Automated structure–activity relationship mining: connecting chemical structure to biological profiles

Mathias J. Wawer\textsuperscript{1}, David E. Jaramillo\textsuperscript{1}, Vlado Dancik\textsuperscript{1,2}, Daniel M. Fass\textsuperscript{1,3}, Stephen J. Haggarty\textsuperscript{1,3}, Alykhan F. Shamji\textsuperscript{1}, Bridget K. Wagner\textsuperscript{1}, Stuart L. Schreiber\textsuperscript{1}, Paul A. Clemons\textsuperscript{1*}

\textsuperscript{1}Center for the Science of Therapeutics, Broad Institute, 7 Cambridge Center, Cambridge, MA 02142

\textsuperscript{2}Mathematical Institute of the Slovak Academy of Sciences, Grešáková 6, Košice, Slovakia (on leave)

\textsuperscript{3}Chemical Neurobiology Laboratory, Center for Human Genetic Research, Massachusetts General Hospital, Departments of Neurology & Psychiatry, Harvard Medical School, 185 Cambridge Street, Boston, MA 02114, USA

* pclemons@broadinstitute.org

KEYWORDS: structure-activity relationships, small-molecule profiling, high-content screening, frequent itemset mining, association rule mining
Abstract

Understanding structure–activity relationships (SARs) of small molecules is important for developing probes and novel therapeutic agents in chemical biology and drug discovery. Increasingly multiplexed small-molecule profiling assays allow simultaneous measurement of many biological response parameters for the same compound, e.g. expression levels for many genes or binding constants against many proteins. While such methods promise to capture SARs with high granularity, few computational methods are available to support SAR analyses of high-dimensional compound activity profiles. Many of these methods are not generally applicable or reduce the activity space to scalar summary statistics before establishing SARs. In this article, we present a versatile computational method that automatically extracts interpretable SAR rules from high-dimensional profiling data. The rules connect chemical structural features of compounds to patterns in their biological activity profiles. We applied our method to data from novel cell-based gene-expression and imaging assays collected on more than 30,000 small molecules. Based on the rules identified for this dataset, we prioritized groups of compounds for further study, including a novel set of putative histone deacetylase inhibitors.
Introduction

Small-molecule profiling, i.e., the characterization of compounds by multiple biological activity measurements, has been shown to capture detailed information about biological effects and mechanisms of action of small molecules.\(^1\) This level of granularity holds great promise for a comprehensive understanding of compound SARs,\(^2,3\) which in turn allows optimizing compounds simultaneously against multiple biological endpoints.\(^4\) To overcome limitations of existing approaches, we developed a computational method to automatically mine small-molecule profiling data for SAR rules.

Profiles can be obtained in parallel by combining separately acquired assay results such as binding constants against different purified proteins,\(^4\) drug sensitivity of different cell lines,\(^5,6\) or high-throughput screening results from different assays.\(^7\) By contrast, multiplexed profiling assays capture complex cell states by simultaneously measuring many features in a single well format as a “signature”, e.g. gene expression, protein levels and modifications, or cell morphology descriptors.\(^1,4\)

Growing public bioactivity databases\(^8,9\) and novel high-throughput experimental methods\(^10,11\) enable researchers to apply both types of profiling data in early stages of drug- and probe-discovery projects when the number of compounds to consider is still very large.\(^1,4\) This application, in turn, requires robust and scalable computational analysis methods to help interpret the large numbers of complex activity profiles. In addition to the inherently multi-dimensional space of chemical structures, the number of possible profiles grows exponentially with each biological measurement dimension.
Most available SAR analysis methods, however, were designed to handle between one and five bioactivity annotations and are not easily extensible.\textsuperscript{12-14} One strategy is therefore to reduce high-dimensional bioactivity data to summary statistics, such as the number of shared targets between two compounds, before connecting this information to chemical structure.\textsuperscript{15}

Here, we present a computational approach for directly deriving interpretable SAR rules from large amounts of biological profiling data. We used frequent pattern mining (FPM; also called frequent itemset mining) and association rule mining (ARM) to find combinations of substructures that are associated with characteristic biological profiles.\textsuperscript{16}

FPM was originally developed for market-basket analysis to find combinations of products that are frequently bought together. ARM builds on these “frequent patterns” (or “frequent itemsets”) to find interesting associations between products. These associations are formulated as rules. For example, the rule \{\textit{paper, envelopes}\} \rightarrow \{\textit{stamps}\} indicates that customers who buy paper and envelopes are likely to also purchase stamps. We applied ARM to small-molecule profiling data to identify SAR rules of the form \{\textit{chemical attributes}\} \rightarrow \{\textit{biological effect patterns}\}. Importantly, rules can be prioritized by measuring the strength of the association between features and their “interestingness”.

FPM has been used to find frequent substructures in sets of compounds and to identify single contiguous fragments that distinguish active and inactive compounds in a bioactivity assay.\textsuperscript{17,18} Fragment \textit{combinations} that separate compounds with distinct binding profiles have been identified using formal concept analysis, a framework that is conceptually related to FPM.\textsuperscript{19} This method captures the precise relations between individual sets of
compounds and fragments, rendering this approach more useful for in-depth analysis of smaller data sets with few activity annotations. The requirement that fragment combinations perfectly separate compound sets is usually too stringent for exploratory SAR studies.

We present here a general approach to analyzing SARs for large numbers of high-dimensional biological profiles using FPM and, in addition, ARM to automatically formulate SAR rules. Rules that connect chemical features to patterns in biological profiles are automatically identified and ranked by interestingness. We evaluated our method on gene-expression and cell-morphology profiles for more than 30,000 compounds. The compound collection contains subsets representative of common screening libraries assembled from various sources as well as planned synthetic libraries of compounds with well-defined structural relationships. We used different chemical and biological descriptors to tailor the general approach to specific requirements of the compound library.
Materials and Methods

Compound sets

We assembled three distinct compound sets with varying structural properties and levels of biological activity annotation. We first selected a collection of 19,637 structurally diverse compounds derived from diversity-oriented synthesis (DOS).\textsuperscript{20} The library was synthesized by the Center for the Science of Therapeutics at the Broad Institute using a build/couple/pair strategy.\textsuperscript{21} The compounds were built around 23 novel chiral core structures by systematically varying the configuration of core stereocenters and decorating them with various side-chains (see Supplementary Information). Whenever possible, complete sets of stereoisomers were included in our set. No biological activity information was used in the selection of the DOS set. Second, we used PubChem (http://pubchem.ncbi.nlm.nih.gov/) assay data to select 10,162 compounds from the Molecular Libraries Small Molecule Repository (MLSMR) that cover many different chemical structures and biological activities. Third, 2,222 known drugs and probes were added as bioactivity landmarks. Again, these compounds were selected to be structurally diverse and cover a wide range of biological activities. The complete compound set is available as Supplementary Material at http://jbx.sagepub.com/.

Gene-expression profiles (GE)

We followed the protocol of Peck et al. (see Supplementary Methods for a detailed protocol).\textsuperscript{10} Briefly, 3,500 U-2 OS cells (ATCC, cat. no. HTB-96) were seeded per well in 384-well plates. After 24h of incubation at 37°C, compounds were added and the cells incubated for another 6h of treatment. Cell lysates were prepared and transferred to oligo-dT capture
plates to immobilize mRNAs. cDNA was then generated via reverse transcription and amplified via polymerase chain reaction (PCR). The cDNA was first annealed with a mix of specific probe pairs designed for each of 978 transcripts. All upstream probes contain a universal 20-nucleotide (nt) primer site (complementary to T7 primer), a specific 24-nt barcode sequence, and a transcript-specific sequence of 24nt complementary to the 3’-end of the transcript. Downstream probes consist of a 5’-phosphorylated 20-nt sequence that is complementary to the sequence contiguous to the binding site of the upstream probe on each transcript, as well as a 20-nt universal primer (T3) site. After annealing and removal of free probes, adjacent probes bound to transcript cDNAs were joined by ligation and amplified by PCR using T3 and 5’-biotinylated T7 primers. The resulting amplicons were hybridized with color-coded Luminex microspheres each of which carries capture probes complementary to one of the barcode sequences in the amplicons. After incubation with streptavidin-phycoerythrin, the captured amplicons were quantified by flow cytometry measuring phycoerythrin fluorescence. At the same time, transcript identity was identified by bead color. Compound effects were measured in triplicate treated wells.

