Shaping a Screening File for Maximal Lead Discovery Efficiency and Effectiveness: Elimination of Molecular Redundancy

Gregory A. Bakken,[Δ] Markus Boehm,[Δ, §] Andrew S. Bell,[•†] Jeremy R. Everett,[•, ‖*] Rosalia Gonzales,[Δ] David Hepworth,[Δ, §] Jacquelyn L. Klug-McLeod,[Δ] Jeremy Lanfear,[•, ≈] Jens Loesel,[•, ‖!*] John Mathias,[•, §] Terence P. Wood[•, ‡]

[Δ]Pfizer Worldwide Research and Development, Groton, CT, USA

[•]Pfizer Worldwide Research & Development, Sandwich, Kent, UK

Current Addresses

[§]Pfizer Research, Cambridge, MA, USA

[†]Imperial College, London UK

[‖]School of Science, University of Greenwich, Chatham Maritime, Kent, UK

[≈]Pfizer Neusentis, Cambridge, Cambridgeshire, UK

[‖!]Peter Fisk Associates Limited, Pean Hill, Canterbury, Kent, UK

[‡]TP & AAW Consultancy, Cliftonville, Kent, UK

ABSTRACT

High Throughput Screening (HTS) is a successful strategy for finding hits and leads that have the opportunity to be converted into drugs. In this paper we highlight novel computational methods used to select compounds to build a new screening file at Pfizer and the analytical methods we used to assess their quality. We also introduce the novel concept of molecular redundancy to help decide on the density of compounds required in any region of chemical space in order to be confident of running successful HTS campaigns.

INTRODUCTION

The pharmaceutical industry is struggling to improve or even maintain its productivity at levels which can sustain current investments in research and development.[1,2,3] Indeed, by some measures, productivity has not changed in 60 years.[4] High Throughput Screening (HTS)[5,6,7,8,9] is a well established and successful method for hit and lead discovery in pharmaceutical and biotechnology companies and increasingly now in academia.[10,11] In our experience and using the existing Pfizer compound file configured in an appropriate manner against selected targets, HTS has delivered hits in $\geq$ 90% of screens, from which new hit to lead projects could be initiated in 50-70% of cases. Other successful methods for hit and lead discovery exist and may be used as alternatives or in parallel to HTS,[12] depending upon the

nature of the target protein whose modulation is required and on the resources available to the researchers. These methods include fragment screening[13,14,15] and virtual screening[16]. In our experience and that of others, these methods are complementary, can be synergistically used together and are not necessarily competitive with each other.[17] Our experience of HTS is that it has successfully led to the drug launch of maraviroc,[18] to the discovery of torcetrapib,[19] and to the enrichment of a significant part of Pfizer's clinical portfolio, as is the experience in other companies.[20]

We identify four key factors that are critical to success in HTS-enabled drug discovery. The first is choice of biological target and its mechanism of modulation (agonism, inhibition etc). Achieving this target modulation must lead to the desired clinical effect in the selected patient population without overt side-effects or toxicities. Secondly, the target must be amenable to modulation by small molecules i.e. it must be druggable.[21,22] Thirdly, the assay(s) chosen to discover hits against that target must be robust and relevant to the physiological environment in which that target functions. Fourth, the screening file used in the assay must be designed to be both effective in producing multiple, high quality hit series and achieve this with an efficient use of resources.

Corporate screening files for HTS consist of collections of molecules from various sources that are built up over a number of years. The sources of these compounds are typically from in-house synthesis, external vendor purchases or strategic collaborations to enrich the file. Compound synthesis and purchase efforts are generally guided by molecular design principles which may include molecular properties, predicted physicochemical properties, predicted safety, predicted molecular stability, molecular diversity and sometimes molecular density considerations. In big pharma, screening files[23,24] are typically made up of multiple, relatively concentrated sets of compounds synthesised for specific medicinal chemistry projects, enhanced by compound or library synthesis, either directed towards specific gene families or to increase chemical diversity. The Pfizer screening file has two distinct components: the library chemistry part of the file, made using library synthesis protocols (mostly as part of our file enrichment efforts) and the legacy chemistry part of the file, made mostly one compound at a time.

Several key concepts are embedded in the molecular design principles mentioned above. Firstly, the Similarity Property Principle states that molecules of similar structure are likely to have similar physical, chemical and biological properties.[25,26] Secondly, whilst chemical space is recognised to be vast, there are regions of chemical space that are desirable because they encompass drug-like or lead-like space, but also regions that are undesirable because they comprise space occupied by potentially reactive, unstable or toxic molecules,[27,28,29] or space occupied by molecules with poor absorption properties. A seminal paper in this latter regard is the work by Lipinski et al[30] which established the now well-known Rule of 5, by defining the ranges of molecular properties that drug-like compounds possessing oral absorption are likely to have. A useful review of methods for assessing drug-likeness or lead-likeness of compounds has recently been published by Oprea et al.[31] Thirdly, molecular diversity is also a key concept in the design of screening files, as it gives a measure of the range of chemical space covered by a particular file. Diversity-based screening files can be

designed using dissimilarity-based algorithms, cell-based approaches such as BCUTs[32] clustering methods or optimisation algorithms.[33,34] Great care must be taken, however, when using dissimilarity-based approaches to select screening compounds, as it is easy to bias the selection towards molecular outliers on the edges of the chemical space of the collection, which may have structures that are unattractive or intractable as chemical leads for optimisation.

The three concepts discussed above allow for the selection of a diverse array of compounds with drug-like or lead-like properties, but a key question still remains: how many compounds should be in a screening collection? Thus, fourthly and finally, we introduce here for the first time the concept of molecular redundancy in order to answer this question. Our work builds on activity probability methodology introduced by Nilakantan et al, [35] which was further developed by Harper et al[36] and Lipkin et al[37] as well as work relating chemical similarity to biological activity by Hajduk and Martin[38] Our computational method enables us firstly to determine how many compounds in a given area of chemical space will be needed in order to generate at least one hit in HTS (see below) and secondly to select redundant compounds that are not required.

The approaches described here were developed to shape the Pfizer screening file in response to the Wyeth acquisition, the merging of the two companies' screening collections and its consequent growth to well in excess of four million compounds. However, the design principles outlined here are of general use and can be applied to the analysis of any existing screening file, or to the creation of a new one. They can also be applied to the design of individual chemical libraries.

The aim of this work was to re-shape the Pfizer screening file so that it could best achieve its primary purpose of finding good quality representatives of each chemical hit series from an HTS in an efficient manner, but without having too many (redundant) representatives of each series present in the file. In other words, we deliberately chose to avoid the generation of dense SAR directly from HTS, with the goal of increased screening efficiency. A more detailed SAR generation on selected hit series would be generated in follow-up experiments, particularly by "exploding" around library hits. This was deemed to be the most efficient way to direct hit discovery activities.

RESULTS

HTS Retrospective Analysis
Our approach to redundancy analysis started with a retrospective, *in silico* analysis of the relationship between the similarity of compounds and the probability of biological activity, in a set of seven recent Pfizer HTS campaigns with a range of target types. In all cases, an identical diversity-based screening subset of 432,000 compounds was utilized. Similar work, using IC50 level data and with a focus on lead-hopping has been reported by Hajduk and co-workers at Abbott.[38] Based on primary HTS data we defined a compound as active at cut-offs of 50% inhibition in the screen (40% for one of the HTS targets). We chose to use primary

HTS data rather than IC50 level data in order to: 1) take a more conservative approach to the estimation of the probability of biological activity at a given Tanimoto similarity and 2) avoid the possibility of compounds being deselected as part of a triage process. A full similarity matrix of all active compounds against all tested compounds was calculated and compounds with Tanimoto values > 0.5 were selected. The compounds were binned according to their Tanimoto similarity values (in 0.05 units). We then derived the probability of activity, p, by determining the percentage of active compounds that had structural similarity to an active compound within a given Tanimoto range (Figure 1). For example, the probability of any compound, which shows a Tanimoto similarity of 0.7 to a known active, to be active itself ranges from approximately 10-18% for our set of HTS targets. We felt reassured that the behavior of the broad range of targets we examined was relatively uniform in the regions of Tanimoto similarity between 0.5 and 0.7: the key region for our analysis (see section on Tanimoto similarity below).
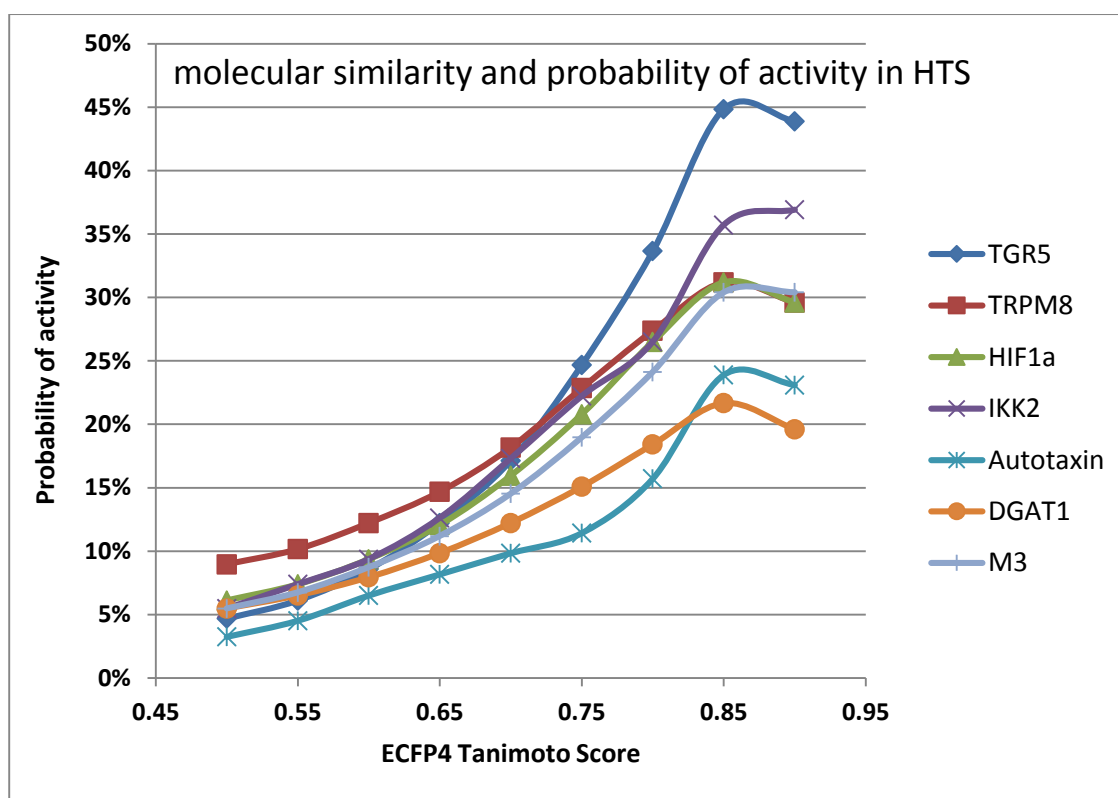

Figure 1: The probability of finding biological activity based on primary HTS data against the Tanimoto similarity of compounds screened for a set of seven recent HTS targets using an identical diversity-based compound set of 432,000 singleton compounds.

