J Comput Aided Mol Des (2009) 23:603–620
DOI 10.1007/s10822-009-9280-5

Lessons for fragment library design: analysis of output from multiple screening campaigns

I-Jen Chen · Roderick E. Hubbard

Received: 8 March 2009 / Accepted: 7 May 2009 / Published online: 3 June 2009

Abstract Over the past 8 years, we have developed, refined and applied a fragment based discovery approach to a range of protein targets. Here we report computational analyses of various aspects of our fragment library and the results obtained for fragment screening. We reinforce the finding of others that the experimentally observed hit rate for screening fragments can be related to a computationally defined druggability index for the target. In general, the physicochemical properties of the fragment hits display the same profile as the library, as is expected for a truly diverse library which probes the relevant chemical space. An analysis of the fragment hits against various protein classes has shown that the physicochemical properties of the fragments are complementary to the properties of the target binding site. The effectiveness of some fragments appears to be achieved by an appropriate mix of pharmacophore features and enhanced aromaticity, with hydrophobic interactions playing an important role. The analysis emphasizes that it is possible to identify small fragments that are specific for different binding sites. To conclude, we discuss how the results could inform further development and improvement of our fragment library.

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

Electronic supplementary material The online version of this article (doi:10.1007/s10822-009-9280-5) contains supplementary material, which is available to authorized users.

I-J. Chen · R. E. Hubbard (✉)
Vernalis (R&D) Ltd, Granta Park, Cambridge CB21 6GB, UK
e-mail: r.hubbard@vernalis.com; rod@ysbl.york.ac.uk

R. E. Hubbard
YSBL and HYMS, University of York, York YO10 5YW, UK

Abbreviations

AK	Adenosine kinase
CDK2	Cyclin-dependent kinase 2
DNAG	DNA gyrase
FAAH	Fatty acid amide hydrolase
HSP70	Human heat shock protein 70
HSP90	Human heat shock protein 90
JNK3	c-Jun N-terminal kinase 3
PDPK1	3-Phosphoinositide-dependent protein kinase 1
PIN-1	Peptidyl-prolyl cis/trans isomerase
PPI	Protein–protein interaction
SeeDs	Structural exploitation of experimental drug startpoints

Introduction

Over the past 10 years, there has been increasing interest in fragment based methods for drug discovery [1, 2]. The excitement (and investment) in the methods has grown over the past few years as a number of compounds have entered clinical trials [1, 2] where fragment derived information has made important contributions. The developments in the area have recently been comprehensively reviewed by Congreve et al. [3]. The central premise is that a small library of compounds can sample a potentially huge chemical diversity. The structures of fragment hits binding to an active site can then guide medicinal chemists to rapidly expand and optimise these hits into leads and then onto clinical candidates. This provides an attractive alternative to High Throughput Screening for identifying tractable starting points for generation of novel lead compounds, particularly for new classes of target, provided sufficient structural data can be generated on binding modes to guide compound optimisation. In

addition, the entry cost to fragment based discovery is relatively low as a small library of fragments (a 1,000 or so compounds) is sufficient to give hits against most targets. For this reason, many of the recent developments have been driven by small, technology focussed companies.

The primary characteristic of fragments is that they are small, weak hits. This has required developments in three main areas—detecting fragments that bind, evolving fragments into hits and the design of fragment libraries.

There has been continuing improvements in the various biophysical methods which can detect fragment binding, which is typically with a K_D between 100 μM and 10 mM. Initially, techniques such as NMR monitoring protein signals (HSQC) [4] and high throughput crystallography [5] were used. These are now augmented with techniques such as NMR that monitors ligand signals (STD, LOGSY, etc. [6]) and Surface Plasmon Resonance [7]. Most practitioners are now converging on a similar approach, where a relatively rapid biophysical method is used to identify which fragments are binding competitively to a target site, with confirmed hits taken into crystallisation trials to confirm binding and characterise binding modes. At Vernalis, our fragment discovery platform is known as SeeDs [8], where fragments are screened in pools of 8–12 compounds for binding to a target through a competitive NMR experiment. A combination of ligand monitoring NMR experiments (STD [9], CPMG [10], Water LOGSY [11]) are measured for the pool of fragments in the presence and absence of a competitor ligand. The main limitation this places on the design of the fragment library is the need for sufficient solubility (typically 0.5 mM) for the NMR experiment to detect binding.