**Multiplexed cytological morphology profiles (MC)**

We followed the protocol of Gustafsdottir et al. (see Supplementary Methods for a detailed protocol). Briefly, 1,500-2,000 U-2 OS cells (ATCC, cat. no. HTB-96) were seeded per well in 384-well clear-bottom imaging plates. After 24h of incubation at 37°C, compounds were added and cells incubated for another 48h of treatment. Six fluorescent dyes were added to stain different cell compartments: nucleus (Hoechst 33342), endoplasmic reticulum (concanavalin A/AlexaFluor488 conjugate), nucleoli (SYTO 14 green fluorescent nucleic acid stain), Golgi apparatus and plasma membrane (wheat germ agglutinin/AlexaFluor594 conjugate).
conjugate, WGA), F-actin (phalloidin/AlexaFluor594 conjugate) and mitochondria (MitoTracker Deep Red). WGA and Mitotracker stains were performed on live cells. Afterwards, cells were fixed with 16% paraformaldehyde and the remaining dyes added. Images in 5 fluorescent channels were captured for 9 sites per well at 20× magnification. Images were measured in quadruplicate and CellProfiler software was used to calculate morphological features for each cell.\textsuperscript{11}

**Chemical descriptors**

**DOS-FP**

We used synthetic history information to define DOS compound core structures. Based on these cores, appendages were determined by R-group decomposition. Information about the attachment position of each appendage to the core was recorded. In addition, the absolute configuration of all stereocenters in the core was determined. This determination was performed on core structures without appendages to keep absolute stereochemistry consistent irrespective of substituent priority changes with different appendages. Concatenating core, appendage, and stereochemistry features into a feature string generated the DOS-FP for each compound.

**ECFP**

Extended connectivity fingerprints\textsuperscript{22} with atom types were calculated using a bond distance of 4 (ECFP4). Structural fragments that correspond to each bit were combined into a feature string for each compound. Only features with a bond diameter of 2 and higher were considered to exclude non-informative single atom fragments. All calculations were performed in SciTegic Pipeline Pilot 8.5.
Biological descriptors

Cluster attributes

We hierarchically clustered the data based on biological profiles (GE or MC) using a complete linkage method on pairwise correlation distances. All nodes in the cluster dendrogram were indexed. Each compound was assigned a set of indices based on cluster membership.

Signatures

GE signatures were obtained from GE profiles by applying a threshold of 2 (-2) to z-scores relative to DMSO to obtain up-regulated (down-regulated) genes. Genes were then combined into a feature string, indicating the direction of regulation.

Frequent pattern mining and association rules

Frequent itemsets were determined using a publicly available implementation (http://mahout.apache.org/) of the Parallel FPgrowth algorithm on a Hadoop platform (version 1.0.4; http://hadoop.apache.org/). Association rules were generated from the list of frequent patterns in R (version using the arules package.

SAR score

We calculated confidence, purity, and p-values (Fisher’s exact test) for each rule and combined them into a single SAR score. P-values were transformed into negative decadic logarithms. The distributions of all three quality measures were first normalized using a Box-Cox transformation to avoid one criterion dominating the others and the resulting values for each quality measure were scaled to a range [0,1]. Purity values were multiplied
with a weighting factor of 0.2. The sum of the normalized quality measures yielded the SAR score for each compound. SAR scores were scaled to range [0,1] by division with the theoretical maximum score of 2.2.

**Replicate-based hit selection**

To determine which compounds led to stable profile changes compared to negative controls (DMSO), we used replicate correlation and connectivity, *i.e.*, the similarity of replicates compared to all other wells on the same set of replicate assay plates. Replicate correlation was calculated as the Pearson correlation coefficient between their GE (or MC) profiles. Replicate connectivity was calculated for each pair of replicates (*e.g.*, R1 and R2) by ranking all wells on the same set of replicate plates by their profile similarity against R1 and identifying the fraction of unrelated wells that rank higher than R2. The reverse calculation (R2 vs. R1) was also performed because connectivity values are not symmetric. To allow individual replicates to fail, we only considered the top 50% of replicate correlation and connectivity values for each compound. We then calculated the two-dimensional distribution of negative-control replicate correlation and connectivity. Compounds were considered hits if they exceeded the 97% confidence interval of this negative-control distribution.
Results

Background

We applied FPM and ARM to small-molecule gene-expression and cell-morphology profiles to derive SAR rules of the form \{chemical fragments\} $\rightarrow$ \{biological effect pattern\}, indicating that compounds containing specific molecular fragments (or chemical substructures) often lead to the same pattern in biological responses (e.g. similar expression changes for a common set of genes). Even though our method can be applied to more abstract chemical attributes (e.g. calculated properties), using fragments makes it easy to interpret the final rules. Rules that connect fragments to biological effects can directly inform synthetic chemistry decisions, especially if these fragments match existing molecular building blocks.

We derived biological attributes from two profiling assays, gene expression (GE) and cell morphology (multiplexed cytological, MC). For GE profiles, the expression levels of 978 protein-coding transcripts were determined by ligation-mediated amplification and Luminex bead-based detection.\(^{10}\) For MC profiles, changes in 812 cell morphology features were captured by automated microscopy and computational image analysis. We used a “cell-painting” assay with 6 fluorescent dyes in five channels to distinguish cellular compartments and organelles.\(^{11}\)

In the following, we follow the notation of Tan, Steinbach, and Kumar\(^{16}\) to introduce FPM and ARM. FPM and ARM operate on collections of binary attribute vectors ("objects"). Let $I = \{i_1, i_2, \ldots, i_d\}$ be a set of binary attributes and $T = \{t_1, t_2, \ldots, t_N\}$ a set of objects where each object is characterized by a subset of all attributes $I$, i.e., $t_i \subseteq I$. The support count $\sigma$ of a set of attributes $X$ is defined as the number of objects that have all attributes in $X$: 
Given a minimum support count threshold \( s \), FPM finds all sets of attributes ("itemsets" or "patterns") that occur at least \( s \) times ("frequent patterns").

For the purpose of SAR rule mining, objects \( t_i \) represent compounds tested in a profiling experiment. \( I \) then is a set of chemical and biological attributes. Each compound \( t_i \) is represented by a specific subset of chemical and biological attributes – the substructures it contains and the biological response parameters it affects (Figure 1). A screening library usually contains only a few compounds with both similar structure and similar biological activity. We therefore used a support count threshold of \( s = 3 \), identifying all patterns that occur at least three times.

From these patterns, ARM derives rules of the form \( X \rightarrow Y \), indicating that the occurrence of attribute set \( X \) (left-hand side, LHS) is associated with the occurrence of attribute set \( Y \) (right-hand side, RHS). We limited chemical attributes to the LHS and biological attributes to the RHS to obtain SAR rules of the form \( \{ \text{chemical fragments} \} \rightarrow \{ \text{biological pattern} \} \). To reduce redundancy, we only consider rules that consist of closed patterns, i.e., sets of attributes for which no superset exists that has the same support count.

