Journal of Medicinal Chemistry

Article

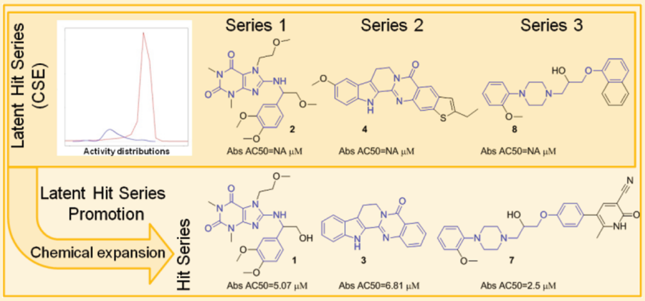
pubs.acs.org/jmc

Latent Hit Series Hidden in High-Throughput Screening Data

Thibault Varin,*,† Marie-Cecile Didiot,† Christian N. Parker, and Ansgar Schuffenhauer

Novartis Institutes for BioMedical Research, Forum 1, Novartis Campus, CH-4056 Basel, Switzerland

Ⓢ *Supporting Information*



ABSTRACT: Recently a novel method termed compound set enrichment (CSE) has been described that uses the activity distribution of a structural class of compounds to identify hit series from primary screening data. This report describes how this method can be used to identify such hit series, even when no hits according to conventional hit-calling methods for a given structural class are present in the data set. Such series, which were called latent hit series, were identified prospectively in a cell-based screening campaign and also in a series of retrospective analyses of publicly available data sets from PubChem. The assay used for the prospective case study was developed to identify compounds modulating protein translation directed from the internal ribosome entry site (IRES) of the encephalomyocarditis virus (EMCV) genomic RNA. The assay was designed with the ability to detect two assay readouts. The first assay readout monitors compound effects on IRES-directed translation, and the second readout monitors the cell viability and general effect on protein expression. By applying CSE separately to both of them, six validated latent hit series with apparently no effects on cell viability were identified. For each of these series, further testing of new compounds enabled identification of additional hits, also apparently with no effect on cell viability. These validated latent hit series would have been missed by a conventional cutoff-based hit-calling approach. This prospective study further supports CSE as a method for the analysis of high-throughput screening experiments.

■ INTRODUCTION

High-throughput screening (HTS) has become firmly accepted as an important tool in small-molecule drug discovery.[1−3] In fact, HTS was shown to have been pivotal in the discovery of a number of recently approved drugs.[4] Nevertheless, despite the increasing size of libraries and the complexity of biological assays available for HTS, this technique is not guaranteed to identify suitable starting points for medicinal chemistry, even if a biologically meaningful and robust assay is used.[4] Given the effort required for development and the cost of executing an HTS campaign, such an outcome is clearly unsatisfactory.

The goal of a HTS campaign is not the discovery of an optimal drug candidate for the target of interest but rather to provide a starting point for exploration by medicinal chemistry. For this reason, the objective of an HTS campaign is not the discovery of a single high-potency hit but, rather, a series of compounds displaying a rational structure−activity relationship (SAR).[3,5] Hits identified during HTS campaigns are often used as just the starting point for the discovery of additional, related compounds with a range of potency and activity. All biological assays, including HTS assays, are limited in sensitivity and the activity range that they can monitor. Even for HTS campaigns conducted with a primary concentration response, such as qHTS campaigns, the assays are limited to detecting compounds with absolute AC_{50} values less than the highest concentration tested.[6] This in turn is usually dependent on the concentration of the dimethyl sulfoxide (DMSO) stock solutions for screening and the DMSO tolerance of the assay. However, it is still possible that there may be series of active compounds within the screening library without any compound passing an arbitrary activity threshold being used to define a hit. Such a series of compounds would not be identified by conventional hit-calling methods and would be missing from a hit list. Weak active compounds were shown to contain meaningful information, as for target-related affinity profiling or SAR extraction, for example.[7−9] Also, recently Mestres and Veeneman[10] highlighted the potential of such weak, but still active, compounds. They suggested that it may be possible to transform, by small modifications of their structure, weakly active compounds into hits; such compounds were called "latent" hits. The authors describe how the use of pharmacophore information can be used to "awaken" such latent hits. But prior knowledge of the active pharmacophore is required for such a transformation. In addition, this pharmacophore-directed transformation must be applied individually to each potentially latent hit.

The concept of latent hit series can be rationalized by the projection of a chemical series in the biological space. An active compound contains an optimal combination of the correct scaffold and side groups for binding to a target. Side groups not optimal for binding result in only weakly active or, in the extreme case, inactive compounds. Even if a molecular scaffold presents the correct size and shape features required for binding, it is still necessary to test a range of compounds based on

Received: September 6, 2011
Published: December 19, 2011

dx.doi.org/10.1021/jm201328e | *J. Med. Chem.* 2012, 55, 1161−1170

this scaffold in order to find active compounds. Therefore, even if a scaffold is biologically validated, it is necessary to synthesize and screen a library of compounds around this scaffold in order to identify compounds with the right substitution patterns to be identified as hits, as reported for the biology-oriented synthesis concept (BIOS).[11] The virtual library space that is covered by current synthesis protocols has a size at least on the order of magnitude of 10^{11}−10^{12} compounds.[12,13] Although in a typical high-throughput screening group up to a few million compounds can be tested against a particular target, this number is small compared to the size of the virtual library space. Thus it is clear that the chance of finding a compound with the optimal combination of scaffold and side chains is limited. Nevertheless, the possibility of finding latent hit series is much higher. This is especially interesting for screening campaigns that have initially failed to discover potent chemical starting points.

Recently Varin et al.[5,14] introduced a method for the analysis of HTS data that is not dependent on an activity cutoff. The method, called compound set enrichment (CSE), was initially developed to identify hit series (defined by a common scaffold) instead of individual hit compounds. Compound series were defined by a common scaffold obtained either from the scaffold tree, as in the initial publication,[5] or from the scaffold network, recently published.[14] The scaffold network is obtained by systematically dissecting the molecular framework in order to obtain a set of smaller parent scaffolds. The parents obtained in this way are iteratively dissected themselves. The relationship between the parent and child scaffolds is recorded, leading effectively to a network of scaffolds. In a second step, a nonparametric statistical hypothesis test (Kolmogorov−Smirnov or KS) was used to evaluate whether the compounds of a given scaffold class present an activity distribution different from the general population of inactive compounds. In the assays previously analyzed with this method,[5] many of the active scaffolds were populated with compounds defined as hits by a standard activity cutoff. Most of these scaffolds were also populated by additional weakly active and inactive compounds. These results highlight the efficiency of the method to identify hit series of compounds with relevant SAR directly from primary HTS data. Nevertheless, these results do not permit one to determine whether these scaffolds would also be predicted as active if no hits according to conventional hit selection criteria were present in the data set. If this is possible, then CSE is suitable to detect latent hit series.

