J Comput Aided Mol Des (2011) 25:621–636
DOI 10.1007/s10822-011-9434-0

Design of a multi-purpose fragment screening library using molecular complexity and orthogonal diversity metrics

Wan F. Lau · Jane M. Withka · David Hepworth · Thomas V. Magee · Yuhua J. Du · Gregory A. Bakken · Michael D. Miller · Zachary S. Hendsch · Venkataraman Thanabal · Steve A. Kolodziej · Li Xing · Qiyue Hu · Lakshmi S. Narasimhan · Robert Love · Maura E. Charlton · Samantha Hughes · Willem P. van Hoorn · James E. Mills

Received: 3 February 2011 / Accepted: 6 May 2011 / Published online: 21 May 2011

Abstract Fragment Based Drug Discovery (FBDD) continues to advance as an efficient and alternative screening paradigm for the identification and optimization of novel chemical matter. To enable FBDD across a wide range of pharmaceutical targets, a fragment screening library is required to be chemically diverse and synthetically expandable to enable critical decision making for chemical follow-up and assessing new target druggability. In this manuscript, the Pfizer fragment library design strategy which utilized multiple and orthogonal metrics to incorporate structure, pharmacophore and pharmacological space diversity is described. Appropriate measures of molecular complexity were also employed to maximize the probability of detection of fragment hits using a variety of biophysical and biochemical screening methods. In addition, structural integrity, purity, solubility, fragment and analog availability as well as cost were important considerations in the selection process. Preliminary analysis of primary screening results for 13 targets using NMR Saturation Transfer Difference (STD) indicates the identification of uM–mM hits and the uniqueness of hits at weak binding affinities for these targets.

Electronic supplementary material The online version of this article (doi:10.1007/s10822-011-9434-0) contains supplementary material, which is available to authorized users.

W. F. Lau (✉) · J. M. Withka (✉) · D. Hepworth · T. V. Magee · G. A. Bakken · M. D. Miller · Z. S. Hendsch · V. Thanabal
Pfizer Global Research and Development (PGRD), Eastern Point Rd, Groton, CT 06340, USA
e-mail: wlau64p@gmail.com

J. M. Withka
e-mail: jane.m.withka@pfizer.com

D. Hepworth
e-mail: david.hepworth@pfizer.com

T. V. Magee
e-mail: thomas.v.magee@pfizer.com

G. A. Bakken
e-mail: gregory.a.bakken@pfizer.com

M. D. Miller
e-mail: michael.d.miller@pfizer.com

Z. S. Hendsch
e-mail: zhendsch@yahoo.com

V. Thanabal
e-mail: venkataraman.thanabal@pfizer.com

Y. J. Du
PTC Therapeutic, Inc., 100 Corporate Court, South Plainfield, NJ 07080, USA
e-mail: yjoshdu@gmail.com

S. A. Kolodziej
Pfizer Global Research and Development (PGRD), Saint Louis, MO, USA
e-mail: steve.a.kolodziej@pfizer.com

L. Xing · M. E. Charlton
Pfizer Global Research and Development (PGRD), Cambridge, MA, USA
e-mail: li.xing@pfizer.com

M. E. Charlton
e-mail: mecharlton@verizon.net

Q. Hu · L. S. Narasimhan · R. Love
Pfizer Global Research and Development (PGRD), La Jolla, CA, USA
e-mail: jerry.hu@pfizer.com

Keywords Fragment screening · Fragment based drug design · Library design · Chemical diversity · Chemical space

Introduction

The identification of novel chemical matter using FBDD has become a broadly accepted alternative to traditional high throughput screening (HTS) methods [1–4]. This alternative multi-discipline paradigm has historically been often been employed in pharmaceutical companies for difficult targets in which large corporate compound files did not provide appropriate chemical substrate for lead development. The major gap in the discovery of novel, viable chemical matter for challenging targets has been attributed to the lack of coverage of appropriate chemical space despite the multi-million compound collections which contain ligands biased toward the Lipinski's Rule of 5 [5]. Over the years corporate files have become saturated with compounds designed and synthesized for well established target families and mechanisms such as GPCRs, kinases and ion channels. As we tackle less established drug targets and pathways, we will need to find creative solutions for the expansion of both target and chemical space while reducing the cost of early discovery. FBDD offers at least one avenue to pursue in the identification of novel chemical substrate for lead development. This approach for lead generation involves screening libraries of low molecular weight fragments that are significantly smaller and functionally simpler than drug molecules. These fragments typically exhibit low affinity binding (uM–mM) and thus require higher sensitivity detection methods to detect the binding [6]. Despite their weak potencies, the fragments often exhibit high ligand efficiency (LE) of binding (LE $\geq$ 0.3), defined as the ratio of free energy of binding to number of heavy atoms [7]. These fragments are typically optimized by carefully adding functionalities to increase binding affinity and selectivity while maintaining high LE in order to maximize chances of obtaining leads with appropriate pharmacokinetic properties as potency is enhanced. In addition to dealing with new and tough targets [2], the FBDD strategy has shown promise in efficiency improvements in the screening and chemical optimization process for conventional drug targets [8, 9] as well as becoming a screening tool to determine druggability [10] for prioritization of more expensive screening methods such as HTS.

The use of fragment based screening for critical decision making regarding project initiation or continuation only underscores the need for a very carefully selected and diverse chemical fragment library. Several other library design strategies have been previously reported [11–15]. Many of these approaches utilize similar property filters to ensure quality but we were aware that the overall strategy taken to generate a fragment library may often depend upon the intended fragment screening method and optimization strategy as well as resources available to create the library. Previous NMR fragment libraries generated at Pfizer taught us that in addition to maximizing diversity, having chemically expandable fragments, availability of solid material to enable hit validation/characterization and access to structurally related analogs is critical to engage chemists and facilitate fragment optimization. In this manuscript we describe the design of a new Pfizer fragment library named Global Fragment Initiative (GFI) library that is suitable for multi-disciplinary screening using Nuclear Magnetic Resonance (NMR), Surface Plasmon Resonance (SPR), X-ray crystallography, Mass Spectrometry (MS) and biochemical techniques. Our consideration of multiple screening methods was necessary to accommodate specific target requirements in terms of availability of stable reagents, sensitivity of developed assays for detection and druggability of binding sites. Biochemical and biophysical fragment screening methods and their respective advantages and disadvantages for high concentration screening have been extensively described [6, 16–18]. Biochemical assays are historically the method of choice for high throughput screening at relatively low ligand concentrations. Although they can be employed in fragment screening, the methods of detection and the possibility of artifacts must be considered at high ligand concentrations. As a consequence, orthogonal biophysical methods, capable of detection in the uM–mM range are often implemented to survey the entire fragment hit landscape for many targets. NMR methods such as STD [19] and Waterlogsy [20] experiments have been shown to be extremely sensitive screening methods [1]. Incorporation of competitive probes in NMR allows higher affinity detection into the nM range, if desired [21]. X-ray

L. S. Narasimhan
e-mail: lakshmi.narasimhan@pfizer.com

R. Love
e-mail: robert.love@pfizer.com

S. Hughes · J. E. Mills
Pfizer Global Research and Development (PGRD), Sandwich, UK
e-mail: samantha.hughes@pfizer.com

J. E. Mills
e-mail: james.e.mills@pfizer.com

W. P. van Hoorn
Accelrys Ltd, 334 Cambridge Science Park, Cambridge CB4 0WN, UK
e-mail: willem.vanhoorn@accelrys.com

crystallography is another excellent method to detect a wide range of fragment affinities in the nM–mM range and is generally done by screening small mixtures or "cocktails" of fragment molecules [22]. This method can be limited by throughput and fragment solubility and does not provide information on target—fragment affinities. SPR methods are becoming efficient fragment screening methods as higher throughput instruments are available. Optimal affinity ranges for detection in SPR are typically in the nM–uM range [23, 24].

Two main classes of compounds are typically used in development of fragment-based screening libraries and methods. The first class of compounds usually called fragments have $\leq$250 MW where useful hits bind with affinities in the high uM to mM range, and is generally anticipated to be screened using high sensitivity biophysical methods [25]. The second class of compounds often called scaffolds, have up to 350 MW and useful hits often have an affinity in the nM–uM range, which may be more suitable for higher throughput biochemical screening [26]. Our goal was to formulate a library incorporating between 3,000 and 5,000 fragments in multiple phases to maximize diversity while enabling efficient screening with medium throughput biophysical screening methods. This library contains the first class of compounds that is, 96% of the compounds are $\leq$250 MW and the remaining 4% of the compounds have a maximum of 300 MW. The planned size of this library is consistent with other fragment libraries that have previously been described, which typically contain 500–5,000 fragments [27]. As chemical space is exponentially smaller for lower MW fragments, significantly smaller libraries are capable of efficiently sampling this space [9, 11]. Additionally, the hit rate is expected to be higher for fragment screening compared to screening full sized and more complex ligands in HTS as predicted by Hann's model for ligand-receptor interactions. Hann and colleagues have described a simple model to predict the probability of finding a hit as the complexity of the compound increases [28]. This model predicts that as complexity increases the probability of measuring the interaction increases; however, the probability of finding a productive match falls exponentially due to an increase in the number of mismatches. According to this model, the probability of detecting a hit binding in a unique manner increases when small, less complex fragments with fewer interactions are screened with high sensitivity. The group at Novartis [13] validated the Hann model with an analysis of their HTS data and comparing that with NMR screening data on low MW fragment libraries using Similog keys [29] as a complexity measure. In this study, the HTS actives contained many more Similog keys, were significantly more complex than NMR actives, and had lower hit rates of 0.001–0.151% in the uM range as compared with hit rates of greater or equal to 3% for NMR hits in the mM range. We reasoned that if these results were reproducible for our fragment library, then we should on average generate $\sim$100 hits for each target screened.