Three measures of rule quality or "interestingness" are then used to filter and prioritize rules. First, the confidence of a rule is defined as the fraction of objects with LHS attributes \( X \) for which the rule is true. In other words, the confidence quantifies how likely it is to observe the biological pattern \( Y \) when the chemical attributes \( X \) are given:

\[
\text{confidence}(X \rightarrow Y) = \frac{\sigma(X \rightarrow Y)}{\sigma(X)}
\]
Second, we defined the purity of a rule as the reverse measure, i.e., the fraction of RHS attributes Y for which the rule is true, which gives a measure of how characteristic a biological pattern is for a certain chemotype:

\[
purity(X \rightarrow Y) = \frac{\sigma(X \rightarrow Y)}{\sigma(Y)} \]

Third, we calculated a p-value for each rule using Fisher's exact test. Low p-values indicate that chemical attributes X and biological pattern Y are unlikely to be independent of each other.

We only kept rules with p-values smaller than 0.01. To rank the remaining rules by their quality, we combined confidence, purity, and p-value into a single SAR score ranging from 0 to 1 (see Methods). A weighting factor can be applied to each individual quality measure to balance their effect on the score. We lowered the influence of the purity with a weighting factor of 0.2 because we considered a high purity less critical for a good rule. This avoids exaggerated down-rating of cases where multiple different structural classes share a biological activity pattern. However, we considered rare and chemotype-specific bioactivities more interesting than common ones and therefore included purity in the score. The set of rules usually contains groups of related rules. We created a “minimal rule set” by removing all rules that consist of attribute subsets or supersets of a rule with a higher SAR score. Unless stated otherwise, all reported rule ranks refer to this minimal set.

We tested our method on GE and MC profiles for two distinct compound libraries. In the following sections, we discuss examples of rules obtained with different chemical structural and biological attributes.
ARM with custom substructure descriptors on planned synthetic libraries leads to intuitive and testable SAR hypotheses

SAR analyses take advantage of structural similarity between compounds. While analog series in advanced projects are tailored to meet this requirement, many screening libraries do not. Early-stage SAR exploration is hence especially useful for planned screening libraries that were designed to include groups of compounds with defined structural relationships.\(^4\)

We therefore included 19,637 compounds in our experiment that were derived from diversity-oriented synthesis (DOS).\(^20\) The compounds represent systematic combinations of three basic structural elements: core structures, appendages, and stereochemistry. Compounds were built by modular synthesis, attaching different appendages to 23 chiral core structures and systematically varying the configuration of core stereocenters.\(^21\) The resulting collection is chemically diverse but, at the same time, contains related structures: groups of compounds that share common cores, stereochemistry, or appendages can be directly compared in SAR analysis. Importantly, this modular design also allows quickly synthesizing structural analogs of identified hits for follow-up studies.

We used as chemical attributes a custom structural fingerprint ("DOS-FP") that reflects the modular design of DOS compounds to obtain rules that directly relate to compound synthesis. Based on the three DOS diversity elements, DOS-FP specifies for each compound the (a) core structure, (b) appendages attached to the core (as a simple occurrence or as “R-groups” that include information about their attachment position in the core), and (c)
configuration of core stereocenters. Stereoisomers of appendages were treated as distinct structures. Note that R-group and stereochemistry features refer to specific positions in the core and can hence only occur in combination with a core feature.

We added a reference set of 2,222 compounds with known bioactivities (BIO) to the DOS library to support the generation of hypotheses about the biological effects of novel DOS compounds. The BIO collection was assembled to contain structurally diverse compounds that cover a wide range of biological activities and often have known direct targets. We measured both GE and MC profiles for most compounds. Exact compound numbers measured with each method differ slightly due to quality-control filters on the experimental data (GE: 17,553 DOS compounds + 1,935 known bioactives; MC: 17,805 DOS compounds + 2,211 known bioactives).

We first aimed to identify SARs on global phenotypic effects. We hierarchically clustered GE and MC profiles to find groups of compounds with similar biological activities. Each node of the resulting dendrogram was considered one cluster. Each compound could therefore be a member of multiple clusters (characterized by superset/subset relationships). Importantly, both DOS and BIO compounds were clustered, but only DOS compounds were subject to ARM for SAR rule mining. While BIO compounds cannot be represented with DOS-FP, this allowed us to evaluate structural rules for DOS compounds in the context of biologically annotated reference molecules.

We found a total of 6,206 rules for GE and 7,098 rules for MC (2,752 and 2,861 for minimal rule sets, respectively). About 70% of the rules for both GE and MC were combinations of core and stereo features and another 17% specified core features alone (Table 1). This
distribution is likely to reflect the fact that the DOS library contains many more compounds that match on these two features than on appendages. However, rules that include R-group information are about 5 to 6-fold enriched among the top 10% of rules (Table 1), indicating their importance for describing specific small-molecule effects.

The majority of rules that include R-group information specify analog series, i.e., compounds that vary in one R-group (70% for GE and 80% for MC). Hence, these rules lead to directly testable hypothesis that are easily accessible through DOS. Analogously, 11% (GE) and 14% (MC) of rules with R-group information specify all R-groups and hence apply to series of stereoisomers. The remaining fraction consists of rules that allow two of three R-groups to vary. Overall, chemical feature combinations occur with highly similar frequencies in the rules for both profiling experiments (Table 1). These frequencies did not change appreciably when only profiles with very good replicate agreement were used (see Methods).

We next examined rules for DOS compounds that co-cluster with known bioactive molecules to find rules with interpretable structural and biological parts. The highest-ranking rule identified a cluster that was strongly enriched for compounds carrying an ortho-aminoanilide (OAA) appendage (Figure 2). 50% of all OAA DOS compounds are contained in this cluster (confidence: 0.5) and hence have similar GE profiles. The OAA motif is known as a biasing element for inhibitory activity against class I histone deacetylases (HDACs) due to its ability to interact with zinc and active site resides chains required for hydrolysis of the acetyl moiety on lysine side chains.27 Indeed, the DOS compounds co-cluster with two known HDAC inhibitors that also contain an OAA residue, mocetinostat (MGCD0103) and tacedinaline (CI-994). Furthermore, the only DOS cluster
member that does not match the structural rule, BRD4805, has recently been identified as a low micromolar HDAC inhibitor.\textsuperscript{28}

Mocetinostat, tacedinaline, and BRD4805 exhibit selectivity for HDAC1, 2, and 3 over other isoforms.\textsuperscript{28,29} Interestingly, less selective inhibitors\textsuperscript{29} contained in our compound library – trichostatin A (TSA), vorinostat (SAHA), and belinostat (PXD-101) – formed a separate cluster, indicating that GE profiling can distinguish compounds with distinct HDAC isoform selectivity patterns. We therefore hypothesize that the clustered OAA DOS compounds show some degree of specificity for HDACs 1-3. Supporting this hypothesis, stereoisomers of structures 1, 2, and 4 have recently been identified as HDAC1-3 selective inhibitors.\textsuperscript{30} Members of this structural class shared the highest profile similarity with the known HDAC inhibitors in our experiment (Table S1).

Several of the OAA DOS compounds have further been tested in abiochemical assay measuring inhibition of HDAC2 deacetylase activity (Figure S2). 90\% (9 of 10) of the OAA DOS compounds that were in the cluster were hits in this screen. By contrast, OAA DOS compounds that were not part of the profile-based cluster achieved a lower hit rate (57\%, 4 of 7). Beyond this enrichment of hits in the cluster, we found no apparent correlation between the profile similarities of DOS compounds to known HDAC inhibitors and their performance in the HDAC2 assay. While HDAC2 serves as a representative class I HDAC, the discrepancies between profile similarity to class I HDAC inhibitors and degree of HDAC2 inhibition may reflect the fact that ortho-aminoanilide-based HDAC inhibitors also target HDAC1 and HDAC3 and/or other in vivo cellular factors that may be influencing the gene-expression signatures.
Taking SAR information into account can increase our confidence in profiling assay results. Only 6 out of the 14 OAA DOS compounds in the cluster had sufficient replicate agreement to be considered ‘hits’ in the GE assay, *i.e.* compounds that induce a reproducible gene-expression change (see Methods). Relying on replicate agreement as an activity criterion alone would hence disregard the remaining 8 compounds, and thus 55% of the HDAC2 hits in the cluster (5 of 9). This result provides an example of the overall utility of our method for guiding the selection of initial hits for follow-up biological studies and chemical synthesis.