This hypothesis can be retrospectively tested: One takes a HTS data set, removes all the compounds more active than an activity cutoff in the range typically used to process a primary screening hit list, and tries to identify nevertheless the scaffolds containing the active compounds in the full data set. The results of such a study using two screening data sets from PubChem are reported here. The assay used are the HSD17B4 [hydroxysteroid (17-β) dehydrogenase 4] inhibitor assay (PubChem assay ID 893)[15] and the HADH2 (hydroxyacyl-coenzyme A dehydrogenase, type II) inhibitor assay (PubChem assay ID 886).[16] The motivation for using these qHTS data sets is that concentration−response curves are available for all compounds tested in these bioassays.[6] For all curves, individual data points, fitting parameters, and IC_{50} values are available. The compounds are annotated in PubChem as active, inconclusive, and inactive. These results allow the simulation of a classical screening process by using the results from the highest concentration tested to simulate the results that would be obtained from a standard primary screen. Then the complete concentration response curve and compound activity annotation derived from it are available for the validation step.

The results of the retrospective analysis were encouraging enough to justify a prospective study on an in-house medium-throughput screening project. The assay chosen for this study was developed both to identify compounds modulating protein translation directed from the internal ribosome entry site (IRES) of the encephalomyocarditis virus (EMCV) genomic RNA and to deprioritize compounds that alter cell viability.[17] To facilitate compound screening, a bicistronic reporter gene construct, containing a neomycin gene (Neo) and the firefly luciferase (Fluc) reporter gene, was used (see Figure S7 in Supporting Information). Following transcription, the neomycin resistance gene is translated by a cap-dependent mechanism, while Fluc expression is directed by the EMCV IRES (pNeo-EMCV) in a cap-independent manner. This assay can then potentially identify compounds that inhibit Fluc expression (cap-independent) without influencing the cap-dependent expression of the neomycin resistance gene. Such differential activity can be detected because the presence of Geneticin in the cell culture means that inhibition of the resistance gene will lead to inhibition of cell viability or growth. In order to monitor the desired compound activity as well as the undesired influence on cell viability from the same sample, the assay was multiplexed via a noninvasive measure of cell viability by resazurin, followed by quantification of the luciferase expression as described by Didiot et al.[17]

The activity of these two readouts was then normalized by use of DMSO as the neutral control (0% residual activity) and benzalkonium chloride as the active control (i.e., −100% residual activity) for both the resazurin and luminescence assay readouts. This slightly artificial control had to be used because at present there are no known control compounds influencing just the EMCV IRES-directed translation. This multiparametric assay was developed so as to allow identification of compounds inhibiting cell growth, as such compounds would also appear as hits for compounds inhibiting IRES-mediated translation. In addition, compounds that stimulate cell proliferation would also appear as compounds increasing IRES mediated translation without this internal control.

■ RESULTS

Retrospective Study on PubChem q-HTS Data Sets. From the complete screening data sets downloaded from PubChem, different reduced data sets were prepared by progressively removing compounds with a given potency threshold. All compounds below the threshold were removed, even if they are annotated as inconclusive or inactive by PubChem. For both assays, four thresholds were used: 10, 20, 30, and 40 μM (40 μM is the lowest potency associated with a compound). However, the definition of a hit is more restricted. Only a compound with a potency less than, or equal to, 10 μM and annotated as active by PubChem was considered as a hit. Compound set enrichment was applied to complete and reduced data sets. For an illustration of reduced data sets and hit definitions, see Figure S6 in Supporting Information.

Three conditions were applied to the definition of validated latent hit series. The scaffold defining the series must be evaluated as active in both the complete (first condition) and the reduced (second condition) data sets according to compound set enrichment. These two conditions assess activity of the series but do not validate its potential to promote any weakly active compounds into hits. Thus a scaffold is only

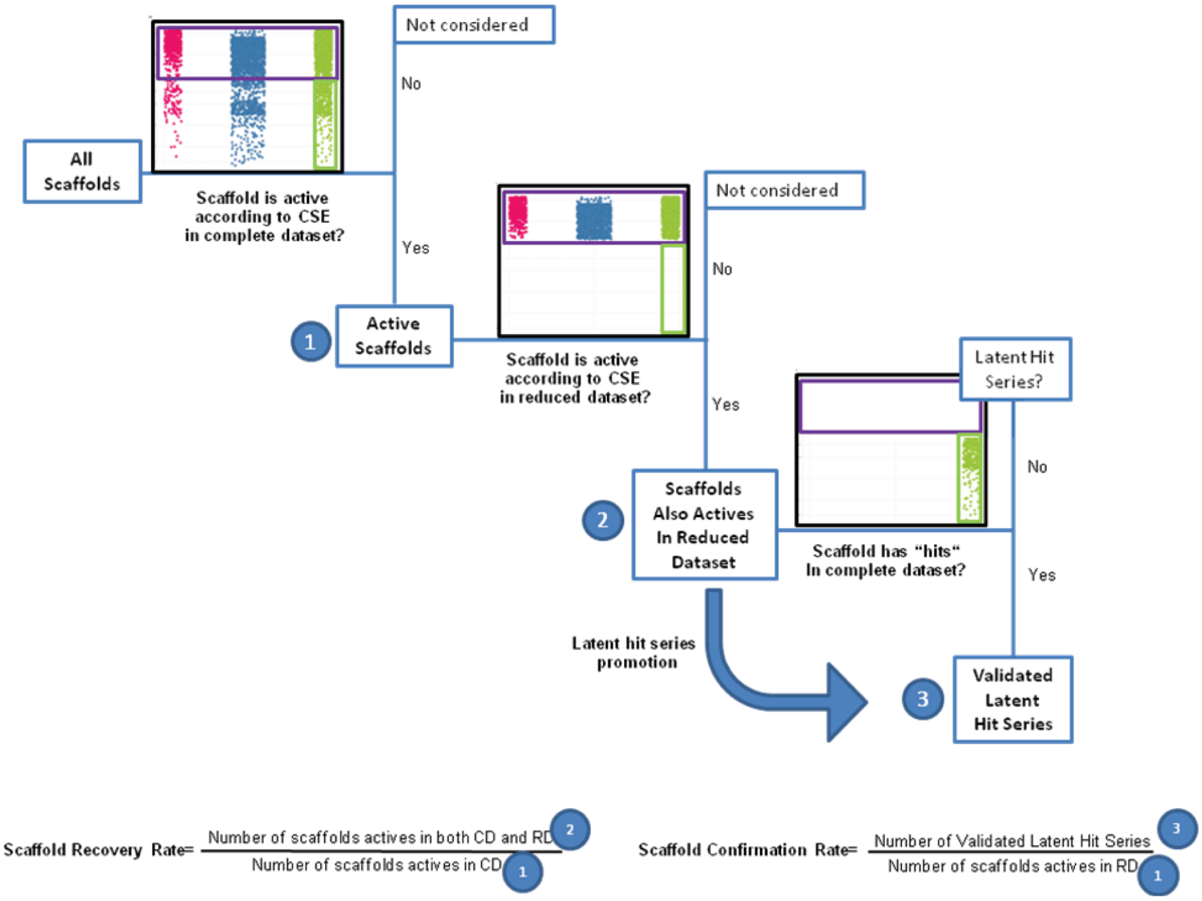

Not considered
No
All Scaffolds
Scaffold is active according to CSE in complete dataset?
Yes
1
Active Scaffolds
Not considered
No
Scaffold is active according to CSE in reduced dataset?
Yes
Latent Hit Series?
No
2
Scaffolds Also Actives In Reduced Dataset
Scaffold has "hits" In complete dataset?
Yes
Latent hit series promotion
3
Validated Latent Hit Series
Scaffold Recovery Rate= Number of scaffolds actives in both CD and RD 2 / Number of scaffolds actives in CD 1
Scaffold Confirmation Rate= Number of Validated Latent Hit Series 3 / Number of scaffolds actives in RD 1

Figure 1. Definitions of validated latent hit series, scaffold recovery rate, and scaffold confirmation rate. A validated latent hit series is defined by three conditions. The scaffold that defines the series must be active according to compound set enrichment in both the complete and reduced data sets. This scaffold must also have hits in the complete data set. This third condition corresponds to a latent hit series promotion and thus confirms that the series is truly latent. The scaffold recovery rate is defined as the number of scaffolds active in both the complete and the reduced data set divided by the number of scaffolds active in the complete data set. The scaffold confirmation rate is defined as the number of validated latent hit series divided by the number of active scaffolds in the complete data set. CD, complete data set; CSE, compound set enrichment; RD, reduced data set.

considered as a validated latent hit scaffold if a third condition is met: the scaffold must have at least one hit in the complete data set (which has been removed in the reduced set). This definition of validated latent hit series is illustrated in Figure 1.