The overall design aims for our multi-purpose fragment library was to identify low MW fragments that have the appropriate complexity for binding and detection, covers as diverse a set of pharmacophoric and structural functionality as possible within the constraints of good molecular properties and are chemically attractive and expandable to ensure suitability for lead development. The intent of these efforts was not to develop a library with known target chemotypes but rather to have a collection that will be suitable for screening against a wide range of pharmaceutical targets and binding sites.

Library design strategy

Our design process was carried out in two phases with the first phase referred to as GFI-I, designed to achieve pharmacophore and framework diversity while the second phase referred to as GFI-II, focused on framework and target diversity. The general process is summarized in Fig. 1, and included four stages in each phase. The first stage was focused on filtering corporate and commercial databases and the Pfizer Global Virtual Library (PGVL) [30] on availability of solids in the corporate file, unwanted functionality, reactivity, chemical complexity, molecular properties and a set of physicochemical property filters based around the Rule of three (RO3) [31]. Physical properties were critical design features and included a more stringent $100 < \text{MW} \leq 250$ and $\text{cLogP} \leq 2.0$ requirement than RO3 criteria. Of particular interest was the identification of fragments with $\text{cLogP} \leq 2.0$ for reasonable solubility and optimal enthalpic interactions with the target to enable follow-up [32]. The second stage involved diversity selection using multiple orthogonal diversity metrics as discussed below. The third stage involved visual inspection by a team of medicinal chemists, NMR spectroscopist and computational chemist with experience in different therapeutic areas to ensure suitability for detection, chemical attractiveness, stability and synthetic accessibility. A library consisting of drug-like fragments which are chemically expandable at more than one vector was an essential component to ensure rapid fragment follow-up and optimization. Chemical expandability was not defined explicitly but done by individual fragment inspection as judged by experienced medicinal chemists. The final stage involved analytical confirmation of structural integrity, purity, solubility and determination of sample logistics for multiple screening methods to enable high concentration screening. To facilitate fragment follow-up and optimization

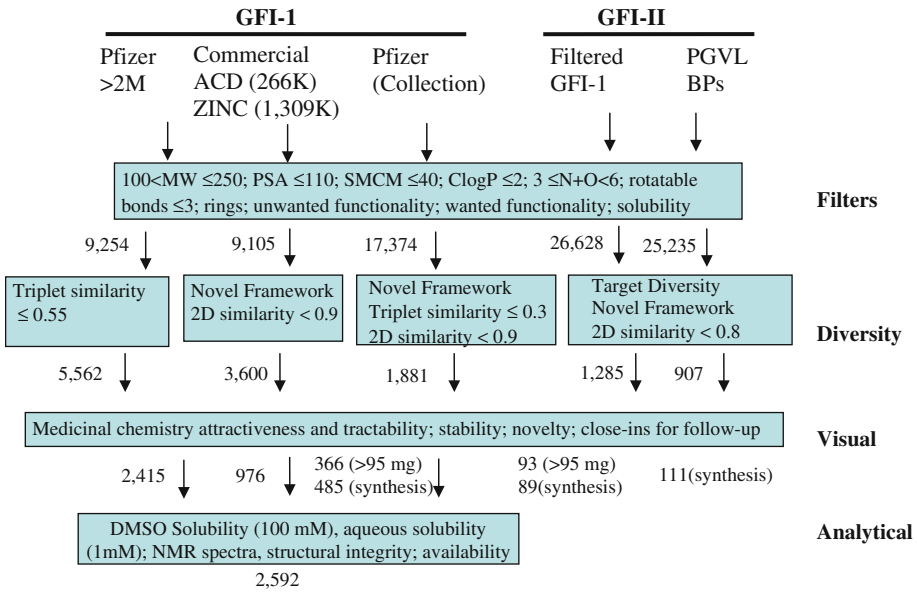

Fig. 1 Summary of GFI fragment library design process

across all research sites, the long term availability and cost of fragments as well as access to close-in analogs in Pfizer's large corporate file or commercial sources were important considerations on the selection of each phase of this library.

The GFI was designed to have a ratio of roughly 4:3:3 for neutral, +1 and −1 charged fragments respectively, similar to the AstraZeneca HCS set of 2,000 compounds which had roughly equal proportion of neutral, basic, and acidic compounds [14, 33]. An analysis of the WOMBAT 2005.1 [34] database found that 80% of the 342 unique structures with MW $\leq$ 200 that gave activities better than 10 nM against 98 targets: 41 enzymes, 42 receptors, 6 ion channels and 9 other proteins were likely to be charged at pH 7.4. There were 269 cations, 22 anions and 15 zwitterions [35]. Although the majority of these are cations, there are many marketed low molecular weight drugs that are anionic (e.g. aspirin, naproxen, barbiturates, AZT, thiopentone, bendazac, mycophenolic acid, penicillins, NSAIDs, antifolates). Twenty two of the top one hundred brand name drugs in 2006 are anionic, excluding acid prodrugs and zwitterionic drugs [36]. The use of charged polar fragments seemed reasonable as many drugs have been optimized from polar leads by increasing molecular weight and lipophilicity to improve potency and selectivity [37]. Hence we opted to follow the AstraZeneca strategy of including a high proportion of charged compounds in the screening set and chose a balanced ratio of charged fragments.

While filters for molecular properties, diversity methods and criteria for synthetic accessibility to some degree, can be defined, metrics for biological activity were less well known. A library of attractive and synthetically expandable fragments covering diverse chemical and pharmacological space would not be very useful if the activity cannot be detected by our multiple screening methods. The work of Hann and Schuffenhauer on the relationship between molecular complexity and activity suggested complexity as a potential metric for biological activity. Hence, we needed some quantitative measures for molecular complexity and to define optimal ranges for these measures where there is an acceptable probability of a match as well as being able to measure it with our different screening methods. To cover the wide range in sensitivities of the different screening methods for the library, HTS screening and NMR screening were selected as references for low sensitivity and high sensitivity screening respectively, and the Synthetic and Molecular Complexity Metric (SMCM) first introduced by Oprea [38] and 3D pharmacophore triplet counts were utilized as measures for determining the optimal molecular complexity ranges for the HTS and NMR screens respectively. In preparation for library selection, a set of Pfizer HTS hits and NMR fragment screening hits were analyzed to obtain metrics for our SMCM and pharmacophore complexity indices. The SMCM method which takes into account structural complexity is fast and efficiently processes millions of compounds, making it ideal as an initial filter. An analysis of HTS actives culled from 23 targets, of which 9 are kinases and 7 are GPCRs indicated actives had a minimal molecular weight (MW $\geq$ 138) and a minimal molecular complexity (SMCM $\geq$ 8). As fragments would be screened at much higher concentrations than HTS compounds, we did not apply a lower cutoff to MW and SMCM. Since only 3.5% of the computationally filtered fragments had MW $<$ 140 and less than 0.5% had SMCM $<$ 8, the

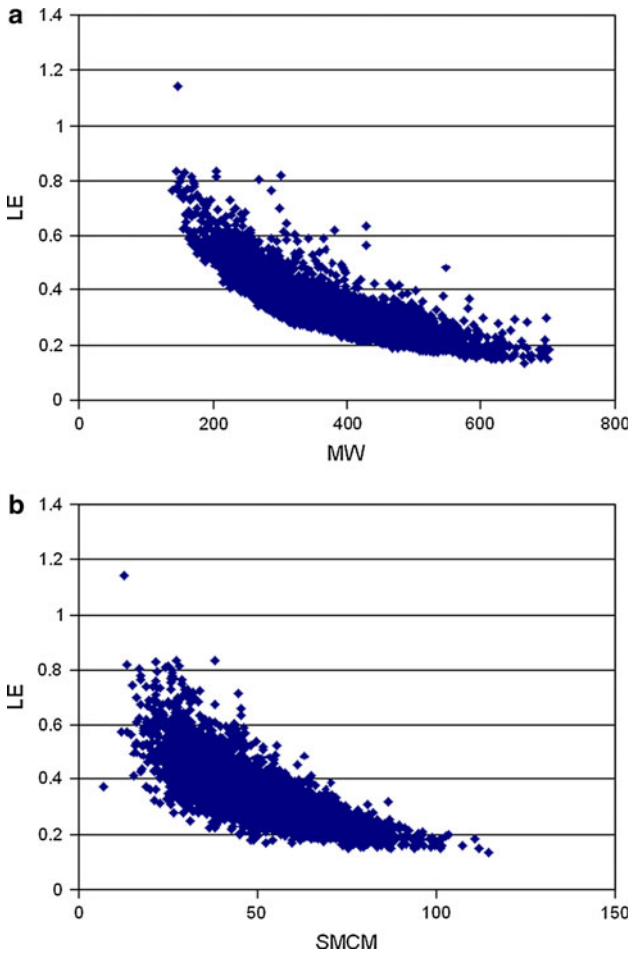

Fig. 2 **a** Correlation of L.E. with molecular weight for Pfizer HTS actives. **b** Correlation of L.E. with SMCM for Pfizer HTS actives

majority of the fragments were kept at a MW $\geq$ 140 and SMCM $\geq$ 8 during the selection process. Plots comparing LE, relative to MW and SMCM for the HTS hits in Fig. 2a, b respectively show a rapid drop in LE as MW and SMCM increases. This trend is similar to that reported previously in which a number of factors including enthalpy, entropy, structural constraints and surface area have been hypothesized to play a role in the reduction of LE with increasing MW and complexity [39]. There was a moderate correlation of SMCM with MW ($R^2 = 0.597$, N = 5,109) in our data set. This partly contributes to the trend seen in the rapid drop in LE with increasing SMCM and the extremely high rate of LE hits with SMCM < 40. The HTS data illustrated that the probability of finding ligand efficient hits increased when the molecular complexity and molecular weight were low, that is SMCM ≤ 70 and MW ≤ 450, respectively. With SMCM alone as a filter, the percentage of ligand efficient hits (LE ≥ 0.3) when SMCM ≤ 40 is 95.6% and rapidly decreases to 72.6, 42.5, 21.8 and 4.5% respectively as SMCM is incremented by 10. This suggested a SMCM maximum of 40 for optimal activity. Furthermore, visual inspection of random subsets of compounds with SMCM scores ranging from 40 to 50 showed that the majority of them were complex and synthetically unattractive. By contrast, random subsets of compounds with SMCM scores between 35 and 40 contained many compounds that were synthetically accessible suggesting the cutoff of 40 was also suitable for synthetic tractability.

