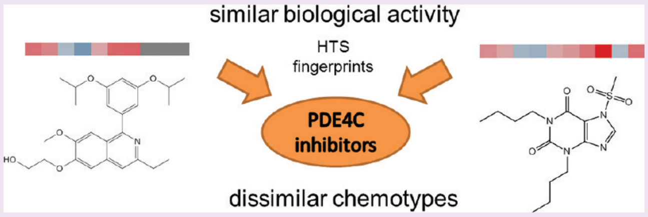
ACS chemical biology

Rethinking Molecular Similarity: Comparing Compounds on the Basis of Biological Activity

Paula M. Petrone,[†] Benjamin Simms,[†] Florian Nigsch, Eugen Lounkine, Peter Kutchukian, Allen Cornett, Zhan Deng, John W. Davies, Jeremy L. Jenkins,*[,‡] and Meir Glick*[,‡]

Center for Proteomic Chemistry, Novartis Institutes for Biomedical Research Inc., 250 Massachusetts Avenue, Cambridge, Massachusetts 02139, United States

Supporting Information



ABSTRACT: Since the advent of high-throughput screening (HTS), there has been an urgent need for methods that facilitate the interrogation of large-scale chemical biology data to build a mode of action (MoA) hypothesis. This can be done either prior to the HTS by subset design of compounds with known MoA or post HTS by data annotation and mining. To enable this process, we developed a tool that compares compounds solely on the basis of their bioactivity: the chemical biological descriptor "high-throughput screening fingerprint" (HTS-FP). In the current embodiment, data are aggregated from 195 biochemical and cell-based assays developed at Novartis and can be used to identify bioactivity relationships among the in-house collection comprising ~1.5 million compounds. We demonstrate the value of the HTS-FP for virtual screening and in particular scaffold hopping. HTS-FP outperforms state of the art methods in several aspects, retrieving bioactive compounds with remarkable chemical dissimilarity to a probe structure. We also apply HTS-FP for the design of screening subsets in HTS. Using retrospective data, we show that a biodiverse selection of plates performs significantly better than a chemically diverse selection of plates, both in terms of number of hits and diversity of chemotypes retrieved. This is also true in the case of hit expansion predictions using HTS-FP similarity. Sets of compounds clustered with HTS-FP are biologically meaningful, in the sense that these clusters enrich for genes and gene ontology (GO) terms, showing that compounds that are bioactively similar also tend to target proteins that operate together in the cell. HTS-FP are valuable not only because of their predictive power but mainly because they relate compounds solely on the basis of bioactivity, harnessing the accumulated knowledge of a high-throughput screening facility toward the understanding of how compounds interact with the proteome.

A central goal of chemical biology is to understand the underlying mechanisms of biological systems by their response to certain compounds. With the advent of high-throughput screening (HTS), relationships between compounds and biological entities have been studied on an enormous scale. Despite the accessibility of large databases of proprietary and public data, there is urgent need for methods that facilitate the interrogation and mining of this information to build hypotheses on the mode of action of compounds (MoA) and also on the cellular processes perturbed in phenotypic screens. Typically, MoA hypotheses are based on the assumption that structurally chemical compounds are likely to share similar properties and will bind to the same group of proteins.[1] Chemometric approaches that rely on the use chemical descriptors to build quantitative structure—activity relationships (QSAR) have been geared toward predicting activity against a target. One reason why these models often do not live up to expectations is the rugged and high dimensional nature of the activity landscape.[2,3] Furthermore, by construction chemical similarity cannot explain the activity of a compound against a specific pathway or groups of pathways that may or may not be known. Compounds that incur similar phenotypes and yet are structurally diverse are therefore often overlooked because the traditional searching methods do not take into account the biological similarity of compounds.

Recently, there has been an increasing awareness that the cellular response of a compound can be described without the chemical structure, focusing instead on the chemical biology of the compound through its interactions with the proteome. This has largely entailed organizing the screening data for a compound obtained against a panel of targets and/or cell lines into a fingerprint that is used to describe that compound. In seminal work performed at the NCI, the growth inhibition of 60 cancer cell lines was evaluated for a panel of compounds and incorporated into a fingerprint that could be used to compare the similarity of compounds.[4] These fingerprints were then employed as input to predict mechanism of action with neural networks[5] and to search for target specific compounds.[6] Kauvar *et al.*,[7] on the other hand, derived "affinity fingerprints" based on a compound's interaction with specific targets and used these fingerprints to predict ligand affinities. In a subsequent study, affinity fingerprints were then assessed for the design of

Received: March 6, 2012
Accepted: May 18, 2012
Published: May 18, 2012

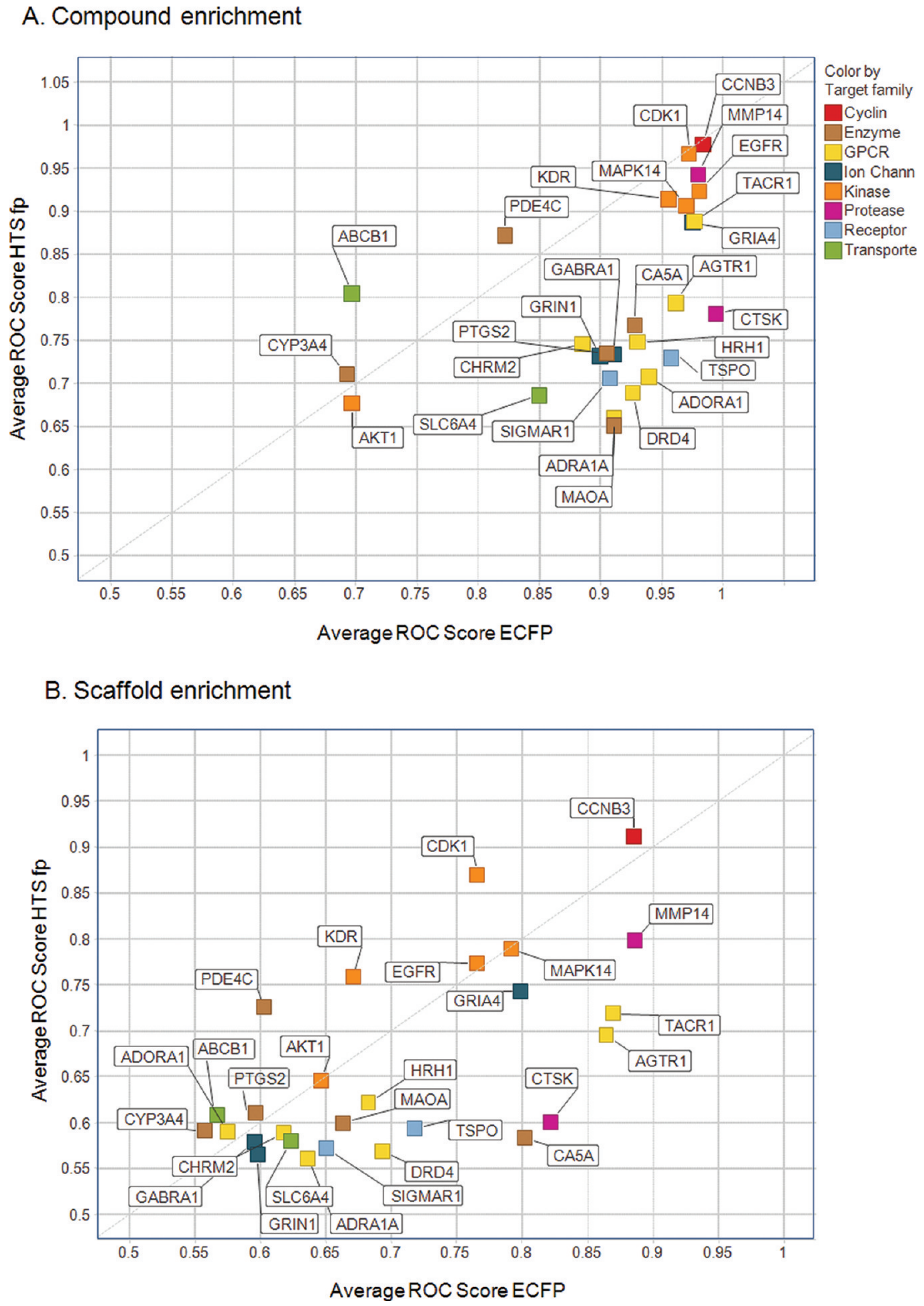


Figure 1. (A) Compound recall and (B) scaffold recall benchmark comparing the performance of HTS-FP and ECFP4 similarity methods by means of ROC scores. For each target, the *y*-axis plots the HTS-FP scores, and the *x*-axis the ECFP4 ROC scores. The color code identifies target families.