Belief Theory

In the next step we used Belief Theory[39] and the results from the HTS retrospective analysis to determine how many compounds at a given Tanimoto similarity need to be screened in order to achieve a threshold probability of finding active compounds from a certain area of chemical space. The chance of finding at least one active compound is given as

$$DC = 1 - (1 - p)^N$$

with: DC = degree of certainty to have at least one active compound; p = likelihood of activity at a given Tanimoto threshold; and N = number of compounds tested with a Tanimoto similarity above the given threshold. For instance, if we want to find at least one active compound in a cluster of 20 compounds above a certain Tanimoto similarity, and with a likelihood p = 10% of finding activity at that Tanimoto value, the degree of certainty of finding that active, DC, is given as follows:

$$DC = 1 - (1 - 0.1)^{20} = 0.88 \text{ i.e. } 88\%$$

In order to reach a degree of certainty of > 95%, at least 30 compounds above a certain Tanimoto similarity need to be considered, assuming the same probability of activity p = 10% (Figure 2).
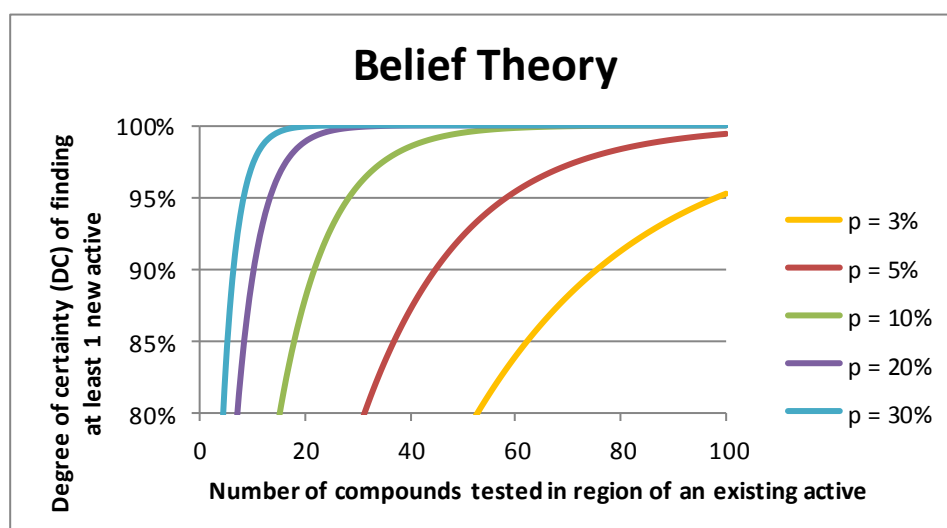

Figure 2: Belief Theory predictions for the likelihood of finding at least one active compound when testing different numbers of similar compounds, calculated for a range of probabilities p of finding biological activity (see Figure 1).

Tanimoto Similarity and Degree of Certainty Thresholds
In the next step we decided on practicable thresholds for the Tanimoto similarity and the degree of certainty for hit finding to use in the design of the re-shaped screening file. Using ECFP4 fingerprints we decided on a Tanimoto similarity threshold of 0.60 as the optimal cut-off. With thresholds below 0.60 we had too many compounds in each similarity sphere that appeared to be unrelated; with Tanimoto similarity thresholds greater than 0.60 we excluded too many compounds inside a similarity sphere of related compounds, according to expert medicinal chemists who extensively reviewed bins of compounds at different Tanimoto cut-offs.

The Degree of Certainty (DC) threshold was deemed to be a business decision. The key question we had to face was how critical it is <u>not</u> to misclassify a chemical series as completely inactive due to insufficient screening. It was crucial from a business perspective

to keep the risk of missing an active series due to redundancy reduction to a minimum. We therefore set the DC threshold as > 95% chance of finding at least one member of an active series in any given region of chemical space. According to our retrospective Pfizer HTS analysis, there is on average an approximately 8% chance that any compound with a Tanimoto similarity of > 0.60 to an active neighbour, will be active itself (see Figure 1). Screening a set of N = 36 such compounds would result in a DC of just over 95% (see Figure 3). We note that this figure is in exact agreement with that from the related work of Lipkin et al[37] who looked at the library size needed for gaining SAR directly from HTS in a small collection containing 85 scaffolds. We decided to go with the slightly more conservative approach of setting N= 40 similar compounds, which results in a DC = 96.4% and more confidence that we would not miss series. This DC value reassuringly also gives $\geq$ 84% chance of finding at least two actives and > 63% chance of finding 3 actives, from that same cluster (Figure 3).
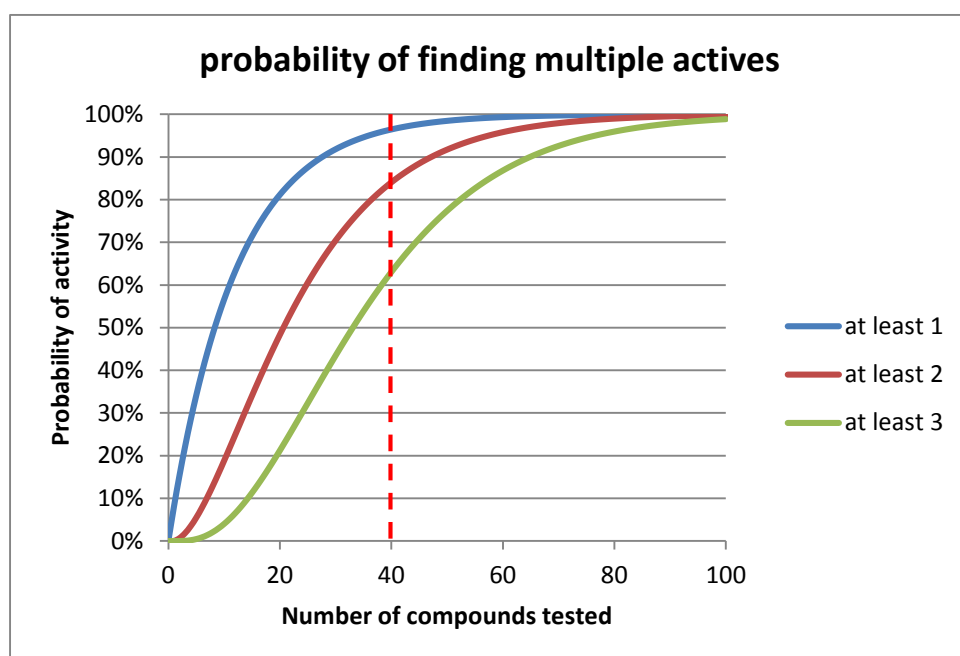

Figure 3: Probability of finding at least 1, 2 or 3 active molecules according to Belief Theory, when screening different numbers of related compounds, with a Tanimoto similarity of > 0.6 to each other, and an average 8% probability of finding activity in any one compound.

Redundancy

The Tanimoto similarity and the probability of activity at a given Tanimoto similarity in combination allowed us to determine the number of compounds that we need to screen in a given area of chemical space in order to achieve a specific Degree of Certainty:

$$N = \frac{\log(1 - DC)}{\log(1 - p)}$$

with, N = the number of compounds deemed sufficient in a given chemical space; DC = Degree of Certainty to find at least one representative of an active series from a cluster of

related compounds within a given Tanimoto threshold; and p = the likelihood of biological activity at a given Tanimoto threshold, as determined by the HTS retrospective analysis. With DC > 95% (actually we set DC = 96.4%, see above) and p = 8%, this results in N = 40 compounds to be screened.

Once N was determined we used it to define a new property called compound Redundancy, R, as the percentage of compounds with a similarity above the Tanimoto cut-off, that are in excess of the minimal number of compounds, N, required to find at least one representative of an active series with a given Degree of Certainty:

$$R = [1 - ([N\text{-}1] / NN)] \text{ x } 100$$

with NN = Number of Neighbouring compounds with a Tanimoto similarity above the Tanimoto threshold relative to another compound. This calculation was carried out for all 4.5 million compounds in the screening file. Note that any compound with NN > 39 is in a redundant region of space as there will be more than 40 compounds in that region. See the Discussion for a greater exemplification.

Selection of Redundant Compounds for Removal from the Screening File

The Redundancy value of a compound, R, as described above can be used directly in any selection/de-selection process to eliminate redundancy in the overall screening file. In the simplest form that we implemented, and the form that we actually used, a random number between 0 and 100 is generated and compared to the percentage Redundancy value of the molecule, which also varies between 0 and 100 (in the limit of infinite similar neighbours: see also the Discussion below) in order to derive compound selection rules:

Redundancy, R < Random => keep compound

Redundancy, R > Random => remove compound

Thus the more redundant a compound is, as measured by its percentage redundancy, R, the more likely it is to be removed in this random selection process. However, the removal of redundancy is more complex than at first sight. Each compound in the screening file is treated as a separate object in the redundancy calculations and the removal of any one compound as redundant reduces the redundancy of all other compounds within its similarity zone. The magnitude of these effects will also vary according to the density of occupation and the shape of the chemical space in that particular region: edge effects can occur. We thus distinguished between Candidate Redundant Compounds (CRC) that at the start of the analysis had NN > 39 and Redundant Compounds (RC) that were actually selected for removal. Note that in all cases CRC > RC. These concepts are illustrated in Figures 4 a to d.
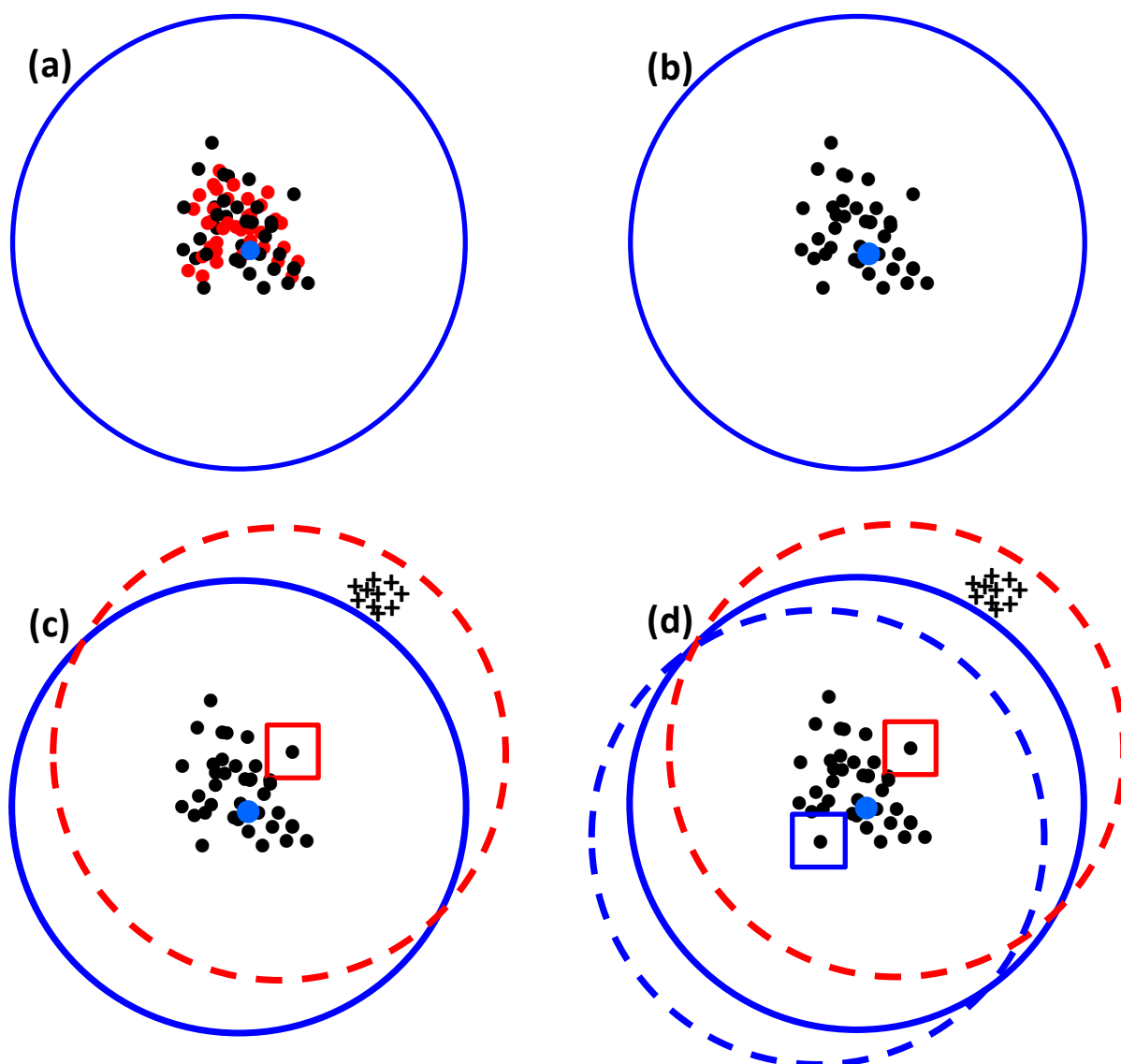
Figure 4: a) a schematic representation of a set of 39 black and 39 red compounds in a molecular neighbourhood with a Tanimoto similarity of > 0.60 to the large central compound in blue: all molecules outside the blue circle have a Tanimoto similarity of < 0.60 and none are show here: the neighborhoods in a) and b) are assumed to be remote from other compounds for simplicity. According to our mathematics and definitions, all 79 compounds in a) (that is blue, red and black compounds) are candidate redundant compounds, the redundancy, R, of each molecule is 50% = $(1 - (39/78))$ x 100% and the redundancy of the set as a whole is 50%. b) after redundancy reduction and removal of 39 redundant compounds shown in red in a), the remaining 39 black compounds within the neighborhood are non-redundant: c) and d) if this neighborhood is actually proximate to other compounds, such as the 10 compounds represented by black + signs, then edge effects will occur. For instance the compound highlighted by the square red box in the north-east quadrant of the neighborhood is now at the centre of a sphere of similarity or neighborhood (dashed red circle) that comprises all 39 of the black circle compounds, the central blue circle compound at the centre of the original neighborhood and the 10 additional black + compounds. All these compounds are now candidate redundant compounds and the compound in the square red box is 20% redundant (R = $(1 - 39/49)$ x 100%). However, as is clear in d) the compound on the other side of the neighborhood, highlighted by the square blue box, still has only 39 neighbors inside the dashed blue circle and is non-redundant (R = 0).

Whilst the above redundancy minimisation procedure can be applied for a set of compounds that is well spread in chemical space, for a highly clustered set or one that is non-uniformly

distributed in chemical space, an iterative redundancy reduction process is preferred. An initial value for N that is significantly higher than the final target number is selected. This first step deselects only compounds from the most densely populated areas of chemical space. Deselected compounds are then removed from lists of neighbours for the remaining compounds. The values for redundancy are recalculated and the value of N is lowered. This process is repeated until the final value for N is achieved and the unwanted redundancy in the screening file is eliminated. We repeated the redundancy reduction process iteratively with N set first at 100, then 80, then 60 and finally at the target value of 40: this worked well in our hands.

In practice however, a further refinement is required before this redundancy reduction methodology can be applied to an existing screening file. Certain compounds which must not be removed e.g. active Research Project compounds, need to be excluded from the removal process by being artificially set as non-redundant and put on so-called Keep Lists.

Keep Lists
To allow preferred handling of important compounds we decided on multiple criteria which would be sufficient to exclude a compound from removal from the file. We called these the 'Keep Lists' as they represented different sets of compounds that needed to be kept in the file irrespective of their redundancy. Our Keep Lists included the following: research project leads, development compounds, Pfizer drugs, members of various active screening subsets, compounds with activity above a certain threshold, new compounds that did not have any biological testing data yet, compounds produced in certain collaborations etc. For simplification, we also included all compounds with less than 39 neighbours on an additional Keep List as we wanted to preserve all non-redundant compounds ($R < 0$).

The use of these Keep Lists had a significant effect, as they reduce the choice of Candidate Redundant Compounds that can be removed in order to reduce the redundancy of the file: see Table 2 in the section on Composition of the Pfizer Screening File below.

Redundancy Reduction in Practice
The overall effect of our iterative, random redundancy reduction process on the library chemistry portion of the screening file, which was the most redundant portion, is shown in Figures 5a to 5d below.
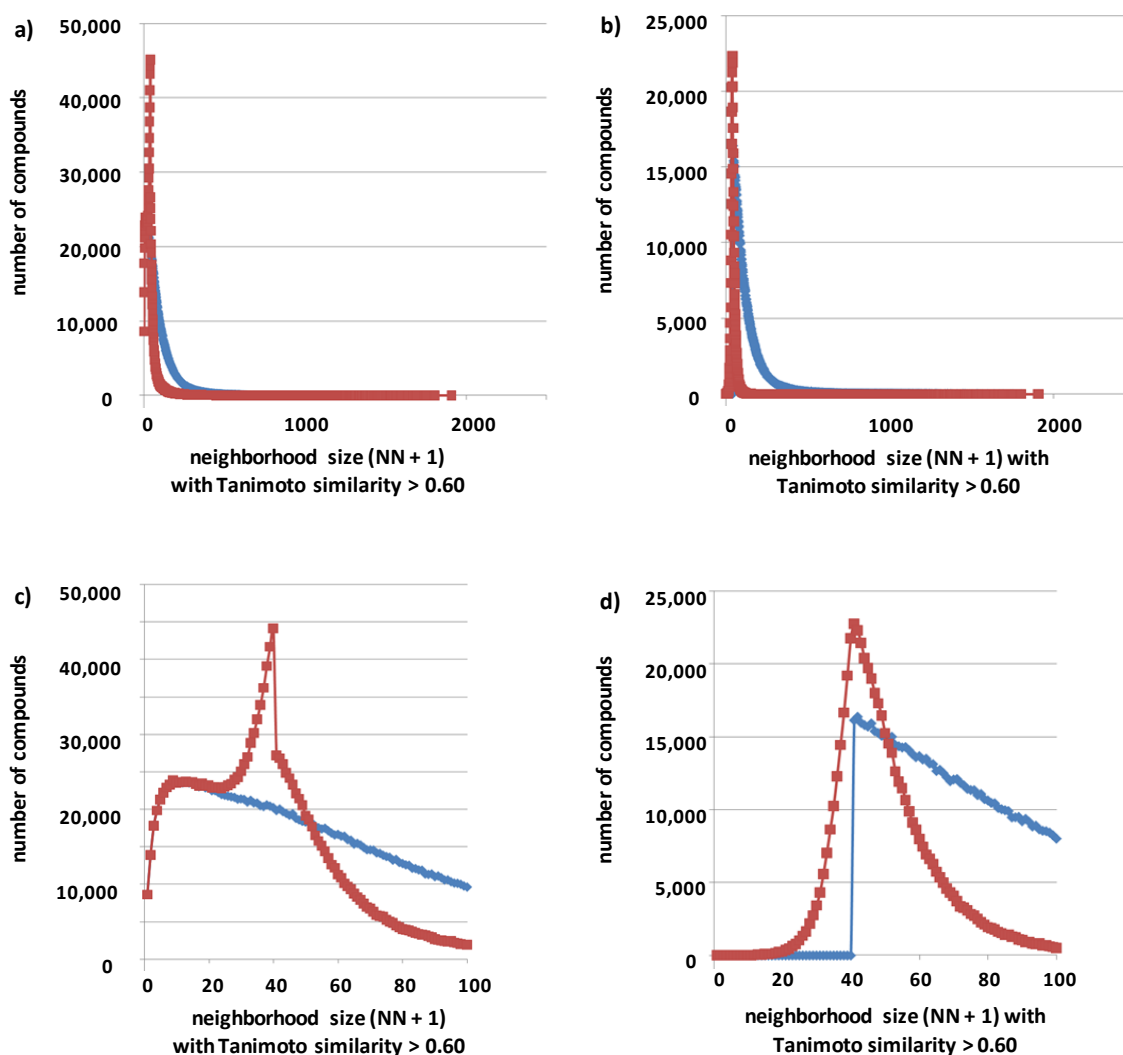
Figure 5: the library chemistry portion of the Pfizer Screening File, before (blue scatter lines) and after (red scatter lines) iterative rounds of random redundancy reduction with N (the redundancy threshold), set successively at 100, 80, 60 and 40. Each plot shows the number of compounds (y-axis) that are present in regions of space with a given neighborhood size (NN + 1, x-axis)), defined with a Tanimoto similarity cut-off of 0.60: a) shown for the entire library chemistry portion of the file with an expansion in c), and b) shown for only those compounds not on Keep Lists and with an expansion at low neighbour numbers in d). Note that because all non-redundant compounds were on a Keep List, there are no compounds in neighborhoods of less than 40 compounds in Figures 5b and 5d, prior to the redundancy reduction. Note also that each neighborhood is defined by a central compound plus its nearest neighbors (NN + 1).