There are three main ways in which fragments have been incorporated into larger hits and optimised into lead compounds. One approach is to link fragments together, as in the pioneering SAR by NMR approach of Abbott [4]. Although there have been some notable successes [12, 13], it is often impossible to develop appropriate chemistry to link fragments while maintaining the orientation and position needed for key interactions. More successful has been the idea of growing fragments, either by structure-guided medicinal chemistry or by using the fragment binding motif to search for similar compounds that can be purchased (so-called SAR by Catalogue). A particularly attractive approach is to use the structure of fragments (and other hit compounds) binding to the target site to design new compounds that combine key interaction features [14]. The ideas about how to use fragments in medicinal chemistry optimisation campaigns continue to evolve, and with experience is influencing the design of fragment libraries.

As with any form of screening, the quality of hits critically depends on the library screened. If the library does not contain appropriate compounds, it may result in no hits in the screening, or finding hits which are inappropriate for chemical optimization. Around the time the first generation of the Vernalis fragment library was published [15], there were few reports describing the design and characterization of fragment libraries [15–17]. Since then, a few more reports have emerged, but with little detailed analysis provided [18–20], perhaps because this information is often regarded as proprietary.

We have now conducted fragment screening campaigns against many different types of targets. The fragment library has been evolving during this time, but it is possible to identify particular trends and characteristics of the hits that provide useful insights into the nature of fragment binding and guidance for further optimisation of the library. In this paper we present a number of computational analyses of the characteristics of the fragment library and the hits obtained against a selected number of protein targets. Only a few reports have appeared previously of such a retrospective analysis. The Abbott group identified certain privileged scaffolds from the first set of screening campaigns [21]. Although AstraZeneca (AZ) summarised the characteristics of their fragment library and hit rates against a collection of targets [19], there was not a full analysis of the characteristics of hits obtained.

This has prompted us to perform a detailed analysis with fragment hits obtained over the past few years. We begin with a description of the design and implementation of our fragment library. The physicochemical properties are then compared across various fragment sets to examine differences or similarities. We conclude with a discussion of how the trends observed from the analysis could be used for future improvements in the library.

Methods

Design of the VER2004 fragment library

The first generation of the fragment library (VER2004) is described in detail by Baurin et al. [15]. The overarching selection criteria were that fragments had to be diverse with suitable physicochemical properties for our screening methodology (ligand-monitoring NMR spectroscopy) and that the compounds were suitable to be taken forward in a medicinal chemistry programme. The main features of the selection process are summarised in Fig. 1.

The primary source of compounds for library enumeration came from the available vendor catalogues. The first

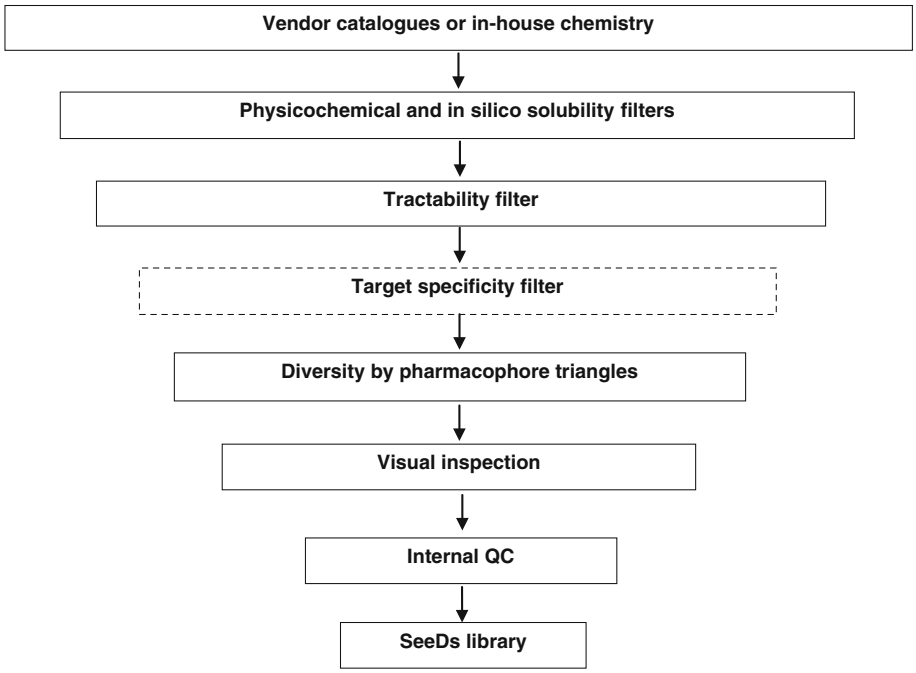


Fig. 1 Vernalis fragment library enumeration process. The *box* with a *broken line* is for an optional procedure