Although more computationally intensive than SMCM, the number of 3D pharmacophore features are more correlated to binding interactions and have the advantage that it can also be used as a measure of structural diversity. An analysis of our legacy fragment collection and the resultant 1,190 NMR hits, when screened against six diverse targets was carried out. The triplet distribution for the NMR hits showed that fewer than 5% of the hits had less than 5 triplets, 75% of the hits had 10–150 triplets and fewer than 5% of the hits had more than 280 triplets. Hence, we attempted to keep the distribution of triplet numbers per molecule and the pharmacophore complexity of the GFI library similar to the NMR actives. In this process, pharmacophore triplet count filtering was applied. Fragments with 5 or fewer triplets were all visually inspected and selected if they contained sufficient chemical features to make them interesting, diverse, and chemically expandable. There were four hits with triplet counts ranging from 350 to 360. Hence the maximum of 400 triplets seen in the three legacy NMR libraries was used as an upper limit and fragments with more than 400 triplets were excluded. If a fragment pair exceeded the pharmacophore Tanimoto similarity cutoff in the pairwise pharmacophore similarity comparison, the more complex fragment with the higher pharmacophore triplet count was eliminated from the selection pool.

Fragment selection methods

Compound selection

In this study, compounds were selected from multiple sources including the Pfizer file collection, commercially available compounds from ACD 2006 version (266 K) and ZINC 2006 database (1,309 K) [40]. For the ZINC database, we looked at compounds from ChemBridge, ChemDiv, Asinex, Maybridge and Comgenex. In addition fragment selection was available from the PGVL, a readily synthesizable virtual compound space of the order of 10^{14}, which represented close to 1,000 synthetic protocols in more than 757 reaction templates. Basis Products (BPs) [41] are the products formed by combining all the reactants for a given reaction component with the simplest set of complementary reactant partners. By this approach, the

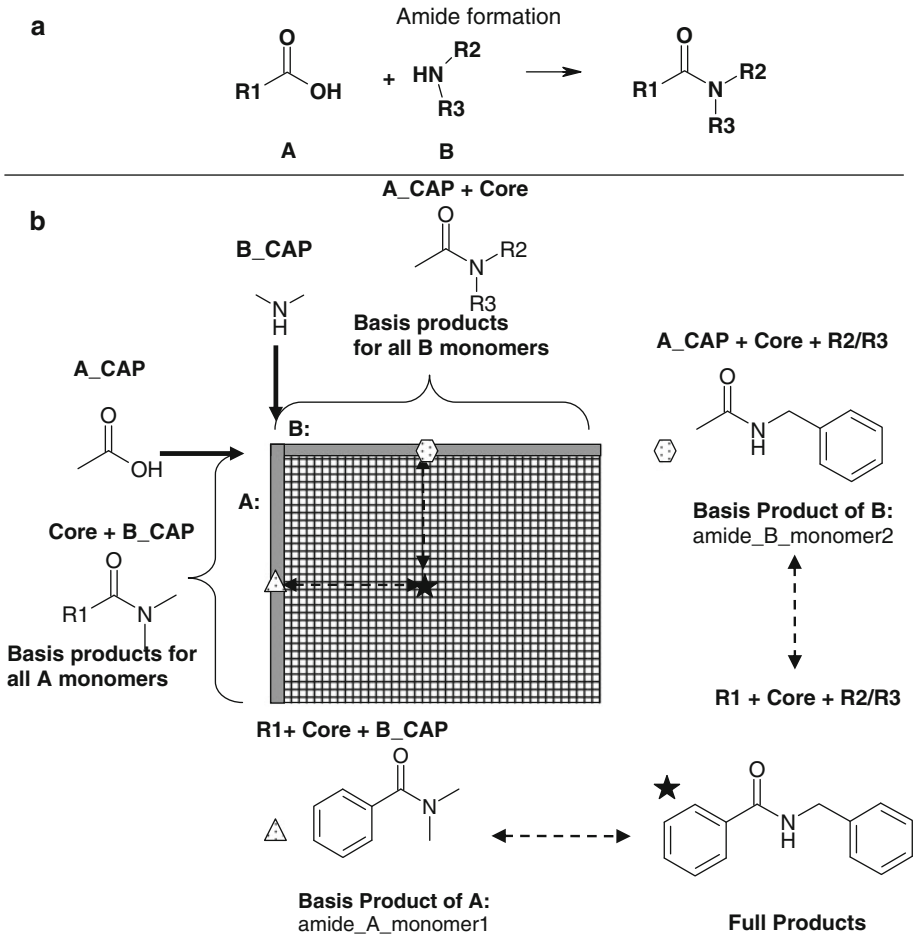

Fig. 3 Illustration of the Basis Products concept using amide formation with carboxylic acids and amines. **a** The reaction scheme shows that all products will share the same amide core. **b** The products matrix for the reaction, with acids A in *rows* and amines B in *columns*. The Basis Products of A are products of B_CAP (dimethyl amine) and all A monomers. The triangle shows an example of Basis Products of A monomer. The Basis products of B are A_CAP (methyl carboxylic acid) and all B monomers. The hexagon shows an example of Basis Products of B. The star is the corresponding full product which can be encoded by the combination of representing monomers from the *triangle* and the *hexagon* respectively

Table 1 List of functional groups, and corresponding molecular weight filters used for molecules containing each group

Exception	MW allowed
1 sulfonamide	300
1 bromine	300
>1 chlorine	292
>1 sulfur	288
>2 fluorines	283
2 fluorines	272
1 chlorine	271
1 sulfur	269
1 fluorine	261

For molecules containing multiple functional groups, the highest molecular weight limit was used

combinatorial explosion is averted, yet BPs provide a robust sampling of the PGVL product space [30] as there is a one to one correspondence between the reactants and the BPs. The current size of BPs is $\sim 10^7$ compounds, far smaller than 10^{14}. A basis product which contains information of R-groups as well as the core is illustrated in Fig. 3.

Computational filters

Molecules were first filtered by ACD ClogP ≤ 2 and atom type. Molecules with elements other than C,N,O,S or halogens were removed. If the total number of nitrogen and oxygen atoms were greater than six, the molecule was removed. A molecular weight filter with a base value of $\leq$250 amu was also applied, with exceptions listed in Table 1. If a single molecule contained multiple exceptions, the highest allowed molecular weight for any exception that was present was used. Additional filters were also applied to remove compounds based on ring systems. Compounds with no ring systems, as well as compounds with more than two separate ring systems or more than 3 rings (single or fused rings) were removed. Molecules with more than two atoms that were part of three ring bonds (fused rings larger than bicyclic ring systems like anthracene) or with more than one atom that was part of 4 ring bonds were removed. Note that spiro ring systems were considered part of the same ring system for this work. Some examples of ring systems that would be removed by this filter are shown in Fig. 4. Filters were used to remove molecules containing reactive or non-drug like functionalities

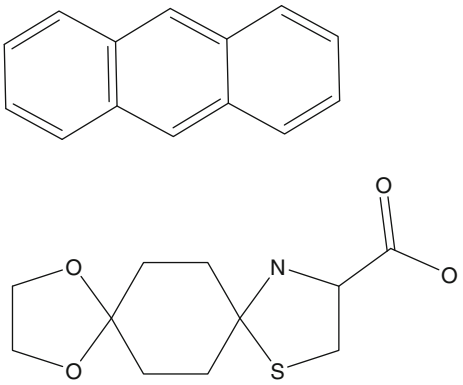

Fig. 4 Examples of ring systems that were filtered out in the design process

Table 2 Desirable chemical groups

Chemical handles		
–COOH	–OH	–NH2
–COOMe	–OMe	–NHMe
–CONH2	–SMe	–NHCOR
–CONHMe	–ArBr	–NRCOR
–CONMe2	–ArCl	–NHSO2R
–SO2NH2	–ArMe	
–SO2NHMe		
–SO2NMe2	R = alkyl group	
Unsubstituted, expandable fragments		
Ar–Ar		
Ar–N–Ar		
Ar–O–Ar		
Imidazoles/benzimidazoles		
Pyrazoles/Indazoles		
Pyrimidines/Quinazolines		
Triazoles		
Oxadiazoles		
Cyclic secondary amines		

that we considered unattractive in lead matter (unwanted functionalities filters) such as aldehydes, acetals, amidines, guanidines, immines, hydrazines, alkynes, alkylators such as epoxides, acrylates and other examples of such functional groups that have previously been published [11, 15]. Molecules containing >4F, >2Cl, >1Br, or >2 S atoms were also removed. Molecules were filtered to remove compounds that did not have at least one of the desirable functional groups listed in Table 2. Molecules were also removed if total polar surface area (TPSA) was greater than 110 $Å^2$. Although a TPSA $\leq$ 60 $Å^2$ has been suggested for a RO3 constructed fragment library, we used 110 $Å^2$ as a cutoff to accommodate the charged fragments. The TPSA calculation used was the topological variant introduced by Ertl et al. [42]. If the total number of nitrogen and oxygen atoms was fewer than 3, or if there were more than three contiguous rotatable bonds, the molecules were removed.