The OAA rule is an example of a single structural motif whose presence is sufficient to elicit a relatively defined bioactivity. Although in principle one could have predicted such compounds to have HDAC inhibitory activity by structure alone, these results help validate the method for prediction of less well-known features. More commonly, combinations of features determine a compound’s biological effect. One such example is a set of related DOS compounds that co-cluster with known microtubule destabilizers (Figure 3). The corresponding rule specifies the core, one of two R-groups, and, importantly, all stereocenters for this compound class, effectively describing a series of analog structures. Stereochemistry is a major determinant of specificity for this compound series. Omitting stereochemistry from the rule causes the confidence to drop from 0.5 (3 rule matches out of 6 structure matches) to 0.065 (3 out of 46). The validity of this rule is further supported by the fact that it occurs at the top of both GE (rank 6) and MC (rank 12) rule lists.

Microtubule inhibitors co-cluster with the DOS compounds in both the GE and MC rule (Figure 3). Unlike the OAA case, the known bioactive compounds are not structurally related to the DOS compounds, making this an interesting rule to study for lead-hopping
purposes. Using structural features that directly match physical DOS building blocks allows rapidly testing such hypotheses as analogs and stereoisomers can be synthesized quickly (if they do not already exist).

**General structure descriptors allow rule generation for arbitrary compound collections**

Not all libraries are designed to include compounds with defined structural relationships like the DOS collection. However, arbitrary libraries can be mined by our method by replacing the chemical structural descriptors. We used this approach to analyze MC profile data for a collection of 10,226 compounds from the Molecular Libraries Small Molecule Repository (MLSMR). We used extended connectivity fingerprints (ECFP4) as a generic structure descriptor that can be used with any compound collection. While any binary fingerprint can readily be used with our method, we chose ECFPs as a reference for this study because they are widely applied, allow direct mapping of features to structural fragments, and were specifically designed for SAR studies. They achieve high structural resolution by encoding a very large number of features and thus present a good test for the ability of ARM to mine large feature spaces.

We used clusters based on MC profiles as biological attributes. Due to the versatility of ECFP4, we were able to include our set of known bioactive compounds in the rule-mining step. We found 14,198 rules (7,328 after redundancy filtering) of which 2,176 contained known bioactive compounds. Unlike the DOS-FP features, ECFP4 substructures can describe arbitrary parts of the molecule through combinations of overlapping
substructures (Figure 4). This description allows a more fine-grained and unbiased mapping of relevant substructures. The downside is a reduction of interpretability as rules cannot be directly mapped to existing analogs or straightforward synthesis strategies. In addition, a large number of features are usually needed to achieve high resolution and generality. Even though ARM can handle such large data sets, the number of rules can increase as a result and complicate downstream analyses. Therefore, each compound collection and project requires balancing generality, complexity, and interpretability when choosing a chemical descriptor.

Direct mining of attributes identifies biological signatures for sets of related compounds

Clustering biological profiles captures global similarities and differences between the biological effects of compounds. However, information on individual features (e.g., genes from a GE profile) is lost in this process. Since ARM is designed to operate on high-dimensional data, it can be used to mine biological features without intermediate aggregation. Useful applications include mining compound-target profiles or GE signatures, i.e., sets of genes that change expression levels in response to compound treatment.

We generated a signature of up- and down-regulated genes for each compound by applying a z-score threshold of 2 and -2 to GE profiles, respectively (relative to DMSO-treated controls). The signatures were combined with DOS-FP features and subjected to ARM. The top-scoring rule is again the occurrence of an OAA residue already found with GE clusters as biological attributes (confidence=0.3, purity=1, p=1.68×10^{-26}, SAR score=0.98).
associated GE signature consists of only three genes: CASP3:down, IER3:up, TERF2IP:down. This relatively low number of common features leads to important differences in specificity compared to the analogous clustering-based rule. In addition to mocetinostat and tacedinaline, compounds that do not target HDACs are also found to regulate these genes (15D-prostaglandin J2, withaferin A, thioridazine, and aminopurvalanol A). However, more specific rules can be found that achieve only slightly lower SAR scores due to their reduced confidence. For example, the signature CASP3:down, IER3:up, TBPL1:up, TERF2IP:down is only found in benzamide-carrying DOS compounds and mocetinostat and tacedinaline (SAR score 0.95, rank 4 in the full set of rules). The second-ranked rule is highly specific for a small set of compounds with a signature of 13 up-regulated genes (Figure 5). A functional analysis strongly related the 13-member gene set to sterol and cholesterol synthesis (Table S2). Interestingly, sirolimus and fulvestrant are also found to change the expression of these genes. Irrespective of their distinct mechanisms of action, each has been linked to cholesterol levels and metabolism. A corresponding rule found by clustering full-length profiles as described above achieved high ranks for both GE (rank 8) and MC (rank 11). However, neither of the known bioactive compounds co-clusters with the DOS compounds in both assays. This finding supports the view that gene-signatures capture subsets of the global biological effects of perturbagens. We therefore hypothesize that the DOS compounds identified by this rule have related biological effects connected to cholesterol metabolism. This example illustrates how the combination of profiling and automated SAR analysis can aid data mining and quickly lead to testable hypothesis.
Discussion

Small-molecule profiling datasets are promising resources for the elucidation of SARs with high granularity. We have illustrated in several examples that association-rule mining can identify interpretable SAR rules from large sets of high-dimensional profiling data.

Our method is designed to support compound selection and synthesis decisions in early-stage drug- and probe-discovery projects. The data in such projects often consist of high-throughput measurements for large compound libraries. Such data are expected to be noisy and contain only small groups of related compounds with similar biological effects. Acknowledging these limitations, we chose a data-mining approach to analyzing such data. Rather than attempting automated classification, our method supports hypothesis generation and human decision making.

These goals further led us to choose interpretable structural fragments as chemical features to obtain rules that can directly inform synthesis decisions. However, our method is not limited to structural descriptors. ARM can be used with any descriptor that can be expressed through binary attributes. Categorical, ordinal, or continuous attributes like physicochemical properties or calculated descriptors can usually be represented by binning and proper encoding without too much loss of information. In any case, a balance between generality and interpretability of chemical attributes needs to be found. Custom synthesis-oriented representations like DOS-FP are highly interpretable but not applicable to all libraries. By contrast, generic fingerprints like ECFP4 rely on more abstract features that usually do not directly relate to synthetically accessible building blocks.
Similar considerations apply to the biological descriptors. First, different experimental methods will likely capture and emphasize different biological effects. We found that, even though several top-scoring rules were identified by both GE and MC, the particular composition of compound groups often differed. However, a comprehensive comparison between such methods has not yet been performed. Second, the computational representation of biological profiles and measures for their comparison will influence the quality of the resulting rules. Global similarity measures like correlation to cluster profiles before SAR mining can be more suitable for compounds that elicit subtle effects across a wide range of features. By contrast, if a phenotype results from strong effects on only a few features, signature-based approaches, *i.e.*, methods that use lists of the most strongly altered features for a perturbation, are likely to perform better than whole-profile similarity methods. The generality of the ARM approach allows using arbitrary combinations of chemical and biological features, and comparing different feature choices can greatly increase the confidence in individual results.