filter includes 2D physicochemical descriptors such as molecular weight ($100 \leq MW \leq 250$, with an upper limit of 350 for compounds containing sulfonamides), number of rotatable bonds ($NRot < 6$) and calculated water solubility ($\leq$2 mM) [15]. The tractability filter contains the lists of preferred and disfavoured functional groups and elements. The list of desirable chemical moieties contains fourteen chemical handles [15] to allow rapid chemistry to be applied to the fragments for evolution of the compounds and the permitted elements are H, C, N, O, F, Cl and S. The list of disfavoured groups was compiled from a number of studies [22–24] and extended by in-house medicinal chemists to ensure the stability and chemical tractability of the fragments for further work. In addition, excluding compounds with such properties should reduce the number of false positives due to aggregation or reaction when screening the fragments as mixtures. All these filters are in the SMARTS [25] format and can be applied in silico to process large libraries of compounds efficiently. Before diversity clustering, there is an optional target pharmacophore matching stage. This stage was introduced to address the needs of internal drug discovery projects, should they request. For example, to enrich library with fragments which satisfy the unique kinase hinge binding motif [26], a subset of fragments were subjected to a kinase pharmacophore match after filtering of functional groups.

The library was designed to be used for a wide range of targets. The goal is to identify fragments that make specific interactions with a given target so it is important that the library contains as much chemical diversity as possible. A measure of diversity was determined by molecular fingerprints based on 2D 3-point pharmacophore triangles (vide infra). The presence or the absence of the encoding fingerprint components (pharmacophore triangles) can be used to assess the uniqueness of a fragment for its putative recognition pattern. The final stage in identifying fragments is inspection by medicinal chemists before purchasing. Although this is a rather subjective selection process, the principal consideration was to incorporate an overall 'expert' view of chemical tractability and suitability of the fragments for their potential to evolve into a lead or drug-like molecule. The fragments which passed the visual inspection were then purchased and subjected to quality control measurements. The internal procedure includes checking compound identity by NMR spectrum and if needed by mass spectrometry, purity, stability, solubility and self-aggregation [15].

Evolution of the Vernalis fragment libraries to give VER2008

The library has evolved continuously over the past 4 years (VER2008) through three main routes (described in more detail in Hubbard et al. [8]). All compounds below MW 250 synthesised in the company are considered for inclusion and new fragments are either identified or designed from templates seen to bind to different targets during a discovery project. Secondly, repeats of the analysis of the changing commercially available compounds have identified new fragments. Thirdly, compounds are removed during regular quality control of the library (visual inspection, consultation with chemists and mass spectrometry) with the general objective of keeping the fragment library about the same

size, for practical reasons. About 555 of the 1,321 fragments identified in VER2004 have been removed during the process of continuous assessment and improvement of the library. Of these 555 fragments, about 40% were removed because the compounds had decomposed in the master plates and about 40% were removed based on the experience of the medicinal chemists in using the fragments in discovery projects. Less than 5% were removed for long term solubility problems and 3% because compounds were no longer available. The remainder were removed for miscellaneous reasons.

Fragment screening campaigns

Table S1 (supplementary material) summarises the conditions used in the screening campaigns against 12 different protein targets. As mentioned above, fragment binding is detected using NMR spectroscopy to measure changes in ligand signal as a fragment binds (for more detail, see description in Hubbard et al. [8]). For a typical screening campaign, protein samples at 10 μM are placed in NMR tubes in an autochanger and mixtures of fragments (up to 12 at 500 μM each fragment) added to each tube. Three spectra are then recorded (STD, Water-LOGSY and CPMG). A competitor ligand is then added and the three spectra recorded again. The NMR experiments are designed to detect signals which change when the ligand binds to the protein. The competitor ligand spectra are recorded to identify non-specific binding. This approach identifies fragments whose binding is disrupted by the competitor ligand, which in all cases reported here is shown to be because the fragment binds to the same binding site (although in principle this could detect binding at an allosteric site).

The resulting spectra are then inspected and a hit is defined as a fragment which binds to and can be displaced by the competitor ligand from the protein. The NMR experiments are dependent on a complex set of exchange phenomena and so it is often seen that a fragment is not a hit in all three experiments. A Class 1 hit is defined as a fragment which shows evidence for binding in all three NMR experiments (STD, Water-LOGSY and CPMG), a Class 2 hit shows changes in two experiments and a Class 3 hit in only one experiment.