Compounds were then filtered using two in-house solubility models to select compounds with predicted aqueous turbidimetric solubility $\geq$10 mg/mL and thermodynamic solubility [43] greater than 100 uM. Any compounds with a SMCM score greater than 40 were filtered out. Compounds with more than one ionizable center were removed and the remaining compounds were separated into three sets based on their charge states of neutral, +1 and −1. Working with a medicinal chemist, the fragments were then filtered using Leadscope Enterprise™ [44, 45] to remove functionalities that were difficult to define with simple SMARTS queries filters without also removing reasonably acceptable compounds. This filtering process gave us 9,254 fragments with availability greater than 200 mg from the Pfizer collection, 17,374 additional new Pfizer fragments where availability was not considered, 9,105 commercial fragments, and 68,293 PGVL BPs.

Fragment diversity methods

2D similarity

Different software packages for the 2D similarity analyses and different similarity cutoffs based on the fingerprints were utilized. PCAT (Pfizer internal software) utilizes Daylight fingerprints [46] and was used for clustering the sets for triage using Wards Hierarchical Clustering [47] with an average distance cutoff threshold of 0.1. PCAT was later used in similarity expansion of the selected fragments to look for close-in analogs with a similarity cutoff of 0.95. After selection, the Pfizer file compounds in GFI-I were compared against the 9,105 filtered commercial fragments with Pipeline Pilot [48] which utilizes MDL public keys [49]. Pipeline Pilot was the most convenient tool for the similarity matrix generation which was needed for this comparison. All commercial fragments with a similarity of 0.9 or greater were manually inspected and removed unless deemed different from visual inspection.

In GFI-II, the 68,293 BPs obtained by applying the above computational filters to the 3,458 K BP virtual library was further filtered to remove BPs that were similar to the GFI-I fragments. A pair-wise similarity analysis to the GFI-I set using MDL public keys were used to remove BPs with greater than 0.8 Tanimoto similarity. As the compounds would require synthesis, we chose a lower similarity cutoff. A Murcko Framework filter [50] was used to filter out BPs with frameworks that are in common with the GFI-I compounds and occur at frequency greater than or equal to 5. A reaction filter was then used to remove

compounds that come from less robust or narrow scope reactions. This gave a final 25,235 BPs for the target profiling.

3D-pharmacophore triplets

An in-house program was used for identifying pharmacophores, computing the pharmacophore triplets and carrying out the conformational search. A conformational search was carried out for each molecule using CORINA-generated [51] structures as starting points (multiple ring conformers) and all possible triplets were tabulated. For each of 5,000 iterations, a random rotatable bond was altered by a random amount (0–360 degrees) and the new conformation unconditionally accepted provided there were no intramolecular van der Waals clashes. All bonds were freely rotatable with the exception of e.g. amide, ester and phenol bonds, which remained within 10 degrees of planarity. The number 5,000 was arrived at empirically by assessing, for a number of trial molecules, the relationship between the number of pharmacophore triplets generated and the number of conformations explored. It was found that the increase in number of triplets was negligible beyond 5,000 conformations, this number being sufficient to generate the same number of triplets as 100,000 conformations for this dataset.

There were six pharmacophore features considered: H-bond donor (D), H-bond acceptor (A), acid (C), base (B), aromatic (R) and hydrophobic (L). H-bonding groups, acid and base were identified by substructure searches. Essentially, all oxygen, nitrogen and sulfur atoms were defined as acceptor atoms, with the exception of trigonal planar nitrogen atoms, ester and ether oxygen atoms, and sulfur atoms bonded to oxygen atoms. All NH and OH moieties were defined as donor groups. Aliphatic amines, amidines, guanidines and 4-amino pyridines were considered to be protonated and were therefore treated as bases, and carboxylic acids, tetrazoles and acyl sulfonamides were deprotonated and treated as acids. For acceptors and acids, the atom position was used as the pharmacophoric point. For H-bond donors, the H atom is extended out to the average H-bond distance (3 Å from the heavy atom) to define the pharmacophoric point. This was where our method differed from the majority of existing methods (e.g. Chem-X [52]) and was built around the assumption that H-bond directionality is stronger about donors than about acceptors (supplementary material Fig. 1a), a conclusion that has been reached through many analyses of H-bond geometries [53]. An aromatic feature has been defined as any 5- or 6-membered planar ring in which the centroid defines the pharmacophoric point. Hydrophobic features were defined as either the centroid of an aliphatic ring or the centroid of any group of three consecutive non-polar atoms, where a polar atom was defined as any N/O/S atom or any atom bonded to an N/O/S atom.

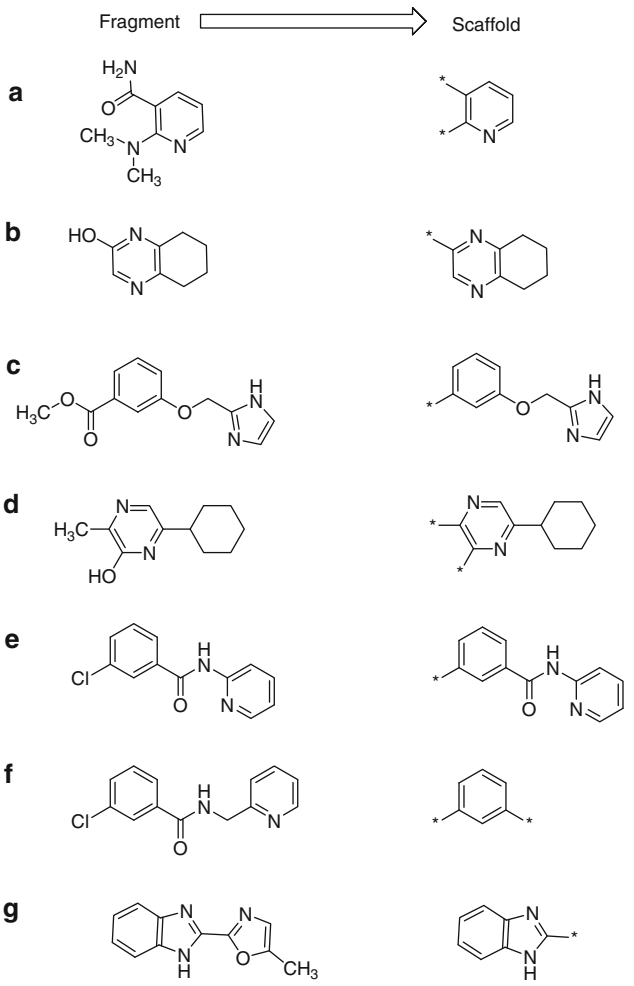

Fig. 5 In-house scaffold rules applied for fragments in Phase I. Attachment points for scaffolds are denoted with an '*'

A conformational search was carried out for each molecule and all possible triplets were tabulated. The distances between three features were binned to complete the definition of each triplet, e.g. CDR213 (supplementary material Fig. 1b). The distance bin boundaries occur at 2.5, 4.0, 6.0, 9.0, 13.0, 18.0 and 28.0 Å, with bin 1 covering the range 2.5–4.0 Å, through to bin 7, which covered distances greater than 28 Å. Distances below 2.5 Å were ignored to prevent substructural features from dominating the triplets. The triplets for each molecule were written out in an ASCII format that could be easily analyzed. Each molecule then was described by a fingerprint, in which each observed triplet sets a bit. Comparison between the pharmacophores set by two molecules can be rapidly calculated using Tanimoto or Tversky similarity coefficients of the binary fingerprints.

Molecular scaffold/frameworks

In GFI-I, a set of molecular framework rules for scaffolds that are summarized in Fig. 5 were used to identify small functionalized rings that are joined with linkers of limited flexibility, characteristics that are desirable in fragments. Heteroatoms and attachment points were defined in the ring systems. Single rings attached to other single rings by two or less atoms defined a scaffold, with the linking atoms retaining their atom type as in molecules C, D and E. When ring systems were connected by more than two atoms, the larger and/or more highly substituted ring system defined the scaffold as in F. All fused ring systems were always defined as the scaffold and other single ring systems that could be attached to the fused rings with less than three atoms were ignored as in molecule G. In-house tools [54] were utilized to construct a series of hierarchical SMARTS queries to identify the ring systems and linkages and to compute the scaffolds for the fragment sets according to these rules. In GFI-II, the Murcko Assemblies [50] as implemented in Pipeline Pilot were used to identify the scaffolds for the fragments.

Target diversity

In the second design phase GFI-II, an orthogonal diversity metric of target diversity was utilized to ensure that the library would be applicable to screening a broad range of pharmaceutically relevant targets. Recently, a comprehensive database of proprietary and published screening data was created from which a set of 515 Bayesian activity models has been derived using Pipeline Pilot, predicting the likelihood of activity of a given compound against human targets from 20 gene families [55]. These models were used to predict the activities of the final GFI-I selection, where a Bayesian score above 0 is interpreted as an "active" prediction. Targets that have fewer than 200 predicted Bayesian actives for the selected GFI-I fragments were considered to be under represented.

To increase target space coverage, the filtered set of 26,628 fragments from the Pfizer collection as well as the 25,235 BPs were then profiled in Pipeline Pilot with the same models. Fragments that were predicted to be active against the underrepresented targets and have novel Murcko Assemblies [50] (scaffold rules as implemented in Pipeline Pilot were utilized for pragmatic reasons), were taken forward for visual inspection.