Four indexes were defined for analysis of complete and reduced data sets:

- **Scaffold recovery rate:** fraction of active scaffolds recovered relative to the number of scaffolds recovered when the complete data set is used.
- **Scaffold confirmation rate:** fraction of validated latent hit scaffolds relative to the number of scaffolds recovered when the reduced data set is used. This is a lower boundary of the true positive rate, as for these scaffolds their active nature is evident from the data; there is a possibility that other scaffolds could be true latent hit scaffolds as well, but the screening data set did not contain enough information to validate them.
- **Hit recovery rate:** Fraction of compounds identified as hits from at least one scaffold predicted as active according to CSE *p*-value, relative to the total number of hits.
- **Active scaffold membership rate:** fraction of compounds covered by at least one active scaffold, relative to the total number of compounds in the data set.

PubChem assays 893 and 886 contain respectively 569 and 227 scaffolds predicted as being active according to the CSE *p*-values, based on the full data sets. Among these scaffolds, 296 (52%) and 123 (54%), respectively, contain hits and can be used to evaluate the ability of this method to identify scaffolds containing latent hits. Figure 2 shows the effect of removal of compounds with an increasing potency cutoff from the complete data set on the prediction performance. For both PubChem data sets the following measures were analyzed: scaffold recovery rate, scaffold confirmation rate, hit recovery rate, and active scaffold membership rate.

The key findings from Figure 2 can be summarized as follows: When the data set is reduced in such a way that less and less active compounds remain, the scaffold recovery rate drops, meaning that fewer active scaffolds are identified. Also the hit recovery rate drops, but to a lesser extent. The active scaffold membership rate remains almost unchanged. These numbers suggest that the remaining "active" scaffolds still contain the majority of the true hits and are scaffolds populated by many compounds in this data set. The activity of such scaffolds can be predicted with high confidence, despite only weakly active compounds being present. What is also remarkable is that the scaffold confirmation rate is relatively high to begin with and drops only slowly when the more active compounds

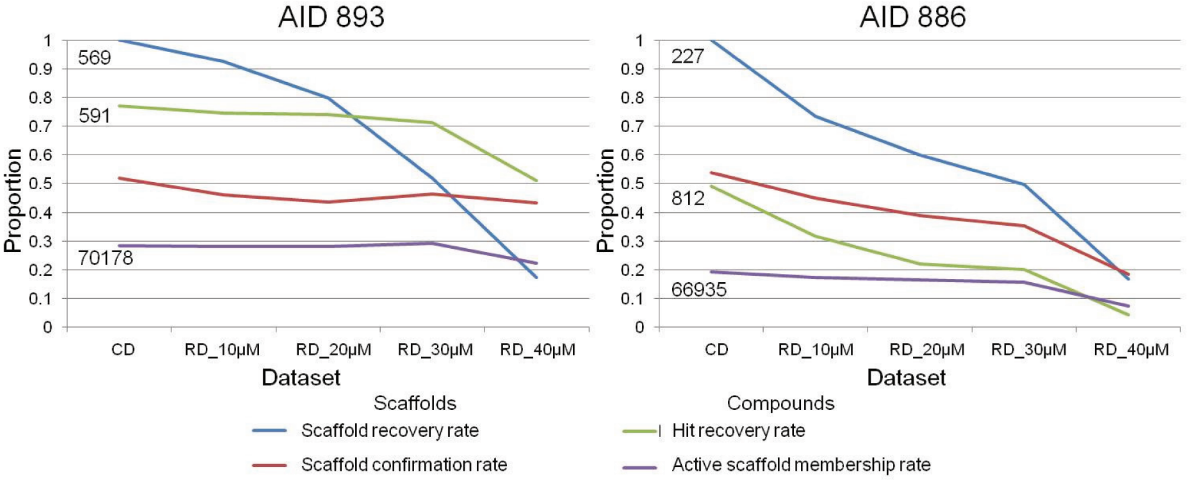


Figure 2. Evaluation of CSE efficacy to identify latent hit series. Compound set enrichment is applied to complete and reduced data sets. These reduced data sets contain only weak and inactive compounds. CD, complete data set; RD_$X\mu$M, reduced data sets obtained by removing all compounds with an activity ≤ X from the data set.

are removed from the data set. In practice, this suggests that such scaffolds, being identified as active by CSE but which lack any highly active hits, are very likely genuine classes of latent hits. A detailed structural description of one example is given in the Supporting Information.

The number of compounds per scaffold predicted as active was evaluated further, to determine whether lacking support in the data set is indeed the reason why the scaffold recovery rate drops when the data set is reduced. Reduction of the data set influences the CSE prediction results in two ways. Some active scaffolds are removed or reduced to singletons in the reduced data sets and their activity p-values cannot be evaluated any more. In addition, the number of compounds per scaffold is an important parameter to compute the p-value with the KS test, and for active scaffolds, decreasing the number of compounds increases the p-value. Progressively removing the most active compounds has almost no influence on the highly populated scaffolds with more than 100 compounds. It has little influence on scaffolds with between 10 and 32 compounds and is dramatic for scaffolds with less than 11 compounds. The detailed distribution of series sizes can be found in Figure S3 in the Supporting Information.

Prospective Study with the IRES Assay. The assay used in the prospective study differs from the simpler assays used in the retrospective study, as there are two readouts to be considered. While activity in the luminescence readout was desired, the compounds should have no effect on the cell viability readout in order to ensure that the observed luminescence effect is not a secondary effect due to changes in cell viability. In the following, this vocabulary will be used: Compounds that inhibit expression of the IRES-directed Fluc gene are termed antagonists and result in negative percentage residual activity, while compounds increasing the amount of Fluc expression are termed agonists and result in a positive percentage residual activity.

This paragraph provides an overview of the protocol used for identification and validation of latent hit series (see Figure 3). In order to favor understanding of the protocol, it has been divided into five distinct steps (a–e). In step a, the primary screen, around 14 000 compounds from a library of purified natural products were tested at a single concentration of 5 μM.