Analysis of SeeDs library and fragment hits

The fragment library used in the screening campaigns analysed here has gradually changed with time and thus different targets were screened with a slightly different library. The changing composition of the fragment library could have made it difficult to compare hit rates and features of the hits across targets. However, the overall features of the library have remained fairly constant (see "Results and discussion") so it is possible to make comparisons. Where appropriate, the analyses presented in this paper have taken the changing nature of the library into account.

As the content of the screening library evolved with time, three versions of the Vernalis SeeDs libraries were prepared for this analysis. The first two versions are VER2004 and VER2008 as described earlier. VER2004 includes the compounds as published in Baurin et al. [15] paper and contains 1,275 unique compounds after removing duplicates. VER2008 is a more current version of the library and contains 1,063 unique compounds. An amalgamated version (VER_ref) was also created by combining fragments from VER2004 and VER2008. This includes 1,605 unique fragments. For analysis of the SeeDs library, a drug reference set derived from The World Drug Index (WDI) [27] was prepared as previously described [15]. The filtered WDI set contains 1,141 drug compounds of molecular weight between 250 and 550 and of O, N, C, H, Br, I, Cl, F, S, or P only.

All fragments from the historical collection of the SeeDs library were retrieved from the internal database as 2D SD format and processed with the Molecular Operating Environment (MOE) software [28]. Following the assignment of bond orders and standard protonation states at pH = 7, 184 2D QuaSAR descriptors [28] were calculated. To assess molecular complexity and diversity, pharmacophore graph triangle fingerprints (GpiDAPH3) [28] were calculated. These are a collection of three-point pharmacophores calculated from the 2D molecular graph. In this pharmacophore fingerprint scheme, each molecule is represented as a set of integers (not as a binary bit-string) where each integer encodes a distinct and unique pharmacophore triangle, and corresponds to a fingerprint component. There are three main stages in the GpiDAPH3 scheme [29]. In the first stage (Fig. 2), pharmacophore features are assigned to the heavy atoms of a molecule. There are 8 pharmacophore features, computed from three atom properties (hydrogen-bond donor, hydrogen-bond acceptor and pi system). The eight features are denoted as (D, A, P, H) or (D=, A=, P=, H=) where D is a donor, A is an acceptor, P is a neutral polar feature which can act as a donor or an acceptor (e.g., a hydroxyl) and H is a hydrophobe. The (D, A, P, H) notation refers to these features when not part of a pi system, while (D=, A=, P=, H=) denotes these features when conjugated to a pi system. For example an sp2 aniline amino group corresponds to D=. An (hetero) aromatic ring is represented as a hydrophobe (H=). No additional feature is defined for anions or cations. In the second stage all possible pharmacophore triangles are formed for the detected features and the graph distances between the features for all triangles are calculated from the shortest path of covalent

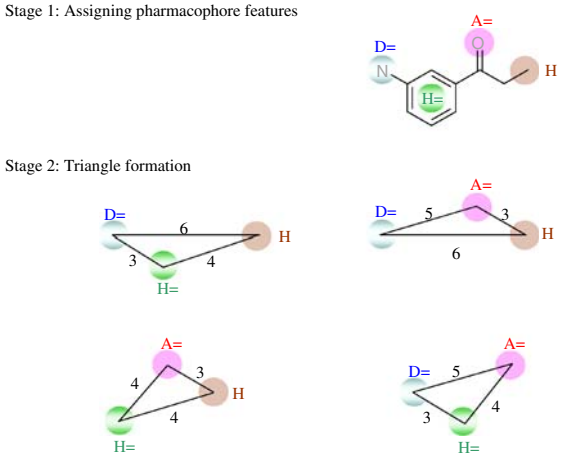
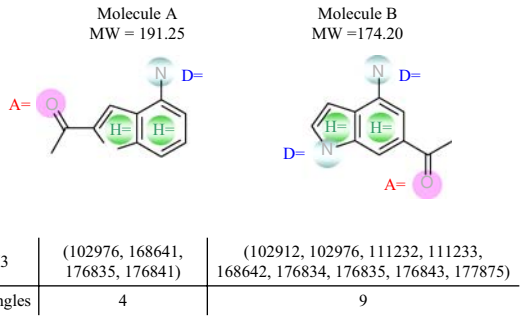


Fig. 2 GpiDAPH3 2D 3-point pharmacophore fingerprint scheme [28]. The fingerprint is generated in three stages [29]. Stage 1 is to assign pharmacophore features to a molecule. Four out of eight pharmacophore features are available for the depicted molecule and they are hydrogen bond donor in the pi system (*D*=), ring (*H*=), hydrogen bond acceptor in the pi system (*A*=) and a hydrophobe (*H*). These four features constitute four unique pharmacophore triangles in Stage 2. The uniqueness of a triangle is assessed by the feature types and the graph distances between them. The numbers in Stage 2 are distances in terms of bonds between features. Stage 3 is to transform each triangle into an integer and save as a component in the fingerprint

bonds between them. The distances are binned to give the final triangles. Only the non-redundant set of triangles is kept. The information associated with a triangle regarding its feature types and interconnecting distances is then transformed into an integer in a deterministic way and assembled as a component into the fingerprint (Stage 3 in Fig. 2).