diverse sets of compounds and sets of compounds with enriched activity for a given target.[8] In a similar vein, Fliri *et al.*[9,10] derived fingerprints based on a compound's interaction with specific targets, which they termed "biospectra," and predicted simultaneous interactions of new molecules with the proteome. The authors used percent inhibition values of 1,597 compounds against 92 targets, and concluded that comparing biological activity profiles of molecules provided an unbiased means for establishing quantitative relationships between chemical structure and broad biological effects. Further studies by Fliri *et al.*[11] showed that drug side effects could also be linked to their activity spectra and therefore clinical effect profiles of drugs could be predicted. Plouffe *et al.*[12] used activity profiles derived from 131 high-throughput cellular *and* biochemical screens to predict targets for antimalarial drugs. In their study, compounds with known and unknown targets

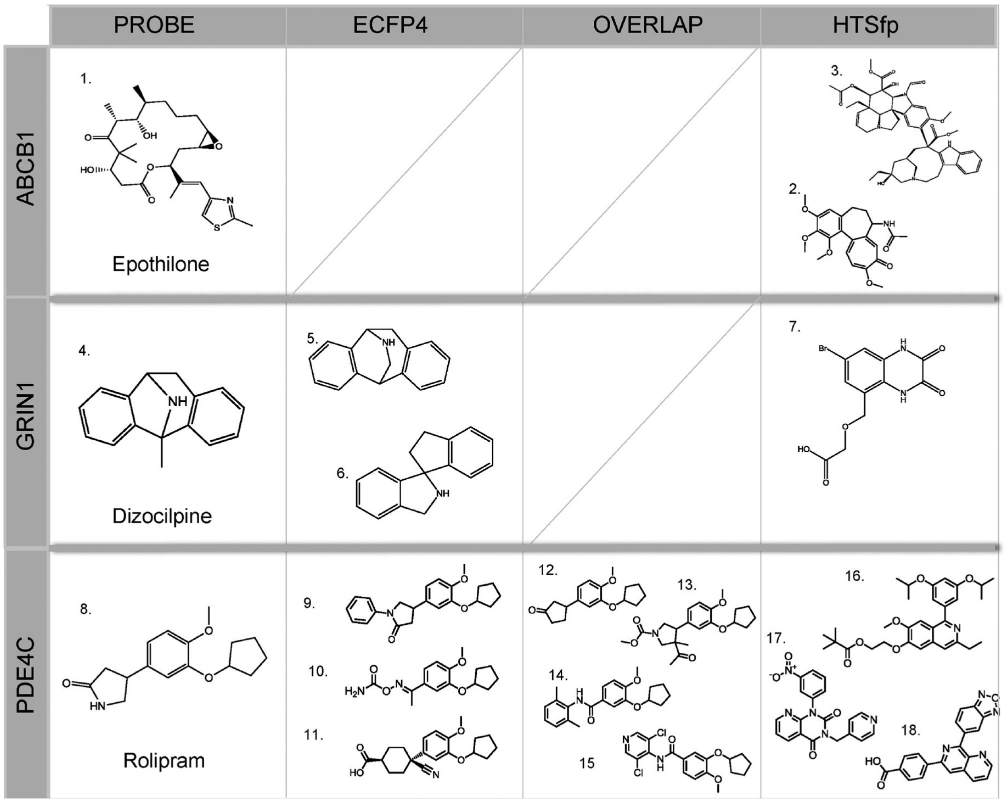


Figure 2. Examples of compounds recalled with ECFP4 and HTS-FP given a set of reference compounds belonging to a particular scaffold. The OVERLAP compounds are those compounds retrieved by both methods. Compounds C1, C4, and C5 represent the probe scaffold sets.

were clustered on the basis of their activity profiles, and a "guilt by association" target prediction method based on target enrichment within clusters was used. In a similar way, Cheng *et al.*[13] recently used the bioactivity profile of compounds to predict their biological targets. Using bioactivity profiles derived against the NCI-60 cell lines, 45% of compound-target associations annotated in public databases could be affirmed. Their results also suggest the possible application of bioactivity similarity profiles in searching for novel chemical matter or "scaffold hopping". At the crux of these reports is the finding that a broad panel of single concentration, HTS data can be used to derive meaningful relationships between compounds without any information about chemical structure.

In our work, we investigate the applicability of bioactivity comparisons between compounds as a means for virtual screening and library design on an unprecedented scale. We have developed a set of biological descriptors, termed "high-throughput screening fingerprint" (HTS-FP), which translates the wealth of HTS data into a form that can be readily interrogated by computational methods. We use data from 195 assays developed at Novartis over a time frame of 10 years, which cover a broad variety of protein families and technologies including fluorescence intensity, radioactivity, and mass spectrometry (Supplementary Tables S1−3). By comparing the similarity of compounds based on their HTS-FP, we elucidate bioactivity relationships among the in-house collection of ~1.5 million compounds. We have intentionally incorporated into our fingerprint both biochemical and cell-based assays for various reasons. The clear advantage of using biochemical assay data is that it adds target-specificity to the fingerprint. This is especially powerful in cases were traditional descriptors would not perform well, as would be the case for two structurally different compounds that inhibit the same enzyme by binding to different pockets. The cell-based assay data, in turn, provide another layer of complexity to the fingerprint. The advantage of utilizing cell-based assays is that many of these assays target an entire functional pathway or high-level phenotype, rather than the ability of a molecule to bind to a specific protein. Therefore, comparing compounds' cell-based activity profiles can lead to the identification of compounds that produce a similar phenotype yet not necessarily operate through the same mode of action. Conversely, comparing activity profiles of independent biochemical assays could not easily lead to such associations.

The objective of this work is to demonstrate the value of the HTS-FP for various applications such as virtual screening, scaffold hopping, and subset design. When applied to virtual screening, HTS-FP has the capability of discovering novel active chemotypes for a phenotype (scaffold hopping)[14] because it uses no structural information when comparing compounds. Also, HTS-FP can be employed to generate subsets of *biodiverse* compounds, allowing for more efficient identification of active compounds when entire libraries cannot be screened. Finally, we demonstrate that clustering compounds by means of HTS-FP is biologically meaningful. By looking at the GO term enrichments within HTS-FP clusters, we show that compounds grouped on the basis of HTS-FP tend to modulate protein targets with related biological function.

