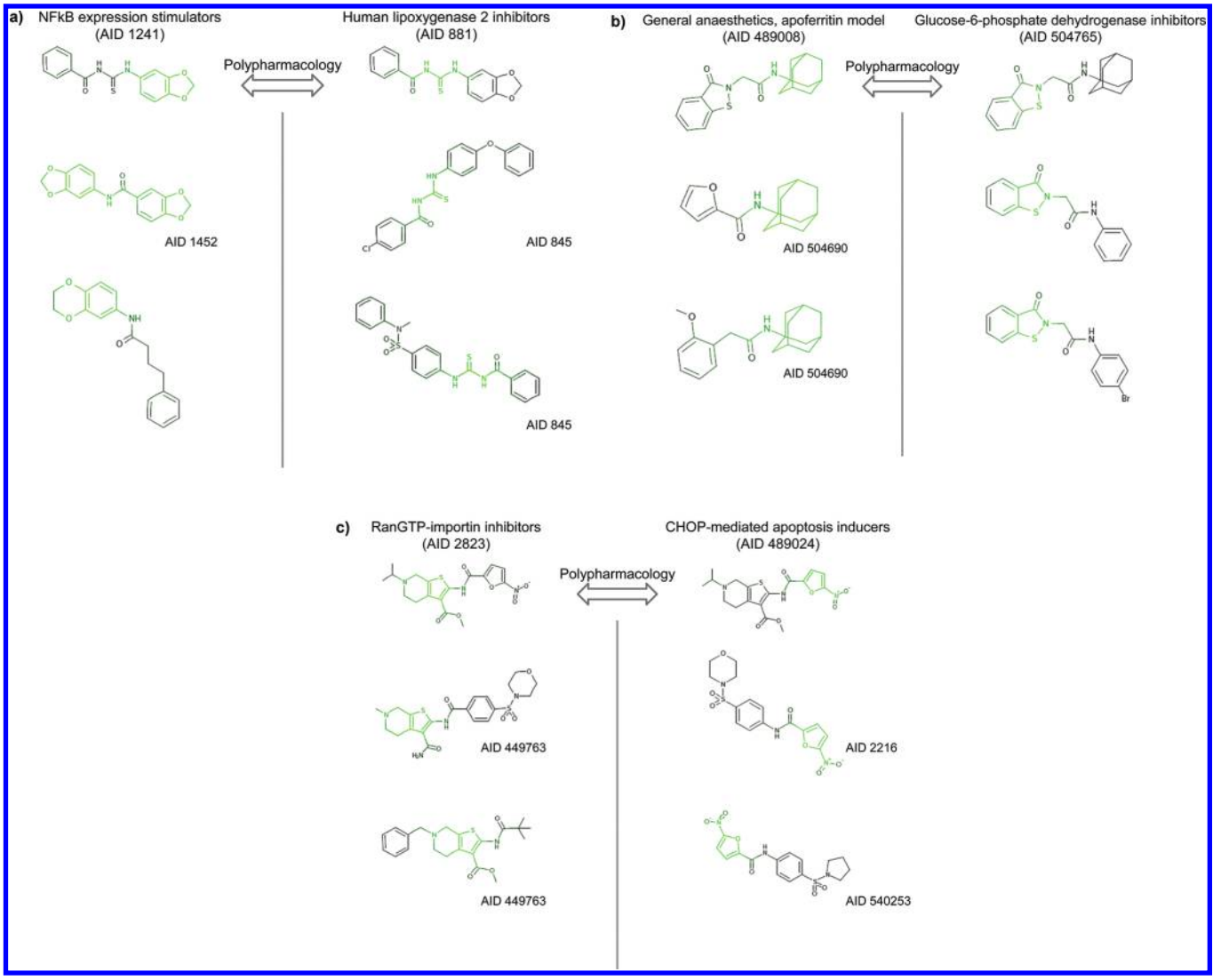


Figure 5. Orthogonal SAR elucidation. (A−C) Representative pylypharmacological compounds with mapped active cores. Two representations of the compound with both activities are shown at the top to highlight the different mapped cores. For each compound and assay, two similar compounds based on weighted fingerprints are shown for each assay. These additional compounds are active in their respective bioassay, but not the other. Additional Bioassay IDs (AID) are reported for compounds that have been tested in parent assays of the other target. For example, both compounds on the right side in A have been tested and found inactive in a primary assay for NFkB expression stimulators (AID 845).

