PNAS RESEARCH ARTICLE | PHARMACOLOGY OPEN ACCESS

High-throughput functional annotation of natural products by integrated activity profiling

Suzie K. Hight[a,1,2], Trevor N. Clark[b,1], Kenji L. Kurita[b,3], Elizabeth A. McMillan[a], Walter Bray[c], Anam F. Shaikh[d], Aswad Khadilkar[c], F. P. Jake Haeckl[b], Fausto Carnevale-Neto[b], Scott La[c], Akshar Lohith[c], Rachel M. Vaden[a], Jeon Lee[e], Shuguang Wei[d], R. Scott Lokey[c], Michael A. White[a,4,5], Roger G. Linington[b,5], and John B. MacMillan[c,d,5]

Edited by Benjamin Cravatt, The Scripps Research Institute, La Jolla, CA; received June 19, 2022; accepted October 19, 2022

Determining mechanism of action (MOA) is one of the biggest challenges in natural products discovery. Here, we report a comprehensive platform that uses Similarity Network Fusion (SNF) to improve MOA predictions by integrating data from the cytological profiling high-content imaging platform and the gene expression platform Functional Signature Ontology, and pairs these data with untargeted metabolomics analysis for de novo bioactive compound discovery. The predictive value of the integrative approach was assessed using a library of target-annotated small molecules as benchmarks. Using Kolmogorov–Smirnov (KS) tests to compare in-class to out-of-class similarity, we found that SNF retains the ability to identify significant in-class similarity across a diverse set of target classes, and could find target classes not detectable in either platform alone. This confirmed that integration of expression-based and image-based phenotypes can accurately report on MOA. Furthermore, we integrated untargeted metabolomics of complex natural product fractions with the SNF network to map biological signatures to specific metabolites. Three examples are presented where SNF coupled with metabolomics was used to directly functionally characterize natural products and accelerate identification of bioactive metabolites, including the discovery of the azoxy-containing biaryl compounds parkamycins A and B. Our results support SNF integration of multiple phenotypic screening approaches along with untargeted metabolomics as a powerful approach for advancing natural products drug discovery.

natural products | pharmacology | metabolomics

Assigning the mechanism of action (MOA) to botanicals, natural products, and synthetic chemicals is an essential step in drug discovery and remains a major challenge in chemical biology. Despite the technological advances in isolation, synthesis, and screening strategies that make many bioactive substances available, in most cases, their biological targets remain unknown (1, 2). This challenge is exacerbated when taking a systems-level approach to gain mechanistic information about entire collections of molecules and complex mixtures, such as encountered in natural product libraries (3).

There has been a concerted effort to return to phenotypic screening approaches in drug discovery efforts (4, 5). This paradigm shift has come with the development of information-rich approaches that provide an unprecedented level of mechanistic understanding. These methods take advantage of gene expression profiling (6–8), high content imaging (9–11), yeast chemical genetics (12), proteomics (13, 14), and others (5, 15). While these platforms are valuable individually, each one is subject to the limitations. Here we test the hypothesis that using computational tools to integrate screening results from orthogonal screening platforms will allow for simultaneous leverage of divergent phenotypic coverage to inform MOA predictions. In this study, we integrate gene expression-based (Functional Signature Ontology; FUSION) and high content imaging-based (Cytological Profiling; CP) screening platforms using Similarity Network Fusion (SNF), and use this fused network to annotate high-resolution mass spectrometric profiling of a library of complex natural product fractions. The result is a novel framework for the functional annotation of natural products that demonstrates the power of leveraging multiple data types.

Our interest in the functional characterization of natural products led our groups to independently develop phenotypic screening strategies to evaluate natural product fraction libraries from marine bacteria. One platform, termed FUSION, utilizes perturbation-induced gene expression signatures coupled with pattern-matching tools to produce verifiable guilt-by-association MOA hypotheses (8). This method has been used to characterize a series of microbially derived molecules with unique mechanisms of action (8, 16–19). In this study, we have adapted the FUSION approach to a non-small-cell lung cancer context using the cell line NCI-H23 and a new set of 14 reporter genes that form the basis for pattern-matching between known and unknown perturbagens. A limitation of this

Significance

New data-driven methods to aid in the discovery and biological characterization of natural products are necessary to advance the field. Assigning the mechanism of action to novel bioactive compounds is an essential step in drug discovery and a major challenge in chemical biology. Advances in metabolomics have provided a better understanding of the constituents present in libraries but are not sufficient to drive the discovery of novel biologically active metabolites. Here, we describe an unbiased, data-driven strategy which integrates phenotypic screening with metabolomics into a single platform that provides rapid identification and functional annotation of natural products. This approach represents a strategy that could significantly accelerate the process of drug discovery.

This article is a PNAS Direct Submission.

[1] S.K.H. and T.N.C. contributed equally to this work.

[2] Present address: Moores Cancer Center, University of California at San Diego, San Diego, CA92037.

[3] Present address: Department of Small Molecule Pharmaceutical Science, Genentech, South San Francisco, CA 94080

[4] Present address: IDEAYA Biosciences, South San Francisco, CA 94080.

[5] To whom correspondence may be addressed. Email: mikewhite3224@gmail.com, rliningt@sfu.ca, or jomacmil@ucsc.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2208458119/-/DCSupplemental.

Published November 30, 2022.

approach, as well as other gene expression-based approaches such as the Connectivity Map (LINCS Consortium) (7), is that the sensitivity and specificity of the signature a bioactive molecule can produce is dependent on the biological context of the assay.

A biologically orthogonal platform, CP, utilizes high-content image analysis of perturbation-treated cells stained with a panel of fluorescent probes to extract sets of cytological features that are then used with pattern-matching tools to predict MOA (20). The CP platform utilizes unsynchronized HeLa cells, which, after treatment with perturbagen are fixed and stained with probes. A total of 251 unique cytological features are then extracted for each perturbagen from automated fluorescence microscopy images. Clustering compounds by their CP fingerprints has revealed both well-established associations among compounds with the same target or MOA, as well as novel or unexpected associations and unique phenotypes of natural products (3, 11). FUSION, CP, and related platforms are subject to limited resolution of bioactive compounds with broad cellular effects, or limited sensitivity to bioactive compounds that engage morphologically silent mechanisms (11).

While these methods have been exploited by our respective groups (3, 8, 11, 16–19, 21), the inherent limitations of these platforms can be especially problematic when exploring large, uncharacterized libraries whose active metabolites may span a wide and divergent range of biological activities. We hypothesized that a bioinformatic approach to integrate the two platforms could expand the biological space covered while retaining the information from both platforms. A challenge with integration of diverse data types is how to handle disparate numbers of features and data scales in individual datasets. To solve this problem, we adopted SNF (22), which overcomes this challenge by constructing similarity networks individually for each available data type and then fusing these into a single network based on shared similarity across both datasets. SNF has been used efficiently in a range of applications, including cell-to-cell heterogeneity (23), drug sensitivity (24), multiomic integrations, COVID-19, and other diseases (22, 25–33).

Natural product screening libraries are typically prepared as complex mixtures. To relate phenotypes to specific components in these mixtures, we required both a detailed description of the chemical constitution of each fraction, and informatics tools to define the associations between constituents and phenotypes. Current mass spectrometry-based methods often yield peak lists with very high false discovery rates, making it difficult to identify biologically relevant features from these large results' files (34). To address this, we developed bespoke acquisition and data processing methods designed to describe the chemical constitution of the natural product fraction library while removing interference signals caused by instrument noise and systemic contaminants from the sample processing workflow. These methods included appropriate replicates, blanks, and sample preparation workflow, employing an ion-mobility spectroscopy-enabled ESI-qTOF (35). Using a modified version of our Compound Activity Mapping platform (21), we then defined activity scores for all analytes based on SNF clustering results and developed a custom data visualization platform to directly relate analytes to specific biological phenotypes.

By integrating orthogonal screening platforms and combining this with next-generation metabolomics analysis of natural product libraries, we have created a unique and powerful framework for natural product biological characterization (Fig. 1*A*).

Results

Integration of Multiple Platforms Retains In-Class Target Classification. As a test-of-concept, we profiled a small collection of 628 randomly selected microbial natural product fractions in the FUSION, CP, and metabolomics profiling platforms (Fig. 1*A* and *SI Appendix*, Supplementary Note 1). For reference benchmarks, we also collected FUSION and CP profiles from a library of 2027 known synthetic small molecules that were selected for known bioactivity and have been annotated for MOA and/or molecular target (*SI Appendix*, Fig. S1, Table S1, and Supplementary Note 1 and Dataset S1). Briefly, all perturbagens were screened in triplicate on both platforms. FUSION gene expression signatures were normalized to nontreated wells, and CP fingerprints were normalized to DMSO-treated wells (*SI Appendix*, Supplementary Note 3). A Z-score transformation was then applied to both normalized datasets. To evaluate differential sensitivity of these orthogonal platforms to detection of chemical activity, we classified as "quiet" any perturbagen where all Z-scored probe values were less than |0.5|. Approximately 33% of perturbagens were quiet in FUSION (n = 889), and ~25% of perturbagens were quiet in CP (n = 686). Notably, natural product fractions comprised 25% of quiet perturbagens in FUSION (n = 226) but only 0.7% of quiet perturbagens in CP (n = 5). In total, intersection of these lists revealed that 12% of all perturbagens were quiet in both datasets (n = 329; three natural product fractions and 326 Selleck chemicals), suggesting that integrating the two datasets will provide active signatures for a larger percentage of the total compounds.

Next, we compared the dispersion of knowns and unknowns in each dataset using two-way hierarchical clustering (Fig. 1*B*). We observed that natural product fractions were interspersed throughout the clustering of each dataset, with more dispersion in FUSION than in CP. This confirmed that the natural product fractions produced sufficiently divergent signatures to allow for similarity analyses with benchmark chemicals that target a broad biological space. Interestingly, comparison of the pair-wise Pearson correlations between natural product fractions and all other perturbagens in FUSION vs. CP revealed that while some correlations trend in the same direction, the overall concordance between the two datasets on a perturbagen-by-perturbagen basis is relatively low (Fig. 1*C*). In fact, there are interactions that are negatively correlated in FUSION but positively correlated in CP, and vice versa. We also observe that the majority of Pearson correlations in CP fall within a relatively narrow range (Pearson r values between −0.5 and 0.5), while correlations in FUSION spread across the full range (Fig. 1*C* and *SI Appendix*, Fig. S2). This suggests that while the two platforms can report on the same biological space, CP may provide less resolution between MOAs than FUSION.