Taken together, we show that HTS-FP can transcend the limitations of molecular structure similarity comparisons, capturing information on the bioactivity of compounds and their impact on cellular pathways regardless of chemical structure. We establish a method by which the vast data

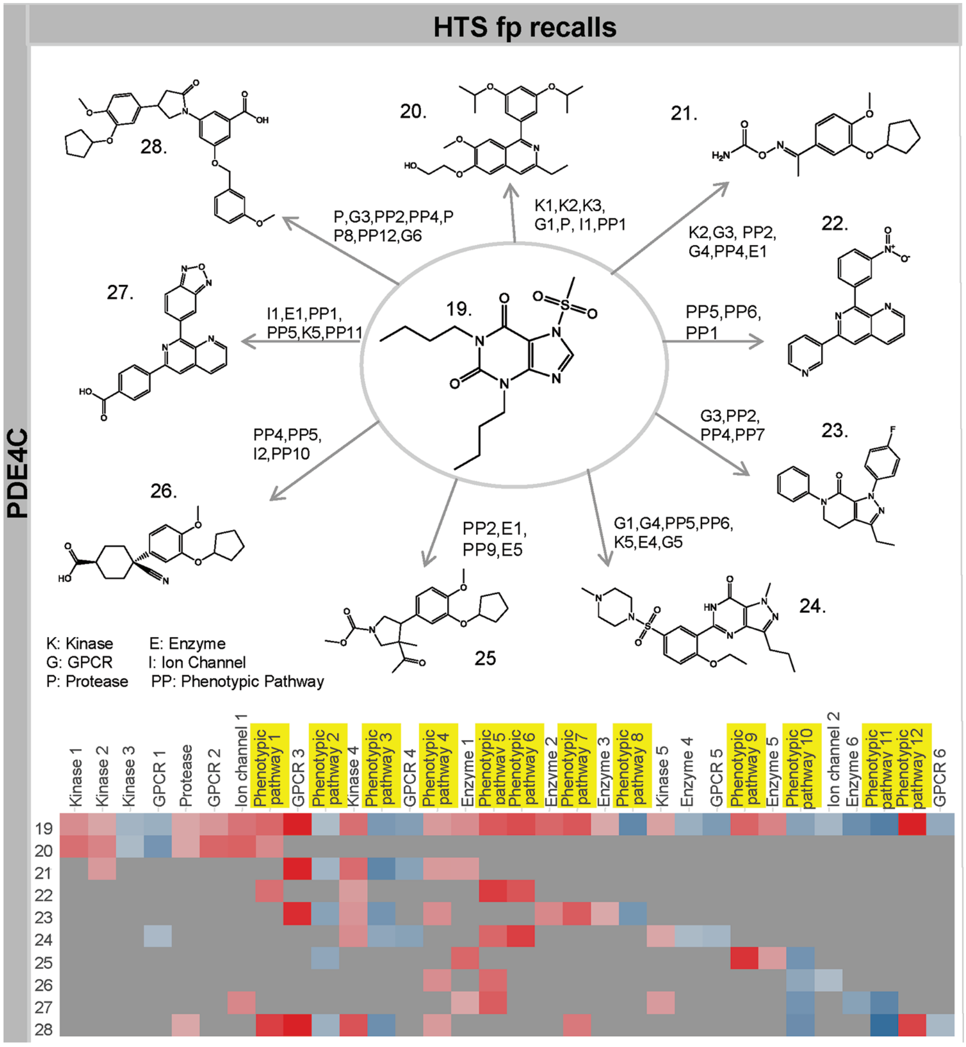


Figure 3. (Top panel) Compounds C20−C28 recalled with a xanthine derivative scaffold, of which C19 is a representative, and HTS-FP similarity. ECFP4 yields no recalls for this scaffold. The recalls were retrieved *via* relevant HTS-FP assays highlighted in the lower panel, and written over the arrows. Highlighted in yellow are the phenotypic pathway assays. The protocol for the selection of relevant assays in the lower panel is detailed in Supporting Information.

generated by an HTS facility can be harnessed to help understand past experiments as well as to generate hypotheses about current and future screening results.

■ RESULTS AND DISCUSSION

We explain the HTS-FP and define a metric for similarity and clustering (Methods). Three applications of HTS-FP are described: (1) virtual screening and scaffold hopping, (2) biodiversity selection of HTS plates, and (3) biological relevance of HTS-FP clusters.

1. HTS-FP Applied to Virtual Screening and Scaffold Hopping. A primary goal of virtual screening is the identification of bioactive molecules against a target given hit compounds used as reference. Traditionally, structurally similar compounds are sought in the assumption that they will also show similar biological properties. By contrast, HTS-FP directly identify compounds that show similar biological profiles.

In Figure 1, we compare the performance of a state-of-the-art structural similarity approach (ECFP4, see Methods) against HTS-FP both in terms of hit rate (compound recall) and diversity of chemotypes retrieved (scaffold recall). Performance is assessed by receiver-operator curves (ROC) scores (Methods). Random selection corresponds to a ROC score of 0.5, whereas a score of 1.0 means perfect recall of active compounds.

Overall, HTS-FP performance depends on the target class but ranges from acceptable (ROC = 0.66) to excellent (ROC = 0.98). In compound recall (Figure 1A), however, ECFP4 performs consistently better than HTS-FP throughout the target families, with kinases having the highest scores, and ABCB1 transporter being a noticeable exception. This is probably due to the fact that the active chemotypes for kinases are extensively studied, well-defined, and populated in the library. Overall, chemical similarity works extremely accurately (most ROC scores > 0.90) when several chemotypes are combined together as probes (Figure 1A) and thus the hit rate in compound recall is higher than that obtained with HTS-FP similarity. On the other hand, in the case of scaffold recall (Figure 1B) both HTS-FP and ECFP4 perform evenly with HTS-FP outperforming for a few targets, especially in the case of kinases and enzymes. Importantly, the performance of HTS-FP is independent of chemical structure by design; therefore HTS-FP excels at recalling remarkably diverse chemical matter as illustrated in Figures 2 and 3.

The ABCB1 transporter is a major efflux pump, involved in multidrug resistance. Figure 2 highlights the use of the natural product epothilone (C1),[15] a known anticancer agent and inhibitor of ABCB1,[16] as a probe to retrieve new scaffolds. While chemical similarity fails to retrieve actives in the top 1% ranked list, HTS-FP instead retrieves active compounds with scaffolds that have no structural overlap with epothilone and may thus be smaller in size and present different physicochemical properties (*e.g.*, drug-like molecule C2). Similarly, dizocilpine (C4), a non-competitive antagonist of the NMDA receptor, retrieves very structurally different inhibitors such as quinoxalines, whereas the scaffolds retrieved by chemical similarity are very similar to each other.

In the case of PDE4C, chemical similarity searches using the anti-inflammatory drug rolipram as a probe retrieve only compounds that have at least one major chemical group in common with rolipram (*i.e.*, the dimethoxy phenyl group, cyclopentane, or pyrrol ring). Even though there is some overlap in the compounds retrieved by chemical similarity and HTS-FP, HTS-FP is able to recall compounds in broadly different patent and chemical spaces. Figure 3 shows active recalls against PDE4C found in the top 1% HTS-FP similarity to probe C19, a xanthine derivative. While chemical similarity yields no active recalls in the top 1%, HTS-FP instead finds compounds C20−28 bearing minimal similarity to the probe. The lower panel shows a visualization of the HTS-FP limited to the assays that contribute the most to the HTS-FP similarity. Based on this heat map, arrows describe the relevant assays in common between the probe and each of the recalls. For example, the probe finds recall C25 through enzymes E1 and E5 and relates to recall C20 through kinases K1−3. This indicates that the recalls participate in different cell mechanisms and probably explains the differences in chemical structure among them and with the probe.

In this PDE4C example, phenotypic pathway assays have a significant contribution to the HTS-FP similarity as compared to any other target class. A result from this is that compounds recalled do not necessarily share the same targets or pathways and range from specific to non-specific inhibitors of PDE4C. Even if many xanthine derivatives tend to behave as non-specific PDE inhibitors,[17] public data shows probe C19 is specific to PDE4C. Recalls C25, C27, and C28 are also specific to PDE4, whereas compound C21 is non-specific (targeting all PDE3−5), and xanthine derivative C24 targets all PDE3−9.

We explore to what extent a bioactivity-based similarity method (HTS-FP) correlates with a chemical structure-based method (ECFP4) (Figure 4) by comparing similarity values for a random sample of ∼10% (100,000) of the pairwise interactions of molecules from the previous study. Figure 4 shows that the linear correlation is poor ($R = 0.022$); however, for some pairs both ECFP4 and HTS-FP similarities are high (top right quadrant). Looking at the conditional distributions (Supporting Information 2) for ECFP4 and HTS-FP similarities, we find that even if the metrics are not linearly correlated, pairs with similar biological fingerprints tend to have similar chemical structures. However, there are a lot of pairs for which ECFP similarity is low and yet the compounds have similar bioactivities. We attribute this to the fact that HTS-FP is able to capture information from a phenotype that consists of multiple targets and therefore multiple chemotypes.