would not have been found using predefined fragmentation approaches, such as scaffold decomposition. Conversely, the thiourea linker was characteristic of lipoxygenase 2 inhibitors, while the benzodioxole fragment could be substituted by other groups (Figure 5A, right panel). Similarly, the adamantane group was characteristic of compounds binding to apoferritin in a general anesthetics model (AID 489008, Figure 5B), consistent with a largely hydrophobic binding pocket of apoferritin.[41] By contrast, another part of the compound, namely, the thiazol substructure, was characteristic of the compound's inhibitory activity at glucose-6-phosphate dehydrogenase (AID 504765). In addition to substructures characteristic of activity at a particular protein, also pathway activity-characteristic substructures could be mapped (Figure 5C). The compound shown in Figure 5C at the top was found both to inhibit the RanGTP-importin β complex (AID 2823) and to induce apoptosis via CHOP (AID 489024), in line with the known pro-apoptotic properties of nitrofurans.[42] While the two compounds shown in the mid panel share a large common moiety, they were selective for their respective assays.

To ensure that the found compounds were indeed selective, and not merely untested in the assay that we analyzed, we searched Pubchem bioassays for primary assay data on these compounds. For nine out of 12 assessed compounds, we found primary assays where they have been tested. In all nine cases, they had been found inactive (Figure 5A−C), corroborating the selectivity-conferring nature of substructures identified by our method.

■ DISCUSSION

Given today's advances of HTS technology, the bottleneck in hit identification has arguably shifted from running the actual screen to building an activity hypothesis and extracting SAR information post-HTS.[2] Manual assessment of hundreds to thousands of hits by medicinal chemists is a daunting and

subjective[37] task that calls for computational methods such as the one presented here to facilitate elucidation of chemical core structures defining active compound series. Polypharmacology considerations, both from a desired activity[4] and pharmacosafety[25] points of view, further complicate hit assessment and require approaches that facilitate identification of suitable medchem strategies to balance selectivity across different targets.

Fingerprint-based molecular similarity approaches have been widely used for various applications including virtual screening[9,11,38] and compound clustering.[6,9,10] Furthermore, Bayesian models, which conceptually differ from molecular similarity, have proved effective in target prediction and identification of active compounds.[24] Here, we have combined these two complementary concepts, allowing for bioactivity-aware clustering of active compounds identified in HTS campaigns. We put emphasis on activity-characteristic molecular cores that were defined in a data-driven manner, tailored toward each individual assay. The resulting clusters were defined by these privileged substructures and facilitated SAR elucidation in these generally diverse data sets. While neither molecular clustering nor Bayesian models are novel approaches in medicinal chemistry, their combination for the elucidation of SAR and medchem strategy guidance, to our knowledge, has not yet been explored. Going beyond individual sets of active compounds, we compared active molecular cores of polypharmacological compounds and identified molecular substructures responsible for differential activity, elucidating medchem strategies to promote selectivity, while retaining primary activity.

Fingerprint Weighting. Molecular fingerprint weighting schemes have been previously explored in the context of activity class-directed virtual screening (VS) and often outperformed generic fingerprints.[11,13,39,40] Individual features can be emphasized on the basis of their occurrence in active compounds[15,41] or their influence on hit retrieval in benchmark calculations.[39] Such reverse-engineering[15] of fingerprints helps assess pharmacophoric features in a data-driven manner. The goal of weighting schemes in a VS context is to enrich active compounds in a large database of virtual compounds by evaluating their molecular fingerprint similarity to a reference set of known actives. Similarly, the standard application of naïve Bayesian models is the prioritization of active compounds from a large database or identification of potential targets of a single compound.[22]