Library construction

GFI-I (Phase 1)

The selection of fragments for GFI-I was initiated with compounds from the Pfizer file with availability of at least 200 mg followed by commercially available compounds. Since the Pfizer corporate file contains a large collection of diverse compounds, which were cost effective to utilize, we opted to use pharmacophore triplet diversity as the first orthogonal method to select the largest portion of the collection. This method has the advantage of being able to select both similar scaffolds with different pharmacophore substituents and different scaffolds which present similar pharmacophores in a different orientation. Within each charged set, a Tanimoto similarity matrix of the pharmacophore triplets for the fragments was used to select fragments that had a maximum similarity of 0.55 between any pair in the set. Another consideration in our design was the availability of close-in analogs either from the Pfizer file or from commercial sources to facilitate fragment hit follow-up and optimization. Each set of Pfizer fragments was grouped by hierarchical clustering to aid the selection process. In addition, a similarity expansion of each of the fragments was carried out relative to commercially available compounds to identify close in analogs that could be purchased for follow-up studies. For this comparison study, commercially available databases were filtered for unwanted functionalities, number of heavy atoms, $2 < N + O \leq 8$ and less stringent property filters, namely $MW \leq 300$, $ClogP \leq 2.5$, $TPSA \leq 110$ Å^2, However, identical solubility and SMCM filters were used. The selected Pfizer fragments were compared against this filtered commercial database and the number of commercial close-ins with Tanimoto similarity coefficient ≥ 0.9 was tabulated for each fragment. The Murcko framework was also calculated using the in-house program. The fragments were then sorted by their cluster membership, number of commercial close-ins and framework IDs to aid in the visualization and selection process. As the availability of close-ins for follow-up studies was desirable, the clusters that had two or more members were inspected first to select one representative from each cluster. However, more than one fragment was selected from large clusters if the design team considered them sufficiently different from a SAR perspective. Fragments that were singletons in the clustering were selected if there were commercial close-ins or other fragments with similar frameworks from visual inspection. The remaining singletons were selected if they were considered novel relative to the other fragments that had close-ins. The pharmacophore diversity filtering gave 5,562 compounds for chemistry triage of which there were 1,276 acids, 1,843 bases and 2,443 neutrals for visual inspection. Visual selection resulted in 684 acids, 770 bases and 961 neutrals for the library selection.

Acquisition of commercial fragments was then considered to supplement our in-house collection and in addition to 2D dissimilarity, framework diversity was utilized as the next orthogonal selection method to evaluate commercial

fragments. Since these fragments required additional funding to procure, we wanted them to provide novel frameworks that were not available in our file. To increase the sampling of similar pharmacophore space with different scaffolds, pharmacophore triplet diversity was not used for the selection criteria. The commercial fragment sets were compared to the corresponding set of Pfizer fragments selected in the first pharmacophore diversity triage to identify fragments with different frameworks. Fragments that had a Tanimoto similarity $\geq$0.9 to the selected Pfizer fragments were visually inspected and removed if they were similar. The novel framework selection resulted in 641 acids, 479 bases and 2,480 neutrals for visual triage which yielded an additional 351 acids, 160 bases and 465 neutrals.

We had up to this point focused our fragment selection on diversity for compounds with significant in-house availability ($\geq$200 mg) or which could be purchased from commercial vendors for immediate library production. However, to increase the overall diversity of the library, we elected to identify additional diverse templates with lower availability or which ultimately could be considered for external synthesis. In the next step, an additional 17,374 Pfizer fragments with low availability and pair-wise Tanimoto similarity of $\leq$0.9 with the fragments already selected were evaluated and any new fragments with pair-wise pharmacophore triplet similarity of greater than 0.3 or identical Murcko framework were removed. This diversity selection based on framework novelty, pharmacophore dissimilarity and 2D-dissimilarity identified 166 acids, 660 bases and 1,055 neutrals for visual triage from which 28 acids, 177 bases and 161 neutrals with availability greater than 95 mgs were selected. Additional candidates with insufficient availability, 99 acids, 359 bases and 27 neutrals, were flagged for potential outsourced synthesis.

X-ray screening subset

An X-ray screening subset of the GFI-I library was created to allow X-ray based screening for a limited set of diverse fragments. X-ray crystallographic screening libraries are typically put in cocktail mixtures of 2–8 [27]. We opted for mixtures of 4 as it was determined that beam time and data analysis requirements were feasible for 125 datasets, which would allow screening a total of 500 fragments, plated in mixtures of 4. Computational procedures have been utilized to minimize the chance of more than one compound in a mixture binding to the protein by maximizing 2D chemical dissimilarity and 2D shape dissimilarity [22] or 3D shape dissimilarity [12]. As the Pfizer fragments are chosen to be diverse in pharmacophores and/or frameworks, we only needed to utilize a combination of 3D shape dissimilarity and structural constraints for the mixture design. Fragments were selected after a pair-wise comparison of the shape (single conformation from CORINA [51, 56]) of each fragment against the others in the same charged class (e.g. acid vs. acid) using ROCS [57], and fragments that had the highest frequency of the ROCS shape Tanimoto less than 0.7 were selected from each set. The selected fragments were then partitioned into two sets, one set with 376 acids and neutrals, and a second set with 400 bases and neutrals. This was done in an effort to avoid strong ionic interactions between oppositely charged fragments which could compete with their interactions with target proteins. The compounds in each set were binned into sets of 4 with the constraints that the difference in the number of heavy atoms was $\leq$2 in each bin, and the pair-wise shape similarity was $\leq$0.8 for the four compounds in each mixture. The heavy atom constraint was used to include molecules of similar size within each mixture such that difference in size would not be a major factor in competing for binding to the target, while the shape similarity constraint was used so that the identity of the fragment would be unambiguous from its electron density in the co-crystal structure. The binning process resulted in a collection of 340 fragments which satisfied both constraints. Each mixture was visually inspected to eliminate symmetric compounds whose binding orientation would be difficult to distinguish from its electron density.

Library preparation

To ensure the highest quality physical library, all identified fragments were evaluated by 1D ^{1}H NMR methods at 500 MHz to assess chemical structure, purity, solubility and possible aggregation in aqueous neutral buffer at 1 mM concentrations. Additionally, the solutions were visually inspected as DMSO stocks (50–100 mM) and in neutral aqueous buffer (1 mM) to assess turbidity and help identify solubility issues. Analysis of this data resulted in the removal of $\sim$15% of the selected compounds with the percent of loss being similar for both in-house and commercial compounds. The final compound number for the GFI-I of this library was 2,592 fragments. Sample logistics and formatting for the entire collection was dependent upon the intended screening method. Samples were supplied as singles at 30 mM in deuterated-DMSO stocks in 384 well plates for bioassay, MS and SPR techniques. For NMR, mixtures of 10 fragments were designed to obtain maximal chemical diversity with resultant chemical shift diversity within each mixture of 10 to facilitate hit deconvolution. These mixtures were prepared at 10 mM in deuterated-DMSO stocks in single use 96 well plates.

GFI-II (Phase II)

One of the goals of the library design is that the collection would be applicable to screening a wide range of targets. Therefore pharmacological space coverage or target diversity was selected as a new orthogonal diversity criteria to be implemented in phase II of our screening library construction. The strategy was to evaluate the target space coverage of GFI-I to identify targets that were sparsely covered (few predicted actives from the Bayesian model) and to identify additional fragments in Phase-II that would be predicted active against these sparsely covered targets. Target coverage was evaluated by profiling the GFI-I collection against the 515 Bayesian activity models in our in-house activity database. This set was predicted to contain actives for all 20 gene families and at least one active for 418 of the 515 targets, with the overall coverage of pharmacological space being just over 80%. On average, fragments in GFI-I were predicted to be active against ~ 3 targets. The predictions of activity were not equally distributed over the gene families. The largest share of active model scores were for protein kinases, while the aspartyl proteases were the least represented. For the target diversity selection, a cutoff of at least 200 predicted actives in an activity model was used as the criteria for target coverage. Using these criteria, 143 targets had more than 200 predicted actives in GFI-I fragment library, with the remaining 367 targets having predicted actives of fewer than 200. In Phase-II, a set of 26,628 Pfizer fragments that passed our filters were profiled against the Bayesian models to identify the fragments that were predicted as actives against any of these 367 targets. These fragments were then compared against the GFI-I fragments to select fragments with novel Murcko frameworks [48]. This diversity filtering identified 186 acids, 499 bases and 600 neutrals for visual selection which provided 9 acids, 53 bases and 31 neutrals that had availability, and 39 acids and 50 bases that had no availability but were interesting fragments.

The same pharmacological space distribution analysis was applied to the Basis Products (BP) set in Phase-II to establish if they could complement the Phase-I set. An advantage of the BP derived fragments was that they would be derived by parallel chemistry; therefore hit follow-up by the same chemistry should be rapid. The BP set was predicted to contain actives against all gene families, and 412 of the 515 targets. In general the coverage trends were the same as the GFI-I set with some additional coverage of GPCRs. Using the same Phase-II diversity selection criteria, BPs that were predicted to be active against any of the above 367 targets were selected for triage. This process identified 7 acids, 39 bases and 65 neutrals from 64 acids, 221 bases and 622 neutrals for outsourced synthesis. Each of the Phase-II fragments and its five closest neighbors in the GFI-I library based on 2D similarity were visually inspected for novelty before requesting them for outsourced synthesis. The Phase-II design process resulted in the selection of 293 fragments of which many fragments required outsourced synthesis. The final distribution of GFI-I and GFI-II selections in target space (summed by gene family) is shown in Fig. 6.