In practice, ECFP and HTS-FP present complementary advantages. ECFP4, on the one hand, is very reliable in both recalling compounds and even scaffold hopping when the set of probes is chemically diverse. HTS-FP, on the other hand, is independent of the chemical structure of the probes and requires *a priori* bioactivity data on the compounds, and its somewhat weaker recall is greatly compensated for by its inherent ability to yield completely novel chemical matter that is inferred only through biology.

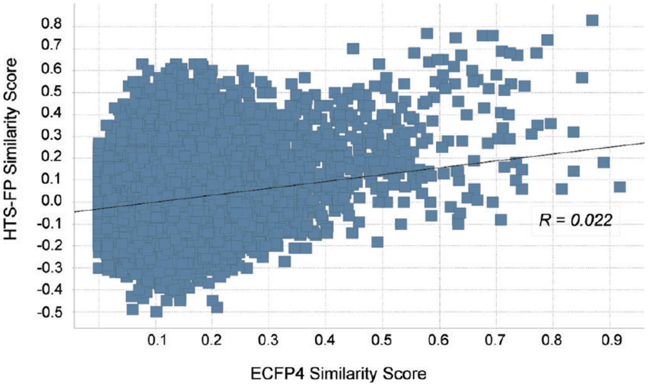


Figure 4. HTS-FP similarity as a function of ECFP4 similarity for 100,000 random pairs of the 10,259 molecules considered in the recall studies.

2. Plate Diversity Selection. State-of-the-art HTS screening systems currently enable screening of 1−5 million compounds in a few weeks.[18] Yet factors other than automation, namely, cost, assay throughput, reagent availability, and time, often limit the size of the compound-screening library. Past studies by Sukuru *et al.*[18] show that screening sets of compounds with increased chemical diversity (*i.e.*, focused screening) present an attractive option to full HTS campaigns. Hit compounds from plates carefully chosen can be expanded using virtual screening approaches into an automated "cherry-picked" set for follow-up tests (Figure 5). We carry out a retrospective validation protocol to evaluate the hit rate of a reduced screening set (plate diversity selection) in comparison to the complete compound collection. The key to this approach, namely, screening a subset of compounds followed by cherry-picked expansion, is to maximize the information content obtained from the initial screen and to exploit this information during expansion. Here we demonstrate the advantages of using HTS-FP to design the initial screen and subsequently expand upon the initial hits.

After clustering the entire collection by their HTS-FPs, we pick a reduced set of 710 384-well HTS plates[19] that maximizes the biologically diversity of compounds, *i.e.*, the number of HTS-FP clusters the compounds fall into. We use this diverse plate set for an initial virtual screening, followed by 10,000-compound cherry-picking expansion using HTS-FP similarity. We compare the performance of this *biodiverse* set of plates to a *chemically* diverse plate set selected instead by maximizing the amount of different scaffolds and expanding the hits using chemical similarity (ECFP4). A random set of HTS plates serves as a control. Random selection of plates yields on average 15% of the collection actives, which is expected since 710 plates represent 14% of the collection (5029 plates).

The performance of each of the plate diversity selection methods was assessed across 13 in-house assays. Figure 6 shows the percentage of active compounds retrieved in the 710-plate diversity library (hereafter called "seeds") and the yield in active hits obtained by expanding the "seeds" with virtual screening by means of 10,000 cherry-picked compounds.

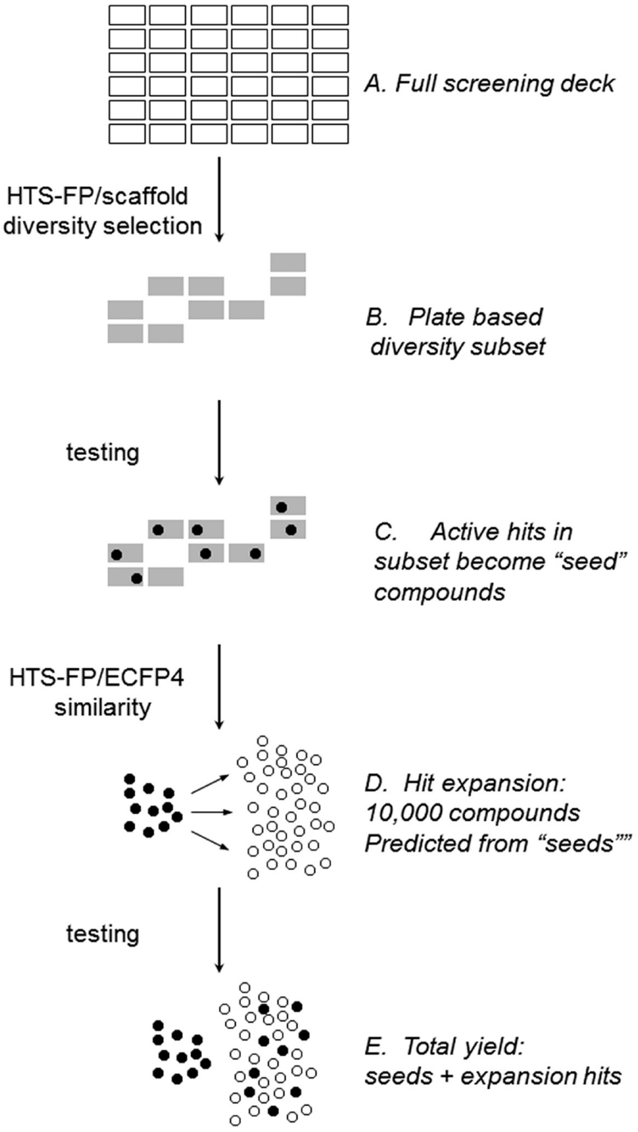

A. Full screening deck
HTS-FP/scaffold diversity selection
B. Plate based diversity subset
testing
C. Active hits in subset become "seed" compounds
HTS-FP/ECFP4 similarity
D. Hit expansion: 10,000 compounds Predicted from "seeds""
testing
E. Total yield: seeds + expansion hits

Figure 5. Flowchart of the protocol for plate diversity selection and hit expansion. Using either HTS-FP or scaffold clusters of the screening collection (A), a 750 diversity subset of plates is selected (B) for initial screening. Actives ("seeds") found in that subset (C) are then expanded *in silico* using either HTS-FP or ECFP4 similarity methods (D), and the top ranking 10,000 compounds are sent for a follow-up screening (E).

Regarding the yield in "seed" compounds (Figure 6A), HTS-FP diversity selection outperforms both ECFP4 and random selection in 10 of 13 assays, with the exception of GPCR, Ion Channel, and the Serine Protease assays. Regarding the total yield in actives after expansion, HTS-FP still proves more efficient in 12/13 of the assays with the exception of Kinase 3 assay. This could be due to the relative overrepresentation in the collection of structurally similar kinase inhibitors, inherently predisposing ECFP4 for high-accuracy retrieval. HTS-FP has outstanding performance in Ion Channel, Kinase 1, and Bacterial Pathway assays.

The diversity of scaffolds obtained from a screen is just as, if not more, important than the total number of hits, as this allows chemists to pursue different avenues during lead optimization. It would be reasonable to expect that plate selection based on chemical diversity is likely to yield more diverse scaffolds than its biodiverse counterpart. Indeed, when considering solely the number of active "seed" scaffolds, chemical diversity outperforms biological diversity in 8 of 13 assays (Figure 6B). Even so, after expansion, this trend is reversed and bioactivity-based methods (HTS-FP plate selection and expansion) outperform random or chemical structure based methods in all but one assay. This superior performance of HTS-FP in terms of scaffold diversity is likely due to its inherent capability of scaffold hopping, which results in the retrieval of a more diverse set of biologically active scaffolds upon expansion of initial hits. In addition, HTS-FP has the ability to avoid the selection of chemotypes that, in general, have no biological relevance, a capability of which ECFP4 is exempt being prone to activity cliffs,[3,20] *i.g.*, pairs of molecules that are structurally similar but exhibit large differences in activity. For hit expansion, HTS-FP consistently outperforms other methods in both compound retrieval and scaffold retrieval regardless of whether the seed compounds were selected on the basis of chemical diversity or biological diversity or randomly (Figure 6 and Supplementary Figure S8).