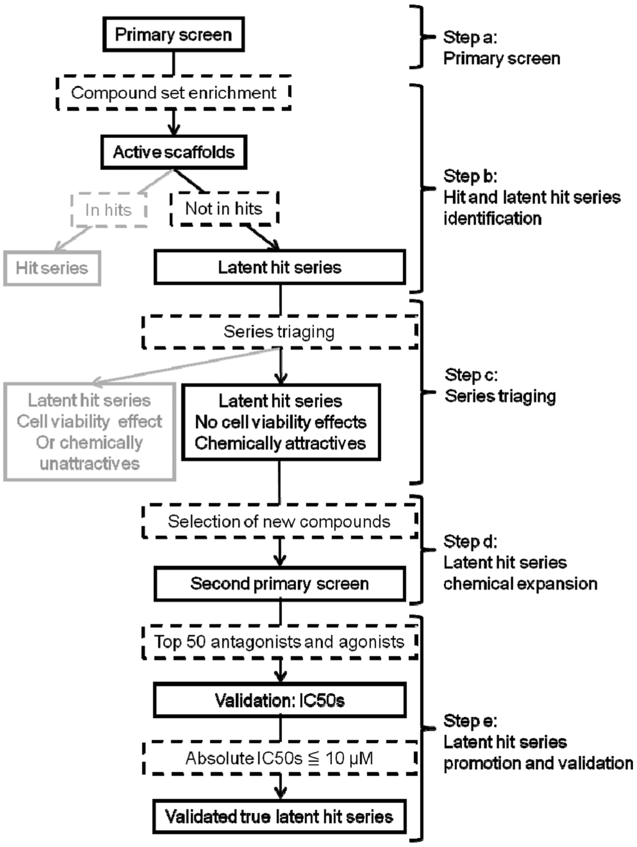


Figure 3. Prospective analysis: overview of the protocol used for identification and validation of latent hit series.

In the data analysis, step b, the activities of the different scaffolds in both luminescence and cell viability readouts were evaluated by applying CSE, in order to identify hit series and latent hit series.[5,14] Primary hits were defined according to the classical cutoff-based method, as compounds with a percentage of activity greater than or equal to 50% for agonist and less than −50% activity in the luminescence readout for antagonists. Then the scaffolds containing at least one hit were defined as hit series and the remaining active scaffolds as latent hit series.

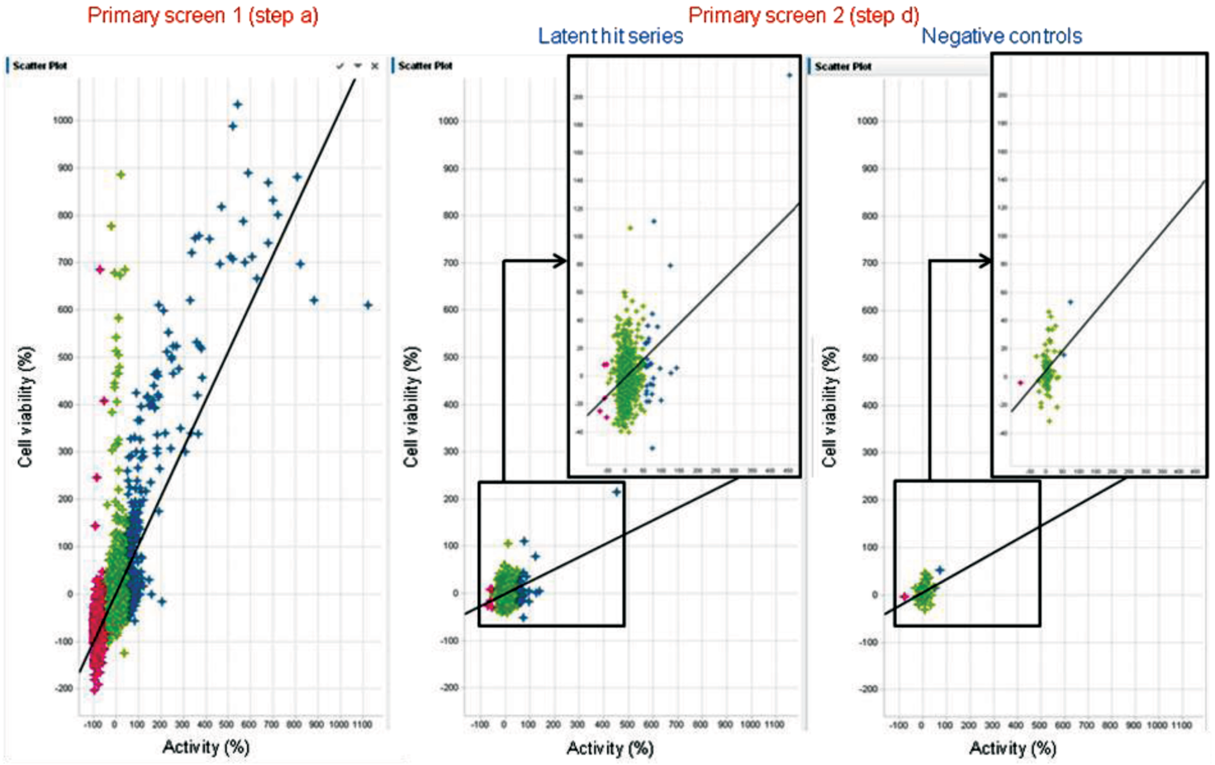


Figure 4. Compound activities (IRES-directed firefly luciferase expression) and cell viabilities in primary screens 1 (step a) and 2 (step d). Red and blue, primary hits respectively for antagonists and agonists; green, inactive compounds.

In step c, latent hit series were selected for further follow-up according to chemical attractiveness and their effect on cell viability. How this has been done in detail is described further below. In the latent hit series expansion, step d, additional compounds containing the chosen latent hit scaffolds were selected from the Novartis compound archive and tested in the same assay format as used for the primary screen. In step e, the activity of the 50 most active antagonists and agonists was confirmed by monitoring their concentration response by AC_{50} measurements. Only series for which at least one compound with an (absolute) AC_{50} less than or equal to 10 μM were considered as validated latent hit series.

The primary activity data of the initially screened ~14 000 compounds shows a strong correlation between activity and cell viability effects (see Figure 4). Activity and cell viability effects of 13 369 scaffolds were evaluated by applying CSE. Moreover, the method was applied to detect both positive and negative effects on both activity (agonists and antagonists) and cell viability (cell division and cytotoxicity). Hence each scaffold is annotated by four p-values (see Figure 5). These annotations enable a scaffold selection prioritizing those series with activity in the IRES readout but with little or no effect on cell viability (p-value > 0.01 for both positive and negative effects on cell viability). The number of scaffolds identified as active with and without the Bonferroni correction (which was used as a multiple hypothesis correction[18]) is given in Table 1. After a Bonferroni correction was applied to account for multiple tests being conducted, 33 scaffolds showing antagonist activity and three scaffolds showing agonist activity (both without significant effects on cell proliferation) were identified. Without Bonferroni correction, these numbers are respectively 393 and 153.

None of the overt hit series were considered for chemical expansion. As mentioned before, primary hits were defined as compounds with an absolute percentage of activity greater than or equal to 50% (619 antagonists and 300 agonists). Almost all active scaffolds after Bonferroni correction are represented by compounds identified as hits in the primary screen (only six latent hit scaffolds are identified for antagonism). However, by considering active scaffolds without Bonferroni correction, 147 antagonist and 76 agonist latent hit series were identified. Interestingly, the proportion of active series that showed little or no effect on cell viability was most pronounced in latent hit series (86% and 93% respectively for antagonists and agonists) than in hit series (35% and 70%). From these latent hit series, 16 and 22 scaffolds respectively were selected for positive (agonists) and negative (antagonists) modulation of activity. The selection of scaffolds for further follow-up was done by visual inspection according to scaffold activity p-values, cell viability p-values, and chemical attractiveness.