In order to assess concordance between the two datasets at the level of molecular target, we selected FUSION and CP signatures from the top 30 largest target classes in the Selleck library and applied k-means clustering (k = 30) to this subset. Using the hypergeometric test with Bonferroni correction for multiple comparisons to score for significant enrichment of target classes within each cluster revealed that FUSION and CP have similar levels of sensitivity in terms of total number of target classes detected (19 and 22 target classes with $P < 0.00167$, respectively); however, there is notable divergence between the two datasets in terms of which target classes are detected (*SI Appendix*, Figs S3 and S4). Comparison of cluster membership between the two datasets revealed that some target classes were robustly clustered together in both platforms (i.e., heat shock proteins (HSP), proteasome, histone deacetylases (HDAC)), while others are clustered more closely in one dataset than another (i.e., Aurora Kinase inhibitors cluster more closely together in CP than in FUSION, and mammalian target of rapamycin (mTOR) inhibitors cluster more closely together in FUSION than CP) (Fig. 1*D* and *SI Appendix*, Table S2 and Fig. S5). Moreover, comparison on a cluster-by-cluster basis of each dataset reveals that while both FUSION and CP are capable of clustering together target classes of similar MOA, the types

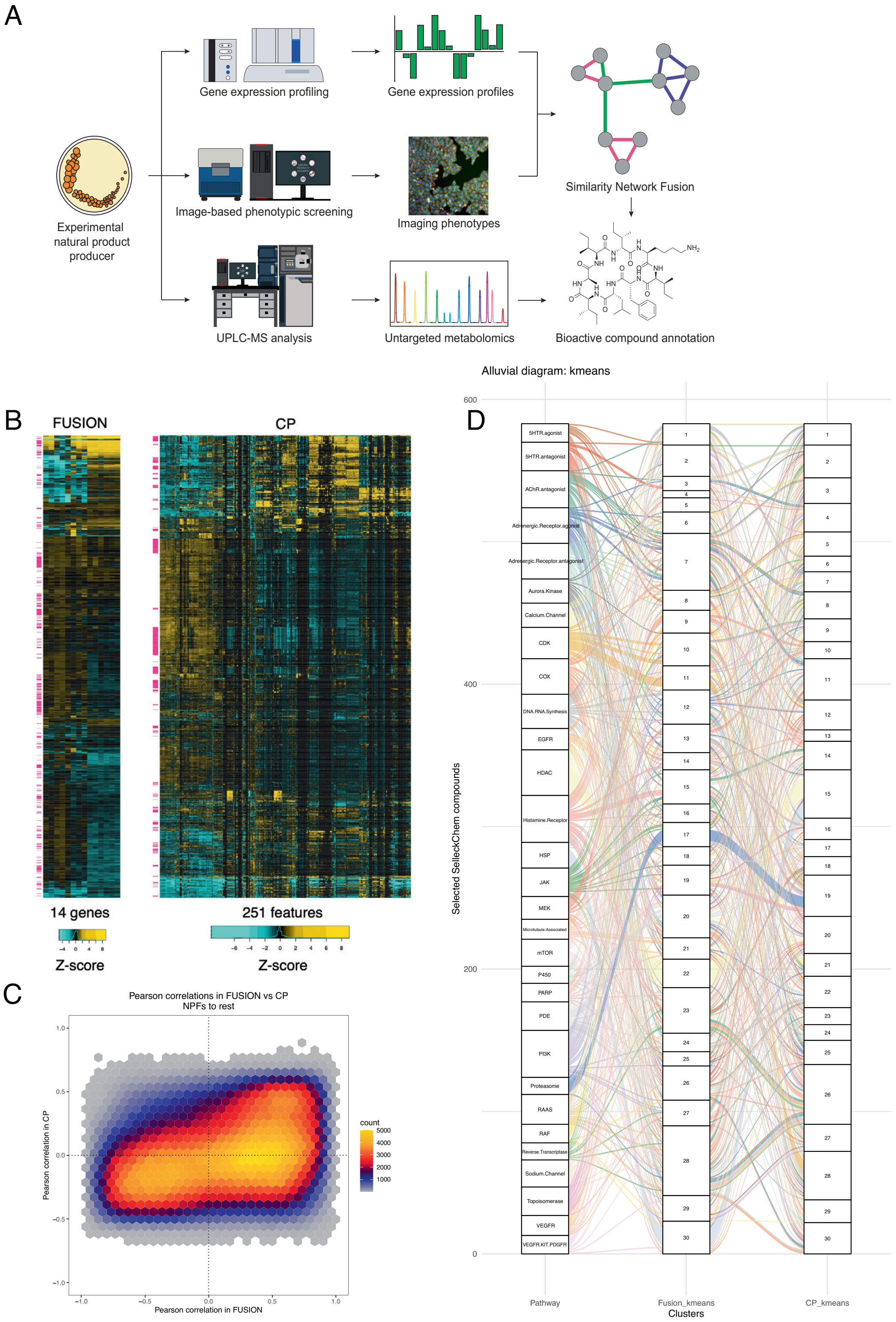

Fig. 1. Overview of screening platforms and initial data collection. (*A*) Experimental outline. Natural product fractions are isolated from marine bacteria, screened through two biological screening platforms (FUSION and CP), and subjected to high-resolution mass spectrometry-based metabolomics profiling. FUSION and CP data are integrated using SNF, which then provides biological annotation on individual metabolites identified. (*B*) Two- way hierarchical clustering of Z-scores from FUSION and CP using Euclidean distance and complete linkage. NPFs are indicated with pink flags. (*C*) Heat-scatter hexplot comparing Pearson correlations between NPFs and all other perturbagens in FUSION vs. CP. (*D*) Alluvial diagram comparing k-means clustering of chemicals in the top 30 largest target classes in FUSION and CP. Each target class is represented by a different color.

of MOAs they pair are different in many cases (*SI Appendix*, Fig. S6). Target class pairings that were observed in FUSION but not CP include mTOR and phosphoinositide 3-kinase (PI3K) inhibitors in cluster 2, RAF and mitogen-activated protein kinase (MEK) inhibitors in cluster 9, topoisomerase andcyclin dependent kinase (CDK) inhibitors in cluster 10, and pan-receptor tyrosine kinase (RTK) and epidermal growth factor receptor (EGFR) inhibitors in cluster 25 (*SI Appendix*, Fig. S4). By contrast, several other target class pairings were observed in CP but not in FUSION, including microtubule and JAK inhibitors in cluster 2, Poly (ADP-ribose) polymerase (PARP) and DNA/RNA synthesis inhibitors in cluster 7, HSP and HDAC inhibitors in cluster 15, and DNA/RNA synthesis and CDK inhibitors in cluster 28. Functional evidence supporting each of these pairings can be found in the literature. Importantly, many target classes were not effectively clustered together by k = 30 clustering, but the identity of these classes were also divergent between the two datasets (*SI Appendix*, Fig. S3). This lack of concordance could reflect differences between gene expression-based versus image-based readouts, low target expression in the cell lines used, misannotation of targets, polypharmacology within target annotated classes, and/or that k = 30 clustering did not offer adequate resolution to discriminate between some target classes. Taken together, these analyses suggest that at this level of resolution, each dataset is able to cover the same biological space with comparable depth, but reports on the same space very differently.

Generation of a fused similarity network across CP and FUSION signatures would allow for the orthogonal information contained in the molecular and morphological phenotypic readouts to be leveraged simultaneously in the annotation of uncharacterized compounds. However, integration of orthogonal datasets is a computational challenge due to inherent differences in experimental collection, measured features, noise, and overall scale between methods (36). In order to test the idea that combining the information from FUSION and CP would lead to an improved platform for MOA assignment, we used a data integration approach called SNF (22). This method addresses challenges associated with differences in scale and feature measurement by first constructing within-sample similarity networks for each data type. A single similarity matrix is then generated by iteratively propagating similarity information simultaneously across all individual networks to generate a single, fused similarity matrix where perturbagens with evidence of similarity across multiple datasets result in higher similarity measures (*SI Appendix*, Fig. S7 and Supplementary Note 4). To optimize for our high-content bioassay data, we adapted SNF by varying the value of *k* nearest neighbors and taking an agglomerate value of similarity across all *k* to generate a final matrix of similarity weights (see *SI Appendix*, Supplementary Note 4). This matrix was then used to calculate a new Euclidean distance or Pearson correlation matrix, and subjected to hierarchical affinity propagation clustering (APC) (37, 38) to group perturbagens based on each metric.

In order to assess the performance of the individual and fused datasets in assigning MOA, we again used our collection of commercial compounds (Selleck) and their target annotations as benchmark references. Among the 195 pre-annotated target classes within this collection, 89 classes contained five or more chemicals. A two-sample, one-sided Kolmogorov–Smirnov (KS) test was applied to each of these 89 target classes to determine whether the pairwise similarities between chemicals, as determined by FUSION or CP, with the same target annotation ("in-class") were significantly closer or more correlated than the pairwise similarities between these chemicals and those from other target classes ("out-of-class"). We compared Euclidean distance and Pearson correlation as similarity metrics. Perturbagens with high Pearson correlation will have signatures whose overall trend is in the same direction, but whose magnitudes may be very different. This can be useful when considering perturbagens which may have similar biological effects but different levels of potency, but will also have the effect of dispersing noise throughout the dataset. Meanwhile perturbagens with small Euclidean distances will have signatures which are closely related in both direction and magnitude. Thus, this metric can be particularly useful to make fine distinctions between different mechanisms of action, but may group noisy signatures together.