We show here that by rationally selecting 15% of the HTS library for initial screening and hit expansion on the basis of diverse biology, we could recover 37% of actives (39% active scaffolds) on average across the 13 assays. A similar protocol based on chemical structure yields only 29% of actives (29% active scaffolds). In conclusion, plate selection and hit expansion based on HTS-FP on average proves more efficient than chemical structure-based methods in terms of retrieving active compounds and diverse scaffolds.

3. Biological Relevance of HTS-FP Clusters. In general, it is accepted that similar chemical structure of compounds corresponds to related biological effect.[21] Correspondingly, by calculating gene enrichments on groups of structurally similar compounds, we observe that compounds that share the same chemotype tend to also modulate the same targets. We show gene enrichments[22] (p-value < 0.05) for scaffold clusters (Supporting Information 7). In a similar manner, if we cluster on the basis of compounds' bioactivity using HTS-FP (Methods), we also find that target genes are enriched in HTS-FP clusters (Supplementary Figure S9), which is expected since the HTS-FP clustering protocol ensures that cluster members modulate similar proteins in the biochemical assays and similar phenotypes in the cell-based assays.

Indeed, the main feature of HTS-FPs is the inclusion of cell-based assays that contribute with information about the compounds' effects on complex cellular systems and processes. In this section, we analyze the extent to which a clustering of compounds by HTS-FP reflects a grouping of compounds by related biology, not just by similar genes. By looking at the enrichment of Gene Ontology (GO) categories[23] (cellular components, biological processes, and molecular functions) within HTS-FP clusters, we want to assess how often bioactively similar compounds also have similar effects on cellular functions and processes.

GO-Term Enrichment. For compounds in each HTS-FP cluster, we identify relationships between compounds and genes using in-house and public chemogenomics data. Given the target genes per cluster, we can query for their associated GO terms[24] and evaluate the frequency of occurrence of a certain GO term within a cluster. GO terms "describe gene products in terms of their associated biological processes, cellular components, and molecular functions in a species-independent manner".[25] The GO term enrichment in HTS-FP

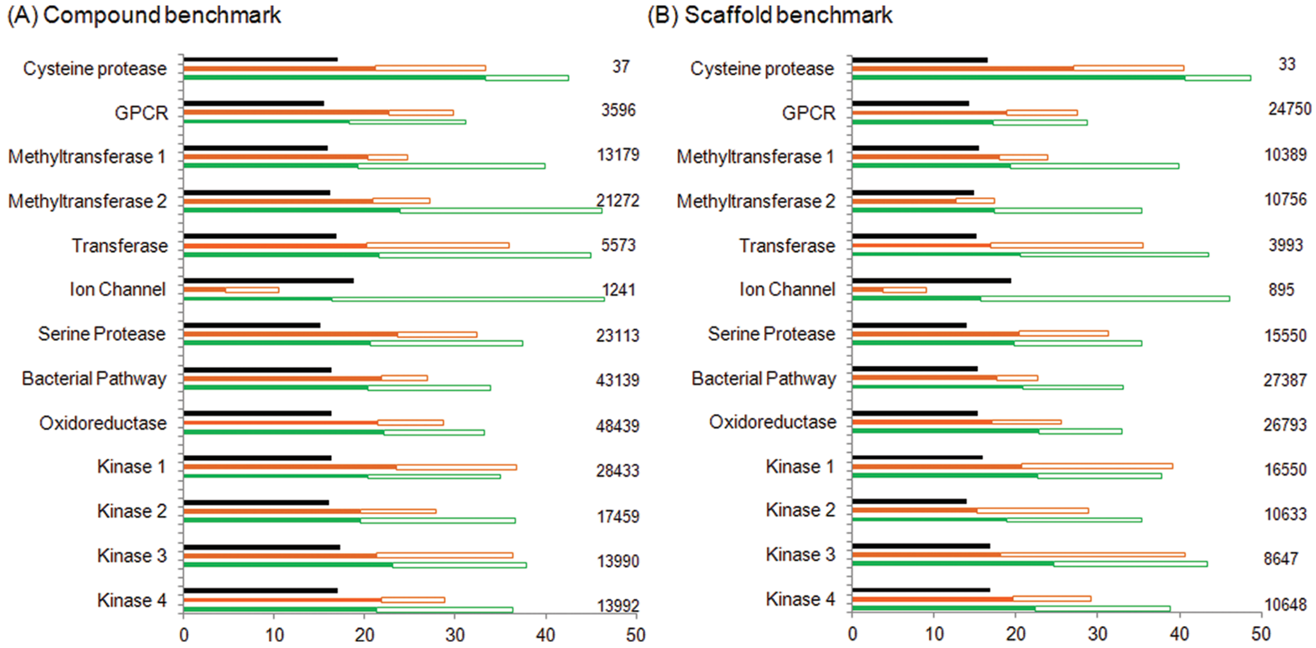


Figure 6. Plate diversity benchmark and hit expansion based on HTS-FP similarity methods (green), chemical structure (orange), and random plate selction (black). The filled boxes correspond to "seed" compounds collected from a diversity set selection corresponding to 13 different assays. The hollow boxes correspond to hit expansion from 10,000 cherry picks selected using ECFP4 similarity (orange) and HTS-FP similarity (green). Numbers on the right-hand column specify total amount of actives per assay.

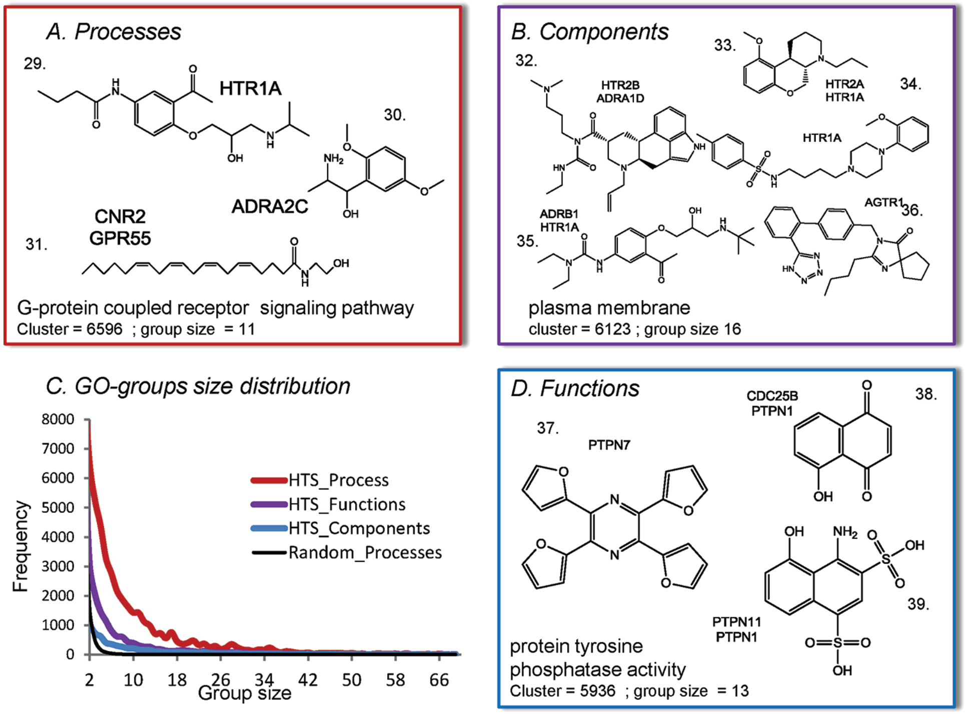


Figure 7. (A, B, D) Groups of compounds that belong to the same HTS-FP cluster and share the same GO term (biological processes (red), cellular components (green), and molecular functions (purple)) defined as GO-groups. (C) GO-group size distribution for each GO-term enriched in HTS clusters. In black, GO-group size distribution for random clusters. The distribution for random functions and components are comparable and smaller respectively and omitted here (Supplementary Figure S10).

clusters is compared to scaffold clusters and random clusters as controls.