To expand chemical space around these scaffolds, additional compounds were selected from the Novartis screening library. For some scaffolds the number of compounds available was too large to include all of them for follow-up testing. For each of the scaffolds, a subset of new compounds was selected according to chemical attractiveness and similarity to the compounds screened in the initial primary screen (for more information, see Materials and Methods). In total, 832 compounds were selected for this chemical expansion. As negative controls, 80 compounds with similar polar surface area (PSA), molecular weight (MW), and log P (ALogP) to those of compounds selected for latent hit promotion were also tested (see Materials and Methods for details). Compounds selected from both latent hit series and negative controls were tested in a second primary screen. Compound activity and cell viability results are shown in Figure 4.

Of these compounds, 30 showed activity greater than 50% and were identified as agonists (compared to only two for the negative controls). A further five compounds with activity less than −50% residual activity were identified as antagonists

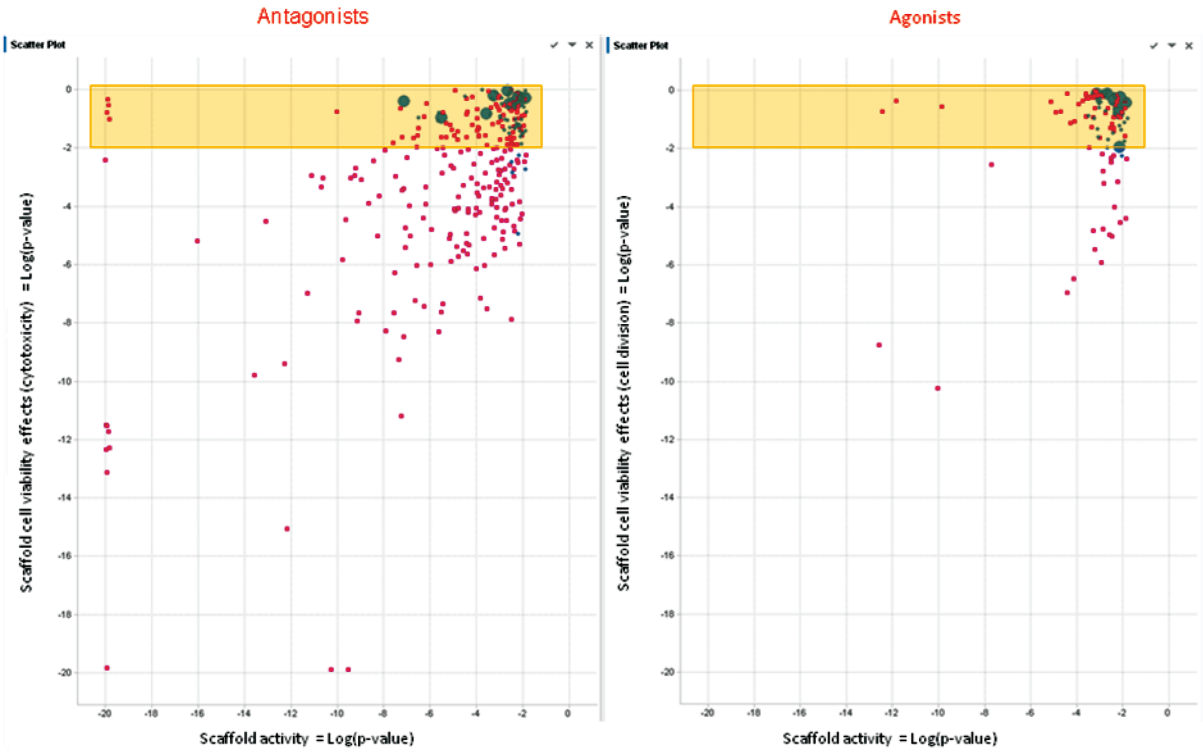


Figure 5. Compound series activities (*x*-axis) and cell viabilities effects (*y*-axis) evaluated by CSE. Only active series are displayed. These active series are divided into hit series (defined by an active scaffold with at least one primary hit) and latent hit series (defined by an active scaffold and the absence of primary hits). They are represented respectively by small red squares and large blue circles. Series with no significant effect (*p*-value > 0.01) on cell viability are highlighted by orange shading. Latent hit series were triaged according to their activity, effects on cell viability, and chemical attractiveness (visual inspection). Latent hit series selected for chemical expansion are represented with a bigger point than the others.

Table 1. Number of Active Scaffolds Identified According to CSE with and without Bonferroni Correction[a]

	active scaffolds with Bonferonni correction, NECV/all (%)			active scaffolds without Bonferonni correction, NECV/all (%)		
	all	HS	LHS	all	HS	LHS
antagonists	33/201 (16)	27/194 (14)	6/7 (86)	393/883 (45)	246/712 (35)	147/171 (86)
agonists	3/7 (43)	3/7 (43)	0/0	153/192 (80)	77/110 (70)	76/82 (93)

[a] Scaffolds were considered to have no effects on cell viability if the corresponding *p*-value was greater than 0.01. NECV, no effect on cell viability; HS, hit series; LHS, latent hit series.

Table 2. Number of Primary Hits Identified in Primary Screens 1 (Step a) and 2 (Step d)[a]

			primary screen 2 (step d)			
	primary screen 1 (step a)		latent hit series (832 cpds)		negative controls (80 cpds)	
	antagonists NECV/T (%)	agonists NECV/T (%)	antagonists NECV/T (%)	agonists NECV/T (%)	antagonists NECV/T (%)	agonists NECV/T (%)
no. of hits	139/619 (22)	93/300 (31)	3/5 (60)	22/30 (73)	1/1 (100)	1/2 (50)

[a] Primary hits are defined as compounds with an absolute percentage of activity greater than or equal to 50%. Nontoxic primary hits are defined as primary hits with an absolute percentage of residual activity in the cell viability assay less than 25%. NECV, number of hits with no effects on cell viability; T, total number of hits (number of hits with no effects on cell viability + number of hits with effect on cell viability).

(compared to only one from the negative controls). Table 2 summarizes the number of primary hits (both with and without effects on cell viability) in both primary screens. The activity distribution of compounds selected for agonist and antagonist activity was compared to the activity distribution of the negative controls via the one-sided two-sample KS test. For both antagonists and agonists, a significant activity shift was observed (with *p*-values respectively of 0.0002 and 0.0069).

The 50 most active agonists (lowest activity at around 40% residual activity) and the 50 most active antagonists (lowest activity at around −20% inhibition) were retested to determine whether they displayed a concentration response. Even if antagonist activity is low, and in order to avoid the missing of interesting hits, the same number of antagonist compounds as for agonist was tested in validation. Activity and cell viability of these compounds were tested for a concentration response at eight concentrations from 0 to 10 μM in quadruplicate (see Figures S4 and S5 in Supporting Information). Cell viability was monitored via the resazurin assay readout; this showed that these compounds are not overtly toxic under the experimental conditions used. As for the retrospective analysis, we defined as a validated hit each compound with an AC_{50} (absolute) value less than or equal to 10 μM. The dose response curves of the primary hits activity show, for 14 of them, an absolute AC_{50} from 0.42 to 8.77 μM (none of them has an absolute AC_{50} for cell viability). No primary hit compounds (neither agonists nor antagonists) were identified from the compounds selected as negative controls. These 14 compounds belong to six chemically different series (some represented by more than one scaffold).