Distinct from virtual screening, we combined Bayesian models and molecular similarity to aid in SAR elucidation. Here, the activity of compounds is known *a priori*, and the goal is to identify activity-characteristic chemotypes that can serve as starting points for a multidimensional optimization strategy. A standard way of identifying common chemotypes is the clustering of active compounds.[6] In traditional approaches, however, information about compound activity is projected onto compound clusters that have been generated in an activity-agnostic fashion. Hence, compounds are often grouped on the basis of chemical features that are not necessarily decisive for activity. This problem can be related to molecular complexity effects in VS:[42] compounds having many decorations may be apparently similar to a large number of diverse compounds.

In our approach, focusing on activity-characteristic molecular features corrected conventional similarity values in cases where they were influenced by substructures equally likely to occur in inactive compounds, like ring decorations or common scaffolds (Figure 2). Although stereochemistry can be important for biological activity, it is often missing or wrongly reported in public databases,[43] thus adding potential noise to the data. In these cases, treating otherwise identical, activity-characteristic features separately might arbitrarily decrease their weight. Therefore, we chose to disregard stereochemistry. However, there might be activity classes, such as steroids, where, given careful annotation, stereoselective clusters could be identified by our approach. In principle, the general form of the Tanimoto coefficient would allow including negative weights as well. We decided to use only positive Bayesian weights, because we wanted to emphasize activity-characteristic molecular substructures and assign a constant weight to the rest of the molecule. The weighting coefficient was devised to also account for noncharacteristic features. An alternative approach could be scaling of the entire range of Bayesian weights. However, this would mix noncharacteristic and activity-characteristic features, making the mapping hard to interpret; conversely, our weighting scheme directly translates to a comparison of activity-characteristic substructures of compounds (Figure 1A).

Data set size might affect the performance on two levels, at the Bayesian weighting step and at the clustering step. It has been noted that naïve Bayesian models might overemphasize features present in small sets of active compounds.[22] However, our fingerprint weighting approach avoids complications typically arising in VS from such overemphasis in two ways. First, only active compounds were compared to each other, and hence only the relative differences of positive weights derived from the activity class were relevant to the similarity calculation. Second, all other features were also taken into account, albeit with a smaller weight. This balancing of activity-characteristic features with other molecular substructures allowed the clustering method to "fall back" onto generic fragments for compounds that did not produce many activity-characteristic fingerprint features. At the clustering step, we adjusted the number of clusters based on the data set size. This ensured comparable cluster size within and between different assay data sets. Combined with a focus on activity-characteristic core structures, this clustering approach yielded stable, biologically interpretable clusters for all assay sets.

Activity-Aware vs Conventional Clustering. Clustering was substantially influenced by the adjusted similarity values—on average, more than half of the compound pairs originally grouped together were separated by weighted clustering and *vice versa*. Weighted clusters consisted of compounds sharing activity-characteristic core structures that could represent ring systems as well as linkers (Figure 4B) or side chains (Figure 4D) and thus were not restricted to any predefined molecular scaffolds.[16,20] Departing from predefined substructures, our method grouped together even small active cores that shared common features but were not exactly the same, such as thiazoles, if they were characteristic of activity (e.g., antimalarials in Figure 3B). It would be very difficult to identify such commonalities using predefined scaffolds. In several cases, the active molecular cores identified by our method could also be found in the literature or patents discussing the compounds' SAR. Feature mapping allowed visual rationalization of weighted clusters in terms of activity-characteristic cores.[32]

A general aim of molecular clustering is to identify common active chemotypes and distinguish them from groups that can be used to diversify compounds. Our methodology explicitly focuses on molecular core structures characteristic of active compounds, without the need of predefined fragmentation rules. Using weighted clustering, compounds were grouped together that normally would not have been directly compared, although they share activity-characteristic features (Figure 3A,B). Conversely, contamination of clusters by compounds with distinct

pharmacophores could be reduced. Integration of both weighted and unweighted clustering into one visualization scheme allowed us to prioritize core structures and further subdivide clusters that were characterized by a comparably small activity-characteristic moiety (such as carboxylic acids in Figure 3B).