We define as "GO-group" a set of compounds that belong to the same cluster and share the same enriched GO term. In Figure 7C and in more detail in Supplementary Figure S10, we show distributions of the observed GO-group sizes for HTS-FP clusters for all three GO categories. The area under the group size-distributions shows the number of GO-terms enriched by each clustering method (Figure 7 and Supplementary Figure S10). Clearly, HTS-FP clusters enrich for far more GO-terms than the random clusters, whose distribution quickly decays to zero for GO-groups larger than 13 compounds. We can see that

GO-groups vary in size, but small groups are very frequent even in random clusters. However, larger GO-groups (size >30) of bioactively similar compounds modulating same cellular processes and biological functions are inherent to HTS-FP clusters. Surprisingly, HTS-FP clustering yields many GO-groups that are larger than 50 compounds for all three GO categories.

Importantly, compounds in HTS-FP GO-groups do not always share the same chemical structure nor target the same genes, yet they participate in related cellular processes and functions (biologically similar). For example, Compounds C29−31 in Figure 7A are known to target the G-Protein Coupled Receptor (GPCR) signaling pathway but the reported targets (HTR1A, ADRA2C, CNR2, GPR55) belong to four different GPCR subfamilies. Methoxamine (C30) and acebutolol (C29) share a common substructure, whereas anandamide (C31) is a fatty acid. Similar chemical diversity can be observed for compounds in Figure 7B and D. These examples show that chemical similarity of compounds is not required for biological similarity and that HTS-FP could be used to target a certain phenotype in a focused screen.

We further explore the relationship between chemical similarity and biological similarity in both HTS-FP clusters and chemical clusters (Figure 8). Biological similarity is calculated on the basis of GO processes and functions modulated by compounds within a cluster; chemical similarity is assessed using ECFP4 (Methods). Chemical similarity in HTS-FP clusters tends to be lower than 0.5, whereas for scaffold clusters it is mostly above 0.5, which is anticipated since by design scaffold clusters have the same chemotype. Most HTS-FP clusters (57%) have higher biological similarity than would be expected in random clusters. However, scaffold clusters tend to have higher biological similarity than HTS-FP clusters. This is expected since many scaffold clusters originate from congeneric series especially designed to test the SAR of specific targets.

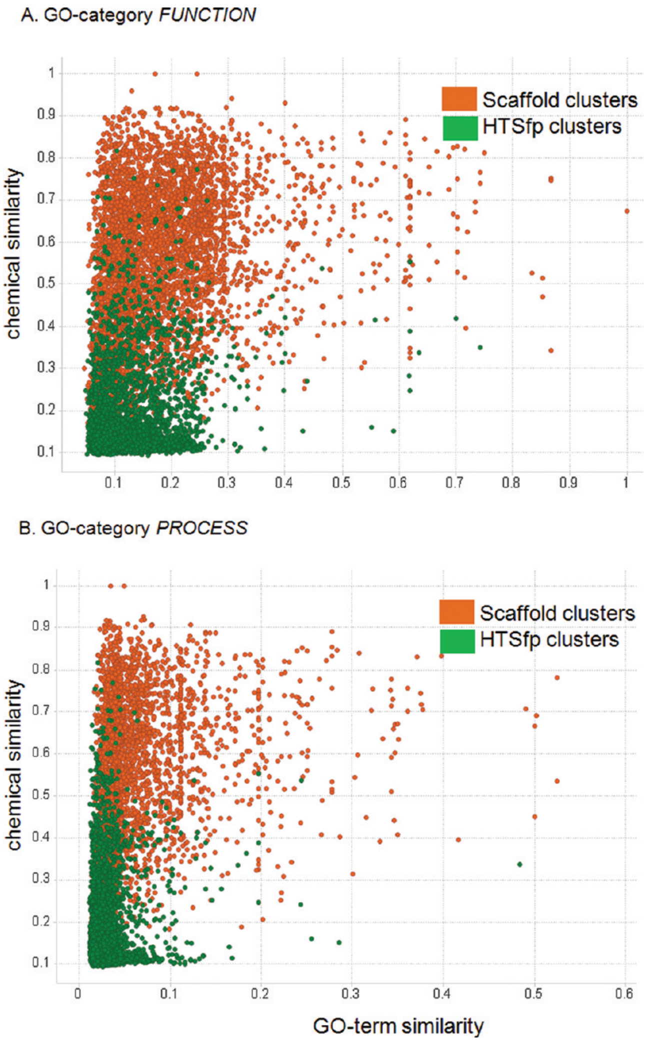


Figure 8. Mean chemical similarity *versus* GO-term similarity in HTS-FP (green) and scaffold clusters (orange), classified by GO categories (A) *functions* and (B) *processes*. Data points correspond to clusters that have higher biological similarity to what would be expected by random (p-value <0.05, Methods).

In short, we have shown that HTS-FP clusters enrich for GO terms, suggesting that HTS-FP groups compounds on the basis of their interactions with targets in the cell. HTS-FP clusters are a useful way to organize compound data because they capture bioactivity associations between molecules at many levels (*e.g.*, target, molecular function, biological process, or cellular component). HTS-FP clusters can be used, for example, to assess a phenotypic screen. Starting from a cell-based assay hitlist, a hypothesis can be generated. Potential pathway targets can be identified by looking at the enriched target genes within HTS-FP clusters where hit compounds are found. The relevant biological processes, functions, and components then can be annotated by looking into the common GO-groups among the hits. To test the hypothesis, additional compounds from the same HTS-FP clusters can be assayed in the phenotypic screen or against the predicted targets.

Conclusion. We have shown that fingerprints based on biological activity profiles can provide effective predictions on the cellular response of compounds without recourse to chemical structure similarity. The robustness of predictions using HTS-FP is verified in virtual screening benchmarks and in the hit expansion of seed compounds for cherry picking. We show that by introducing biodiversity in HTS libraries we can increase not only the hit rate but also the chemical diversity in hit compounds identified. With a workflow based on biodiversity selection and expansion using HTS-FP, we capture ~37% of an assay's actives by only screening 15% of the collection. Furthermore, the GO-term enrichment of HTS-FP clusters of compounds indicates that bioactively similar compounds tend to target genes that operate jointly in the cell. Previous studies by Keiser *et al.*[1] show that protein targets may be quantitatively related by the chemical similarity of their ligands. Here we show that the reverse is also true: even in the absence of chemical similarity, ligands may be quantitatively grouped by the biological closeness of their targets. In this context, HTS-FP therefore become a powerful way of finding those compounds that are bioactively but not structurally related. All in all, focusing on the bioactivity of compounds, HTS-FP present an alternative approach to similarity searches, revealing additional chemotypes that open new opportunities in chemical and patent space.

In summary, HTS-FP represent a heuristic way to systematize assay data from various sources into a machine learning protocol to make valuable predictions for compound similarity, selection, and target prediction. Its success relies on making the most of the accumulated knowledge of a screening facility, creating value out of both active and inactive compounds, and incorporating the differential activity of molecules across targets and pathways.

■ METHODS

1. HTS Fingerprints Definition. Z-scores of percent inhibition values are calculated for each of the 195 assays in the HTS-FP. The vector of a compound's normalized percent inhibition values across all assays creates a profile of its bioactivity as its HTS fingerprint. Because the in-house compound collection is always expanding, it is natural that the earlier assays have fewer compounds tested and the newer compounds have missing activity data in their fingerprints. A distribution of the effective HTS fingerprint sizes for compounds in the collection is provided in Supplementary Figure S1. Because of the sparse nature of the HTS-FP matrix, we used all possible screening data, even if data redundancies could arise from members of similar target families (Supporting Information 1.2).