Comparison of the KS-test *P*-values for each target class across all datasets revealed that that SNF using Euclidean distance identified 20 of the 24 target classes identified in FUSION (83%) and 27 of the 38 target classes identified in CP (71%). We also observed that SNF-Euclidean identified significant self-association between members of 19 additional target classes not identified in either dataset alone (Fig. 2 *A* and *D*; CDF plots for each target class are included in *SI Appendix*, Fig. S8). By contrast, SNF-Pearson identified 29 of the 36 target classes identified in FUSION (81%) and 44 of the 52 target classes identified in CP (85%), and 4 additional target classes that were not identified in either dataset (Fig. 2 *B* and *D* and *SI Appendix*, Fig. S8). Notably, there also was a high degree of overlap in target classes identified by SNF-Euclidean and SNF-Pearson (Fig. 2*C*).

Taken together, these analyses demonstrated that valuable associations can be found in each dataset using either similarity metric, and that SNF retains at least 70% of the information found in individual datasets. Thus SNF is a platform in which the biological associations in both datasets can be leveraged together to provide functional annotation of compounds with unknown MOA.

SNF Integration Drives Clustering of Natural Product Fractions. We next used SNF values to construct a relational network among the reference compounds and natural product fractions using hierarchical affinity propagation clustering (APC) as described previously (38) (Fig. 3). This clustering method was chosen as it is a deterministic method that defines, in a data-driven fashion, both the number and membership of clusters emerging from a given similarity matrix (37). Binning edges based on the contribution from each individual dataset revealed that more than half of associations are supported by both datasets (~54%), while ~14% and ~32% of associations are supported primarily by FUSION and CP, respectively (see *Methods*; Fig. 3*A*). When these edge annotations are quantified on a per cluster basis, we observed that while most clusters are supported by both datasets, some clusters are driven by one dataset (e.g., Clusters 2 and 91 are driven by CP, while Clusters 120 and 125 are driven by FUSION; Fig. 3*B*). Notably, most of the perturbagens that were flagged as "dead" in either platform clustered together in the SNF-Euclidean network, and this list included compounds for which cytotoxicity would be expected at the doses used in these assays (e.g., topoisomerase inhibitors; *SI Appendix*, Fig. S9). In general agreement with our KS-test results, many clusters were significantly enriched for chemicals with the same target annotation, as assessed by a hypergeometric test (Fig. 3*C* and *SI Appendix*, Table S3). We also observed that some clusters were significantly enriched for multiple classes, which may reflect similar mechanisms of action and/or convergence of downstream signaling effects (e.g., Cluster 123 is significantly enriched for PI3K, mTOR and EGFR target classes). It is also possible that overlap of multiple target classes in the same cluster may reflect a limitation of the gene set, cytological features, or the cell lines selected for profiling in both platforms, in that these reporters may not be sufficient to distinguish between those mechanistic classes. A comparison between the SNF-Euclidean and SNF-Pearson APC maps revealed that some target classes which can either be classified as closely related to other classes or be divided into subclasses have more separation in the SNF-Euclidean APC map compared to SNF-Pearson. For example, several chemicals annotated as "epigenetic reader domain" inhibitors cluster together with HDAC inhibitors in the SNF-Pearson APC map (*SI Appendix*, Fig. S10, Cluster 64), but separately from HDAC inhibitors in the SNF-Euclidean APC map (Fig. 3*C*, Cluster 76). The SNF-Euclidean map is also able to

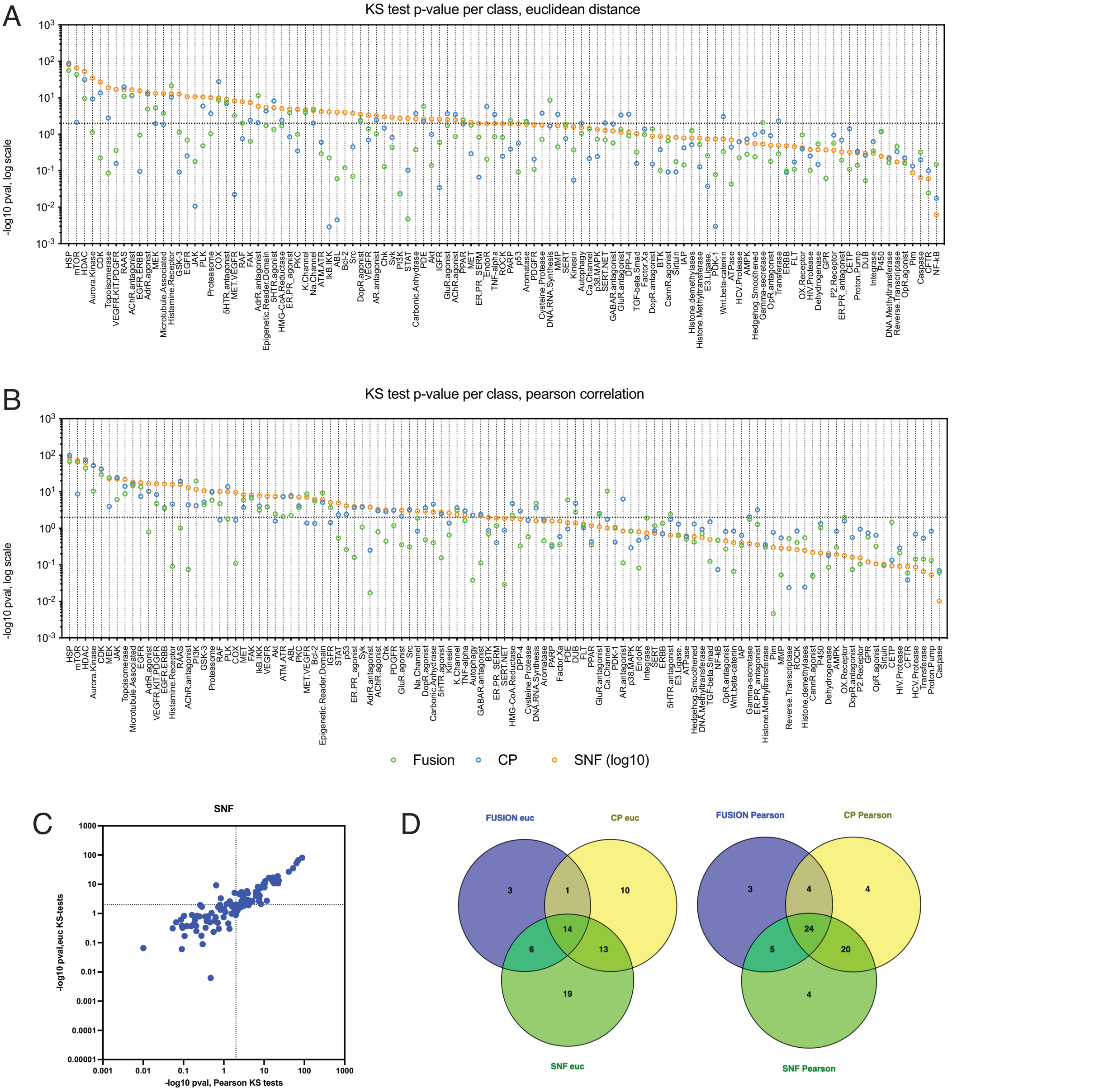

Fig. 2. Comparison of KS-test *P*-values for in-class vs out-of-class target annotation in FUSION, CP, and SNF. Dot plots of $-\log_{10}$ KS-test *P*-values in FUSION, CP and SNF datasets, for each target annotation class with at least five members, using (*A*) Euclidean distance or (*B*) Pearson correlation as the similarity metric. Significance threshold is represented by the horizontal line ($P = 0.01$). (*C*) KS-test *P*-values for every target class in SNF-Pearson vs. SNF-Euclidean. (*D*) Venn diagrams illustrating the overlap in target classes scoring as significantly self-associated be KS-test ($P < 0.01$) using Euclidean distance or Pearson correlation.

cluster pan-CDK inhibitors separately from other more specific CDK inhibitors (Fig. 3*C*, Cluster 51). We also observe that there are more clusters in the SNF-Pearson APC map that contain multiple members of different classes than in the SNF-Euclidean map (Fig. 3*C* and *SI Appendix*, Fig. S10). The Pearson networks do clearly contain valuable associations (Fig. 2*B*), which are likely to be informative across different biological contexts compared to the Euclidean distance networks. Our comparison of the networks suggests that SNF-Euclidean may have superior power to distinguish between related target classes than the SNF-Pearson network, and thus we chose to use this similarity metric in downstream analyses.

Untargeted Metabolomics Relates Chemical Constitution to Functional Signatures via the SNF-Similarity Score. The chemical complexity of natural product fractions increases the difficulty in relating phenotypes to specific molecules or sets of molecules for a given sample. However, in most cases biological activities are driven by a single compound or a small subset of compounds in each extract (39). By determining the distribution of secondary metabolites across the full sample set, it is possible to test the hypothesis that extracts with similar phenotypes contain the same or similar bioactive species. In order to create a clear picture of chemical constitution across the sample set, we performed untargeted metabolomics on the full set of natural product extracts using a UPLC-IMS-qTOF instrument operating in data-independent acquisition mode (DIA) (*SI Appendix*, Supplementary Note 2). Inclusion of ion mobility spectrometry affords an additional axis of separation over standard LCMS systems that improves separation of complex mixtures and provides an additional physicochemical measure (collisional cross-sectional area) for matching analytes between samples. Use of DIA increases the percentage of analytes that are subjected to MS fragmentation compared to traditional data-dependent acquisition. These fragmentation patterns are useful for comparing