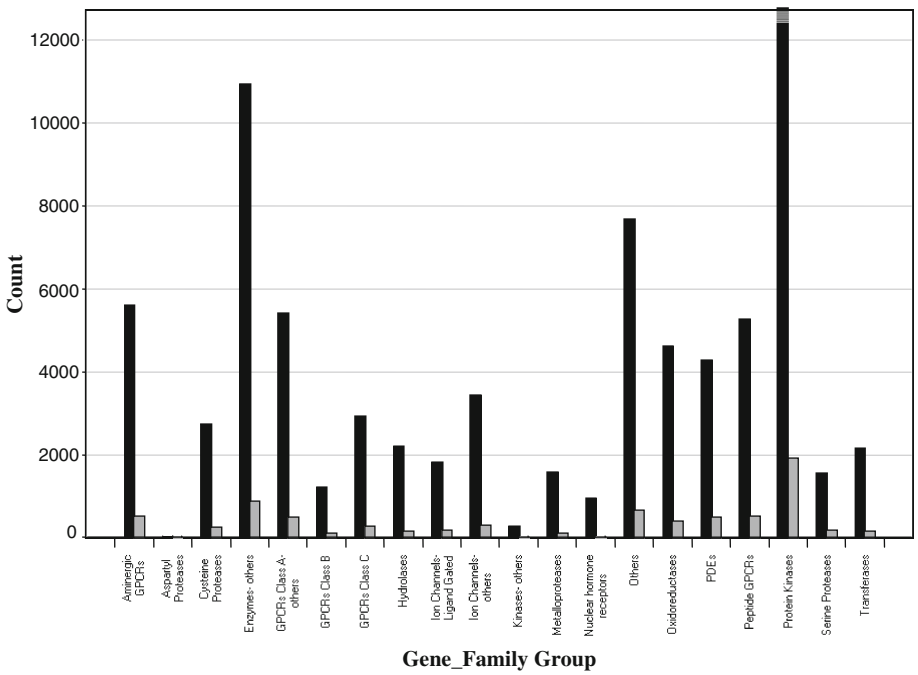

Fig. 6 Distribution of predicted actives from 515 Bayesian activity models for human targets from 20 gene families. Phase-I (*black*); Phase-II (*grey*)

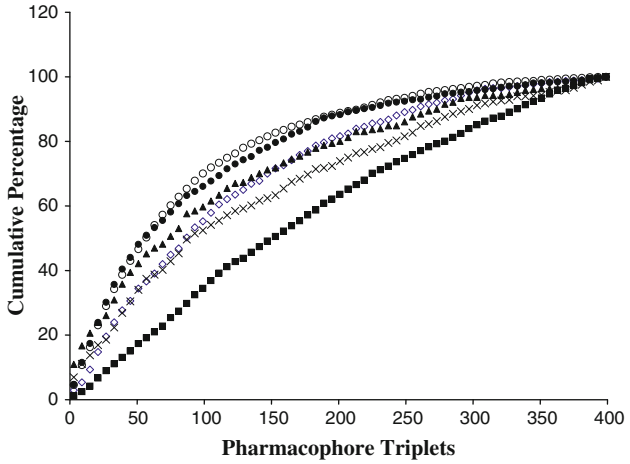

Fig. 7 Cumulative percentage of molecules as total pharmacophore triplets increases. *Empty circle* GFI Phase; *Filled circle* NMR hits legacy; *Filled triangle* Drugs _MW300; *Times* Drugs _MW500; *Diamond* Hits _MW300; *Filled squares* Hits _MW500;

Library characteristics

We had set out to design a new fragment library that has good molecular properties, molecular complexity, chemical diversity, synthetic tractability and predicted activity against diverse pharmaceutical targets. The distribution of physical properties including molecular weight, cLogP, hydrogen bond donors and acceptors, rotatable bonds and total polar surface area for GFI-1,II is shown in supplementary material, Fig. 2. Less than 4% of the combined GFI-I and II libraries have $100 \leq MW \leq 140$, and only 5% of the combined libraries have $MW \geq 250$. Only 22 fragments in the GFI-I collection had an SMCM index of less than 12. The majority of the fragments have a pharmacophore triplet count in the 10–150 range, and the library as a whole has a pharmacophore count distribution similar to the legacy NMR collection hits as shown in Fig. 7. The drugs and HTS actives with $MW \leq 300$ have comparable distributions though shifted further to the right in complexity. Similar to Schuffenhauer's observations, the NMR hits have fewer triplets than the drugs or HTS hits with $MW \leq 300$, with 80% of the NMR hits having <156 triplets compared to 80% of the drugs and HTS actives with <198 and 190 triplets, respectively. As the MW range increases, the average number of triplets increases.

Maximizing the coverage of chemical and target space throughout the design process was a key consideration in the development of this generic library and is a distinguishing factor, relative to earlier library design strategies. 2D-dissimilarity, pharmacophore triplet coverage, framework coverage and target pharmacological space coverage were used as diversity metrics in our design. Multiple metrics were used as no single metric can adequately measure different aspects of diversity. One of our goals was to cover at least 50% of pharmacophore space theoretically accessible to these low molecular weight molecules. The lower limit for the accessible pharmacophore space for the fragments was estimated to be 5,600 triplets as described below, and our target was at least 2,800 triplets. Willett has shown that atom pair distances above 15 Å are infrequent [58]. Given that distances between features of greater than 15 Å would be rare in these low molecular weight fragments, especially when long flexible chains have been filtered out, there would be only 5 distance bins to consider for each distance. With 56 different combinations of three out of six available pharmacophore features and five distance bins, this gave a lower limit to the accessible pharmacophore triplet space as $56 \times 5 \times 5 \times 5$. Davies and Briant have estimated that 20% of these geometries do not form triangles as the sum of the length of two sides is less than the length of the third side [59]. This gave approximately a lower limit of 5,600 triplets as being geometrically accessible. This estimate seems reasonable as a file random subset of 100 K compounds with MW of

Table 3 Comparison of GFI fragment libraries with oral drugs and HTS actives sets

Sets/libraries	No. of molecules	No. of unique pharmacophoric triplets	No. of Murcko assemblies[a]	No. of Murcko assemblies (+Alpha atoms)[a]
GFI-I	2,592	2,888	1,048	1,865
GFI-II Pfizer	182	2,068	133	171
GFI-II PGVL	111	1,621	70	100
GFI-I,II	2,885	2,919	1,251	2,136
Drugs-MW250	229	2,593	93	115
Drugs-MW300	380	3,711	183	232
HTS-MW250	352	2,633	240	304
HTS-MW300	928	3,652	599	778

[a] Murcko assemblies as implemented in Scitegic Pipeline Pilot

Table 4 Pharmacophore triplets in common between GFI libraries, oral drugs and HTS actives sets

	GFI-II Pfizer file	GFI-II PGVL	Drugs-MW250	Drugs-MW300	HTS-MW250	HTS-MW300
GFI-I	2,049	1,601	2,068	2,687	2,340	2,752

Table 5 Murcko assemblies in common between GFI libraries, oral drugs and HTS actives sets

	Drugs-MW250	Drugs-MW300	HTS-MW250	HTS-MW300
GFI-I	27 (21)	38 (24)	40 (33)	55 (43)
GFI-I,II	30 (22)	42 (25)	42 (34)	63 (44)

Murcko assemblies as implemented in Scitegic Pipeline Pilot, number without brackets are Murcko assemblies, number in brackets are Murcko assemblies + Alpha atoms

Table 6 Molecular features in GFI libraries, drugs-MW300 and HTS-MW300

Features	GFI-I	GFI-II Pfizer	GFI-II PGVL	Drugs-MW300	HTS-MW300
Aromatic atom	92.6	93.9	96.4	68.4	97.4
1 chiral center	19.6	30.2	24.3	29.5	12.7
2 chiral center	5.2	14.8	12.6	7.4	2.0
3 chiral center	0.7	5.5	–	4.2	1.3
4 chiral center	0.1	–	–	3.9	0.3
>4 chiral center	–	–	–	2.9	0.3
Br	2.8	1.1	–	0.3	2.2
F	5.7	4.9	2.7	5	5.3
1 ring	22.0	1.7	–	31.3	10.9
2 ring	67.5	57.7	45.9	30.5	39.3
3 ring	10.5	40.6	54.1	14.7	41.5

Count of occurrence of features, expressed as percentage of each set

250 or less has only 5,080 triplets, while 250 K compounds of MW 300 or less has 5,839 triplets (after removing triplets with distance bin 7). Note that these compounds have not gone through any of our unwanted functionalities and ring filters which would have removed many triplets that would not be desirable for the fragments.

A comparison of pharmacophore and framework coverage of the GFI library relative to actives from oral drugs and HTS as summarized in Tables 3, 4, and 5 suggests that this library has accessed similar pharmacophore space with more synthetically tractable compounds and often by more than one scaffold. The number of triplets in GFI-I,II is 2,919 triplets which is within the range covered by the drug and HTS actives for molecules between MW 250 and 300. The triplets that are unique to the drug and HTS active compounds appear to be contributed by compounds that fail the unwanted functionalities and ring filters. There are 1,048 Murcko assemblies in GFI-I and many of these are unique to this library (Tables 3, 5).

Although the library fragments are designed to be simple and synthetically tractable, they have some complexity in terms of chirality and number of rings (Table 6). At least 25% of the collection has one or more chiral centers. By design, the computational filters selected only compounds that had at least one ring and no more than three rings. The final GFI-I library has a composition of 31:28:41 for base:acid:neutral fragments, while GFI-II has a ratio of 47:19:34. The total GFI-I,II library would have a ratio of 33:27:40 assuming the selected compounds in GFI-II could all be acquired and passed QC. The GFI-II collection consisting of in-house and externally outsourced compounds is being assembled for library production.