When comparing the HTS-FP of two compounds, it is only relevant to examine the subset of assays that both compounds share. Missing data from incomplete assays is incorporated as a factor weighting the similarity score.

A similarity metric was derived combining both the numerical correlation of the activity *z*-scores, using the Pearson correlation coefficient, and the number of assays in common between the compounds. This is done to prevent bias resulting from compounds having a smaller number of assays defined being more able to achieve higher numerical correlations. This "Sim Score" is defined as

$$\text{Sim Score}(x, y) = \left[\left(\frac{\text{coverage}(x, y)}{2}\right) + 0.5\right] \times \text{Pearson}(x, y) \quad (1)$$

where coverage(*x*,*y*) is defined as the number of assays in common by the probe (*x*) and test compound (*y*) relative to the number of assays for the probe compound:

$$\text{coverage}(x, y) = \frac{N_{\text{probe}} \cap N_{\text{test}}}{N_{\text{probe}}} \quad (2)$$

This scoring system balances the linear correlation of the activity values by the fractional overlap of the assays in both fingerprints. If a test compound shares every assay with the probe, the Sim Score will equal the Pearson correlation. Otherwise the Pearson correlation will be penalized according to the length of the probe fingerprint. The approach we have taken to attend to the sparseness of the data make the HTS-FP still very robust in its applications as shown in the scaffold recalls and plate diversity selection. More details can be found in Supporting Information 1.3.

2. HTS-FP Clusters. From the ∼1.5 million compounds in the corporate collection, only the 360,105 compounds with less than 60% missing values in the HTS-FP are considered in the initial clustering protocol. The rest of the compounds are later assigned to their corresponding cluster on the basis of maximum similarity to a cluster center. Compounds with missing data are similar to many compounds, and therefore they tend to become cluster centers and prevent the clustering structure from converging. The procedure we choose, therefore, biases uncertainty toward similarity, for the purposes of the applications we describe in Section 3.

Clustering is done using the K-means algorithm implemented in R.[24] The distance metric is Euclidean, adapted to missing data. Once the optimal number of clusters is selected, the best cluster set is chosen on the basis of minimal clustering error, from 100 independent K-means runs with 100 iterations each. HTS-FP clusters are used in the applications C.1 and C.3. Further details on the clustering protocol and choice of cluster centers can be found in Supporting Information 4.

3. Applications. *3.1. HTS-FP Applied to Virtual Screening and Scaffold Hopping.* A benchmarking set is constructed with public domain compounds that have more than 40% non-zero values in the HTS-FP. Compounds with activities <5 μM against a target are considered active, and the rest are used as background sets. Scaffolds are calculated according to the definition of Bemis and Murcko.[26] Twenty-six targets are chosen to cover a broad variety of target families including GPCRs, kinases, ion channels, receptors, proteases, transporters, and enzymes. Importantly, none of these benchmark targets is present in the HTS-FP. A complete list of target families is shown in Supporting Information 4. Virtual screening using HTS-FP similarity is compared to an established 2D similarity searching method, Extended Connectivity Atom Environment Fingerprint with radius 4 (ECFP4) using Tanimoto similarity,[27] following identical protocols of active compound and scaffold retrieval. The one nearest neighbor (1-NN) similarity search strategy[28] is applied to sort the database by similarity score.

The active compound enrichment protocol assesses the capability of HTS-FP to recall active compounds from a background of inactive compounds. Compounds in active train and test sets have non-overlapping scaffolds. To benchmark scaffold hopping we follow the protocol proposed by Bajorath *et al.*[29] For each target, every active scaffold is used as a set of probes to retrieve other active scaffolds. For each target, we calculate active compound enrichment (ACE) as the number of actives found in the top 1% of the sorted database (10,259 compounds) and ROC scores as the area under the ROC curve. Results are reported in Figure 1 and Supplementary Table S4. Details on the benchmark database and protocol can be found in Supporting Information 4.

3.2. Plate Diversity Selection. HTS-FP are used to cluster the Novartis collection, grouping compounds that share similar bioactivity profiles across 195 assays. Under the assumption that compounds that belong to different clusters are likely to target different sets of genes, we assess the biodiversity of an HTS plate according to the number of distinct HTS-FP clusters associated with the plate's compounds. The Novartis HTS compound collection consists of 5029 384-well plates. The objective is to select a bioactively diverse and non-redundant set of 710 plates[19] that captures most of the biological diversity of the collection. Plate selection is then followed by hit expansion using HTS-FP similarity (Figure 5).

We carry out a retrospective validation benchmark to assess whether using biological diversity to select plates is more effective than chemical diversity or random selection. The goals of the validation protocol are (i) to predict the differential yield in active compounds obtained by screening a subset of plates as compared to screening the full collection and (ii) to determine which diversity selection method yields more active compounds.

Plate Selection. We identified the HTS-FP clusters for all compounds of each plate and subsequently sorted the plates by decreasing number of HTS-FP clusters present. The top ranking most diverse plate is selected for the biodiverse plate collection. The remaining plates are resorted according to the number of unseen clusters they contribute to the biodiverse library and the top plate is selected. Every time the plates are sorted, the top ranking diverse plate is incorporated and the process iteratively continues until there are no unseen clusters to be added to the reduced set. Plate selection is thus carried out in a cumulative way, ensuring that selected plates are not only biologically diverse but also differ in content among each other.

This biodiverse plate selection was compared against a similar selection that relies on the chemical structure of compounds instead. For this chemical diversity selection of plates, clustering of compounds was based on identical Murcko scaffolds as previously described.[18]

Hit Expansion. The study has two parts: discovery of seeds and expansion by cherry picking. "Seeds" are those active compounds discovered in the 710-plate diversity set. "Seeds" are then expanded *in silico* to the full compound collection. In practice, hit expansion involves a computational prediction of compounds using virtual screening methods. A hit list of "seeds" from the plate diversity set is compared to each compound in the remaining plates of the collection using a similarity metric. The top 10,000 scoring compounds are reported for each assay as hit expansion library and are individually "cherry picked" from the screening deck for a second round of testing (Figure 5). Biodiverse library "seeds" are expanded using HTS-FP similarity. Alternatively, chemical diversity "seeds" are expanded using ECFP4 similarity.

Validation and Results. Thirteen assays are selected covering different target families and assay technologies (Figure 6). Importantly, the selected 13 screens correspond to full HTS campaigns (>1.3 million compounds). Because primary or more accurate data is not available for the totality of the screens, we resort to the use of *z*-scores as described in Methods above, considering active only those compounds with *z*-score ≤ -3. Furthermore, the 13 validation targets have been removed from both the HTS clustering protocol and HTS-FP to avoid a putative bias of the diversity set. The incorporation of cumulative sorting of plates, increases approximately 6.5% the performance of biodiversity plate selections and 1.6% for chemical diversity selections as compared to sorting of plates based only on the similarities among compounds without taking into account interplate redundancy (data not shown).

3.3. GO-Term Enrichment Studies. HTS-FP clusters are described in Methods. Scaffold clusters are defined by compounds sharing identical Murcko chemical scaffold.[26] Random clusters are chosen by random sampling from the collection and have the same size distribution as the HTS-FP clusters.

For each cluster, bioactivity of its compounds is determined using data from GVK and ChEMBL as well as in-house data. Compounds with activity <5 μM are considered active, and the set of target genes is identified. Genes for all species are considered, and orthologs are grouped together under the same gene symbol.

For each target, Gene Ontology (GO) terms (biological processes, molecular functions, and cellular components) are identified.[23] The enrichment of GO terms in a cluster can be calculated using Fisher's exact test (one-sided, function in R Stats package[25]). The contingency table describes the amount of times that a certain GO term occurs in a cluster as compared to its general occurrence in the available data from the entire collection. *P*-values are corrected for multiple testing using the Bonferroni correction. A particular GO term is considered to be enriched in a cluster if its *p*-value <0.05. Groups of compounds that share the same cluster and the same GO term are counted (hereafter GO groups) and distributions of GO group sizes are calculated for components, functions, and processes. Clusters with less than 10 compound-GO term relationships are excluded.