Two key results will be noted from Figure 5: firstly, significant molecular redundancy existed in the library chemistry portion of the Pfizer screening file and secondly, this redundancy was considerably reduced but not eliminated by our methodology. A further perspective is given in Figure 6, which shows the distribution of compounds on the Keep Lists with respect to neighborhood size. This cohort of compounds is dominated by all compounds with ≤ 39 neighbours, as they are on a Keep List by default ($R < 0$).
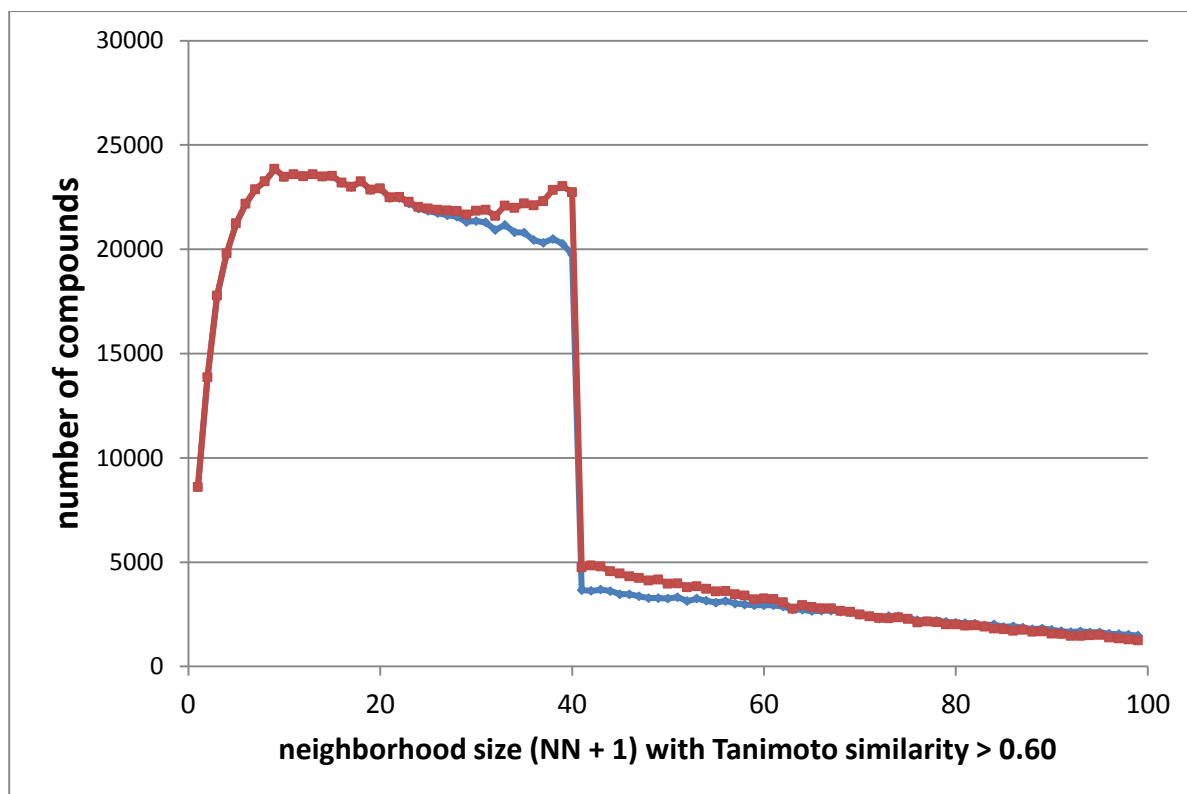
Figure 6: the distribution of library chemistry compounds on Keep Lists against neighborhood size (NN + 1), both before (blue scatter line) and after (red scatter line) iterative, random redundancy reduction. This set includes all non-redundant compounds with NN $\leq$ 39, which were automatically put on a Keep List, hence the discontinuity in the graph at that point. Each neighborhood is defined by a central compound and its nearest neighbours within the similarity boundary T > 0.6.

Validation of Redundancy Removal: Retrospective In Silico HTS Analysis

It was important for us to determine whether the removal of redundant compounds would have any impact on the effectiveness of our HTS hit discovery operations, and to validate that our methodology was sound. To this end, we conducted a retrospective *in silico* analysis of the number of hit series found by 13 recent full file HTS campaigns against a range of target types, using the screening file prior to redundancy removal. We then calculated the number of hit <u>series</u> that would have been missed after the removal of redundant compounds according to the methodology described above (Table 1).

HTS target	Series with IC50 actives	Number of active series removed	% of active series removed
ALK	29	4	14%
Alpha2a	238	4	2%
D1	6	0	0%
HCV	724	4	1%
LEDGF	11	0	0%
OTR	8	0	0%
P2X3	183	7	4%
P2X4	10	2	20%
PDE1b	84	4	5%
PI3Kd	158	8	5%
PI3Kg	363	1	0%
TrpV1	111	4	4%
TrpV4	866	3	0%
average	215	3	4%

Table 1: total number of hit series (with all members of each series having confirmed IC50 activity) found by HTS for 13 recent and varied targets, and the number and percentage of series that would be missed after the redundancy removal from the screening file. Five of these HTSs were biochemical; eight were cell-based.

It was reassuring that the number of hit series that would be missed by redundancy removal was generally very small, with the % of hit series removed generally between 0% and 5%, and with only two of the thirteen targets having more than 10% of hit series removed. The average of 4% of hit series removed is consistent with our mathematics which predicts that the redundancy removal would give ca 96% confidence of still finding a single hit in a cluster of compounds with a Tanimoto similarity of > 0.60.

Structural Filters
In addition to compounds that we deemed necessary to retain (see Keep Lists above) there are conversely compounds that should be preferentially removed due to unfavourable structural features. Hence, a file exclusion filter set was implemented comprising of around 540 separate structural filters and these were applied to eliminate compounds with structural flaws ahead of the redundancy reduction. A full description of these filters is well beyond the scope of this present publication. However, these included established proprietary filter sets from the legacy Pfizer and Wyeth organizations, similar to those that have been published by others,[40,41,42] together with published structural queries to remove structures shown to be problematic and give rise to false positive screening hits,[43, 44] or contain reactive features. [45,46] Additionally, some new filters were added which, for example, removed compounds containing multiple unattractive features/ toxicophores, [47] in a similar fashion to those applied by others.[48] The process used is described in the Experimental section. The outcome is given in Table 2 below.

Composition of the Pfizer Screening File

The initial composition of the Pfizer screening file and the number of compounds that were deemed redundant or that failed the structural filters are summarised in Table 2.

origin	number	number of redundant compounds	% redundant	number of filter fail compounds	% filter fail
Pfizer legacy	1,102,179	87,760	8.0%	108,990	9.9%
Pfizer library chemistry (file enrichment / parallel medicinal chemistry)	2,679,867	932,573	34.8%	8072	0.3%
Pharmacia	452,314	18,570	4.1%	56,778	12.6%
Wyeth	536,000	N/A	N/A	N/A	N/A
Totals	**4,770,360**	**1,038,903**	**21.8%**	**173,840**	**3.6%**

Table 2: Size and origin of the initial dataset of screenable compounds, including the number of compounds in each segment, plus the number of compounds that failed either the redundancy filters or the file entry filters (unwanted structures). Note that the filters were developed and applied prior to the incorporation of the Wyeth compounds so that no further pruning of that segment of the file was needed: hence all entries are N/A = not applicable.

The implementation of the Keep lists (see above) caused an issue for the removal of the redundant compounds. Table 1 indicates that 8.0% of the Pfizer legacy and 34.8% of the Pfizer library chemistry portions of the file were redundant. However, the scope of the Keep Lists was so extensive i.e. so many compounds were protected from removal, that for the Pfizer legacy, Pfizer library chemistry and Pharmacia portions of the file, we needed to remove very high percentages of the candidate redundant compounds (CRC) with more than 39 neighbors. For instance, in the Pfizer library chemistry portion of the file (2,679,867 compounds), 34.8% or 932,573 compounds were redundant. In fact, of the 1,673,611 candidate redundant compounds (more than 39 neighbors) over a quarter of a million were on Keep Lists and protected from removal. We therefore needed to remove 932,573 compounds from the remaining 1.4 MM available: see Table 3 below.

origin	number	number of candidate redundant compounds (CRC) with > 39 neighbors and available for selection	% of candidate redundant compounds requiring selection
Pfizer legacy	1,102,179	158,695	55.3%

Pfizer library chemistry (file enrichment / parallel medicinal chemistry)	2,679,867	1,404,196	66.4%
Pharmacia	452,314	31,339	59.3%
Wyeth	536,000	N/A	N/A
Totals	**4,770,360**	**1,594,230**	**65.2%**

Table 3: Size and origin of the initial dataset of screenable compounds, together with the number of candidate redundant compounds (CRC) with NN > 39, available for redundancy reduction <u>after</u> application of the Keep Lists. Note that the filters were developed and applied prior to the incorporation of the Wyeth compounds, so no reduction of compounds was required from that portion of the file: N/A = not applicable.

Attractiveness and Tiering of the File

An efficient screening file for HTS should deliver novel and attractive compounds as starting material. Early on in this work we developed an algorithm called Medicinal Chemistry Attractiveness (MCA)[49] that assesses whether a compound is sufficiently attractive for medicinal chemistry follow-up, without the involvement of any medicinal chemists. In tests against expert medicinal chemists it performed as well as the best human experts inside Pfizer. The method has similarities to the subsequent QED method published by Hopkins et al.[50]

Analysis of the screening file with the MCA algorithm, after the elimination of both the redundant and the potentially reactive and toxic molecules, revealed a small but significant fraction of the file as unattractive and less likely to be followed up as hits for oral therapeutic targets. Closer inspection of this section of the file revealed that such compounds fall into a number of areas of chemical space, including compounds with known activity against specific targets. For example, steroid-like structures with a tendency to interact with nuclear hormone receptors are likely to be of less interest to projects seeking leads for targets in other gene families.

Including such compound classes in a general screening file for HTS will dilute an efficient screening campaign and potentially unnecessarily deplete those compounds for no reason. We therefore generated a two-tiered file. The Tier 1 set includes all compounds that were deemed as attractive by the MCA algorithm and suitable for standard targets where oral drugs are sought. The Tier 2 set contains compounds tranched in special classes such as macrocyclic compounds, steroids, polypeptides, high-end physicochemical property fails (e.g. Ro6 fails with molecular weight > 600, AlogP > 6.0 etc), generally less attractive compounds e.g. > 6 rings, highly rigid compounds, highly flexible compounds, fragment-like compounds, and complex natural product-like compounds.