Finally, the scoring function can be tailored toward individual needs. A number of objective measures of interestingness have been suggested that can be used with any attributes. More domain-specific measures like intra-cluster distances, for example, can easily be integrated into the score as addends. Weighting individual features gives control over their relative importance to each other and allows different rule types to be prioritized. For example, putting high weight on the purity of a rule would lead to preferential selection of rare bioactivity patterns over common ones. We believe that these features make ARM an efficient and versatile approach for automated SAR mining of biological profiling data.
Acknowledgements

Cheminformatics and data-mining work was supported by National Institute of General Medical Sciences (P50-GM069721 awarded to S.L.S.), as part of the Center of Excellence for Chemical Methodology and Library Development. Profiling measurements were supported as part of the NIH RoadMap Molecular Libraries Initiative (U54-HG005032 awarded to S.L.S.). Associated data can be accessed at http://www.broadinstitute.org/scientific-community/science/programs/csoft/therapeutics-platform/mlpcn/accessing-mlpcn-data. D.M.F. and S.J.H. were supported through funding from the NIH (R01DA028301) and the Stanley Medical Research Institute. S.L.S. is an investigator at the Howard Hughes Medical Institute.

References


Tables

Table 1. GE and MC show comparable distributions of rule types. Shown are absolute and relative frequencies for rules that involve different combinations of chemical structural features. Combinations that are theoretically impossible to occur are omitted. rgroup:

appendages with attachment position to core tracked; app: simple occurrence of appendages irrespective of position

<table>
<thead>
<tr>
<th>structural features in rule</th>
<th>freq. – all rules</th>
<th>freq. – top 10% of rules</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GE</td>
<td>GE[%]</td>
</tr>
<tr>
<td>core stereo rgroup app</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>core rgroup app</td>
<td>1</td>
<td>0.02</td>
</tr>
<tr>
<td>core stereo app</td>
<td>22</td>
<td>0.35</td>
</tr>
<tr>
<td>core app</td>
<td>9</td>
<td>0.15</td>
</tr>
<tr>
<td>core stereo rgroup app</td>
<td>472</td>
<td>7.61</td>
</tr>
<tr>
<td>core rgroup</td>
<td>251</td>
<td>4.04</td>
</tr>
<tr>
<td>core stereo</td>
<td>140</td>
<td>2.26</td>
</tr>
<tr>
<td>core</td>
<td>4252</td>
<td>68.51</td>
</tr>
<tr>
<td></td>
<td>1059</td>
<td>17.06</td>
</tr>
<tr>
<td></td>
<td>6206</td>
<td>100</td>
</tr>
</tbody>
</table>
Figure Legends

Figure 1. ARM identifies and scores rules that relate chemical features of compounds to biological profile attributes. In this schematic example, three out of five compounds match a rule that relates the occurrence of two structural features to the up-regulation of a gene (CASP9). Confidence and purity measure the overlap between compounds that match the structural rule and compounds that elicit the biological phenotype. This information is used to quantify the quality of a rule.

Figure 2. The best-scoring rule for GE profiles identified a known biasing element for HDAC inhibitory activity. The quality measures for this rule indicate that compounds with an ortho-aminoanilide (OAA) residue are strongly enriched in cluster A. All compounds in cluster A are shown. The cluster contains two known OAA HDAC inhibitors (6 and 7) and all DOS compounds except one carry the OAA residue. The remaining DOS compound BRD4805 has recently been identified as an HDAC inhibitor. For simplicity, stereoisomers are shown as combinations of generic structures and configuration tables.

Figure 3. Both GE and MC identify the same rule for DOS compounds that co-cluster with microtubule inhibitors. A series of three DOS analogs was identified based on both GE and MC profile clusters, yielding high-scoring rules in both cases. Interestingly, the rule is highly stereo-specific, constraining all four stereocenters in these compounds to one configuration. Known microtubule inhibitors were present in both clusters, though the exact set of bioactives differed. For clarity of presentation, structures for bioactive compounds are omitted (see Figure S1 for detailed structure information).
Figure 4. ECFP4 features map arbitrary parts of molecules with overlapping substructures. The highest-scoring ECFP4-based rule that involves known-bioactive compounds describes a set of structurally and functionally related kinase inhibitors. While ECFP4 fragments correctly identify their overlapping substructures, feature redundancy and abstractness makes them harder to interpret than synthesis-oriented substructures like DOS-FP (Figure 2 and 3).

Figure 5. ARM can identify biological signatures for groups of structurally related compounds. A set of DOS stereoisomers was found to regulate 13 genes enriched for functions in cholesterol metabolism. The same genes are regulated as part of the signatures of sirolimus and fulvestrant, two mechanistically unrelated compounds that have both been linked to cholesterol metabolism.³¹,³²
Figures

Figure 1

![Chemical structures and SAR rule](image1)

\[
\begin{align*}
\text{chem}(\text{rule}) & = 3 \\
\text{chem}(\text{bio}) & = 4 \\
\text{CASPS}(\text{up}) & = 4 \\
\end{align*}
\]

Figure 2

**SAR RULE**

confidence: 0.5
purity: 0.93
p-value: 2.2x10^-4

![Cluster A](image2)

**COMPOUNDS IN CLUSTER A**

14 DOS rule matches

2 known biactive compounds

6 | mocetinostat
7 | tcedinaline

1 other DOS compound

8 | BRD-4805
**Figure 3**

**SAR RULI**

<table>
<thead>
<tr>
<th>GE</th>
<th>MC</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>0.6</td>
<td>1</td>
</tr>
<tr>
<td>$2.3 \times 10^{-10}$</td>
<td>$2.1 \times 10^{-11}$</td>
</tr>
</tbody>
</table>

**COMPOUNDS IN CLUSTERS B\textsubscript{GE} AND/OR B\textsubscript{MC}**

3 DOS rule matches (B\textsubscript{GE} and B\textsubscript{MC})

14 known bioactive compounds (B\textsubscript{GE})

- colchicine
- nocodazole
- KF 38789
- vinblastine
- parbendazole
- LY-2183240
- vincristine
- beta-peltatin
- ceforanide
- mebendazole
- podophyllotoxin
- MK2a inhibitor
- methazole
- D-64131

2 other DOS compounds (B\textsubscript{GE})

- 12
- 13

4 known bioactive compounds (B\textsubscript{MC})

- albendazole
- colchicine
- D-64131
- nocodazole

**Figure 4**

**SAR RULI**

- confidence: 1
- purity: 0.75
- p-value: $8.5 \times 10^{-43}$

**COMPOUNDS IN CLUSTER C**

3 rule matches

<table>
<thead>
<tr>
<th>14</th>
<th>PP1</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>PP2</td>
</tr>
<tr>
<td>16</td>
<td></td>
</tr>
</tbody>
</table>

1 other MLSMR compound

| 17 |
**Figure 5**

**SAR RULE**

| Confidence: 0.75 | Purity: 1 | P-value: 7.2 x 10^{-22} |

**Gene Signature**

- ABH4
- BAG3
- DDIT
- DNAJ1
- EBP
- SLC2A6
- HERPUD1
- TMEV
- XBP1

**Compounds with Matching Gene Signature**

**6 Rule Matches**

- **18**

**19** sirolimus

**20** fulvestrant

2 known bioactive compounds
SUPPLEMENTARY INFORMATION

Supplementary Methods

A. Detailed references for DOS compound synthesis

We selected the DOS compound collection (19,637 compounds) as a structurally diverse subset representing 23 DOS libraries. The compounds in each library are based on the same chiral core; structural diversity was achieved by varying side chains and the configurations of core stereocenters. For most compounds, all stereoisomers were synthesized.