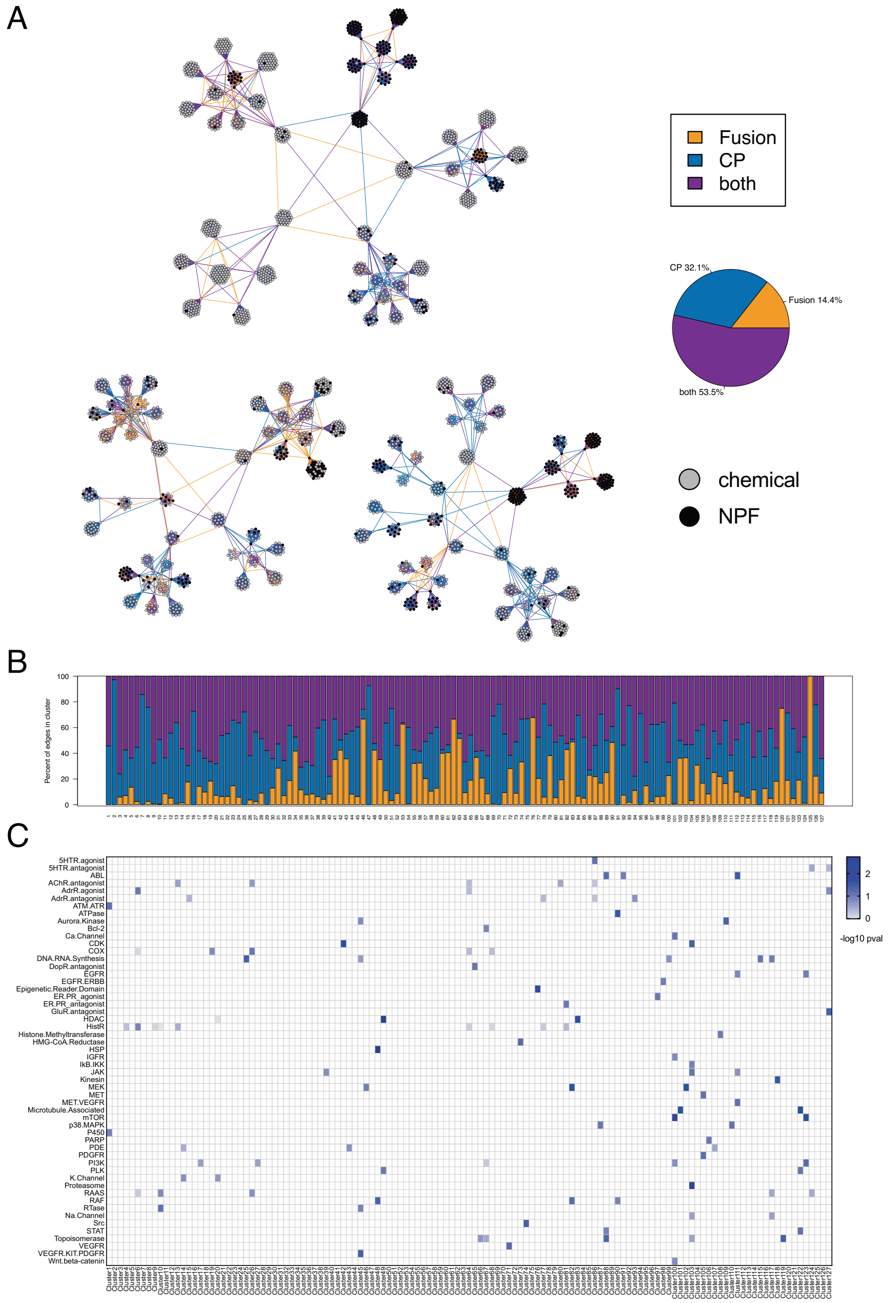

Fig. 3. Affinity propagation clustering map of the SNF network preserves in-class target associations. (*A*) Hierarchical affinity propagation clustering map of the SNF network using Euclidean distance as the similarity metric. Edges are colored based on contribution from individual datasets: Orange, supported by FUSION; blue, supported by CP; purple, supported by both datasets. Perturbagen type is indicated by node color: black, NPF; gray, pure chemical. (*B*) Bar plot showing the percent of total edges in each APC cluster that are supported by FUSION, CP, or both datasets. Clusters are labeled by cluster number. (*C*) Heatmap showing minus $\log_{10}$ *P*-values calculated by hypergeometric test for each target annotation class, per APC cluster. Target classes without significant enrichment in any cluster are omitted (Bonferroni-corrected alpha = 0.0016).

analytes between samples, and for comparing to external reference libraries for compound identification (40, 41).

Samples were analyzed as three independent technical replicates, and consensus feature lists generated for each sample using a suite of in-house data processing scripts. Mass spectrometric features were required to appear in at least two of three replicates to be included in the consensus feature list. These sample-by-sample feature lists were then 'basketed' to produce a single list of unique mass spectrometric features across the full sample set. This feature list included information about mass spectrometric properties (e.g., retention time, mass to charge ratio, collision cross-sectional area) as well as sample distribution (*SI Appendix*, Supplementary Note 2).

For this initial study, a small set of 75 randomly selected strains of marine-derived Actinobacteria and Firmicutes from the MacMillan and Linington culture collections were grown in large-scale liquid culture, extracted using our standard extraction protocol, and pre-fractionated over C18 to afford 628 natural product fractions. Mass spectrometric analysis of these fractions identified a total of 8,108 mass spectrometric features, of which 3,498 appear only once in the sample set (43%). To examine the relationship between individual features and the SNF network, we employed a variation of our previously developed Compound Activity Mapping method to score predicted mass spectrometry feature activities (21). For each unique feature in the metabolomic dataset, we identified the subset of natural product fractions containing the feature and calculated the average of the SNF similarity scores within this set (see *SI Appendix*, Fig. S11 and Supplementary Note 4). This score provides a numerical evaluation of how closely the presence of a specific mass feature is correlated with the presence of a specific biological phenotype in the APC network. In cases where a given feature is responsible for an observed activity, it is expected that the phenotypes of the associated set should be similar, and that the average SNF similarity score (SNF score, with a score range from 0 to 1) should be correspondingly high. By contrast, compounds that do not impart a biological response should not correlate with a specific biological signature, and the SNF score should be correspondingly weak. Using a 95th percentile cutoff of all Euclidean distance-based SNF scores revealed 229 features with high correlation to biological activity (*SI Appendix*, Fig. S12).

SNF scoring is feature-independent, meaning that a high score for one mass spectrometric feature has no impact on the scores of other features in the sample. This is important because the mass spectrometry data are not deconvoluted by either adduct (e.g., $[M+H]^+$ vs. $[M+Na]^+$) or in-source fragments (e.g., $[M-H_2O+H]^+$). It is therefore common to identify a suite of mass spectrometry features with the same retention time that all possess strong SNF scores. These features can be used in concert to determine the correct accurate mass for the active component (which aids in dereplication) and to reconstitute mass spectrometry fragments (which can help with metabolite identification).

Calculating SNF scores for every mass spectrometric feature provides a metric to quickly identify bioactive compounds and prioritize them for subsequent isolation. A valuable visualization for these data is the Compound Activity Map (Fig. 4*A*). In this network, extracts are represented by large nodes, while individual mass spectrometric features are represented by small nodes, color-coded by SNF score. Only mass spectrometric features with SNF scores above a set threshold are included, with edges added between extracts and the features they contain. The network is therefore arranged based on shared bioactive chemical features. Using this visualization, it is possible to identify sets of extracts with the same mechanistic prediction based on SNF annotations. Selection of clusters with similar SNF scores can be used to prioritize target molecules with shared biological properties. For example, fractions SW218953, SW218954, and SW218955 (Fig. 4*B*) share a suite of related mass spectrometric features, including molecular ions, adducts and in source fragments, that possess similar extracted ion chromatograms (Fig. 4*C*). These features also possess strong SNF scores (dark green nodes in Fig. 4*B*), and identical activity predictions as shown by assignment to the same SNF APC cluster, suggestive of a single bioactive compound family in these samples. Conversely, in situations where clusters contain several classes of bioactive compounds, SNF scoring can be used to subdivide these clusters by chemical family. For example, samples SW218928, SW218929, SW218930, and SW218931 (*SI Appendix*, Fig. S13) divide into two groups based on differences in mass spectrometric features, suggesting the presence of two separate compound classes within these related extracts.

Compound Activity Maps thus provide a powerful strategy to prioritize candidate compounds for isolation. However, this visualization is centered on biological attributes, and provides less information about chemical properties. As a complement to Compound Activity Mapping we used the open-source Bokeh server library to create a data visualization tool that enables direct

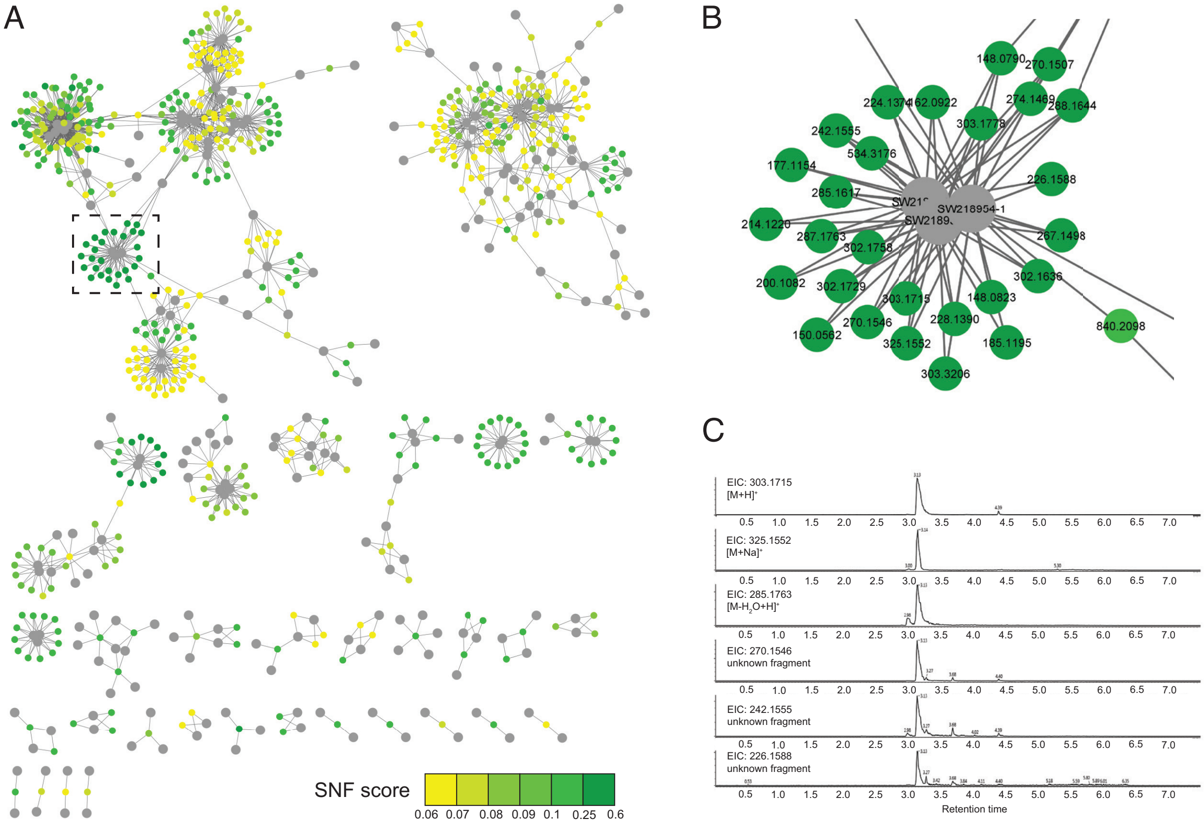

Fig. 4. Compound Activity Map for combined SNF profiles and untargeted metabolomics features. Large nodes represent extracts. Small nodes represent mass spectrometry features. Edges represent presence of mass spectrometric features in connected extracts. Only mass spectrometric features with predicted SNF scores >0.06 are included. (*A*) Full Compound Activity Map. Small nodes are color coded by SNF score. (*B*) Expansion of a representative region of the map with large nodes colored by SNF score. (*C*) EIC of mass spectrometric features present in adjacent fractions with similar SNF scores.