Our method facilitates the identification of privileged substructures that can serve as starting points for diversification. The archetypical question of medicinal chemistry, "What compound should I synthesize next?", thus is answered in an activity-aware way: "Keep the activity-characteristic molecular core of found hits, and change other groups." This strategy will be particularly useful for the identification of "warheads", i.e., well-defined chemical groups crucial for activity.

Orthogonal SAR of Polypharmacologic Compounds. In addition to the primary activity, other molecular properties usually need to be optimized. Of special concern is selectivity against other targets that may cause unwanted effects and need to be balanced with the primary activity.[4,5,8] Early focus on only the primary target may result in undesirable medchem strategies that cannot avoid off-target activity because of parallel SAR: a structural change that leads to more potent compounds may also increase activity at the off-target. Conversely, chemotypes with an orthogonal SAR allow for changes of the two activities independently and thus increase the chance of designing selective compounds.

We compared activity-characteristic cores derived for individual compounds with multiple activities. This allowed selection of chemotypes that can be used in selectivity-increasing medchem strategies. Distinct molecular regions were characteristic of different activities, and only a combination of both distinct features led to activity in both assays (Figure 5A−C). By contrast, even compounds that looked similar, but had very dissimilar activity-characteristic cores, were found to be selective for their respective assay (Figure 5C). Without feature mapping, it would be very difficult to identify and prioritize chemotypes that allow orthogonal SAR exploration for different biological activities. Conversely, using our approach, chemotypes can be prioritized that have a high chance of yielding compounds with a desired selectivity profile. This method can be easily applied beyond HTS hit lists. Going beyond known polypharmacological compounds, the identified structural determinants of selectivity and/or polypharmacology may be combined to form compounds with novel, heterogeneous activity profiles.

CONCLUSIONS

We have introduced an activity-aware clustering approach by combining Bayesian modeling with molecular similarity. Balancing activity-characteristic features with more generic fragments yielded clusters containing compounds with common, activity-characteristic cores. These cores were visualized using feature mapping that allowed us to easily rationalize how activity-aware clusters were formed. Compounds sharing activity-characteristic core structures that were highly dispersed among several conventional clusters were grouped together by our method. Our method presents a new way of elucidating SAR in large and diverse high throughput screening data sets and can help guide medchem strategies that use activity-privileged chemotypes as starting points of diversification.

Furthermore, comparison of active molecular cores for polypharmacological compounds led to the identification of distinct molecular regions that were specific for single activities, while overall similar compounds differing in these regions were found to be selective. Thus, feature mapping identifies chemotypes suitable for medchem strategies aimed at increasing selectivity.

ASSOCIATED CONTENT

Supporting Information. Supplementary methods (parametrization of the weighting scheme), Table S1, and Figures S1−S3. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Tel.: +1 617 871 4953. E-mail: eugen.lounkine@novartis.com.

ACKNOWLEDGMENT

E.L. and F.N. are presidential postdoctoral fellows supported by the Education Office of the Novartis Institutes for Biomedical Research. The authors thank Douglas Auld for helpful discussions of the manuscript.

REFERENCES

(1) Mayr, L. M.; Bojanic, D. Novel trends in high-throughput screening. *Curr. Opin. Pharmacol.* **2009**, *9*, 580–588.

(2) Glick, M.; Jacoby, E. The role of computational methods in the identification of bioactive compounds. *Curr. Opin. Chem. Biol.* **2011**, *15*, 540–546.

(3) Merino, A.; Bronowska, A. K.; Jackson, D. B.; Cahill, D. J. Drug profiling: knowing where it hits. *Drug Discovery Today* **2010**, *15*, 749–756.

(4) Zhang, X.; Crespo, A.; Fernández, A. Turning promiscuous kinase inhibitors into safer drugs. *Trends Biotechnol.* **2008**, *26*, 295–301.