It is important to note that a single compound could enhance the occurrence of a specific GO term in a cluster, *e.g.*, the scenario where a compound hits several genes that belong to an infrequent biological process. To control for such common but uninteresting events, enrichment results obtained with HTS-FP clusters are compared to those from random clusters. The protocol is repeated for 10 different random cluster structures. A mean GO group size distribution is calculated with error bars from standard deviation.

We calculated for each cluster its mean pairwise similarity (mps) in both the chemical and biological domains. Each compound or target (in the following referred to as entity) is expanded to a set of features. In the case of chemical similarity we used Scitegic's circular atom environments (ECFP4) as features, whereas for biological similarity we used the GO terms of known targets of compounds. The similarity measure used in both cases was the Tanimoto coefficient. The mps is then the average similarity of all pairs of entities in a cluster. For a given cluster size a null distribution of the mps was obtained by random sampling (1000 times) of the same number of targets as in the cluster under assessment. This null distribution was used to determine empirical, one-sided *p*-values.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

This material is available free of charge *via* the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: jeremy.jenkins@novartis.com; meir.glick@novartis.com.

Author Contributions

[†]These authors contributed equally to this work.

Author Contributions

[‡]These authors contributed equally to this work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank the NIBR Education Office for funding. P.M.P. thanks L. Whitehead and G. Paris for insightful conversations and support.

■ REFERENCES

(1) Keiser, M., Roth, B., Armbruster, B., Ernsberger, P., Irwin, J., and Shoichet, B. (2007) Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol. 25*, 197–206.

(2) Maggiora, G. (2008) On outliers and activity cliffs—why QSAR often disappoints. *J. Chem. Inf. Model. 48*, 646–658.

(3) Stumpfe, D., and Bajorath, J. A. (2012) Exploring activity cliffs in medicinal chemistry. *J. Med. Chem. 55* (7), 2932–2942.

(4) Paul, K. D., Shoemaker, R. H., Hodes, L., Monks, A., Scudiero, D. A., Rubinstein, L., Plowman, J., and Boyd, M. R. (1989) Display and analysis of patterns of differential activity of drugs against human tumor cell lines: development of mean graph and COMPARE algorithm. *J. Natl. Cancer Inst. 81* (14), 1088–1092.

(5) Weinstein, J. N., Kohn, K. W., Grever, M. R., Viswanadhan, V. N., Rubinstein, L. V., Monks, A. P., Scudiero, D. A., Welch, L., Koutsoukos, A. D., Chiausa, A. J., et al. (1992) Neural computing in cancer drug development: predicting mechanism of action. *Science 258* (5081), 447–451.

(6) Weinstein, J. N., Myers, T. G., O'Connor, P. M., Friend, S. H., Fornace, A. J., Jr., Kohn, K. W., Fojo, T., Bates, S. E., Rubinstein, L. V., Anderson, N. L., Buolamwini, J. K., van Osdol, W. W., Monks, A. P., Scudiero, D. A., Sausville, E. A., Zaharevitz, D. W., Bunow, B., Viswanadhan, V. N., Johnson, G. S., Wittes, R. E., and Paull, K. D. (1997) An information-intensive approach to the molecular pharmacology of cancer. *Science 275* (5298), 343–349.

(7) Kauvar, L. M., Higgins, D. L., Villar, H. O., Sportsman, J. R., Engqvist-Goldstein, A., Bukar, R., Bauer, K. E., Dilley, H., and Rocke, D. M. (1995) Predicting ligand binding to proteins by affinity fingerprinting. *Chem. Biol. 2* (2), 107–118.

(8) Dixon, S. L., and Villar, H. O. (1998) Bioactive diversity and screening library selection via affinity fingerprinting. *J. Chem. Inf. Comput. Sci. 38* (6), 1192–1203.

(9) Fliri, A., Loging, W., Thadeio, P., and Volkmann, R. (2005) Biospectra analysis: model proteome characterizations for linking molecular structure and biological response. *J. Med. Chem. 48*, 6918–6925.

(10) Fliri, A. F., Loging, W. T., Thadeio, P. F., and Volkmann, R. A. (2005) Biological spectra analysis: Linking biological activity profiles to molecular structure. *Proc. Natl. Acad. Sci. U.S.A. 102*, 261–266.

(11) Fliri, A., Loging, W. T., Thadeio, P. F., and Volkmann, R. (2005) Analysis of drug-induced effect patterns to link structure and side-effects of medicines. *Nat. Chem. Biol. 1*, 389–397.

(12) Plouffe, D., Brinker, A., McNamara, C., Henson, K., Kato, N., Kuhen, K., Nagle, A., Adrian, F., Matzen, J. T., Anderson, P., et al. (2008) In silico activity profiling reveals the mechanism of action of antimalarials discovered in a high-throughput screen. *Proc. Natl. Acad. Sci. U.S.A. 105*, 9059–9064.

(13) Cheng, T., Li, Q., Wang, Y., and Bryant, S. H. (2011) Identifying compound-target associations by combining bioactivity profile similarity search and public databases mining. *J. Chem. Inf. Model. 51*, 2440–2448.

(14) We refer to scaffold hopping in a general sense as a new chemotype that leads to a desired phenotype.

(15) Compounds are indexed in increasing order C1–C39 throughout the paper and their sources are reported in Supporting

Information. All compounds in this article are from publicly available databases.

(16) Lee, J. J., and Swain, S. M. (2008) The Epothilones: Translating from the laboratory to the clinic. *Clin. Cancer Res. 14*, 1618−1624.

(17) Essayan, D. (2001) Cyclic nucleotide phosphodiesterases. *J. Allergy Clin. Immunol. 108*, 671−680.

(18) Sukuru, S. C., Jenkins, J. L., Beckwith, R. E. J., Scheiber, J., Bender, A., Mikhailov, D., Davies, J. W., and Glick, M. (2009) Plate-based diversity selection based on empirical HTS data to enhance the number of hits and their chemical diversity. *J. Biomol. Screening 14*, 690−699.

(19) Most lead finding projects request on average ∼250,000 compounds for exploratory screens. In a 384-well plate-based library, these amounts to 710 plates. Even if the Novartis collection continues to expand, this consensus number remains as it empirically seems to maximize the benefit for the cost.

(20) Vogt, M., Huang, Y., and Bajorath, J. A. (2011) From activity cliffs to activity ridges: informative data structures for SAR analysis. *J. Chem. Inf. Model. 51*, 1848−1856.

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

(22) "Enrichment" here means that the occurrence of a certain feature (*e.g.*, gene) in a population (*e.g.*, activities of compounds that have a specific scaffold) is more frequent than what it would be readily expected by chance.

(23) The Gene Ontology Consortium and Ashburner, M. (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet. 25*, 25−29.

(24) http://wiki.geneontology.org/index.php/GO_FAQ

(25) Ihaka, R., and Gentleman, R. (1996) R: A language for data analysis and graphics. *J. Comp. Graph. Stat. 5*, 299−314.

(26) Bemis, G., and Murcko, M. (1996) The properties of known drugs.1. Molecular frameworks. *J. Med. Chem. 39*, 2887−2893.

(27) Rogers, D., and Hahn, M. (2010) Extended-connectivity fingerprints. *J. Chem. Inf. Model. 50*, 742−754.

(28) Hert, J. A., Willett, P., Wilton, D. J., Acklin, P., Azzaoui, K., Jacoby, E., and Schuffenhauer, A. (2004) Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comp. Sci. 44*, 1177−1185.

(29) Vogt, M., Stumpfe, D., Geppert, H., and Bajorath, J. A. (2010) Scaffold hopping using two-dimensional fingerprints: true potential, black magic, or a hopeless endeavor? Guidelines for virtual screening. *J. Med. Chem. 53*, 5707−5715.