B. Experimental procedures

Gene-expression assay: cell plating and compound treatment

1. U-2 OS cells (ATCC, cat. no. HTB-96)
2. 384 well plates (Corning, cat. no. 3712)
3. culture medium
   - DMEM (Fisher Scientific, cat.no. MT10017CV)
   - 10%FBS (Life technologies, cat.no. 10437028)
   - 1% penicillin-streptomycin (Fisher Scientific, cat. no. MT30002CI)
4. TCL buffer (Quiagen, cat. no. 1031576)
5. cold-storage adhesive sealing foil (VWR, cat. no. 89049-034)

U-2 OS cells were plated in 384-well plates (3500 cells/well in 50µL culture medium). Cells were incubated at 37°C. After 24h, compounds were added, followed by another 6h of incubation at 37°C. 40µL of medium were removed. The cells were lysed by adding 30µL TCL buffer. Plates were sealed and incubated for 30min at room temperature (RT). The plates were subsequently stored at -80°C.

Gene-expression: sample preparation and measurement

We followed the protocol by Peck et al. The main differences to the published protocol were:

1. Luminex microspheres: we used MagPlex instead of COOH
2. bead coupling preparation: we used a different volume of beads and wash solutions
3. probe hybridization: we introduced a ramp-down of annealing temperatures
4. polymerase: we used HotStarTaq Plus instead of HotStarTaq
5. PCR cycling conditions were different (29 cycles, 1min step times)
6. to measure all transcripts, we used 2 different bead mixes per sample during detection; furthermore, a different volume of each mix was used
7. bead/amplicon hybridization time: increased from 1h to 16h-20h
8. we introduced bead washes before and after streptavidin-phycoerythrin addition during detection, including new wash solutions
9. Luminex detection instrument: we used FlexMap 3D instead of Luminex 100
**General Notes**

All liquid transfer steps were performed on the Agilent Bravo Liquid Handling Platform. Unreacted products were removed after each step (inverting reaction plates onto laboratory towel and centrifuging for 1 min at 1000g). Non-room-temperature incubations were performed using a Thermo Electron MBS 384 Satellite Thermal Cycler.

**Luminex microsphere (bead) preparation**

Materials:

1. Luminex xMAP MagPlex microspheres
2. bead binding buffer
   - 0.1M 2-(N-morpholino)ethansulfonic acid (pH 4.5)
3. EDC solution
   - 10mg/mL 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide hydrochloride in water.
4. 0.02% Tween
5. 0.1% sodium dodecyl sulfate
6. TE (pH8)
   - 10 mM Tris-HCl (pH 8.0)
   - 1 mM EDTA

We first coupled anti-barcode capture oligonucleotides to Luminex xMAP MagPlex microspheres. For each of 490 distinct microsphere species, approximately $12.5 \times 10^6$ microspheres per well were transferred to 800μL round-bottom deep well 96 plates. After centrifugation and magnetic separation, storage buffer was removed, and the pulled-down beads re-suspended in 130μL binding buffer. 500pmol of each of 490 capture probes were then added to each well, i.e. each microsphere species was paired with a different probe. After adding 5μL of freshly prepared EDC solution, plates were incubated for 30 min to perform the coupling reactions. This step was repeated a second time. Microspheres were then captured by magnetic pull-down. Three successive washing steps followed with 500μL of 1) 0.02% Tween, 2) 0.1% sodium dodecyl sulfate, and 3) TE (pH 8). Coupled microspheres were re-suspended in TE (pH 8) to 50,000 microspheres/μL and pooled. This pool ("dp52") formed the first coupled microsphere set.

This process was repeated with a different set of capture oligonucleotides (coupled to the same 490 microsphere species), yielding the second coupled microsphere set “dp53”.

Finally, 80 distinct oligonucleotides were coupled to another 10 microsphere species such that each microsphere bound 8 oligonucleotides. These 10 sets of 8 oligonucleotides were used to measure 10 invariant “meta-genes” as normalization references.

**Ligation Mediated Amplification**

*RNA extraction*
Materials:

1. Turbocapture 384 mRNA kit (Qiagen, cat. no. 72271)

Frozen cell lysates were thawed at room temperature for 1h. 20μL of lysate per well were transferred to a 384-well oligo-dT capture plate. The capture plates were incubated for 1h at room temperature to immobilize mRNA to the plate via binding of the poly-A tail.

**cDNA Generation**

Materials:

1. M-MLV Reverse Transcriptase kit (Promega, cat. no. M1705)
2. 100mM dNTP set (Invitrogen, cat. no. 10297018)

The immobilized mRNA was reverse-transcribed into cDNA by adding 5μL M-MLV reaction mix to the capture plate and incubating for 1.5 h at 37°C.

**Probe Hybridization**

1. pool of custom up- and down-stream probes
2. Taq DNA Ligase Reaction Buffer (NEB, cat. no. B0208S), used to prepare probe-pool working dilution

Sets of custom up- and down-stream probes were then hybridized to the cDNA. The upstream probes contained a 20nt gene-specific sequence, a 24nt FlexMap barcode, and a T7 primer site. Each downstream probe contained a 20nt gene-specific sequence that hybridized adjacent to the upstream probe and a T3 primer site. Probes were hybridized to the cDNA by a 5min denaturation at 95°C followed by a ramp-down from 70°C to 40°C, 12min per degree, to anneal the probes. The plates were kept at 4°C overnight.

**Probe Ligation**

Materials:

Taq DNA ligase kit (NEB, cat. no. M0208L)

Upstream and downstream probes adjacently bound to the same cDNA were ligated by adding 5μL Taq DNA ligase to each well and incubating at 45°C for 1h, followed by 65°C for 10min.
**PCR Amplification**

**Materials:**

1. HotStarTaq Plus DNA polymerase kit (Qiagen, cat. no. 203603)
2. 100mM dNTP set (Invitrogen, cat. no. 10297018)

Ligated probes were amplified via PCR using T3 and biotin-T7 universal primers. After adding 15μL HotStarTaq Plus reaction mix each well of the plate, the following PCR program performed:

   a. initial denature: 92°C/15min
   b. 29 amplification cycles: 92°C/1min – 60°C/1min – 72°C/1min
   c. final extension: 72°C/5min

**Hybridization and Detection**

**Microsphere Hybridization**

**Materials:**

1. dp52 coupled bead set
2. dp53 coupled bead set
3. invariant-gene bead set
4. 1.5x TMAC hybridization solution
   - 4.5 M tetramethylammonium-chloride
   - 0.15% N-lauryl sarcosine
   - 75mM Tris-HCl (pH 8)
   - 6mM EDTA

Each of the two previously prepared microsphere sets dp52 and dp53 was diluted in 1.5x TMAC hybridization solution to obtain a final concentration of about 200 of each microsphere species per well. Invariant-gene microspheres were added to each of the two bead mixes to a similar concentration. The two dilutions were transferred to 384-well plates at 23μL per well. Amplified LMA product was then added to the prepared dp52 and dp53 plates at 5μL per well. Samples were distributed on detection plates such that pairs of dp52 and dp53 wells for the same sample were found on the same plate. Each resulting detection plate therefore contained 192 samples assayed on both bead sets (dp52: left half, dp53: right half of the plate). The detection plates were denatured at 95°C for 2min, and incubated at 45°C for 16h-20h overnight to hybridize the amplified probes to the microspheres.
**Microsphere Detection**

**Materials:**

1. 1x TMAC hybridization solution
   - 3M tetramethylammonium-chloride
   - 0.1% N-lauryl sarcosine
   - 50mM Tris-HCl (pH 8)
   - 4mM EDTA
2. low-stringency wash buffer
   - 6x SSPE
   - 0.01% Tween-20
3. high-stringency wash buffer
   - 0.1x MES
   - 25mM NaCl
   - 0.01% Tween-20
4. reporter mix
   - 3% streptavidin-phycoerythrin (Invitrogen, cat. no. S866) in 1X TMAC hybridization solution

Microspheres were pulled down by centrifugation at 1000 rpm for 1min and magnetic pull-down. Two successive washing steps followed with 1) low-stringency wash buffer and 2) high-stringency wash buffer. 10μL of reporter mix were added to each well and samples incubated at 45°C for 10min to allow streptavidin-phycoerythrin to bind the biotinylated amplicons. The plates were then centrifuged and microspheres magnetically pulled down. Samples were washed once with low-stringency wash buffer and 3 times with 1X TMAC wash solution. The samples were analyzed using the Luminex FlexMap 3D detector to identify microsphere species and measure phycoerythrin levels.