(5) Giacomini, K. M.; Krauss, R. M.; Roden, D. M.; Eichelbaum, M.; Hayden, M. R.; Nakamura, Y. When good drugs go bad. *Nature* **2007**, *446*, 975–977.

(6) 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 Discovery Today* **2010**, *15*, 630–639.

(7) Schnur, D. M.; Hermsmeier, M. A.; Tebben, A. J. Are target-family-privileged substructures truly privileged? *J Med. Chem.* **2006**, *49*, 2000–2009.

(8) Mencher, S. K.; Wang, L. G. Promiscuous drugs compared to selective drugs (promiscuity can be a virtue). *BMC Clin. Pharmacol.* **2005**, *5*, 3.

(9) Bender, A.; Glen, R. C. Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.* **2004**, *2*, 3204–3218.

(10) Varin, T.; Bureau, R.; Mueller, C.; Willett, P. Clustering files of chemical structures using the Székely-Rizzo generalization of Ward's method. *J. Mol. Graphics Modell.* **2009**, *28*, 187–195.

(11) Geppert, H.; Vogt, M.; Bajorath, J. Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. *J. Chem. Inf. Model.* **2010**, *50*, 205–216.

(12) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

(13) Shemetulskis, N. E.; Weininger, D.; Blankley, C. J.; Yang, J. J.; Humblet, C. Stigmata: an algorithm to determine structural commonalities in diverse datasets. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 862–871.

(14) Chen, X.; Rusinko, A., III; Tropsha, A.; Young, S. S. Automated pharmacophore identification for large chemical data sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 887–896.

(15) Williams, C. Reverse fingerprinting, similarity searching by group fusion and fingerprint bit importance. *Mol. Divers.* **2006**, *10*, 311–332.