A design team of medicinal, computational and NMR chemists assessed the fragments throughout the construction of the libraries to have a consistent view of what is detectable by various screening methods, synthetically tractable and drug-like. As suggested by Hubbard [23], perhaps the most important step is the manual inspection and assessment of the fragments by medicinal chemists. This process can be time consuming, and thus we had to be as efficient as possible in the filtering and diversity triage to minimize the number of fragments that had to be manually inspected and to maximize the probability that a fragment would be selected for the final library. We were able to manually select 3,757 fragments out of 11,043 fragments and 69% of the selected fragments made it to the final library, to give us an overall efficiency of 23.5% for GFI-I (see Fig. 1). We were unable to acquire about 15% of the selected fragments and the remaining 15% did not pass QC using NMR due to the presence of impurities (>10%), incorrect structures or poor solubility in 50–100 mM DMSO or aqueous solution at 1 mM.

Discussion and conclusion

In summary, we have fulfilled our design goals of low but sufficient molecular complexity for activity detection, structural and target diversity, chemical expandability and good molecular properties for the GFI fragment library. The fragments have molecular complexities similar to known actives as judged by their SMCM values and pharmacophore triplets composition. The libraries cover at least 50% of accessible pharmacophore triplet space and contain over 1,200 Murcko assemblies. While the majority of the pharmacophore triplets that are found in the drug and HTS sets used in this study are also found in the fragment libraries, the libraries do possess many more unique

scaffolds than the drug and HTS sets. The novelty of these additional scaffolds and more importantly, their ability to productively bind to diverse pharmacological targets will be assessed as this library is screened against an increased number of diverse targets. We estimate the overall coverage of pharmacological space for the libraries at over 80%, based on profiling the libraries against Bayesian activity models for 515 targets from 20 gene families.

To date, the GFI-I library has been screened across many target types including kinases, proteases, protein–protein interactions, polymerases, chaperones, acetyltransferases, acetylases, and reductases within Pfizer. Depending upon the resources and the target requirements and restraints, the best primary screening and orthogonal characterization methods are chosen. Our philosophy is to identify and confirm the binding and functional activity of fragments by several orthogonal methods prior to chemistry follow-up. The majority of GFI-I fragment screening thus far, has been carried out using NMR STD [19] binding methods and traditional biochemical assays. Additionally, we routinely develop functional NMR screening assays for all appropriate enzyme systems to generate a functional readout, using an orthogonal detection method [60]. These functional methods can often detect activity in the nM–mM range by direct observation of reactants and products at atomic resolution in NMR and have been shown in-house to have fewer screening artifacts at high ligand concentration than traditional biochemical detection methods. More recently, SPR and X-ray screening methods have been employed in-house and appear very promising as alternative primary screening techniques. For some targets, we had the luxury of carrying out the primary screen by multiple methods which included a biophysical and biochemical component. As we have observed in the past, coupled bioassays or those which rely on fluorescence, UV and colorimetric readouts can produce false negative and positive results at high ligand concentrations. On the other hand, biophysical binding studies can also result in false positives due to non-specific binding interactions at high ligand concentrations. For all of these situations, appropriate control experiments can be run to eliminate these effects. However, it is clear that fragment hits must be validated by several orthogonal methods to correlate binding and functional activity, and to enable follow-up to achieve any improvement in efficiency in chemical optimization.

For the majority of targets screened thus far, fragment hits have been identified in the uM–mM range resulting in L.E. $\geq$0.3. Fragments receiving the most chemical attention within teams are those ligand efficient hits which (1) have productive vectors for optimization based upon structural information, (2) are most easily modified synthetically for optimization, (3) have available analogs in

Table 7 Primary screening results of the GFI-I library against 13 targets using the NMR STD protein–ligand binding method

Target	Protein class	Therapeutic area	# Binders	Hit rate (%)
1	Protein–protein	Oncology	217	8.37
2	Plastidic enzyme	Metabolic Dx	130	5.02
3	Protease	CNS	232	8.95
4	Chaperone	Oncology	153	5.9
5	Peroxisomal enzyme	CNS	73	2.82
6	Reductase	Anti-bacterial	128	4.94
7	Glycosyltransferase	Anti-bacterial	204	7.87
8	Kinase	Oncology	218	8.41
9	Kinase	Inflammation	169	6.52
10	Kinase	Oncology	176	6.79
11	Kinase	Anti-bacterial	142	5.48
12	Polymerase	Antiviral	165	6.4
13	Deacetylase	Metabolic Dx	326	12.58

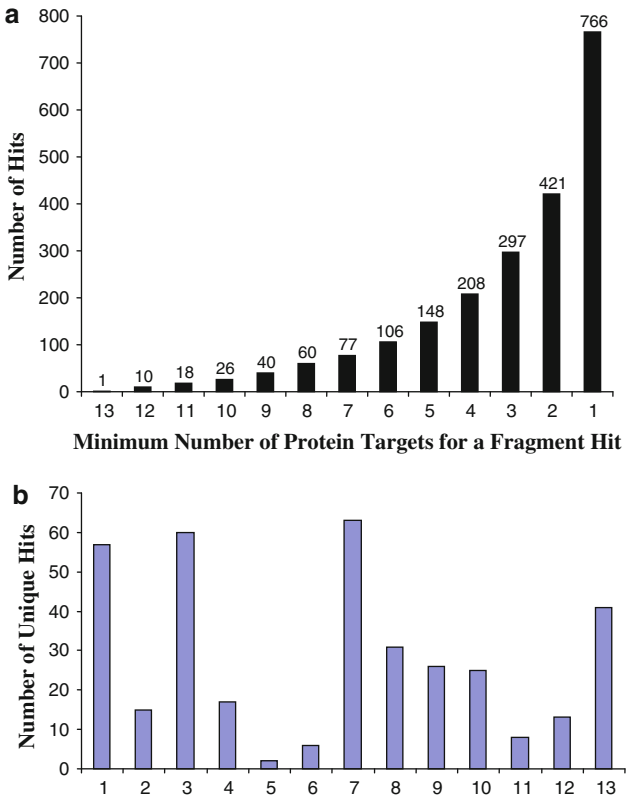

Fig. 8 **a** GFI Phase-I primary hit distribution against 13 diverse targets using NMR based STD screening method. Only 1 fragment hits all 13 targets while 766 fragments hit at least 1 target in this study. **b** Number of primary screening hits that are unique to each of the targets in Table 7. Majority of targets belong to diverse protein families with the exception of targets 8–11, which are kinases

the corporate file or commercially for rapid assessment and (4) have the possibility of a high resolution structure. In addition, creating the fragment library with compounds that have sufficient liquid and solid in-house supplies for rapid confirmation has been critical to the success in using this library. An early comparison of primary GFI-I hits from NMR STD screening was carried out for 13 diverse targets in 8 different protein families to assess the potential specificity of this library. In this work, protein and ligand concentrations for the NMR STD assay were optimized for each target protein but typically contain 1–5 uM protein and 240 uM fragments in mixtures of 10. The primary screening hit rates using the NMR STD experiments ranged between 2.8 and 13% for this set of targets (Table 7). Interestingly, only 1 out of 2,592 fragments hit all 13 targets and less than 1% of the total collection hit 11 or more of all targets evaluated in this study (Fig. 8a) indicating a very limited number of promiscuous binding fragments in the collection. Additionally, a number of unique hits were obtained for each of the 13 targets in this analysis (Fig. 8b). These results provided confidence that there is not significant overlap of chemical hit space across diverse targets despite fragment binding affinities in the uM–mM range. Also encouraging was the fact that unique fragment hits were obtained for 4 kinases, targets 8 though 11, where specificity could be difficult to achieve for weak binding at the fragment level [61].

The availability of related analogs within the Pfizer corporate file or commercial sources was one of the library design criteria to enable rapid follow-up and optimization. Indeed the ability to rapidly access and screen a number of near neighbors has been invaluable on a number of projects. In these cases, an initial round of nearest neighbor file mining around multiple confirmed fragment hits resulted in improved potencies, efficiencies or properties in many instances. The resulting SAR from this and further rounds of file mining has allowed medicinal chemists to assess and prioritize fragment series to take forward, as well as identifying novel proprietary fragment hits with different vectors for elaboration. Fragment optimization by mining the corporate file prior to synthesis is both rapid and cost-effective. Fragments from commercial sources are less often used in the file mining process, due to the extra time and cost associated with obtaining and screening these; however, mining and screening commercially available fragments still represents a more time- and cost-efficient solution than custom synthesis.

Despite the identification of ligand efficient hits for most targets, several observations have been made for the GFI-I library, which may require adjustments with future library additions. In consideration of results for the 13 NMR-STD screens, only $\sim$33% of the total GFI-I collection hit at least one target, consistent with other studies [62]. Preliminary analysis of this limited dataset indicate that hitters have more aromatic character as shown in an increase in the median number of aromatic rings and ring assemblies relative to non-hitters. Hitters are more hydrophobic and possess slightly higher cLogP and a decrease in the number of H-bond donors. Additionally, the rotatable bond distribution is higher for non-hitters, which indicates an entropic penalty may be required for these fragments to bind. For this set of targets, the charged ratio of the hits was 13:20:67 for base:acid:neutral fragments while that of the GFI-I library was 31:28:41. The differences observed, particularly for basic fragments may be a reflection of the targets screened and these ratios may change as targets such GPCRs are considered for fragment screening.