examination and filtering of the untargeted metabolomics data with a range of data display options (*SI Appendix*, Fig. S11). This platform can display metabolomics data as plots of retention time vs. *m/z* ratio, filtering based on SNF score, presence in extract list, or both. This viewpoint on the data is valuable for selecting lead compounds that not only score well based on bioactivity predictions, but also display robust chemical signatures for subsequent isolation and structure elucidation.

In order to evaluate the efficiency of this new platform for de novo bioactivity prediction from complex mixtures, we tested two different query approaches; 1) querying the SNF network for natural product clustering predominately near other reference compounds, (biology-first discovery), and 2) filtering for metabolites with highly correlated biological activity as assessed by SNF score (chemistry-first discovery). The first case highlights compounds that possess mechanisms in line with known bioactives while the second approach identified sets of compounds with mechanisms that are not covered by the training set of known bioactives. This second approach attempts to address one of the largest challenges in natural products, that is, identification of new natural products with novel mechanisms of action. These strategies were selected to test the platform under different conditions, from simple situations where the annotations were unanimous, to complex situations with multiple reference compound types and multiple natural product fractions. In each approach, we highlight the contribution of SNF and metabolomics toward identification of both the natural product driving the signature and its biological MOA.

Identification of Trichostatin A from an HDAC-Inhibitor-Enriched Cluster Validates the Integrated SNF Platform. We first sought to validate the SNF network by querying the dataset for natural product fractions that clustered mainly with reference compounds from a single target class. In the SNF-Euclidean APC, there were 6 clusters that were highly enriched (P < 1e-10) for chemicals belonging to the same target class: Cluster 48 (HSP), Cluster 49 (HDAC), Cluster 76 (Epigenetic Reader Domain), Cluster 100 (mTOR), and Cluster 103 (Proteasome) (Fig. 3*C* and *SI Appendix*, Table S4). Of these, Clusters 49, 100, and 103 contained natural product fractions (3 in Cluster 49, 1 in Cluster 100, and 1 in Cluster 103), thus identifying readily testable MOA hypotheses for the bioactive natural products present in each case. A KS-test confirmed that association between chemicals in the HDAC inhibitor target class is preserved in the full dataset, and that these associations were still significantly improved in SNF compared to FUSION or CP (P = 1.8e−61; Fig. 5*A*).

We observed that the HDAC inhibitor cluster contained three sequentially isolated natural product fractions (SW218953, SW218954, and SW218955) (Fig. 5*B*). The presence of multiple natural product fractions (NPFs) from the same series suggests the presence of common metabolite profiles. Filtering the metabolomics data for features that are only present in these three natural product fractions revealed a vertical "stripe" of mass spectrometry features at 3.13 min with a precursor mass feature of 303.1712*m/z* (Fig. 5*C*; red box). This pattern of signals is indicative of both a precursor mass and associated in-source fragments from the LCMS analysis (extracted ion chromatograms for the NP fractions are included in *SI Appendix*, Fig. S14). Subsequent chromatographic optimization, purification and NMR analysis from SW218953 identified this product as the known bacterial metabolite trichostatin A (Fig. 5 *D*–*E*). Trichostatin A has been extensively studied for its activity as an HDAC inhibitor (42, 43). Notably, Cluster 49 also contained pure trichostatin A from the Selleck library (*SI Appendix*, Table S5). An analysis of the top 50 nearest neighbors to trichostatin A in the SNF-Euclidean network also confirms that these three natural product fractions are tightly associated with HDAC inhibitors (Fig. 5*F*). Collectively, the natural product fractions containing trichostatin A also had a very high SNF score (0.537; 99.9th percentile). Therefore, our integrated platform successfully assigned known MOA to natural product fractions in an agnostic manner, confirming the power of this annotation strategy.

SNF-Score-Driven Identification of Surugamide A as Modulator of CDK Activity. Using a top 98th percentile cutoff of SNF scores also identified a single natural product (precursor mass-to-charge ratio of 912.6266) that was present across multiple extracts from multiple bacterial species (SW218824, SW218835, SW218858, SW218859, and others; Fig. 6*A*). Purification and structure elucidation identified this metabolite as the cyclic octapeptide surugamide A (Fig. 6*B* and *SI Appendix*, Fig. S15 and Supplementary Note 5) (44). The surugamides are a recently discovered class of cyclic peptides that appear to be widely distributed in *Streptomyces* sp. Initial biological activity reports for surugamide A show weak activity as protease inhibitors with an IC_{50} of 21 μM in an enzymatic assay for inhibition of bovine cathepsin B (44). Querying the SNF-Euclidean APC network showed several surugamide-containing fractions clustering near each other and in close proximity to CDK inhibitors (*SI Appendix*, Fig. 15*C*). The function of the retinoblastoma tumor suppressor (Rb) is tightly controlled by CDK complex proteins and the phosphorylation state of Rb is indicative of cell-cycle progression (45). Based on this APC clustering, we evaluated surugamide A (10 μM) for its ability to inhibit Rb phosphorylation, compared to the CDK inhibitor dinaciclib (100 nM). Western blot analysis of surugamide A treatment clearly shows strong suppression of Rb phosphorylation on Ser 807/811 relative to untreated cells (Fig. 6*C*).

The identification of surugamide A in our dataset confirms that the use of the SNF score with untargeted metabolomics can also be used to quickly identify single bioactive metabolites in fractions that share a phenotype. In the case of surugamide A, clustering with a subset of CDK inhibitors provided a testable and validated biological hypothesis. Taken together, these data demonstrate that the bioinformatic integration of FUSION, CP, and metabolomics datasets can effectively drive the rapid discovery and characterization of bioactive natural products.

SNF-Guided Discovery of Parkamycins A and B. Finally, several clusters in the APC map contained exclusively NP fractions, suggesting the presence of bioactive metabolites with mechanisms not represented in the training set. One of these clusters contained two isobaric species with strong SNF scores (*m/z* = 455.2604 and 455.2627, rt = 4.06 and 4.40 min Fig. 6*D* and *SI Appendix*, Fig. S16), suggestive of a bioactive compound family. Refermentation and isolation yielded two molecules with matching UV spectra, one of which (parkamycin B) was highly unstable, rapidly converting to the more stable isomer (parkamycin A) upon exposure to light (Fig. 6*E*). Structure determination of the more stable isomer using a full suite of NMR and spectroscopic techniques (*SI Appendix*, Supplementary Note 6, NMR spectra as *SI Appendix*, Figs. S17–S30) identified parkamycin A as a natural product containing the highly unusual biphenylazoxy core pharmacophore. While azoxy motifs have some precedent in medicinal chemistry there are very few examples of this motif in nature (46). As the parkamycins did not correlate to any molecules in the informer set by SNF, the effect of parkamycin A on H23 cells after a 6-h treatment was evaluated by profiling 748 gene transcripts that are part of the Nanostring curated metabolic pathways probe set (*SI Appendix*, Supplementary Note 3, Fig. S31, and Table S6 and Dataset S2). Genes were assigned to pathways through the nSolver Advanced Analysis system 4.0 and Advanced Analysis