(16) Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. A.; Waldmann, H. The scaffold tree--visualization of the scaffold universe by hierarchical scaffold classification. *J. Chem. Inf. Model.* **2007**, *47*, 47–58.
(17) Wetzel, S.; Klein, K.; Renner, S.; Rauh, D.; Oprea, T. I.; Mutzel, P.; Waldmann, H. Interactive exploration of chemical space with Scaffold Hunter. *Nat. Chem. Biol.* **2009**, *5*, 581–583.
(18) 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.
(19) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.
(20) Agrafiotis, D. K.; Wiener, J. J. M. Scaffold explorer: an interactive tool for organizing and mining structure-activity data spanning multiple chemotypes. *J. Med. Chem.* **2010**, *53*, 5002–5011.
(21) Nidhi; Glick, M.; Davies, J. W.; Jenkins, J. L. Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *J. Chem. Inf. Model.* **2006**, *46*, 1124–1133.
(22) Bender, A. Bayesian methods in virtual screening and chemical biology. *Methods Mol. Biol.* **2011**, *672*, 175–196.
(23) Glick, M.; Klon, A. E.; Acklin, P.; Davies, J. W. Enrichment of extremely noisy high-throughput screening data using a naïve Bayes classifier. *J. Biomol. Screen.* **2004**, *9*, 32–36.
(24) Bender, A.; Jenkins, J. L.; Glick, M.; Deng, Z.; Nettles, J. H.; Davies, J. W. Bayes affinity fingerprints improve retrieval rates in virtual screening and define orthogonal bioactivity space: when are multitarget drugs a feasible concept? *J Chem. Inf. Model.* **2006**, *46*, 2445–2456.
(25) Bender, A.; Scheiber, J.; Glick, M.; Davies, J. W.; Azzaoui, K.; Hamon, J.; Urban, L.; Whitebread, S.; Jenkins, J. L. Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure. *ChemMedChem* **2007**, *2*, 861–873.
(26) Scheiber, J.; Jenkins, J. L.; Sukuru, S. C. K.; Bender, A.; Mikhailov, D.; Milik, M.; Azzaoui, K.; Whitebread, S.; Hamon, J.; Urban, L.; Glick, M.; Davies, J. W. Mapping adverse drug reactions in chemical space. *J. Med. Chem.* **2009**, *52*, 3103–3107.
(27) Li, Q.; Cheng, T.; Wang, Y.; Bryant, S. H. PubChem as a public resource for drug discovery. *Drug Discovery Today* **2010**, *15*, 1052–1057.
(28) Stein, S.; Heller, S.; Tchekhovski, D. An Open Standard for Chemical Structure Representation - The IUPAC Chemical Identifier. *Nimes International Chemical Information Conference Proceedings*; Nimes, France, October 19−22, 2003; Infonortics: Tetbury, U.K., 2003; pp 131−143.
(29) Cormack, R. M. A Review of Classification. *J. R. Stat. Soc. Ser. A (Gen.)* **1971**, *134*, 321–367.
(30) The R Project for Statistical Computing (version 2.13). http://www.r-project.org (accessed Nov. 2011).
(31) *Multivariate Analysis*; Mardia, K. V., Kent, J. T., Bibby, J. M., Eds.; Academic Press: Waltham, MA, 1980.
(32) Lounkine, E.; Batista, J.; Bajorath, J. Mapping of activity-specific fragment pathways isolated from random fragment populations reveals the formation of coherent molecular cores. *J. Chem. Inf. Model.* **2007**, *47*, 2133–2139.
(33) Auld, D. S.; Southall, N. T.; Jadhav, A.; Johnson, R. L.; Diller, D. J.; Simeonov, A.; Austin, C. P.; Inglese, J. Characterization of chemical libraries for luciferase inhibitory activity. *J. Med. Chem.* **2008**, *51*, 2372–2386.
(34) Hasson, S. A.; Damoiseaux, R.; Glavin, J. D.; Dabir, D. V.; Walker, S. S.; Koehler, C. M. Substrate specificity of the TIM22 mitochondrial import pathway revealed with small molecule inhibitor of protein translocation. *Proc. Natl. Acad. Sci. U.S.A* **2010**, *107*, 9578–9583.
(35) Boxer, M. B.; Jiang, J.-k.; Vander Heiden, M. G.; Shen, M.; Skoumbourdis, A. P.; Southall, N.; Veith, H.; Leister, W.; Austin, C. P.; Park, H. W.; Inglese, J.; Cantley, L. C.; Auld, D. S.; Thomas, C. J. Evaluation of substituted N,N′-diarylsulfonamides as activators of the tumor cell specific M2 isoform of pyruvate kinase. *J. Med. Chem.* **2010**, *53*, 1048–1055.
(36) Hoelder, S.; Naumann, T.; Schoenafinger, K.; Will, D.; Matter, H.; Mueller, G.; Le Suisse, D.; Baudoin, B.; Rooney, T.; Halley, F.; Tiraboschi, G. Pyridazone Derivatives as GSK-3beta inhibitors. WO 2004046117, Jun. 3, 2003.
(37) Lajiness, M. S.; Maggiora, G. M.; Shanmugasundaram, V. Assessment of the consistency of medicinal chemists in reviewing sets of compounds. *J. Med. Chem.* **2004**, *47*, 4891–4896.
(38) Willett, P. Similarity searching using 2D structural fingerprints. *Methods Mol. Biol.* **2011**, *672*, 133–158.
(39) Wang, Y.; Bajorath, J. Bit silencing in fingerprints enables the derivation of compound class-directed similarity metrics. *J. Chem. Inf. Model.* **2008**, *48*, 1754–1759.
(40) Vogt, I.; Bajorath, J. Analysis of a high-throughput screening data set using potency-scaled molecular similarity algorithms. *J. Chem. Inf. Model.* **2007**, *47*, 367–375.
(41) Xue, L.; Godden, J. W.; Stahura, F. L.; Bajorath, J. Profile scaling increases the similarity search performance of molecular fingerprints containing numerical descriptors and structural keys. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1218–1225.
(42) Flower, D. R. On the Properties of Bit String-Based Measures of Chemical Similarity. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 379–386.
(43) Williams, A. J.; Ekins, S. A quality alert and call for improved curation of public chemistry databases. *Drug Discovery Today* **2011**, *16*, 747–750.