Out of the six series for which we observed and validated a latent hit promotion, scaffold and representative compounds of three series are represented in Figure 6. Nine latent hit promotions were observed for the first series and one for the second and the third series. Series 1 is clearly confirmed as active, as nine validated hits were identified (see Table S3 in Supporting Information). Series 4 (structures not disclosed) gave two validated hits. For each remaining series, only one validated hit was found. Nevertheless, a weak activity was also validated for other compounds within these series (except series 6). This weak but consistent activity indicates that these hits are not false positives but truly active in the assay. For each series, we provide AC_{50} values (concentration at which the activity is equal to half the AC_{50} curve activity range), absolute AC_{50} values (concentration at which the activity is equal to −50% for antagonists and 50% for agonists), and A_{max} values (activity measured at the highest concentration) of all compounds as well as distance with best compounds in the series in Supporting Information (Tables S3−S6) and in Figure 6. Series 1, 2 and 5 were agonists and series 6 was antagonist in the first primary screen (according to CSE and so based on their activity distribution). In confirmation stage, these series also show compounds with the same activity direction. However, series 3 and 4, which were both antagonist in the first primary screen, show compounds with both agonist and antagonist activity at the confirmation stage after chemical expansion. These scaffolds can deliver both agonist and antagonist compounds. For the latent hit series 2, the active scaffold itself was identified as a hit. This result is of interest for using CSE to identify active fragments, which could further be confirmed by fragment-based screening technologies.

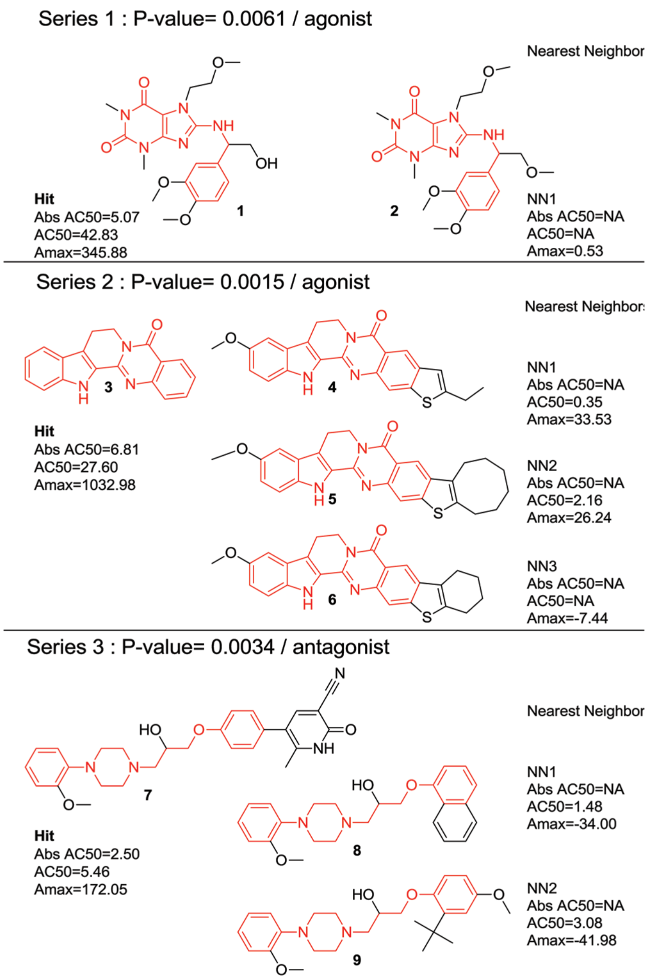


Figure 6. Activity and structures of compounds selected for validation for three out of six of validated latent hit series. Active scaffolds have been determined by analysis of primary screening data with CSE. A chemical expansion around active scaffolds defining latent hit series has been done. Activity of these new compounds was evaluated in a second primary screen. Best compounds were validated in a dose response activity measurement at height concentrations in quadruplicate. For each series, the active scaffold identified by CSE is highlighted in red (series definition) and its p-value is indicated. For each series, the compound on the left is a hit (defined as a compound with an absolute AC_{50} less than or equal than 10 μM) representative of the series. Other compounds are nearest neighbors (NN) from this hit. All compounds tested for series 2 and 3 are represented. However, for the first series, eight other hits and seven latent hits have been identified. Absolute AC_{50}, AC_{50}, and distance from the hit can be found in Supporting Information (Tables S3−S6).

■ DISCUSSION

This report demonstrates again the utility of the "similarity principle" first articulated by Johnson and Maggiora,[19] that similar compounds will have similar activity. In the context of compound set enrichment, the similarity principle also integrates weak active compounds, that is, that compounds with structural similarity to an active or a set of active (even weak) compounds will tend to be active themselves. This work classifies similar compounds by their chemical scaffold rather than by some measure of similarity. This approach has the advantage that the navigation in the scaffold space is intuitive to medicinal chemists and also tends to highlight how activity is associated with particular scaffolds, which can then help in pharmacophore identification.[20]

One of the important aspects of CSE is that the method does not use a "hard" cutoff to define compound activity during analysis of the compound classes; rather, it monitors the activity distribution of the compound class. This helps to remove the uncertainty introduced in selecting a suitable activity cutoff to define active or inactives. In fact, while there have been other methods for hit identification (or even for rescue of false negatives) from screening results using structural information, these methods have all relied on the use of an activity cutoff to first select "active" compounds.[21−23] This is highlighted by the observation that even if the most liberal definition of active compounds were removed from the test data set, the method was still able to identify classes of compounds that would have acted as starting points for latent hit discovery. This shows that the method could be used to "rescue" starting points from screening campaigns that may appear to have failed.

The results of the retrospective study presented here also have implications for the design of compound libraries for screening. There have been a number of compound library design strategies that argue for the selection of as diverse a set of compounds as possible to cover as wide a range of compound space as possible, or alternatively to screen as many compounds as possible in a random manner.[24] These methods all result in uneven coverage of chemistry space with the presence of "singletons" for which no related compounds are also present in the library. While these can be valuable starting points, if they display sufficient activity in the primary screen, it is not possible to use such compounds as starting points for latent hit series identification. In fact, there have been a small number of studies that suggest that screening library design might be better focused to include groups of related compounds.[25] This study would support such strategies as a means to maximize the ability of CSE to identify latent hits. Analysis of the class size distributions suggests that a screening collection should contain at least 10 members of a compound scaffold class but that testing more than 100 compounds from a scaffold class will not help identify active classes. This observation is in accordance with previous studies from Schreyer et al.[26] and Nilakantan et al.[27]

In the prospective study, CSE has been applied to a complex cell-based assay monitoring a range of physiological processes. Even in such a complex screening model, CSE was able to identify classes of compounds that, when expanded, yielded additional active compounds that gave a validated, concentration-responsive activity, confirming that this method has the potential to identify latent hit series hidden within HTS results even without prior knowledge about SAR hits or target structure. This is in contrast to a set of compounds that were chosen to purely reflect the physical−chemical properties of the test set of compounds. In this control, no active compounds were identified.