**Gene-expression assay: data processing**

Expression values for each measured transcript in a sample were calculated using a peak detection algorithm on raw fluorescence intensity curves. The binary logarithms of expression values were then normalized based on the 10 “invariant meta-genes”. Each meta-gene is a composite of 8 genes that were pre-selected to have similar and relatively stable expression levels (coefficient of variation < 10%) across a large number of distinct reference samples. The 10 meta-genes span a wide range of expression levels. A calibration curve was computed for each sample using the expression levels of these invariant genes. Sample values were then rescaled based on reference curves computed from a large collection of expression profiles. The value range was subsequently limited to [0, 15] ([http://lincscloud.org/the-data-set/how-data-were-prepared/](http://lincscloud.org/the-data-set/how-data-were-prepared/)). Detailed descriptions of the data-collection and data-processing pipeline will be published separately by the NIH LINCS project, and are summarized online ([http://lincscloud.org/](http://lincscloud.org/)).
Gene-expression array: data normalization and correction

The distributions of gene-expression levels for each well on a plate were quantile normalized. For each transcript, plate medians were subtracted from each well profile. Positional (well) effects for each gene were corrected using GeneData Screener Assay Analyzer. Robust Z-scores were calculated by dividing the resulting values by 1.4826 * plate median absolute deviation (MAD). Finally, we combined replicates into per-sample profiles using Stouffer’s Z-score.

Multiplex cytological imaging assay: materials

1. U-2 OS cells (ATCC, cat. no. HTB-96)
2. Aurora 384-well black/clear bottom plates, imaging quality (Brooks, cat. no. 1022-11330)
3. culture medium
   - DMEM (Fisher Scientific, cat.no. MT10017CV)
   - 10%FBS (Life Technologies, cat.no. 10437028)
   - 1% penicillin-streptomycin (Fisher Scientific, cat. no. MT30002Cl)
4. Mitotracker Deep Red (Invitrogen, cat. no. M22426)
5. wheat germ agglutinin Alexa 594 conjugate (Invitrogen, cat.no. W11262)
6. paraformaldehyde 16%, methanol free (Electron Microscopy Sciences, cat.no. 15710-S)
7. Hank’s Balanced Salt Solution, HBSS (Invitrogen, cat. no. 14065-056)
8. Triton X-100 (Sigma, cat. no. T8787)
9. phalloidin 594 (Invitrogen, cat. no. A12381)
10. concavalin A 488 (Invitrogen, cat. no. C11252)
11. Hoechst 33342 (Invitrogen, cat. no. H3570)
12. SYTO 14 green fluorescent nucleic acid stain (Invitrogen, cat.no. S7576)
13. sodium biocarbonate (HyClone, cat. no. SH30033.01)
14. methanol (BDH, cat. no. 67-56-1)
15. bovine serum albumin
16. REMP blue thermo heat seal (REMP/Nexus Biosystems, cat. no. 1800336)
17. ImageXpress Micro (Molecular Devices)

Multiplex cytological imaging assay: assay protocol

1500-2000 U-2 OS cells with 50μL culture medium per well were plated on 384-well clear bottom plates and incubated for 24 at 37°C. Compounds were added and the samples incubated for another 48h at 37°C. A staining solution was prepared in pre-warmed medium with a concentration of 500nM Mitotracker and 60μg/mL WGA. 40μL of media were aspirated and 30μL of the staining solution added to each well. After 30min of incubation at 37°C, cells were fixed for 20min at RT with 10μL paraformaldehyde. Cells were washed once with 70μL HBSS and permeabilized with 30μL of a 0.1% solution of Triton X-100 in 1x HBSS for 10-20min. Two washing steps with 70μL 1xHBSS followed. A 1mg/mL solution of Concanavalin A was prepared in 0.1M sodium bicarbonate solution. Phalloidin was dissolved in 1.5 mL methanol per vial. A staining mix was prepared from 0.025 μL phalloidin/μL, 100 μg/mL Concanavalin, 5 μg/mL Hoechst, and 3μM SYTO staining solution in 1x HBSS 1% BSA. 30μL of the staining mix were added to each well and incubated for
30min. After staining, cells were washed three times with 70 μL 1×HBSS (no final aspiration). Plates were thermally sealed at 171°C for 4 seconds.

**Multiplex cytological imaging assay: image capture**

All images were automatically captured using an ImageXpress Micro epifluorescent microscope. We recorded 9 sites per well at 20x magnification in 5 fluorescent channels, DAPI (387/447 nm), GFP (472/520 nm), Cy3 (531/593 nm), TexasRed (562/642 nm), Cy5 (628/692 nm). A laser-based auto focus was performed on the first site of each well in the DAPI channel.

**Multiplex cytological imaging assay: image analysis and data processing**

We segmented cells and measured morphological profiles for each individual cell using the CellProfiler\textsuperscript{13} software version 2.0.9925. The pipelines used to correct for uneven illumination and segment cells into nuclei and cytoplasm are described and provided by Gustafsdottir et al.\textsuperscript{14} Size, shape, texture, intensity statistics, and local density were measured for nuclei, cytoplasm, and entire cells.\textsuperscript{14} Each feature was normalized per plate by linearly scaling the 1st and 99th percentiles of the cell-level value distributions of DMSO control wells to 0 and 1, respectively. Plate medians for each feature were subtracted from each sample on a plate and positional effects corrected with GeneData Screener Assay Analyzer.\textsuperscript{11}
Supplementary Figures

Figure S1. Structures for microtubule inhibitors and other bioactive compounds from Figure 3.

- Mebendazole
- Beta-peltatin
- Vincristine
- Nocodazole
- Podophyllotoxin
- Vinblastine
- Methiazole
- Colchicine
- Albendazole
- MK2a inhibitor
- Parbendazole
- LY-2183240
- D-64131
- KF 38789
- Cefoxanide
Figure S2. HDAC2 screening results for ortho-aminoanilide DOS compounds.

Using a trypsin-coupled, HDAC2 kinetic assay, two-point kinetic data (n=2) were collected for 17 ortho-aminoanilide-containing DOS compounds at 10 min and 30 min time points with compounds tested at ~1.6 μM. All data was normalized to the % inhibition relative to 8.3 μM Cpd-60, a benzamide-containing control compound. A range of potencies was observed for the 17 compounds tested. Hits were considered to be those compounds with % inhibition greater than 25%, but only those with 60% inhibition or greater were selected for follow-up dose responses. Only compound 32 was chosen for follow-up validation and was confirmed as an HDAC2 inhibitor and to not effect trypsin itself in a counter screen for false positive (data not shown). Bar labels shown in bold indicate numbers of compounds that are part of cluster A (see Figure 2 and Table S1).
Supplementary Tables

Table S1. Several stereoisomers of DOS compounds with ortho-aminoanilide and tertiary amide substituents show the highest profile correlations to HDAC inhibitors (HDACi) mocetinostat and tacedinaline. Highlighted compounds represent stereoisomers of recently identified HDAC1-3 inhibitors\textsuperscript{15} that share the highest GE profile similarity with mocetinostat and tacedinaline. Depictions of parent structures are shown in Fig. 2 of the manuscript.