Screening File Sample Quality Analysis

We used the large number of analytical quality control (QC) results gathered as part of our on-going HTS activities to assess the physical quality of different sectors of the screening file, as a key adjunct to our redundancy reduction and file filtering activities. All compounds synthesised in Pfizer have their identity and purity confirmed at the point of synthesis. in addition, all compounds undergoing IC50 measurement as part of HTS follow-up in Pfizer have analytical QC measurements made on them (see below) to re-check their quality. While samples from both the Pfizer legacy and the more recent Pfizer library chemistry (internal and outsourced) sectors of the file typically have a modest ~20% failure rate (either wrong compound, impure or very low sample quantity), compounds from certain external sources acquired prior to the application of our QC criteria had a > 50% failure rate. Consequently, all compounds from these sources were removed from the screening file, as their follow-up was known to be wasting resources.

In a separate exercise, a data set containing 130,500 sample QC results, collected from June 2005 to October 2011 of 109,000 Pfizer library chemistry compounds, were analysed to determine if we could confidently remove individual compound/batches based on QC fails, as opposed to whole sets of compounds as described above. We found 3,767 compound/batches that had N ≥ 3 QC results and another 12,536 with N = 2 QC results. These two subsets of data were closely examined for consistency of QC results to assess our degree of confidence in removing these compound/batches from the file. However, disappointingly, issues with the accumulated QC data prevented us from confidently removing suspect library chemistry compound/batches based on past QC data. These issues included: inconsistent definitions of QC 'pass', 'fail', and 'suspect' between different Pfizer groups and experimenters running the QC, as well as lack of repeatability of QC result data when the same compound/batch was tested. Based on this analysis, no individual Pfizer compound/batches were removed from the file based on <u>past</u> QC results. However, due to these issues, process changes were made to ensure that analytical QC could be effectively used in the future: see Discussion and Experimental sections below.

DISCUSSION

Our mission was to reshape the Pfizer screening file so that it was both effective in finding attractive hit series that would be likely to be taken up by therapeutic area project teams, and also efficient in its use of resources. In particular, we made the critical decision that we did not want to generate SAR within hit series as part of the HTS process: that can be more efficiently done as part of the hit follow-up process post-HTS, once the priority hit series are selected, especially via 'explosion' of the chemical space around parallel chemistry hits, using the known synthetic protocols. We did however, want to make sure that we found the vast majority of attractive hit series that would be present in the file in any given HTS.

 Four key approaches were taken in this work: 1) reducing redundancy in the screening file i.e. identifying and removing compounds that are so similar to one another that they add little

or no value to hit series identification and if found would just contribute to SAR understanding; 2) reducing the number of compounds in the file that are unlikely to be selected for follow-up in a medicinal chemistry project, typically due to expected reactivity or toxicity, by the implementation of a battery of molecular filters; 3) partitioning the file into two distinct Tiers: Tier 1 compounds for standard oral delivery projects and Tier 2 compounds for hit identification against non-standard targets such as protein-protein interactions, or targets where non-oral delivery is required and therefore different molecular characteristics will be required; 4) the removal of sectors of the file with particularly high QC failure rates. We will discuss each of these in turn.

Molecular Redundancy

An intensive File Enrichment program[51] in Pfizer in the late 1990s and early 2000s resulted in some large libraries of similar compounds being made by combinatorial or non-combinatorial parallel chemistry i.e. the libraries exhibit various degrees of molecular redundancy. Molecular redundancy was also seen in the Pfizer collection in regions where medicinal chemistry teams have worked within a similar series for some time, leading to many close-in analogues entering the corporate screening file. These issues arose because there were no good methods for determining how many compounds are needed in any given region of chemical space and no clear definition of molecular redundancy. The aim of this present work was to generate the methodology to derive a minimally redundant screening file.

We developed the concept of Redundancy using activity probabilities and Belief Theory, building on the concepts of Nilakantan et al,[35] Harper et al,[36] Lipkin et al[37] and Hajduk and Martin et al,[38] to encompass an analysis of our entire screening file and determine how many compounds with a given similarity we need in the chemical space region of an active compound in order to provide a > 95% confidence level of finding at least one of those compounds by HTS.

An analysis of seven recent singleton HTS against a variety of targets using a subset of the Pfizer screening collection called the Plate-Based Diversity Screening subset (PBDS)[52,53,54] of 432,000 compounds gave us the probability of finding biological activity in a group of compounds with a given Tanimoto similarity coefficient, if a hit was already present. The PBDS is a diverse subset of the screening file, selected according to the chemical diversity and properties of the compounds on a screening plate, rather than an individual compound basis. It is a workhorse screening set of the Pfizer HTS group, is now in its third generation and it will be the subject of a future publication. To establish a Tanimoto threshold that best incorporated the understanding of a chemical series, two expert medicinal chemists visually inspected various Tanimoto bins of multiple compounds series. After extensive medicinal chemistry and computational chemistry analysis, a Tanimoto similarity threshold of 0.60 was chosen as the lowest cut-off below which the concept of a chemical series was lost. The optimal Tanimoto similarity cut-off of 0.60 thus reflected best our in-house medicinal

chemistry judgment; a decision that was ratified by the Pfizer medicinal chemistry leadership and broader community. This collective buy-in was critical since medicinal chemists are generally the prime customers for triaging and assessing hits from HTS campaigns.

With this similarity level, and an average probability of 8% of finding biological activity from the HTS PBDS data, Belief Theory showed us that a neighborhood of 40 similar compounds would give a > 95% chance of finding at least one active molecule in that neighborhood (Figures 1, 2). We used this 40 molecule neighborhood size as the guide to re-shaping the screening file with minimisation of molecular redundancy. Reassuringly, this neighborhood size also gave a > 80% chance of finding 2 hits and a > 60% chance of finding three hits (see Figure 3 above). Note that these neighborhoods are not at all like clusters, as each compound in the file is treated as being at the centre of its own unique neighborhood and thus neighborhoods can massively overlap. A cluster of X compounds isolated in chemical space within a certain similarity, represents, in the context of this work, X neighborhoods, where each of the X compounds has X-1 nearest neighbors. This is an important concept and distinction to make, as if clusters are treated as isolated objects, and no consideration given to their overlapping in chemical space, the results may be quite different[37].Note that in contrast to others[37] we deliberately chose not to make the neighborhood size big enough to generate SAR directly from the HTS, as we felt that this was not the most efficient way to operate. We also note that our selection of an equal number of molecules in each neighborhood is consistent with the conclusions of Harper et al.[36]

So, Redundancy as we defined above is a compound specific property. The original starting file of 4.77 million compounds was treated as having 4.77 million distinct chemical neighborhoods, each centred on a different compound. For example, given our value of N = 40 to have a DC > 95%, a compound with 78 neighbors within the set Tanimoto similarity of > 0.60, would have a redundancy R = 50% i.e. half of its neighbors (39) could be removed without decreasing the probability of finding an active by HTS in that cluster below 95% (see Figure 4). Note that all 79 compounds are Candidate Redundant Compounds (CRC). A compound with 39 similar neighbors has R = 0 and is on the threshold between redundancy and non-redundancy. A negative value of R indicates that further compounds can be added into an area of chemical space without generating redundancy. For the purpose of this work we put all compounds with zero and negative redundancy onto a Keep List as we wanted to protect all these compounds from removal by redundancy reduction.

Redundancy values of compounds can be turned into a Redundancy value for the whole screening file (or subsets thereof) by calculating the average Redundancy of all compounds in a given set. This number represents the percentage of compounds in the screening file which do not contribute positively to finding compounds representative of hit series (as opposed to all members of a hit series within a file) in an efficient way.

The random redundancy de-selection process that we used treats every single compound equally. This works very well for a set of novel compounds – like a virtual combinatorial library. The Pfizer screening file on the other hand has a long history and additional information is known for many compounds. For example, a potent lead compound in an

active program should be treated differently from a compound inside the similarity threshold for which no activity information is available and the lead compound must be protected from removal for redundancy elimination. We therefore implemented the concept of Keep Lists to protect compounds in certain categories from being selected as redundant and removed (see Results section above). The protection of a significant numbers of redundant compounds in Keep Lists caused issues for our redundancy reduction campaign. For example, in the legacy Pfizer portion of the screening file, 8.0% of the 1,102,179 compounds (87,760) needed to be removed to eliminate redundancy, based on our calculations. However, of the total number of 254,094 compounds formally classified as candidate redundant compounds in this portion of the file, only 158,695 were available for removal, after the application of the Keep Lists. That meant that a large percentage (55.3%) of the candidate redundant compounds required actual removal and this caused issues. For example, the removal of a compound X, from a given region of chemical space in the file will reduce the redundancy or eliminate it for the remaining compounds within a given similarity value of compound X (see Figure 4). If compound X has NN = 60, each of those 60 close neighbors will have their redundancy reduced in this single step. The removal of compound X may even have the effect of reducing the number of neighbors for a related compound Y, to less than the redundancy threshold. This secondary effect is difficult or impossible to control with a random redundancy reduction approach. That this effect does occur to a minor extent can be clearly seen for the Pfizer library chemistry portion of the file in Figure 7 below, where there are a minority of compounds which were in neighborhoods of > 40 compounds before, but in neighborhoods of less than 40 compounds after redundancy reduction.
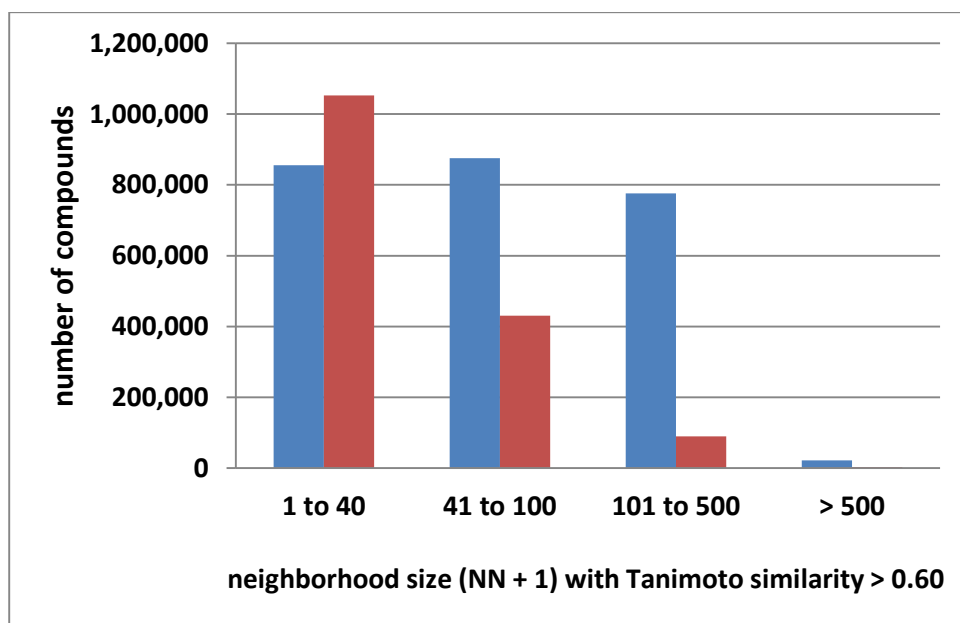

Figure 7: the distribution of compounds in neighborhoods of 1 to 40, 41 to 100, 101 to 500, and > 500 compounds within a Tanimoto similarity of 0.60, both before (blue bars) and after (red bars) the iterative, random, redundancy reduction process in the library chemistry portion of the Pfizer screening file.