This prospective study also highlighted another application of CSE in that the method was used to deprioritize compound classes showing effects on cell viability as well as the desired activity. This resulted in almost all of the compounds selected for CSE having a markedly reduced activity profile in the cell viability readout, while the method was still able to select compounds acting on IRES-directed gene expression. However, such use of CSE for multiparameter compound optimization is not completely automated. The observation in the prospective study that CSE resulted in compounds showing agonist as well as antagonist activity also highlights the strength of using the KS statistic to identify classes of compounds showing activity different from the bulk of inactive compounds. The observation that classes of compounds can show both agonist and antagonist activities has been reported for nuclear hormone receptors as well as for ligands binding G protein-coupled receptors (GPCRs), so while unexpected, it is not without precedent.[28,29]

■ CONCLUSION

The ability of compound set enrichment to identify valid latent hit series has been demonstrated both retrospectively and in a prospective study. Six nontoxic latent hit series were identified in the prospective study. These series display interesting SAR that can be exploited to expand the series or to derive relevant pharmacophores that can be used to discovered new active scaffolds. It also supports the objective that latent hit series can be identified without prior knowledge about SAR hits, about target structure, and even without knowing the target itself.

■ MATERIALS AND METHODS

Compound Set Enrichment. *Compound Series Definition: Scaffold Network.* The molecules represented in the data sets used in this study were processed to remove salts and to standardize charges and stereochemistry as described previously.[5] The preprocessed data sets were classified by the scaffold network classification.[14] The scaffold network is a scaffold-based compound classification. To derive this network, all side chains from compounds are removed to derive their molecular framework. Then peripheral rings are progressively removed to generate smaller and smaller rings. This process generates a parent/child relationship. Each parent scaffold is a substructure of its largest child scaffolds. Children can have several parents and vice versa. Then all scaffold activities were predicted by applying compound set enrichment to simulated (retrospective analysis) or real (prospective analysis) primary screening data.[5,14]

Compound Series Activity Evaluation: Kolmogorov−Smirnov Hypothesis Test. Compound set enrichment uses a nonparametric test followed by a multiple hypothesis test correction to evaluate scaffold activities. The Kolmogorov−Smirnov (KS) hypothesis test used here does not use any activity cutoff. Rather, it evaluates how likely it is that the activity distribution of the set of compounds sharing a given scaffold can be obtained by random sampling of the overall activity distribution of the assay. If that is the case, then the null hypothesis is true and the scaffold is not considered to be active. For each scaffold, the probability that the null hypothesis is true (p-value) is computed. If this p-value is smaller that a critical level of significance ($\alpha \leq 0.01$), the null hypothesis is rejected and the scaffold is considered as "active". This definition works only if a single scaffold activity is evaluated. When multiple scaffolds are tested, a correction that decreases the level of significance needs to be applied in order to decrease the number of false positives. We used the Bonferroni correction. The new level of significance for the analysis is obtained by dividing the level of significance for an individual scaffold by the number of scaffolds being considered. This correction was applied separately for each level in the scaffold network, as initially proposed by Varin et al.[5]

Retrospective Study. For the retrospective study, the activity readout at the highest concentration of the qHTS data has been used to simulate a conventional, single-concentration HTS. Details are given in Supporting Information.

Prospective Study. *Compound Libraries.* The primary screen was performed on a proprietary collection of purified natural products; at the time of screening it was about 14 000 compounds. Compounds have mainly been isolated from actinomycetes, myxobacteria, and plant sources. For the latent hit series expansion, additional compounds were selected from the Novartis compound archive. The archive contains more than one million compounds, mainly synthetics.

Assay Format of EMCV IRS Assay. Compound screening was performed in a stable cell line expressing a bicistronic reporter construct. This transcript contains a neomycin gene (Neo), which is translated in a cap-dependent mechanism, and the firefly luciferase (Fluc) reporter gene under the control of the EMCV IRES (see Figure S7 in Supporting Information). This assay was designed to potentially identify compounds that modulate Fluc expression and thus the EMCV IRES activity in a cap-independent manner, with the ability to detect two assay readouts with each well monitoring cell viability as well as compound effects on IRES-directed translation. Cell viability was monitored via noninvasive reduction of resazurin dye to evaluate the effect of compounds on viability. The compound activity on the IRES element was monitored by quantification of luciferase expression via the Steady-Glo assay. In order to identify compounds inhibiting or activating IRES-directed expression of the firefly luciferase, a proprietary library of natural products was tested in a primary screen at a concentration of 5 μM. As a positive control inhibition of the Fluc expression, the cells were killed by use of 500 μM benzalkonium chloride. After 24 h of incubation of cells with or without compounds, the test plates were incubated with the resazurin reagent, and a cell viability measurement was taken. The cells were then lysed and EMCV-IRES activity was determined by measuring the activity, and thus the expression, of luciferase. Data were then analyzed by use of a software package developed in-house. The Z'-factor for each assay plate was calculated for the activity and cytotoxic end points from the control wells, and this factor was used to monitor assay quality.[30] All plates gave a Z'-factor value above 0.5, indicating acceptable assay quality.[30] More details on materials and methods are given in Supporting Information.

Compound Data Sets for Prospective Study. In the initial primary screen, a library of around 14 000 compounds, mostly natural products, was tested. Other compounds selected for chemical expansion and negative controls were selected in a library of nonnatural compounds. Compound purities were not established for compounds tested in primary screens but were evaluated for all compounds tested for AC_{50}. Purity was determined by HPLC−MS at 214 nm. Analytical reverse-phase (Ascentis Express C18; 30 mm ×2.1 mm, particle size 2.7 μm) high-performance liquid chromatography coupled to mass spectrometry (HPLC−MS) was performed with an Agilent 1100 binary gradient module G1312A equipped with Agilent G1315B DAD, Waters Acquity ELSD, and Waters micromass ZQ. Mass spectra were acquired in both ESI^+ and ESI^- modes, scanning from m/z 100 to 1600 Da. All final compounds were analyzed employing a linear gradient from 2% to 98% methanol (+0.04% formic acid) in water (+0.05% formic acid) over 1.5 min and a flow rate of 1.1 mL/min, and unless otherwise stated, the purity level was >95%.

Active Scaffold Chemical Expansion. For each of the selected active scaffolds, available compounds in the Novartis compound library were identified and sorted by similarity to the best compounds tested in the primary screen within the corresponding series. These compounds were then triaged according to chemical attractiveness (by eye determination). This process is not automated.

Negative Controls. To select these negative controls we derived a binned 3D matrix (three axes for PSA, MW, and ALogP) from the 832 compounds selected for latent hit series chemical expansion. Bins of 50 Å^2, 100, and 1 used, respectively, for PSA, MW, and ALogP. These values were computed by use of the following components of the Pipeline Pilot software:[31] "Surface Area and Volume" (PSA), "Molecular Weight", and "AlogP". Then a set of 80 new compounds with a similar distribution of physicochemical properties as this binned matrix were selected in the Novartis compound archive.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Additional text, seven figures, and six tables as described in the text. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail thibault.varin@novartis.com; tel +41 61 3249297; fax +41 61 3243357.

Author Contributions

†These authors contributed equally to this work.