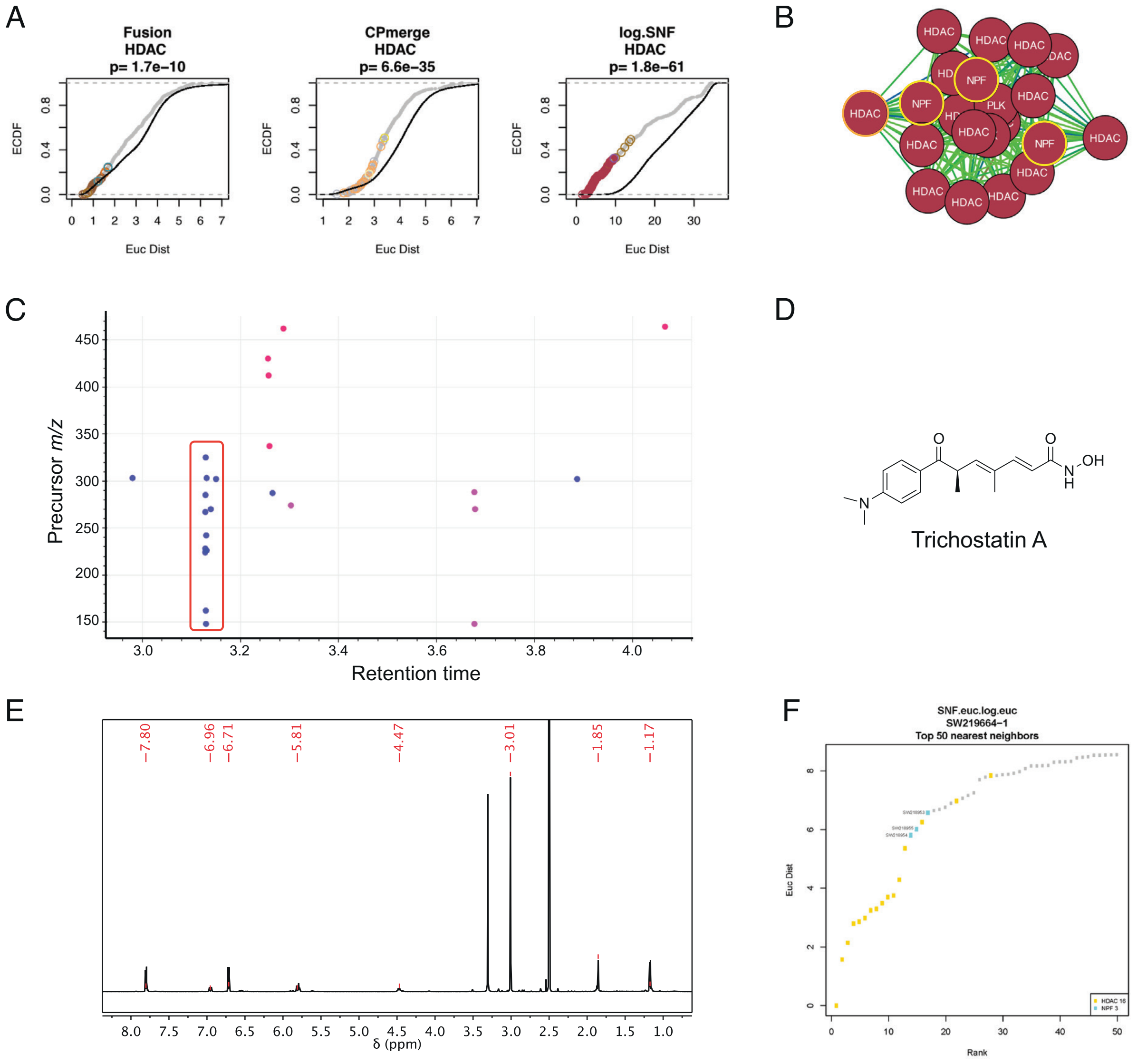

Fig. 5. SNF correctly assigns MOA to the major metabolite in a series of natural product fractions. (*A*) CDF plots comparing pairwise Euclidean distances between chemicals annotated as HDAC inhibitors in the full dataset, versus out-of-class associations. Gray, in-class; Black, out-of-class. Colored circles correspond to cluster membership in the associated APC map. The *P*-value was calculated by KS-test. (*B*) Cluster 49 from the SNF-Euclidean map, drawn using a spring-embedded layout. (*C*) Retention plot showing common mass spec features present in the NPFs highlighted panel *B*. (*D*) Chemical structure of trichostatin A. (*E*) NMR spectra confirming trichostatin A. (*F*) The top 50 nearest neighbors to trichostatin A in the SNF-Euclidean network, with the three natural product fractions found in Cluster 49 labeled.

module 2.0, and scores were generated that showed upregulation or downregulation of metabolic pathways as a whole (*SI Appendix*, Fig. S32; 47). Several genes related to AMPK, endocytosis, KEAP1, p53, and Myc signaling were down-regulated, while cell-cycle signaling, epigenetic and cell-cycle regulation were up-regulated (*SI Appendix*, Fig. S33 *A–H*). These results suggest that the parkamycins have complex biological activity that requires further investigation.

The isolation of parkamycin compounds demonstrate the value of 'chemistry-first' prioritization methods for discovering novel natural product scaffolds with biological activities not represented in the reference compound training set.

Discussion

The natural product literature contains thousands of examples of novel compounds with biological activities reported from simple end-point assays, such as cytotoxicity or antimicrobial growth inhibition assays. While this provides a handle for further investigation, the lack of detailed mechanistic information means that the majority of these molecules are never followed up on for biological characterization. This is due to the aforementioned challenges associated with characterizing the mode of action of pharmacological agents. Previous biological screening platforms developed by our laboratories (CP and FUSION) have been successful at characterizing new natural products with detailed mechanistic assignments (3, 8, 11, 16–19). While powerful, both platforms encountered scenarios where no prediction for a natural product fraction was possible due to weak signatures. Differences in both resolution and sensitivity between platforms can limit their utility, either because a given mechanism is not reported on by the assay system, or because the resolving power of the platform is insufficient to differentiate between mechanistic classes. In order to maximize the amount of information used to predict MOA, we applied an adapted version of SNF to integrate data from both CP and FUSION.

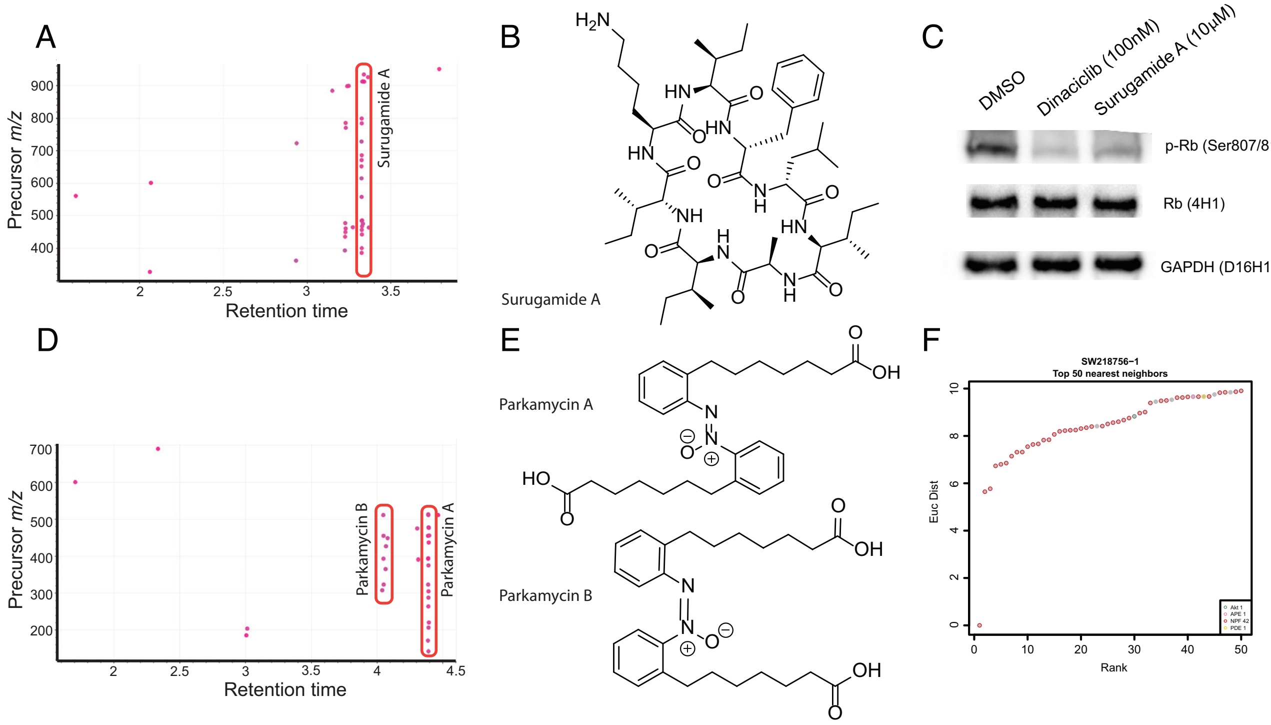

Fig. 6. Filtering for highly correlated SNF scores can identify single metabolites with biological activity. (*A*) Retention plot showing common mass spectrometry features associated with SNF scores above 0.5 for an APC region. (*B*) Chemical structure of surugamide A. (*C*) Immunoblot showing changes in phospho-Rb in H23 cells after treatment with Dinaciclib or Surugamide A for 24 h. (*D*) Retention plot showing common mass spectrometry features associated with SNF scores above 0.5 for a second APC region. (*E*) Chemical structure of parkamycin A and B. (*F*) The top 50 nearest neighbors to SW218756-1 in the SNF-Euclidean network. Points are colored by either target annotation or perturbagen type.

The SNF network retained the associations of many pure chemicals to their annotated target class that were observed in either FUSION or CP, delivering the capacity to leverage both datasets simultaneously for untargeted mode of action prediction. We validated the utility of this network in assigning MOA by demonstrating that natural product fractions containing trichostatin A were clustered with pure trichostatin A and other HDAC inhibitors. We then further developed a robust pipeline to assign the mechanistic annotation that the SNF network provides to specific natural product structures. Using untargeted metabolomic profiling of the full natural product fraction library and creating a scoring method (SNF score) to relate these mass spectrometry features to defined phenotypes, it is possible to directly predict the contributions of all mass spectrometry features to the biological landscape of the sample set. Development of the Bokeh server visualization suite (*SI Appendix*, Fig. S11) also provides a facile platform for data filtering and visualization that enables the rapid exploration of these data using a range of different viewpoints. Using this approach, we were able to link surugamide with novel biological activity against CDKs. Thus, SNF scores provide a rich perspective on chemical and functional interpretation from the natural product library. For example, in situations where two different compound classes cause the same biological phenotype (i.e., one cluster in the APC network), SNF scores can correctly identify these two compounds as high-priority candidates, even though neither compound is present in all members of the biological cluster, provided that each molecule is predominantly found within that cluster. Similarly, in situations where extracts contain many mass spectrometric features, most features will be quickly deprioritized because their distributions throughout the sample set do not correlate to specific biological phenotypes. Finally, SNF scores can be used to prioritize fractions that have biologically active natural products even in the absence of benchmark compounds, as demonstrated by the identification of parkamycins from clusters primarily enriched in other NPFs. The discovery of a new natural product with no clear associations to the diverse biological space represented in our chemical training set confirms that our integrated approach can quickly identify high priority bioactive compounds even without biological annotation. This mechanism for compound prioritization is therefore a robust and powerful strategy for directly targeting biologically relevant compounds from large, complex, natural product libraries, and can greatly accelerate the discovery of both novel chemistry and biology. An important aspect of new technologies is the ability to identify minor components in complex mixtures, which the integrated biological/metabolomics signatures are able to do.