<table>
<thead>
<tr>
<th>compound number</th>
<th>parent structure</th>
<th>stereo ((C_2C_3C_6))</th>
<th>avg. GE correlation to mocetinostat and tacedinaline in cluster A</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>2</td>
<td>SRS</td>
<td>0.57</td>
</tr>
<tr>
<td>22</td>
<td>2</td>
<td>RRS</td>
<td>0.51</td>
</tr>
<tr>
<td>23</td>
<td>1</td>
<td>SRS</td>
<td>0.51</td>
</tr>
<tr>
<td>24</td>
<td>4</td>
<td>RSS</td>
<td>0.44</td>
</tr>
<tr>
<td>25</td>
<td>1</td>
<td>RRS</td>
<td>0.43</td>
</tr>
<tr>
<td>26</td>
<td>2</td>
<td>RSR</td>
<td>0.40</td>
</tr>
<tr>
<td>27</td>
<td>2</td>
<td>RRR</td>
<td>0.37</td>
</tr>
<tr>
<td>28</td>
<td>3</td>
<td>SRS</td>
<td>0.32</td>
</tr>
<tr>
<td>29</td>
<td>2</td>
<td>SSR</td>
<td>0.31</td>
</tr>
<tr>
<td>30</td>
<td>2</td>
<td>SSS</td>
<td>0.30</td>
</tr>
<tr>
<td>31</td>
<td>5</td>
<td>SSR</td>
<td>0.28</td>
</tr>
<tr>
<td>32</td>
<td>5</td>
<td>RSR</td>
<td>0.25</td>
</tr>
<tr>
<td>33</td>
<td>1</td>
<td>SRR</td>
<td>0.24</td>
</tr>
<tr>
<td>34</td>
<td>3</td>
<td>RRR</td>
<td>0.23</td>
</tr>
<tr>
<td>35</td>
<td>2</td>
<td>SRR</td>
<td>0.20</td>
</tr>
<tr>
<td>36</td>
<td>1</td>
<td>SSR</td>
<td>0.19</td>
</tr>
<tr>
<td>37</td>
<td>2</td>
<td>SSR</td>
<td>0.18</td>
</tr>
<tr>
<td>38</td>
<td>5</td>
<td>SRS</td>
<td>0.18</td>
</tr>
<tr>
<td>39</td>
<td>4</td>
<td>RRS</td>
<td>0.18</td>
</tr>
<tr>
<td>40</td>
<td>4</td>
<td>SRS</td>
<td>0.17</td>
</tr>
<tr>
<td>41</td>
<td>1</td>
<td>SRR</td>
<td>0.16</td>
</tr>
<tr>
<td>42</td>
<td>5</td>
<td>RRS</td>
<td>0.13</td>
</tr>
<tr>
<td>43</td>
<td>1</td>
<td>RRR</td>
<td>0.12</td>
</tr>
<tr>
<td>44</td>
<td>3</td>
<td>SSR</td>
<td>0.11</td>
</tr>
<tr>
<td>45</td>
<td>3</td>
<td>RRS</td>
<td>0.10</td>
</tr>
<tr>
<td>46</td>
<td>5</td>
<td>SRR</td>
<td>0.10</td>
</tr>
<tr>
<td>47</td>
<td>3</td>
<td>RSR</td>
<td>0.08</td>
</tr>
</tbody>
</table>
Table S2. Signature mining identifies compounds whose gene signature is enriched for terms associated with sterol/cholesterol synthesis and metabolism. A functional analysis of the 13-member gene signature shown in Fig. 5 of the manuscript was carried out using the DAVID\textsuperscript{16,17} web service. The output of this analysis that identifies a cluster of cholesterol metabolism-related genes is shown below (DAVID enrichment score ES= 5.87). Low false-discovery rate (FDR) values indicate enrichment of genes associated with several cholesterol-related terms.

<table>
<thead>
<tr>
<th>category</th>
<th>term</th>
<th>count</th>
<th>%</th>
<th>p-value</th>
<th>genes</th>
<th>number of input genes</th>
<th>genes that match term</th>
<th>total no. of genes in category</th>
<th>fold enrichment</th>
<th>FDR (Benjamini)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SP_PIR_KEYWORDS</td>
<td>sterol biosynthesis</td>
<td>5</td>
<td>38.46</td>
<td>6.52E-08</td>
<td>EBP, HMGCR, HMGCS1, FDFT1, NSDHL</td>
<td>13</td>
<td>5</td>
<td>977</td>
<td>75.15</td>
<td>2.93E-06</td>
</tr>
<tr>
<td>SP_PIR_KEYWORDS</td>
<td>cholesterol biosynthesis</td>
<td>5</td>
<td>38.46</td>
<td>6.52E-08</td>
<td>EBP, HMGCR, HMGCS1, FDFT1, NSDHL</td>
<td>13</td>
<td>5</td>
<td>977</td>
<td>75.15</td>
<td>2.93E-06</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0006695 cholesterol biosynthetic process</td>
<td>5</td>
<td>38.46</td>
<td>1.28E-07</td>
<td>EBP, HMGCR, HMGCS1, FDFT1, NSDHL</td>
<td>13</td>
<td>5</td>
<td>826</td>
<td>63.54</td>
<td>1.70E-05</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0016126 sterol biosynthetic process</td>
<td>5</td>
<td>38.46</td>
<td>1.28E-07</td>
<td>EBP, HMGCR, HMGCS1, FDFT1, NSDHL</td>
<td>13</td>
<td>5</td>
<td>826</td>
<td>63.54</td>
<td>1.70E-05</td>
</tr>
<tr>
<td>SP_PIR_KEYWORDS</td>
<td>steroid biosynthesis</td>
<td>5</td>
<td>38.46</td>
<td>8.94E-07</td>
<td>EBP, HMGCR, HMGCS1, FDFT1, NSDHL</td>
<td>13</td>
<td>8</td>
<td>977</td>
<td>46.97</td>
<td>2.01E-05</td>
</tr>
<tr>
<td>SP_PIR_KEYWORDS</td>
<td>lipid synthesis</td>
<td>5</td>
<td>38.46</td>
<td>2.65E-06</td>
<td>EBP, HMGCR, HMGCS1, FDFT1, NSDHL</td>
<td>13</td>
<td>10</td>
<td>977</td>
<td>37.58</td>
<td>3.97E-05</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0006694 steroid biosynthetic process</td>
<td>5</td>
<td>38.46</td>
<td>3.11E-06</td>
<td>EBP, HMGCR, HMGCS1, FDFT1, NSDHL</td>
<td>13</td>
<td>9</td>
<td>826</td>
<td>35.30</td>
<td>2.07E-04</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0008203 cholesterol metabolic process</td>
<td>5</td>
<td>38.46</td>
<td>3.11E-06</td>
<td>EBP, HMGCR, HMGCS1, FDFT1, NSDHL</td>
<td>13</td>
<td>9</td>
<td>826</td>
<td>35.30</td>
<td>2.07E-04</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0016125 sterol metabolic process</td>
<td>5</td>
<td>38.46</td>
<td>5.15E-06</td>
<td>EBP, HMGCR, HMGCS1, FDFT1, NSDHL</td>
<td>13</td>
<td>10</td>
<td>826</td>
<td>31.77</td>
<td>2.28E-04</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0008202 steroid metabolic process</td>
<td>5</td>
<td>38.46</td>
<td>4.26E-05</td>
<td>EBP, HMGCR, HMGCS1, FDFT1, NSDHL</td>
<td>13</td>
<td>16</td>
<td>826</td>
<td>19.86</td>
<td>1.41E-03</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0008610 lipid biosynthetic process</td>
<td>5</td>
<td>38.46</td>
<td>7.05E-05</td>
<td>EBP, HMGCR, HMGCS1, FDFT1, NSDHL</td>
<td>13</td>
<td>18</td>
<td>826</td>
<td>17.65</td>
<td>1.87E-03</td>
</tr>
</tbody>
</table>
References