■ ACKNOWLEDGMENTS

We thank the NIBR Education Office for the NIBR postdoctoral fellowship for T.V. We thank Meir Glick for bringing to our attention the possibility of CSE to identify latent hits. We gratefully acknowledge Juergen Reinhardt, Hanspeter Nick, Felix Freuler, Karoline Boesch, Jérôme Giovannoni, and the cell-based, medium-throughput screening, and compound management groups at Novartis for helpful discussions and experimental help. We gratefully thank Kristin Coan, Gerry Maggiora, and Edgar Jacoby for careful reading and suggestions to improve the study. We thank Steffen Renner and the whole MLI group for discussion. Finally, we also thank reviewers for their constructive comments, which were helpful for improving our manuscript.

■ REFERENCES

(1) Frearson, J. A.; Collie, I. T. HTS and hit finding in academia: from chemical genomics to drug discovery. *Drug Discovery Today* **2009**, *14*, 1150−1158.

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

(3) Mayr, L. M.; Fuerst, P. The future of high-throughput screening. *J. Biomol. Screening* **2008**, *13*, 443−448.

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

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

(6) Inglese, J.; Auld, D. S.; Jadhav, A.; Johnson, R. L.; Simeonov, A.; Yasgar, A.; Zheng, W.; Austin, C. P. Quantitative high-throughput screening: A titration-based approach that efficiently identifies biological activities in large chemical libraries. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 11473−11478.

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

(8) Hsu, N.; Cai, D. Y.; Damodaran, K.; Gomez, R. F.; Keck, J. G.; Laborde, E.; Lum, R. T.; Macke, T. J.; Martin, G.; Schow, S. R.; Simon, R. J.; Villar, H. O.; Wick, M. M.; Beroza, P. Novel cyclooxygenase-1 inhibitors discovered using affinity fingerprints. *J. Med. Chem.* **2004**, *47*, 4875−4880.

(9) Beroza, P.; Damodaran, K.; Lum, R. T. Target-related affinity profiling: Telik's lead discovery technology. *Curr. Top. Med. Chem.* **2005**, *5*, 371−381.

(10) Mestres, J.; Veeneman, G. H. Identification of "latent hits" in compound screening collections. *J. Med. Chem.* **2003**, *46*, 3441−3444.

(11) Wilk, W; Zimmermann, T. J.; Kaiser, M.; Waldmann, H. Principles, implementation, and application of biology-oriented synthesis (BIOS). *Biol. Chem.* **2010**, *391*, 491−497.

(12) Andrews, K. M.; Cramer, R. D. Toward general methods of targeted library design: topomer shape similarity searching with diverse structures as queries. *J. Med. Chem.* **2000**, *43*, 1723−1740.

(13) Boehm, M.; Wu, T. Y.; Claussen, H.; Lemmen, C. Similarity searching and scaffold hopping in synthetically accessible combinatorial chemistry spaces. *J. Med. Chem.* **2008**, *51*, 2468−2480.

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

(15) National Center for Biotechnology Information. PubChem Bioassay Database; AID=893, source=Scripps Research Institute Molecular Screening Center, http://pubchem.ncbi.nlm.nih.gov/assay/assay/assay.cgi?aid=893 (accessed June, 17, 2010).

(16) National Center for Biotechnology Information. PubChem Bioassay Database; AID=886, source=Scripps Research Institute Molecular Screening Center, http://pubchem.ncbi.nlm.nih.gov/assay/assay/assay.cgi?aid=886 (accessed January, 11, 2010).

(17) Didiot, M. C.; Serafini, S.; Pfeifer, M. J.; King, F. J.; Parker, C. N. Multiplexed reporter gene assays: monitoring the cell viability and the compound kinetics on luciferase activity. *J. Biomol. Screening* **2011**, *16*, 786−793.

(18) Abdi, H. The Bonferonni and Sidak corrections for multiple comparisons. In *Encyclopedia of Measurement and Statistics*, 1st ed.; Salkind, N. J., Ed.; Sage Publications: Thousand Oaks, CA, 2007; Vol. *1*, pp 103−107 (available at http://www.utdallas.edu/ξ~herve/Abdi-Bonferroni2007-pretty.pdf, accessed Aug 13, 2008).

(19) Johnson, M. A., Maggiora, G. M., Eds. *Concepts and Applications of Molecular Similarity*; Wiley: New York, 1990.

(20) Renner, S.; van Otterlo, W. A.; Dominguez Seoane, M.; Möcklinghoff, S.; Hofmann., B.; Wetzel, S.; Schuffenhauer, A.; Ertl, P.; Oprea, T. I.; Steinhilber, D.; Brunsveld, L.; Rauh, D.; Waldmann, H. Bioactivity-guided mapping and navigation of chemical space. *Nat. Chem. Biol.* **2009**, *5*, 585−592.

(21) Yan, S. F.; Asatryan, H.; Li, J.; Zhou, Y. Y. Novel statistical approach for primary high-throughput screening hit selection. *J. Chem. Inf. Model.* **2005**, *45*, 1784−1790.

(22) Glick, M.; Jenkins, J. L.; Nettles, J, H.; Hitchings, H.; Davies, J. W. Enrichment of high-throughput screening data with increasing levels of noise using support vector machines, recursive partitioning, and Laplacian-modified naive Bayesian classifiers. *J. Chem. Inf. Model.* **2006**, *46*, 193−200.

(23) Yan, S. F.; King, F. J.; He, Y.; Caldwell, J. S.; Zhou, Y. Y. Learning from the data: Mining of large high-throughput screening databases. *J. Chem. Inf. Model.* **2006**, *46*, 2381−2395.

(24) Lajiness, M.; Watson, I. Dissimilarity-based approaches to compound acquisition. *Curr. Opin. Chem. Biol.* **2008**, *12*, 3.

(25) Lipkin, M. J.; Stevens, A. P.; Livingstone, D. J.; Harris, C. J. How large does a compound screening collection need to be? *Comb. Chem. High Throughput Screening* **2008**, *11*, 482−493.

(26) Schreyer, S. K.; Parker, C. N.; Maggiora, G. M. Data shaving: A novel strategy for analysis of high throughput screening data. *J. Chem. Inf. Comput. Sci.* **2003**, *44*, 470−479.

(27) Nilakantan, R.; Immermann, F.; Haraki, K. A novel approach to combinatorial library design. *Comb. Chem. High Throughput Screening* **2002**, *5*, 105−110.

(28) Grover, G. S.; Turner, B. A.; Parker, C. N.; Meier, J.; Lala, D. S.; Lee, P. H. Multiplexing nuclear receptors for agonist identification in a cell-based reporter gene high-throughput screen. *J. Biomol. Screening* **2002**, *8*, 239−246.

(29) Lyer, P.; Stumpfe, D.; Bajorath, J. Molecular mechanism-based network-like similarity graphs reveal relationships between different types of receptor ligands and structural changes that determine agonistic, inverse-agonistic, and antagonistic effects. *J. Chem. Inf. Model.* **2011**, *51*, 1281−1286.

(30) Zhang, J. H.; Chung, T. D.; Oldenburg, K. R. A simple statistical parameter for use in evaluation and validation of high throughput screening assays. *J. Biomol. Screening* **1999**, *4*, 67−73.

(31) Pipeline Pilot, version 8.0. Accelrys Software Inc., San Diego, CA, 2010. http://accelrys.com/products/pipeline-pilot/.