Notwithstanding the value of this approach, there are several situations which remain difficult to resolve. Currently, the SNF score is not weighted by relative intensity of each MS feature. This is because determining relative concentrations of unknown analytes in complex samples remains an unsolved challenge in mass spectrometry-based metabolomics. In situations where a bioactive metabolite is present both above and below its EC_{50}, the SNF score will be reduced, as there will be no measurable phenotype in extracts where the concentration is low. Secondly, the system cannot differentiate between active and inactive metabolites if they are always co-expressed. Review of our metabolomics dataset suggests that this circumstance is rare; however, in these cases both metabolites would be scored as active candidates, requiring downstream deconvolution. Finally, in situations where bioactive compounds are frequently encountered with other unrelated bioactives, the resulting phenotypes could bear little relationship to one another. Review of the dataset suggests that this situation is also unusual; however, in these cases SNF scores will also deteriorate because of the reduced similarity scores between samples with different phenotypic signatures.

The use of SNF to merge orthogonal data is limited by the breadth of space covered in the input datasets, both in terms of the reference training set and the number of features used for readout. The proliferation of information-rich screening technologies, such as L1000 and cell painting, provide enhanced opportunities for chemical/ biological associations in natural product and other libraries (6, 8). While we chose to use FUSION and CP in this study, the approaches and methodology outlined here would be amenable to these platforms as well. The field of natural products will be greatly enhanced by the adoption of more complex screening platforms, as one of the major limitations in the field is the lack of mechanistic understanding of the large majority of isolated molecules.

Natural product research brings with it a number of challenges, such as the chemical complexity of extracts, re-isolation of known compounds and characterization of biological activity. These challenges limit the pace of natural product research and leave knowledge gaps around the value of a given natural product structure or class. Recent initiatives to develop resources to better understand the genomics of natural product biosynthetic gene clusters (48) and the development of the Global Natural Products Social molecular networking platform (41) have fundamentally changed how natural product research is conducted, but the field as a whole is far behind in leveraging 'Big Data' to address outstanding challenges. The approach we have detailed here provides an unbiased, data-driven platform that can be used to integrate biological assay and metabolomics results to provide a comprehensive viewpoint on chemical/biological relationships in the natural product sphere.

Materials and Methods

Chemical Libraries.

Selleck chemical library. The reference library is a set of 2027 molecules spanning 196 compound classes, with 789 compounds not belonging to any annotated class. The library was purchased from Selleck as a premade library. Further details are included in *SI Appendix*, Supplementary Note 1.

Natural product libraries. Two natural product libraries were utilized in this study. The MacMillan lab collection used in this study was comprised of ~500 fractions derived from 25 marine-derived bacterial strains. The Linington natural product fraction library contains >5,000 microbial fractions. The library is comprised of extracts of marine sediment-derived bacterial strains isolated by the Linington laboratory over the past 10 y. Further details are included in *SI Appendix*, Supplementary Note 1.

Metabolomics.

DIA UPLC-MS/MS data acquisition. All measurements were performed with an Acquity UPLC H-Class (Waters) using an HSS C18, 100-mm × 2.1-mm, 1.7-µm column (Waters). The LC flow was directly infused into a Synapt G2-Si operated in positive ion mode. Mass spectra were acquired from 50–1500 *m/z* at a 2-Hz scan rate in continuum mode without lockmass correction. Details for data processing are included in *SI Appendix*, Supplementary Note 2.

Cell Culture Conditions. NCI-H23 cells were cultured in RPMI supplemented with 5% FBS (Gibco). HeLa cells were cultured in DMEM supplemented with 10% FBS (Difco, MT35015CV). All cells were cultured at 37°C under 5% CO_2.

CP Assay. Briefly, HeLa cells were seeded into 384-well at 2,500 cells/well. After a 24-h incubation, cells were treated with test fractions or pure compounds using a Janus MDT robot (PerkinElmer). Staining procedures are described in *SI Appendix*, Supplementary Note 3. Natural product fractions were screened at either 10 µg/mL (Macmillan library), or 1000x dilution (Linington library). Selleck chemicals were screened at 10 µM. If a CP fingerprint was flagged as inactive at 10 µM, then the chemical was re-screened at 50 µM and that resulting fingerprint was used in downstream analyses. Image capture, processing and analysis follows previously published methodology (11; *SI Appendix*, Supplementary Note 3).

Functional Signature of Ontology Assay. The FUSION assay concept was described previously (8). We extended this concept to a lung cancer context by selecting a new set of genes that can report on the physiological state of lung cancer cell lines. Expression of 14 dynamic reporter genes (*DUSP6, FAM3C, GCNT3, GRHL2, HSD17B7, KIAA0922, LCN2, LTBR, RRM2, SIRPA, TLE2, TMEM30B, WSB2, YAP1*) and two static reporter genes (*EEF1A1, SIRT6*) were detected using a 16-plex QuantiGene Plex 2.0 Assay (ThermoFisher). Further details on sample processing, data acquisition and analysis are included in *SI Appendix*, Supplementary Note 3.

Data Integration and Statistical Analysis

SNF Methodology. SNF is a novel similarity metric designed to aggregate information across multiple datasets and assign a similarity score to perturbations based on evidence from multiple datasets. SNF was performed as previously described, with some modifications (*SI Appendix*, Fig. S7; 22). Similarity matrices used for input were calculated using either the dist2 function in the 'SNFtool' R package for Euclidean distance, or the distance matrix function from the 'ClassDiscovery' R package to generate Pearson distance similarity matrices. In the step where k-nearest neighbors are chosen, we chose to vary k from k = 2 to k = n/2, where n is the total number of perturbations in each dataset, and use an agglomerate value of similarity across all k. Networks were then fused by prorogation information from each dataset until a final fused matrix is calculated. This procedure results in n/2-1 fused matrices total. Each matrix is then normalized by dividing by the maximum nondiagonal value, and then the average value between all matrices is calculated to result in a final, fused aggregate similarity matrix. This matrix was then $\log_{10}$-transformed and clustered by APC using either Euclidean distance or Pearson correlation as the similarity metric. A full description of the algorithm is provided in *SI Appendix*, Supplementary Note 4.

Clustering and Statistics. The similarity between Z-scored perturbagen profiles was measured using either Euclidean distance or Pearson correlation. To assess the ability of each dataset to identify significant associations within Selleck target classes, we used a two-sample, one-sided KS test to determine whether in-class associations were significantly smaller (Euclidean distance) or larger (Pearson correlation) than out-of-class associations. Only target classes with at least five members were considered for this analysis. Hypergeometric tests were performed to assess for statistical enrichment of target classes in clusters, and threshold for significance was corrected for the total number of classes considered. K-means clustering was performed using the kmeans function in the 'stats' R package. Alluvial diagrams were generated using either the 'ggalluvial' or 'alluvial' R packages.

Hierarchical APC was performed as previously described using either Euclidean distance or Pearson correlation as the similarity metric (38). APC was chosen because it is a deterministic clustering method. In addition, APC will determine, in a completely automated fashion, not only the number of clusters arising from the data but also the exemplar member of each cluster. APC performs clustering by passing messages between the data points (37), taking a square matrix representing pairwise similarity measures between all data points as input. Each data point is treated as a node in a network and is initialized by connecting all the nodes together, where edges between nodes are proportional to the distance between them. Messages are then iteratively transmitted along the edges, which are pruned with each iteration until a set of clusters and exemplars emerges. In our implementation, we clustered the exemplars identified by APC. This was repeated until no more clusters emerged, thereby identifying a hierarchical structure of clusters. A full description of our implementation is described in ref. 38.

FUSION data were clustered using the final 14 gene signatures, CP data were clustered using the final 251 features, and SNF data were clustered using the aggregate W_{fused} matrix of weights over all k values. Networks were visualized in Cytoscape (49) with edge lengths drawn using the Allegro Spring-Electric layout.

SNF APC network edges were colored according to dataset contribution as described below. For each pair-wise association, the $\log_{10}$ ratios of the affinity values in each dataset (i.e., Aff_{FUS}:Aff_{CP}) were calculated. Ratios ≤ -0.5 were flagged as being supported by FUSION, ratios ≥ 0.5 were flagged as being supported by CP, and all other associations were flagged as being supported by both. All data processing and statistical analyses for FUSION and SNF were carried out using the statistical platform R (http://www.R-project.org).

Integration of Metabolomics to SNF, CP, and FUSION Data. Integration of the basketed metabolomics data was performed similarly to previous studies (21). This approach treats every observed MS feature or basket as an individual chemical entity and asks the question, on average, what biological phenotype is expected when cells in the high-content assays are treated with compound? Each basket is treated as an object and assigned five numeric descriptors or attributes specific to the biological data acquisition: CP Cluster Score, CP Activity Score, FUSION Cluster Score, FUSION Activity Score, and SNF Cluster score. The Cluster Score is computed using the NXN similarity matrices from each assay the combined SNF similarity matrix and is simply the average of the nondiagonal values of the sub NXN matrix consisting of all the natural product fractions containing that basket. These descriptors are then exported as a table that can be used for discovery and visualized using tools such as the custom Bokeh server (*SI Appendix*, Supplementary Note 4 and Fig. S11).

Compound Isolation and Bioactivity Assays. Procedures for extraction, isolation, and structural elucidation of natural products from bioactive fractions are described in *SI Appendix*, Supplementary Notes S5 and S6. Immunoblotting and gene expression assay procedures are described in *SI Appendix*, Supplementary Note 3.

Data, Materials, and Software Availability. Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contacts, John B. MacMillan (jomacmil@ucsc.edu) and Roger G. Linington (rliningt@sfu.ca).

ACKNOWLEDGMENTS. This research was supported by NIH grants U41 AT008718 to R.G.L, M.A.W and J.B.M (NCCIH and ODS) and R01 CA149833 to J.B.M (NCI). E.A.M was supported by a NIH training grant 5T32GM8203-27. R.M.V was supported by CPRIT training grant RP140110 and NIH training grant 5T32 CA124334-09. A.F.S was supported by NSF GRF#2017247469. We would like to thank the UC Santa Cruz Chemical Screening Center (RRID: SCR 021114) for support of this work.