Due to this secondary effect of random redundancy reduction, we investigated alternative, non-random algorithms to deselect compounds, such as choosing only those compounds whose removal would <u>not</u> cause other compounds to have less than 39 neighbours. Some of

these showed a better overall profile of redundancy elimination in the final selected file i.e. the process resulted in a smaller percentage of the file having its neighborhood size reduced to fewer than 40 compounds. However, after the application of the Keep Lists, the number of compounds with greater than 39 neighbours whose removal would not cause other compounds to have less than 39 neighbours was lower than our targets for removal. The criterion was that a compound would only be selected for removal if it had more than 39 neighbours, and also if it was only present on neighbour lists that were in redundant space. So, even if a compound had > 100 neighbours, if it was the neighbour of a compound with only 20 other neighbours, it would not be removed. We therefore discarded all non-random approaches and used only a random approach.

A final consequence of the utilisation of Keep Lists is that the redundancy in a screening file cannot be eliminated: it can only be minimised. For the Pfizer library chemistry portion of the file, the redundancy was approximately 35% prior to redundancy reduction and approximately 11 % after redundancy reduction i.e. redundancy was reduced by two-thirds, but not eliminated. Of this residual redundancy, over two-thirds was itself due to the retention of redundant compounds with NN > 39 on the Keep Lists.

The overall effectiveness of the redundancy reduction can be assessed in Figure 7 which shows the large decrease in the number of compounds in either large (41 to 100) or very large (101 to 500) neighborhoods. Note that the compounds shown as having a neighbourhood size of > 40 after redundancy reduction in Figure 7 are a mixture of candidate redundant compounds (more than 39 neighbours and NOT on a Keep List) and compounds on Keep Lists which cannot be removed. The number of these compounds exceeds the stated redundancy because of the fact that each candidate redundant compound with NN > 39, will be causing redundancy to some extent in many other compounds.

 We verified that the random redundancy removal methodology shown above was effective by a retrospective *in silico* analysis that showed that only a small percentage of hit series would have been missed as a result of the removal of the redundant compounds in 13 recent HTS campaigns across a variety of target types (Table 1). We deemed this minimal loss of hit series as insignificant relative to the benefits and efficiencies gained by redundancy removal and proceeded on this basis. It is also worth noting that the average HTS in this set generated 215 novel hit series, each comprising of many compounds confirmed at the IC50 level. On average, only 3 of these series would be lost due to redundancy reduction.

Our analysis showed that the library chemistry portion of the Pfizer screening file was approximately 35% redundant, the legacy Pfizer portion of the file was around 8% redundant and the legacy Pharmacia portion ca 4% redundant. The excessive redundancy in the library chemistry (file enrichment/parallel medicinal chemistry) portion of the screening file is a lesson to all teams involved in parallel/combinatorial chemistry and file enrichment. It is all too easy to synthesise large libraries of redundant compounds, most of which will make no contribution to hit series discovery. In total, over 1 million compounds in the file were declared redundant. These compounds have now been removed from the screening file as far as possible.

The type of analysis that we illustrate above can be 'turned on its head' to highlight areas of screening files which are scantily populated and can therefore be targeted for future file enhancement activities. The analysis can also be used to build new, non-redundant screening files effectively from scratch.

Finally, our new methodology is already incorporated into the molecular design tools used by all Pfizer chemists, so that library chemistry resources are optimally used and chemists are alerted if their design will result in molecular redundancy. The chemist designing the library still has the option of synthesising a redundant library but this is now a positive choice rather than something being done unawares.

Structural Attractiveness Filter

'Ugly' compounds are those with molecular features that make them unattractive to medicinal chemists when they come to select compounds to follow up from HTS hit lists. These compounds may possess unwanted elements, reactive features, potential toxicophores or have undesirable molecular properties such as too high molecular weight or too lipophilic etc. The issue of less desirable compounds in the screening file arose in spite of concerted efforts to eliminate these unwanted compounds over more than a decade in Pfizer. Prior to the start of our work, over 170 computational filters were in place to stop reactive or overtly toxic compounds entering the file. In addition, the file had also been purged of unattractive compounds twice since the mid-1990s in exercises called the 'good, bad and ugly' and the 'beautiful' file splits. The fact that significant numbers of ugly compounds managed to remain, or enter the screening file anew, reflects the great difficulty of creating universal and comprehensive structural filters that do not simultaneously remove desirable compounds.

Wyeth had a similar set of filters in place to protect their file. Applying both filters showed a surprising number of compounds picked up by only a single set of the filters. The merging of the Pfizer and Wyeth screening file gave us the opportunity to revisit the existing filter sets, to combine the experiences from both companies, and incorporate also new findings of structural filters from the literature,[40-48] in addition to the creation of a small number of new proprietary filters to remove for example, compounds with very low feature content (no heteroatoms and low structural complexity) and compounds with multiple unattractive features, such as several nitro groups.[48] This knowledge integration resulted in the generation of a combined set of around 540 molecular filters,[55] which were applied to eliminate compounds with structural flaws ahead of the redundancy reduction.

As we did for the redundancy reduction methodology, the filter set was systematically reviewed by medicinal chemists and computational chemists in all Pfizer sites worldwide, including groups working on animal health and human health, in order to make sure filters were functioning correctly and to gain cultural acceptance for their implementation.

Application of the combined bank of structural filters resulted in just 0.3% of the Pfizer library chemistry compounds failing but approximately 10% and 13% of the legacy Pfizer and Pharmacia compounds failing (see Table 2). These file filter fail compounds have also now been removed from our screening file as far as possible. The high level of file filter

passes (99.7%) for our library chemistry compounds is a testament to the huge design efforts that went into these libraries over the course of a decade, in which every chemist in Pfizer Discovery was involved. This is especially remarkable because this filter pass rate was achieved with hundreds of new filters put in place <u>well after</u> the compounds were designed and synthesised.

File Tiering

We have used additional filter criteria to partition the screening file into two Tiers: Tier 1 is the main screening file for standard, oral drug projects and Tier 2 contains specialized compounds in separate tranches: macrocyclic compounds, natural products, peptides, steroids, fragments, high-end physicochemical property fails, generally unattractive compounds (> 6 rings, highly rigid compounds, highly flexible compounds (> 5 contiguous rotatable bonds), compounds in no other category but low overall structural attractiveness expressed by the MCA score, low featured compounds (one functional group) etc. These Tier 2 compounds may be selected, tranche-by-tranche, as required, for hit identification against non-standard or less druggable targets such as protein-protein interactions, or targets where non-oral delivery is required. However, going forward, Tier 2 compounds will not be selected for general-use screening sets including diversity-based subsets such as the global diverse representative subset (GDRS)[56] or targeted subsets such as the kinase targeted library subsets.[57]

Screening File Sample QC Improvements

The lack of analytical reproducibility found in our retrospective analysis of QC data triggered a campaign to standardise our sample QC workflow and capitalize on this aligned QC data to further improve the overall quality of the file. Firstly, globally consistent definitions of compounds that were 'Pass', 'Fail', and 'Suspect' in analytical QC were proposed and vetted with medicinal and analytical chemists. Analytical methods across different groups were also aligned, right down to the column types, instrument settings, injection parameters, and data analysis algorithms (see Experimental). These aligned methods have since made it easier for us to operate consistently internally and to be able to transfer and enable external partners with equivalent sample QC workflows. Secondly, we defined a QC Discard criterion: this is when a compound has three consecutive QC Fail results. When a compound/batch meets the QC Discard criteria, it triggers the removal of all samples of said compound/batch from existing liquid stocks, thereby preventing the follow-up of the compound/batch post primary screening. A newly synthesized compound/batch meeting the QC Discard criteria is also prevented from entering the screening collection. To date, close to 10,000 compound/batches have met the QC Discard criteria and have been eliminated from the file. Finally, it is worth noting that the strategy of first developing and then using automated purification[58] that operated at a capacity of hundreds of thousands of compounds per year, as part of our very large, outsourced, file enrichment process, has led to those compounds having a quality at least equivalent to that of standard medicinal chemistry singletons. This is a remarkable achievement and critically important for those compounds to be accepted by medicinal chemists as desirable substrate for HTS follow-up activities.

EXPERIMENTAL

Dataset

Determining the final dataset for an analysis as described above is a non-trivial task. The screenable file, or those compounds available for screening in any particular HTS, is not easily defined, since new compounds are registered on a daily basis, and at the same time, a small number of other compounds get depleted. For this reason we defined a set cut-off date (October 25th 2009) at which a snapshot of the file was taken for this analysis.

Further complications arose due to file discrepancies and inconsistencies caused by the existence of multiple versions of the screening file, including compounds stored in compressed format (screening plates with multiple compounds mixed in a single well),[59] in singleton format (screening plates with only a single compound in each well), in deep freeze storage (compound master plates), or as solid material. This, coupled with the large numbers of compounds involved, and with geographical considerations (the Pfizer file was stored and replicated in two main locations) further exacerbated the complexities of the task.

The dataset was generated by starting with the complete Pfizer file of all registered compounds. From this collection a subset was generated that comprised all 4,770,360 screenable compounds. Screenable compounds have a defined chemical structure, with sufficient physical material available for screening and are available from the corporate material management centres. This set included multiple compound sources: Legacy Pfizer (including Warner Lambert), Pfizer library chemistry (File Enrichment, external, combinatorial compounds), Pfizer parallel medicinal chemistry (PMC, internal, combinatorial and non-combinatorial), legacy Pharmacia and Wyeth. Compounds with the same parent (non-salt form) structure were selected only once i.e. a complete de-duplication and de-replication of the file was effected, removing a total of 236,937 compounds: itself a remarkable figure. Note that duplicates are different forms of the same parent compound e.g. a sodium salt as well as a potassium salt, whereas replicates are multiple instances of the same form of the compound i.e. multiple batches of the sodium salt of a compound. This process resulted in a total dataset of approximately 4.53 million non-replicated and non-duplicated compounds. All calculations and % values in Table 2 above are against the original file containing both duplicates and replicates.