As additional screens are run, evaluation of the complexity and 3D pharmacophore triplet diversity for hitters versus non-hitters will be determined to guide future additions to the library. Another key issue is aqueous solubility for X-ray crystallographic follow-up. Solubility for all fragments was experimentally assessed by NMR at 1 mM in PBS, pH 7.4 and any fragment that appeared insoluble at this concentration was removed from the final collection. In NMR, STD based screening is generally carried out at micromolar concentrations at 200–500 uM and fragment hits using this method can be identified well below the binding K_D. However, we have recently found that ligand concentrations in the range of 5–20 mM are often required for observation of fragment density in crystallization soaking experiments. This large discrepancy in solubility has created a lack of success in obtaining the crystal structure for some NMR based fragment hits and will also be a limitation on X-ray screening. For future fragment additions, the computational filter for aqueous solubility will be increased, which may result in a larger number of lower MW and/or more hydrophilic fragments. This change does require careful attention to ensure that lower MW fragments can still be detected in less sensitive assays and that fragment hydrophobicity is sufficient to facilitate molecular interactions [62].

Our global library was designed to be a living library which would ideally evolve with additional chemical matter and screening experience for many diverse target families. As we screen more and more targets, chemical gaps in the library are being observed and efforts to fill them are planned for the future. As we consider the use of fragment screening in areas such as understanding new targets and pathways and assessing druggability, we may need to reconsider the incorporation of some fragments with less drug like functionalities. Although these may not be as chemically attractive or as synthetically expandable as medicinal chemists would like, they may serve as good chemical tools to investigate less established targets.

Acknowledgments Tudor Oprea for the SMCM program; Marty Marx, Kim Daoust, James Forman and Bob Mecca for RI support; Parag Sahasrabudhe, Hong Wang, Diana Omechinsky, Kris Borzilleri, Cathy Moore and Jiangli Yan for NMR support; Gaia Paolini and Zhengwei Peng for computational input; Bob Chambers, Kim Matus, Kyle Blair, Lisa Thomasco, Jan Snape, Shirell Gray, Jola Nowakowski, Maria Anhalt, Steve Curioso, Craig Hines, Diane Johnson, Bernadette Udasco, Jason Harraden, Erin Cyr, Betsy Poe, Monica Gorny, Elizabeth Mostowy, Holly McKeith, Jed Morris, Tim Britt, Frank Girardi, C.K. Chan for library preparation and sample logistics support; G. Tim Benson, Mike Clark, Tony Wood, Ron Wester and Suvit Thaisrivongs and Alan Mathiowetz for initiative support.

References

1. Congreve M, Chessari G, Tisi D, Woodhead A (2008) J Med Chem 51:3661
2. Albert J, Blomberg N, Breeze A, Brown A, Burrows J, Edwards P, Folmer R, Geschwindner S, Griffen E, Kenny P, Nowak T, Olsson L, Sanganee H, Shapiro A (2007) Curr Topics Med Chem 7:1600
3. Shuker S, Hajduk P, Meadows R, Fesik S (1996) Science 274:1531
4. Lepre C, Peng J, Fejzo J, Abdul-Manan N, Pocas J, Jacobs M, Xie X, Moore J (2002) Comb Chem High Throughput Screen 5:583
5. Lipinski C, Lombardo F, Dominy B, Feeney P (2001) Adv Drug Deliv Rev 46:3
6. Carr R, Congreve M, Murray C, Rees D (2005) Drug Discov Today 10:987
7. Hopkins A, Groom C, Alex A (2004) Drug Discov Today 9:430
8. Howard N, Abell C, Blakemore W, Chessari G, Congreve M, Howard S, Jhoti H, Murray C, Searvers L, van Montfort R (2006) J Med Chem 49:1346
9. Saxty G, Woodhead S, Berdini V, Davies T, Verdonk M, Wyatt P, Boyle R, Barford D, Downham R, Garrett M, Carr R (2007) J Med Chem 50:2293
10. Hajduk P, Huth J, Tse C (2005) Drug Discov Today 10:1675
11. Baurin N, Aboul-Ela F, Barril X, Davis B, Drysdale M, Dymock B, Finch H, Fromont C, Richardson C, Simmonite H, Hubbard RE (2004) J Chem Inf Comp Sci 44:2157
12. Blomberg N, Cosgrove DA, Kenny PW, Kolmodin K (2009) J Comput Aided Mol Des 23:513
13. Schuffenhauer A, Ruedisser S, Marzinzik AL, Jahnke W, Blommers M, Selzer P, Jacoby E (2005) Curr Topics Med Chem 5:751
14. Leach A, Hann M, Burrows J, Griffen E (2006) Structure-based drug discovery. Royal Society of Chemistry, Cambridge
15. Lepre C (2001) DDT 6:133
16. Erlanson D, McDowell R, O'Brien T (2004) J Med Chem 47:3463
17. Fattori D, Squarcia A, Bartoli S (2008) Drugs R D 9:217
18. Barker J, Courtney S, Hesterkamp T, Ullmann D, Whittaker M (2006) Expert Opin Drug Discov 1:225
19. Mayer M, Meyer B (1999) Angew Chem Int Ed 38:1784
20. Dalvit C, Fogliatto G, Stewart A, Veronest M, Stockman B (2001) J Biomol NMR 21:349
21. Wang Y, Liu D, Wyss D (2004) Magn Reson Chem 42:485
22. Hartshorn M, Murray C, Cleasby A, Frederickson M, Tickle I, Jhoti H (2005) J Med Chem 48:403
23. Hubbard R, Davies B, Chen I, Drysdale M (2007) Curr Topics Med Chem 7:1568
24. Neumann T, Junker H, Schmidt K, Sekul R (2007) Curr Topics Med Chem (Sharjah, United Arab Emirates) 7:1630
25. Hubbard R, Chen I, Davies B (2007) Curr Opin Drug Discov Dev 10:289
26. Card G, Blasdel L, England B, Zhang C, Suzuki Y, Gillette S, Fong D, Ibrahim P, Artis D, Bollag G, Milburn M, Kim S, Schlessinger J, Zhang K (2005) Nat Biotechnol 23:201
27. Jhoti H, Cleasby A, Verdonk M, Williams G (2007) Curr Opin Chem Biol 11:485
28. Hann M, Leach A, Harper G (2001) J Chem Inf Comp Sci 41:856
29. Schuffenhauer A, Floersheim P, Acklin P, Jacoby E (2003) J Chem Inf Comp Sci 43:391
30. Hu Q, Peng Z, Kostrowicki J, Kuki A (2010) In: Zhou Z, Walter J (eds) Chemical library design in methods in molecular biology (MiMB) series. Humana Press, New York, pp 253–276
31. Congreve M, Carr R, Murray C, Jhoti H (2003) Drug Discov Today 8:876
32. Ladbury J, Klebe G, Freire E (2010) Nat Rev Drug Discov 9:23
33. Burrows J (2004) Soc Med Res Trends Drug Discov
34. WOMBAT (2005) Santa Fe available at: http://www.sunsetmolecular.com/
35. Oprea T, Blaney J (2006) In: Jahnke W, Erlanson DA (eds) Fragment-based approaches in drug discovery, pp 91–111
36. Njardarson Group, Kwon L, Rogers E, McGrath N, Brichacek M, Njarðarson J (2006) Available at: http://cbc.arizona.edu/njardarson/group/homepage
37. Teague SJ, Davis AM, Leeson PD, Oprea TI (1999) Angew Chem Int Ed 38:3743
38. Allu T, Oprea T (2005) J Chem Inf Model 45:1237
39. Reynolds C, Tounge B, Bembenek S (2008) J Med Chem 51:2432
40. Irwin J, Shoichet B (2005) J Chem Inf Model 45:177
41. Zhou J, Shi S, Na J, Peng Z, Thacher T (2009) J Comput Aided Mol Des 23:725
42. Ertl P, Rohde B, Selzer P (2000) J Med Chem 43:3714
43. Gao SVH, Lee P (2002) Pharm Res 19:497
44. Leadscope, Inc., 1393 Dublin Road, Columbus, OH 43215
45. Roberts G, Myatt G, Johnson W, Cross K, Blower P (2000) J Chem Inf Comput Sci 40:1302
46. Daylight Chemical Information Systems Inc In: Aliso Viejo, CA 92656, USA
47. Ward J (1963) J Am Stat Assoc 58:236
48. PipelinePilot In: 10188 Telesis Court, Suite 100, San Diego, CA 92121-4779
49. Durant J, Leland B, Henry D, Nourse J (2002) J Chem Inf Comput Sci 42:1273
50. Bemis G, Murcko M (1996) J Med Chem 39:2887
51. Corina Inc In: molecular-networks GmbH
52. Murral N, Davies E (1990) J Chem Inf Comp Sci 30:312
53. Mills J, Dean P (1996) J Comput Aided Mol Des 10:607
54. Bakken G, Du J, Li D, Lu J, Schulte G, Sridaharan S, Tinniswood A, Miller M (2006)
55. Paolini G, Shapland R, van Hoorn W, Mason J, Hopkins A (2006) Nat Biotechnol 24:805
56. Gasteiger J, Rudolph C, Sadowski J (2004) Tetrahedron Comput Methodol 3:537
57. ROCS OpenEye Scientific Software In: Santa Fe, New Mexico, USA
58. Jakes S, Willet P (1986) J. Mol. Graphics 4:12
59. Davies K, Briant C (1995) In: MGMS meeting, Leeds
60. Stockman B, Lodovice I, Fisher D, McColl A, Xie Z (2007) J Biomol Screen 12:457
61. Hu Q, Yan J, Withka J, Sahasrabudhe P, Moore C, Na J, Narasimhan L (2009) Abstracts of papers, 238th ACS National Meeting, Washington, DC, USA
62. Chen I, Hubbard R (2009) J Comput Aided Mol Des 23:603