Author contributions: S.K.H., T.N.C., K.L.K., E.A.M., A.F.S., R.M.V., R.S.L., M.A.W., R.G.L., and J.B.M. designed research; S.K.H., T.N.C., K.L.K., E.A.M., W.B., A.F.S., A.K., F.P.J.H., F.C.-N., S.L., A.L., R.M.V., J.L., and S.W. performed research; E.A.M. and J.L. contributed new reagents/analytic tools; S.K.H., T.N.C., K.L.K., E.A.M., W.B., A.F.S., A.K., F.P.J.H., F.C.-N., A.L., M.A.W., R.G.L., and J.B.M. analyzed data; and S.K.H., T.N.C., E.A.M., R.S.L., M.A.W., R.G.L., and J.B.M. wrote the paper.

Author affiliations: [a]Department of Cell Biology, University of Texas Southwestern Medical Center, Dallas, TX 75390; [b]Department of Chemistry, Simon Fraser University, Burnaby, BC V5A 1S6, Canada; [c]Department of Chemistry, University of California Santa Cruz, Santa Cruz, CA 95064; [d]Department of Biochemistry, University of Texas Southwestern Medical Center, Dallas, TX 75390; and [e]Department of Bioinformatics, University of Texas Southwestern Medical Center, Dallas, TX 75390

Competing interest statement: The authors declare a competing interest. Michael White is employed at IDEAYA Biosciences and Elizabeth McMillan at Odyssey Therapeutics, Inc. Rachel Vaden is a current employee of Treeline Biosciences, Inc. Kenji Kurita is a current employee of Genentech, Inc.

1. G. T. Carter, Natural products and Pharma 2011: Strategic changes spur new opportunities. *Nat. Prod. Rep.* **28**, 1783–1789 (2011).
2. B. K. Wagner, S. L. Schreiber, The power of sophisticated phenotypic screening and modern mechanism-of-action methods. *Cell Chem. Biol.* **23**, 3–9 (2016).
3. C. J. Schulze *et al.*, Function-first lead discovery: Mode of action profiling of natural product libraries using image-based screening. *Chem. Biol.* **20**, 285–295 (2013).
4. W. Zheng, N. Thorne, J. C. McKew, Phenotypic screens as a renewed approach for drug discovery. *Drug. Discov. Today* **18**, 1067–1073 (2013).
5. J. G. Moffat, F. Vincent, J. A. Lee, J. Eder, M. Prunotto, Opportunities and challenges in phenotypic drug discovery: An industry perspective. *Nature* **16**, 531–543 (2017).
6. J. Lamb *et al.*, The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929–1935 (2006).
7. A. Subramanian *et al.*, A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* **171**, 1437–1452.e17 (2017).
8. M. B. Potts *et al.*, Using functional signature ontology (FUSION) to identify mechanisms of action for natural products. *Sci. Signal.* **6**, ra90 (2013).
9. M. A. Bray *et al.*, Cell painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nat. Protoc.* **11**, 1757–1774 (2016).
10. M. A. Bray *et al.*, A dataset of images and morphological profiles of 30 000 small-molecule treatments using the cell painting assay. *Gigascience* **6**, 1–5 (2017).
11. M. H. Woehrmann *et al.*, Large-scale cytological profiling for functional analysis of bioactive compounds. *Mol. Biosyst.* **9**, 2604–2617 (2013).
12. A. B. Parsons *et al.*, Exploring the mode-of-action of bioactive compounds by chemical-genetic profiling in yeast. *Cell* **126**, 611–625 (2006).
13. M. Schenone, V. Dančík, B. K. Wagner, P. A. Clemons, Target identification and mechanism of action in chemical biology and drug discovery. *Nat. Chem. Biol.* **9**, 232–240 (2013).
14. U. Rix, G. Superti-Furga, Target profiling of small molecules by chemical proteomics. *Nat. Chem. Biol* **5**, 616 (2009).
15. G. E. Croston, The utility of target-based discovery. *Expert Opin. Drug Discov.* **12**, 427–429 (2017).
16. Y. Hu *et al.*, Discoipyrroles A-D: Isolation, structure determination, and synthesis of potent migration inhibitors from bacillus hunanensis. *J. Am. Chem. Soc.* **135**, 13387–13392 (2013).
17. M. B. Potts *et al.*, Mode of action and pharmacogenomic biomarkers for exceptional responders to didemnin B. *Nat. Chem. Biol.* **11**, 401–408 (2015).
18. R. Vaden, N. Oswald, M. Potts, J. MacMillan, M. White, FUSION-guided hypothesis development leads to the identification of N6, N6-dimethyladenosine, a marine-derived akt pathway inhibitor. *Mar. Drugs* **15**, 75 (2017).
19. B. Das *et al.*, A functional signature ontology (FUSION) screen detects an AMPK inhibitor with selective toxicity toward human colon tumor cells. *Sci. Rep.* **8**, 3770 (2018).
20. Z. E. Perlman *et al.*, Multidimensional drug profiling by automated microscopy. *Science* **306**, 1194–1198 (2004).
21. K. L. Kurita, E. Glassey, R. G. Linington, Integration of high-content screening and untargeted metabolomics for comprehensive functional annotation of natural product libraries. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 11999–12004 (2015).
22. B. Wang *et al.*, Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* **11**, 333–337 (2014).
23. M. H. Rohban, H. S. Abbasi, S. Singh, A. E. Carpenter, Capturing single-cell heterogeneity via data fusion improves image-based profiling. *Nat. Commun.* **10**, 2082 (2019).
24. N. El-Hachem, *et al.,* Integrative cancer pharmacogenomics to infer large-scale drug taxonomy. *Cancer Res.* **77**, 3057–3069 (2017).
25. Y. Xiao *et al.*, Comprehensive metabolomics expands precision medicine for triple-negative breast cancer. *Cell Res.* **32**, 477–490 (2022).
26. M. Sinkala, N. Mulder, D. Martin, Machine learning and network analyses reveal disease subtypes of pancreatic cancer and their molecular characteristics. *Sci. Rep.* **10**, 1212 (2020).
27. S. Bhalla *et al.*, Patient similarity network of newly diagnosed multiple myeloma identifies patient subgroups with distinct genetic features and clinical implications. *Sci Adv.* **7**, eabg9551 (2021).
28. COvid-19 Multi-omics Blood ATlas (COMBAT) Consortium, *et al., A blood atlas of COVID-19 defines hallmarks of disease severity and specificity*. **185**, 916–938.e58 (2022).
29. Y. Raita *et al.*, Integrated omics endotyping of infants with respiratory syncytial virus bronchiolitis and risk of childhood asthma. *Nat. Commun.* **12**, 3601 (2021).
30. M. M. Aogáin *et al.*, Integrative microbiomics in bronchiectasis exacerbations. *Nat. Med.* **27**, 688–699 (2021).
31. P. Skowron *et al.*, The transcriptional landscape of Shh medulloblastoma. *Nat. Commun.* **12**, 1749 (2021).
32. G. R. Jacobs *et al.*, Integration of brain and behavior measures for identification of data-driven groups cutting across children with ASD, ADHD, or OCD. *Neuropsychopharmacologyl* **46**, 643–653 (2021).
33. G. Ramaswami *et al.*, Integrative genomics identifies a convergent molecular subtype that links epigenomic with transcriptomic differences in autism. *Nat. Commun.* **11**, 4873 (2020).
34. T. N. Clark *et al.*, Interlaboratory comparison of untargeted mass spectrometry data uncovers underlying causes for variability. *J. Nat. Prod.* **84**, 824–835 (2021).
35. F. C. Neto, T. N. Clark, N. P. Lopes, R. G. Linington, Evaluation of ion mobility spectrometry for improving constitutional assignment in natural product mixtures. *J. Nat. Prod.* **85**, 519–529 (2022).
36. P. Kirk, J. E. Griffin, R. S. Savage, Z. Ghahramani, D. L. Wild, Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics* **28**, 3290–3297 (2012).
37. B. J. Frey, D. Dueck, Clustering by passing messages between data points. *Science* **315**, 972–976 (2007).
38. J. Kim *et al.*, XPO1-dependent nuclear export is a druggable vulnerability in KRAS-mutant lung cancer. *Nature* **538**, 114–117 (2016).
39. T. S. Bugni *et al.*, Marine natural product libraries for high-throughput screening and rapid drug discovery. *J. Nat. Prod.* **71**, 1095–1098 (2008).
40. J. A. van Santen *et al.*, The natural products atlas: An open access knowledge base for microbial natural products discovery. *ACS Cent. Sci.* **5**, 1824–1833 (2019).
41. M. Wang *et al.*, Sharing and community curation of mass spectrometry data with Global natural products social molecular networking. *Nat. Biotechnol.* **34**, 828–837 (2016).
42. Y. Hoshikawa H. J. Kwon, M. Yoshida, S. Horinouchi, T. Beppu, Trichostatin A induces morphological changes and gelsolin expression by inhibiting histone deacetylase in human carcinoma cell lines. *Exp. Cell Res.* **214**, 189–197 (1994).
43. M. Wood, S. Rymarchyk, S. Zheng, Y. Cen, Trichostatin A inhibits deacetylation of histone H3 and p53 by SIRT6. *Arch. Biochem. Biophys.* **638**, 8–17 (2018).
44. K. Takada *et al.*, Surugamides A -E, Cyclic octapeptides with four d-amino acid residues, from a marine streptomyces sp.: LC–MS-aided inspection of partial hydrolysates for the distinction of d- and l-amino acid residues in the sequence. *J. Org. Chem.* **78**, 6746–6750 (2013).
45. F. A. Dick, S. M. Rubin, Molecular mechanisms underlying RB protein function. *Nat. Rev. Mol. Cell Biol.* **14**, 297–306 (2013).
46. M. Wibowo, L. Ding, Chemistry and biology of natural azoxy compounds. *J. Nat. Prod.* **83**, 3482–3491 (2020).
47. R. Rathore *et al.*, Metabolic compensation activates pro-survival mTORC1 signaling upon 3-phosphoglycerate dehydrogenase inhibition in osteosarcoma. *Cell Rep.* **34**, 108678 (2021).
48. M. H. Medema *et al.*, Minimum Information about a biosynthetic gene cluster. *Nat. Chem. Biol.* **11**, 625–631 (2015).
49. P. Shannon *et al.*, Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).