Similarity calculations

Computational work on the dataset was carried out using Pipeline Pilot from Accelrys.[60] ECFP4 fingerprints[61] were used for all similarity calculations of compounds. The Tanimoto coefficient[62] was used as the measure of similarity between compounds. In an initial step a complete similarity matrix for two random subsets of 50,000 compounds (one from legacy Pfizer, one from Pfizer library chemistry compounds) against a 3.8 million compound set (full set without the Wyeth compounds) was calculated. All compounds with a Tanimoto similarity of > 0.5 to each member of the 50,000 compound subset were selected. For the final analysis a complete similarity matrix of 4.53 million by 4.53 million non-duplicated and

non-replicated compounds was generated. The calculations were executed by calculating a series of 50,000 by 4.53 million compound matrices in parallel using all available Pipeline Pilot servers from Sandwich and Groton, with several jobs on each server and then combining the results. This overall computation took several days.

File Filters

Structural filters to remove molecules with undesirable functional groups were implemented as a file exclusion workflow using Pipeline Pilot from Accelrys.[60] This workflow allowed the convenient processing of any given compound through several sets of structure filters from different sources and in various formats (MDL mol format,[63] Daylight SMARTS queries[64]). Pre-existing legacy Pfizer filters consisted of 170 substructure queries (MDL mol). These were originally developed to prevent compounds with undesirable reactivity entering the HTS screening file. An additional set of 230 substructure queries (MDL mol) were added that were derived from the legacy Wyeth structure filter set. The combined set of Pfizer and Wyeth filters were further refined based on the analysis of a diverse set of 200,000 compounds from the legacy Pfizer and Wyeth portion of the file. A random subset of compounds that either passed or failed the filters were visually inspected by expert medicinal chemists. Compounds that failed the filters but were deemed to be desirable by the chemists were tagged, and the according substructure query was modified to pass the molecule. Conversely, any undesirable compounds still passing the filters were captured and new substructure queries were defined (Daylight SMARTS) and added to the set of already existing filters. In parallel, surveying the recent literature describing structural filters allowed us to include additional substructure filters.[40-48] Overall, around 140 filters were newly created and combined with the existing filter set, resulting in 540 unique substructure queries.

Sample QC Fails

A number of analytical methods have been used at Pfizer over many years to both confirm the identity and measure the concentration of compounds in the Pfizer file, usually as part of normal compound synthesis or screening activity, or as a protocol for HTS, which in Pfizer employs an automated and integrated analytical quality control (QC) check of compounds that are selected for IC50 measurements. The standard analytical methodology for the QC of IC50 compounds includes automated LC-MS with UV and evaporative light scattering detection (ELSD) so that molecular identity (molecular weight via positive and negative ion electrospray ionisation (ESI) MS), purity (via UV profile) and concentration (ELSD) are all determined from a single sample injection.

A Waters Acquity UPLC instrument was interfaced with a Waters PDA detector and Waters ZQ mass spectrometer. UV was scanned from 210 to 400 nm and extracted at 215 nm. The ESI data was acquired in positive and negative ion mode by scanning the mass range of m/z 150-1000 every 0.18 sec with a scan delay of 0.1 sec. Other mass spectrometer parameters included an applied capillary voltage of +3.5 kV, a sample cone potential at 25 V, and source and desolvation temperatures of 120 and 400 ºC, respectively. The ELS Detector nebulizer

temperature was set at 30ºC while the evaporator was at 55 ºC, with gas flow at 1.6 L/min. Data rate was fixed at 10 Hz.

Samples in 96- or 384-well plates at 1 mM in DMSO were injected onto a 2.1 x 50 mm CSH C18 1.7 μm column (Waters) at a flow rate of 1.0 mL/min. The mobile phase solvents were: A = 0.1% Formic Acid in Water, B = 0.1% Formic Acid in Acetonitrile. Elution was achieved via the delivery of a 1.2 min 5-100% B gradient followed by a 1.5 min hold at 100% B prior to re-equilibration.

Sample QC Result Definitions: the identity PASS criteria are when: a protonated molecular ion, adduct, or simple fragment are found e.g., M-H, M+H, M+Na, M+DMSO, M+H+ACN, M-H + formic acid, multiple charges such as (M+2H)/2, (M+3H)/3, (M+4H)/4, dimers such as 2M+H, 2M + Na, 2M-H, or simple fragments M+H-H20, M+H-NH3. The purity PASS criteria are when: (i) the signal from the ultraviolet detector at 215 nm wavelength is greater than or equal to 80% of the expected signal (UV215 ≥ 80%) with no ELSD signal, or (ii) ESLD ≥ 80% with no UV215, or (iii) UV215 ≥ 80% and ELSD ≥ 80%. The identity FAIL criteria are when the expected mass ions are not found by positive or negative ionization mode using ESI or APCI methods. The purity FAIL criteria are when: (i) UV215 < 80% with no ELSD, or (ii) ELSD < 80% with no UV215, or (iii) UV215 < 80% and ELSD < 80%. A sample must meet both identity and purity PASS criteria in order to get a QC Result of PASS. If results are ambiguous; for example, due to a method failure or the absence of the required ionization method, then the sample gets a QC Result of SUSPECT. This is a conservative approach.

Conclusions

We developed novel methodologies and used a novel combination of existing methods to remove compounds from the Pfizer screening file on the basis of molecular redundancy in addition to molecular attractiveness. These methods will also be of use to those who wish to build a screening file, avoiding redundancy in the collection and will also find application in compound and library design initiatives for on-going projects, where the density of chemical space coverage is varied according to stage of lead optimization e.g. avoiding overly dense sampling of chemical space in the more exploratory phase of lead development.

Overall this work resulted in the reduction of the Pfizer screening collection by approximately 1,449,680 compounds, including redundant, replicated, duplicated and filter fail compounds, thus allowing for significant efficiencies in material handling and storage with, we believe, minimal impact on our ability to identify attractive and viable lead series from HTS screens.

Recent work has demonstrated that whilst corporate screening files have undoubtedly improved in quality over the past 10 years, significant differences in the approaches taken to optimise HTS hits in different companies can lead to widely diverging properties in the leads and drugs derived from those hits.[65] Thus, the work described here can help provide a good

starting point for HTS-based lead discovery file design but it is important to ensure that good drug design principles and practices are adopted in the hit-to-lead-to-drug phases that follow the initial hit identification by HTS.

AUTHOR INFORMATION

Corresponding Authors

*E-mail: j.r.everett@greenwich.ac.uk , Jens.Loesel@ecotoxchem.co.uk

Notes

The authors declare no competing financial interest.

ABBREVIATIONS

APCI, atmospheric pressure chemical ionisation; CRC, candidate redundant compound; DMSO, dimethyl sulfoxide; ECFP4, Scitegic/Accelrys' level 4 extended connectivity fingerprints; ELSD, evaporative light scattering detector; ESI, electrospray ionisation; FE, file enrichment; HTS, high throughput screening; LC, liquid chromatography; MCA, medicinal chemistry attractiveness; MS, mass spectrometry; PBDS, plate-based diversity screening; PDA, photodiode array; PMC, parallel medicinal chemistry; QC, quality control; RC, redundant compound; SAR, structure activity relationships; UPLC, ultra high pressure liquid chromatography; UV, ultraviolet

TABLE OF CONTENTS GRAPHIC

probability of finding multiple actives

[Chart: Probability of activity (y-axis, 0% to 100%) vs Number of compounds tested (x-axis, 0 to 100), with three curves: "at least 1" (blue), "at least 2" (red), "at least 3" (green)]
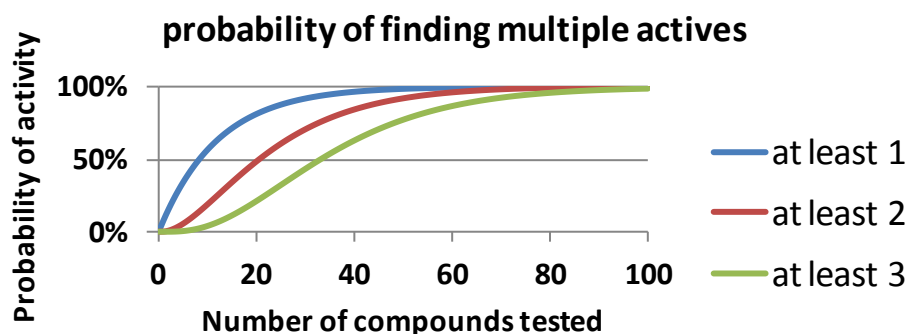
REFERENCES

1 Scannell, J. W.; Blanckley, A.; Boldon, H.; Warrington, B. Diagnosing the decline in pharmaceutical R & D efficiency. *Nat. Rev. Drug Discovery* **2012**, *11*, 191-200.

2 Pammoli, F., Magazzini, L.; Riccaboni, M. The productivity crisis in pharmaceutical R&D. *Nat. Rev. Drug Discovery* **2011**, *10*, 428-438.

3 Mullard, A. 2011 FDA Drug Approvals. *Nat. Rev. Drug Discovery* **2012**, *11*, 91-94.

4 Munos, B. Lessons from 60 years of pharmaceutical innovation. *Nat. Rev. Drug Discovery* **2009**, *8*, 959-968.

5 Pereira, D. A.; Williams, J. A. REVIEW: HISTORICAL PERSPECTIVES IN PHARMACOLOGY. Origin and evolution of high throughput screening, *Br. J. Pharmacol.* **2007**, *152*, 53-61.

6 *High Throughput Screening: Methods and Protocols (Methods in Molecular Biology)*; Janzen, W. P., Ed.; Humana Press: New Jersey, 2002.

7 *High-Throughput Screening in Drug Discovery (Methods and Principles in Medicinal Chemistry)*; Huser, J., Ed.; Wiley-VCH: Weinheim, 2006.

8 A *Practical Guide to Assay Development and High-Throughput Screening in Drug Discovery (Critical Reviews in Combinatorial Chemistry)*; Chen, T., Ed.; CRC Press: Boca Raton, 2010.

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

10 Frye, S.; Crosby, M.; Edwards, T.; Juliano, R. US Academic Drug discovery. *Nat. Rev. Drug Discovery* **2011**, *10*, 409-410.

11 Brenk R.; Schipani, A.; James, D.; Krasowski, A.; Gilbert, I. H.; Frearson, J.; Wyatt, P. G. Lessons Learnt from Assembling Screening Libraries for Drug Discovery for Neglected Diseases. *ChemMedChem*. **2008**, *3*, 435-444.

12 Leach, A. R.; Hann, M.H. Molecular complexity and fragment-based drug discovery: ten years on. *Curr. Opin. Chem. Biol*. **2011**, 15, 489-496.

13 Hajduk, P.; Greer, J. A decade of fragment-based drug design: strategic advances and lessons learned. *Nat. Rev. Drug Discovery* **2007,** *6*, 211-219.

14 Coyne, A. G.; Scott, D. E.; Abell, C. Drugging challenging targets using fragment based approaches, *Curr. Opin. Chem. Biol*. **2010**, *14*, 299-307.

15 Larsson, A.; Jansson, A.; Åberg, A.; Norlund, P. Efficiency of hit generation and structural characterization in fragment-based drug discovery. *Curr. Opin. Chem. Biol*. **2011**, *15*, 482-488.

16 Schneider, G. Virtual screening: an endless staircase? *Nat. Rev. Drug Discovery* **2010**, *9*, 273-276.

17 Hajduk, P.J.; Galloway, W. R. J. D.; Spring, D. R. A question of library design, *Nature*, **2011**, *470*, 42-43.

18 Dorr, P.; Westby, M.; Dobbs, S.; Griffin, P.; Irvine, B.; Macartney, M.; Mori, J.; Rickett, G.; Smith-Burchnell, C.; Napier, C.; Webster, R.; Armour, D.; Price, D.; Stammen, B.; Wood, A.; Perros, M. Maraviroc (UK-427,857), a Potent, Orally Bioavailable, and Selective Small-Molecule Inhibitor of Chemokine Receptor CCR5 with Broad-

Spectrum Anti-Human Immunodeficiency Virus Type 1 Activity. *Antimicrob. Agents and Chemother.* **2005**, *49*, 4721-4732.

[19] Schmid, E. F.; Smith, D. A. R&D technology investments: misguided and expensive or a better way to discover medicines? *Drug Discovery Today* **2006**, *11*, 775–784.

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

[21] Hopkins, A.L.; Groom, C. R. The Druggable Genome. *Nat. Rev. Drug Discovery* **2002**, *1*, 727-30.

[22] Overington, J. P.; Al-Lazikani, B.; Hopkins, A. L. How many drug targets are there? *Nat. Rev. Drug Discovery* 2006, *5*, 993-996.

[23] Schuffenhauer, A.; Brown, N. Chemical diversity and biological activity, *Drug Discovery Today: Tech*. **2006**, *3*, 387-395.

[24] Jacoby, E.; Schuffenhauer, A.; Popov, M.; Azzaoui, K.; Havill, B.; Schopfer, U.; Engeloch, C.; Stanek, J.; Acklin, P.; Rigollier, P.; Stoll, F.; Koch, G.; Meier, P.; Orain, D.; Giger, R.; Hinrichs, J.; Malagu, K.; Zimmermann, J.; Roth, H- J. Key Aspects of the Novartis Compound Collection Enhancement Project for the Compilation of a Comprehensive Chemogenomics Drug Discovery Screening Collection. *Curr. Top. Med. Chem*. **2005**, *5*, 397-411.

[25] Willett, P. Chemoinformatics: a history. *Wiley Interdisc. Rev.: Comp. Mol. Sci.* **2011**, *1*, 46-56.

[26] Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood behaviour: a useful concept for validation of molecular diversity descriptors. *J. Med. Chem*. **1996**, *39*, 3049-3059.

[27] Dobson, C. M. Chemical space and biology. *Nature* **2004,** *432*, 824-828.

[28] Lipinski, C.; Hopkins, A. Navigating chemical space for biology and medicine. *Nature* **2004**, *432*, 855-861.

[29] Stepan, A. F.; Walker, D. P.; Bauman, J.; Price, D. A.; Baillie, T. A.; Kalgutkar, A. S.; Aleo, M. D. Structural Alert/Reactive Metabolite Concept as Applied in Medicinal Chemistry to Mitigate the Risk of Idiosyncratic Drug Toxicity: A Perspective Based on the Critical Examination of Trends in the Top 200 Drugs Marketed in the United States. *Chem. Res. Toxicol*. **2011**, *24*, 1345-1410.

[30] Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, *Adv. Drug Deliv. Rev*. **1997**, *23*, 3-25.

[31] Ursa, O.; Rayan, A.; Goldblum, A.; Oprea, T. I. Understanding drug-likeness, *Wiley Interdisc. Rev.: Comp. Mol. Sci.* **2011**, *1*, 760-781.

[32] Pearlman, R. S.; Smith, K. M. Novel software tools for chemical diversity, *Perspect. Drug Discov. Des.* **1998**, *9-11*, 339-353

[33] Gillet, V. J. Diversity selection algorithms. *Wiley Interdisc. Rev.: Comp. Mol. Sci.* **2011**, *1*, 580-589.

[34] Huggins, D.; Venkitaraman, A.; Spring, D. R. Rational methods for the selection of diverse screening compounds. *ACS Chem. Biol.* **2011***, 6*, 208-217.

[35] Nilakantan, R.; Nunn, D. S. A fresh look at pharmaceutical screening library design. *Drug Discovery Today,* **2003**, *8*, 668-672.

[36] Harper, G.; Pickett, S. D.; Green, D. V. S. Design of a Compound Screening Collection for use in High Throughput Screening. *Comb. Chem. High Throughput Screening* **2004**, *7*, 63-70.

[37] Lipkin, M. J.; Stevens, A. P.; Livingstone, D. J.; Harris, C. J. How Large Does a Compound Collection Need to Be? *Comb. Chem. High Throughput Screening* **2008,** *11*, 482-493.

[38] Muchmore, S. W.; Debe, D. A.; Metz, J. T.; Brown, S. P.; Martin, Y. C.; Hajduk, P. J. Application of Belief Theory to Similarity Data Fusion for Use in Analog Searching and Lead Hopping. *J. Chem. Inf. Model.* **2008**, *48*, 941-948.

[39] Dempster, A.P. Upper and Lower Probabilities Induced by a Multivalued Mapping. *Ann. Stat*. **1968**, *28*, 325-339. See also Glenn Shafer at http://www.glennshafer.com/ for further references and information (accessed October 1st, 2012).

[40] Sink R.; Gobec S.; Pečar S.; Zega A. False positives in the early stages of drug discovery. *Curr. Med. Chem*. **2010**, *17*, 4231-55

[41] Nadin, A.; Hattotuwagama, C.; Churcher, I. Lead-Oriented Synthesis: A New Opportunity for Synthetic Chemistry. *Angew. Chem. Int. Ed*. **2012**, *51*, 1114–1122.

[42] Jadhav, A.; Ferreira, R. S.; Klumpp, C.; Mott, B. T.; Austin, C. P.; Inglese, J.; Thomas, C. J.; Maloney, D. J.; Shoichet, B. K.; Simeonov, A. Quantitative Analyses of Aggregation, Autofluorescence, and Reactivity Artifacts in a Screen for Inhibitors of a Thiol Protease. *J. Med. Chem.* **2010**, *53*, 37–51

[43] Pearce, B. C.; Sofia, M. J.; Good, A. C.; Drexler, D. M.; Stock, D. A. An empirical process for the design of high-throughput screening deck filters. *J. Chem. Inf. Model.* **2006**, *46*, 1060-1068.

[44] Baell, J. B.; Holloway, G. A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J. Med. Chem.* **2010** *53*, 2719-2740

[45] Hann, M.; Hudson, B.; Lewell, X.; Lifely, R.; Miller, L.; Ramsden, N. Strategic pooling of compounds for high-throughput screening. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 897-902.

[46] Huth, J. R.; Mendoza,R.; Olejniczak, E. T.; Johnson, R. W.; Cothron, D. A.; Liu, Y.; Lerner, C. G.; Chen, J.; Hajduk, P.J. ALARM NMR: A Rapid and Robust Experimental Method To Detect Reactive False Positives in Biochemical Screens*. J. Am. Chem. Soc.* **2005,** *127*, 217-224.

[47] Blagg, J. Structure–Activity Relationships for In vitro and In vivo Toxicity. *Ann. Rep. Med. Chem.* **2006***, 41*, 353-368.

[48] Lajiness, M.; Watson, I. Dissimilarity-based approaches to compound acquisition. *Curr. Opin. Chem. Biol.* **2008**, *12* , 366-371.

[49] Loesel, J. Medicinal Chemistry Attractiveness (MCA): A new algorithm for selecting HTS hits. Personal communication*.* **2011**.

[50] Bickerton, G. R.; Paolini, G. V.; Besnard, J.; Muresan, S.; Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nat. Chem*. **2012**, *4*, 90-98.

[51] Milne Jr, G. M. Pharmaceutical productivity — the imperative for new paradigms. *Ann. Rep. Med. Chem*. **2003**, *38*, 383-396.

[52] Everett, J. R.; Loesel, J. Unpublished methodology in use in Pfizer since 2006, but related to the methods disclosed by Novartis in references 53 and 54

[53] Crisman, T. J.; Jenkins, J. L.; Parker, C. N.; Hill, W. A.; Bender, A.; Deng, Z.; Nettles, J. H.; Davies, J. W.; Glick, M. "Plate cherry picking": a novel semi-sequential screening paradigm for cheaper, faster, information-rich compound selection*. J. Biomol. Screen*. **2007**, *12*, 320-7.

[54] Sukuru, S. C.; Jenkins, J. L.; Beckwith, R. E.; Scheiber, J.; Bender, A.; Mikhailov, D.; Davies, J. W.; Glick, M. Plate-based diversity selection based on empirical HTS data to enhance the number of hits and their chemical diversity. *J. Biomol. Screen*. **2009,** *14*, 690-9.

[55] Boehm, M.; Hepworth, D. Personal communication, **2011**.

[56] Yeap, S. K.; Walley, R. J.; Snarey, M.; Van Hoorn, W. P.; Mason, J. S. Designing Compound Subsets: Comparison of Random and Rational Approaches Using Statistical Simulation. *J. Chem. Inf. Model*. **2007**, *47*, 2149-2158.

[57] Xi, H.; Lunney, E.A. The design, annotation, and application of a kinase-targeted library. *Methods Mol. Biol*. **2011**, *685*, 279-91.

[58] Everett, J.R.; Gardner, M.; Pullen, F.; Smith, G.F.; Snarey, M.; Terrett, N. The Application of Non-Combinatorial Chemistry to Lead Discovery. *Drug Discovery Today*. **2001**, *6*, 779-785.

[59] Kainkaryam, R.M.; Woolf, P.J. Pooling in high-throughput drug screening. *Curr. Opin. Drug Discovery Dev*. **2009**, *12*, 339-350.

[60] *Pipeline Pilot,* version 7.5; Accelrys: San Diego, CA, **2008.**

[61] Gardiner, E. J.; Holliday, J. D.; O'Dowd, C.; Willett, P. Effectiveness of 2D fingerprints for scaffold hopping. *Fut. Med. Chem*. **2011**, *3*, 405-414.

[62] Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today*. **2006**, *11*, 1046-1053.

[63] *MDL molfile*, Accelrys: San Diego, CA, **2008.**

[64] *Daylight Toolkit*, version 4.94; Daylight Chemical Information Systems Inc.: Laguna Niguel, CA, **2008**

[65] Leeson, P. D.; St-Gallay, S. A. The influence of the organizational factor on compound quality in drug discovery. *Nat. Rev. Drug Discovery* **2011**, *10*, 749-765.