GpiDAPH3 fingerprints have been used in two ways in this study. As shown in Fig. 3, the first approach is to examine the *identity* of the encoding components (i.e., pharmacophore triangles). When a set of common integers between two fingerprints is collected, this can be used to calculate the similarity of two fingerprints (e.g., in the Tanimoto coefficient calculation). The novelty can also be calculated by the same token for a given set of unique integers (pharmacophore triangles). The second approach is to just *count* the number of components (triangles) separately for each fingerprint. The more components a molecule has, the more complex a molecule is suggested to be. This is what is referred to as the number of pharmacophore triangles (NumPh4Triangles) in the remainder of the study. As NumPh4Triangles is only a count number



GpiDAPH3	(102976, 168641, 176835, 176841)	(102912, 102976, 111232, 111233, 168642, 176834, 176835, 176843, 177875)
NumPh4Triangles	4	9

Fig. 3 GpiDAPH3 fingerprint usage. The available pharmacophore features are labelled for the two depicted molecules *A* and *B*. *A* Similarity or novelty calculation based on common or novel components (triangles)—2 common triangles between Molecule A and Molecule B are: 102976, 176835. Molecule A has 2 novel pharmacophore triangles (168641, 176841); Molecule B has 7 novel pharmacophore triangles (102916, 111232, 111233, 168642, 176834, 176843, 177875) and *B* molecular complexity estimation based on count of fingerprint components (NumPh4Triangles, irrespective of the identity of the triangles)—Molecule B is more complex than Molecule A, even though Molecule B is lighter than Molecule A. The average NumPh4Triangles is 7

(irrespective of the nature of the underlying triangles), NumPh4Triangles can be used to estimate the relative complexity of compound sets, but does not provide information regarding the identity of the pharmacophore triangles between fingerprints. All fingerprint operation and analysis was facilitated by MOE and Scitegic Pipeline Pilot [30].

To provide a measure of the diversity in a fashion which would appeal to medicinal chemists, the diversity of the SeeDs library (VER_ref) and the fragment hits were also determined by clustering analysis based on the Jarvis-Patrick clustering method [31] as implemented in MOE [28] with the public MACCS structural keys [32, 33] and a Tanimoto similarity metric of 0.7.

Characterization of protein surface cavities

There have been various attempts to characterize the features that contribute to the ability to identify drug molecules for a particular target binding site. Here, we have used a computational method called SiteMap [34, 35] to scan the entire protein surface for potential binding cavities. The location and the druggability of each cavity was detected and ranked blindly without any information of bound ligand. A SiteMap calculation is done in three stages. The first stage is to locate the sites by setting up a grid over the protein surface and site points are grouped into sets according to various criteria to define the sites. Second, the sites are mapped onto another grid to generate maps for visualization. The final stage calculates a druggability score (Dscore) for each site by evaluating the identified site points and the mapping grids. Prior to

SiteMap calculations, the X-ray structures of the 12 targets were prepared by Protein Preparation Wizard [36] which included adding hydrogens, assigning bond orders, optimizing tautomers and rotamers of protein sidechains when appropriate and energy minimization. The energy minimization was performed for hydrogens only. After protein preparation, all non-protein atoms were removed and the bare protein was used for SiteMap calculations.

Results and discussion

Library evolution and analysis

The primary design principle is to have a single fragment library that contains compounds that are compatible with our screening approach and which will provide suitable starting points for hit discovery for a broad range of targets. Therefore, our design strategies took into account physicochemical properties, calculated solubility, medicinal chemistry tractability and diversity. The methods used to generate the first version (VER2004) of the library were published in 2004 [15]. Routine screening and quality control of the fragment library since then have led to quite high turnover (555 fragments removed, primarily due to long term stability concerns). There have been new fragments selected using the same procedure and added to the library. The version as in 2008 (VER2008) has 1,063 compounds. Table 1 summarises the properties of the different fragment libraries. Compared to VER2004, VER2008 has maintained a very similar physicochemical profile, both in overall distribution of properties (Fig. S1) and in average values (Table 1). One notable difference is chemical complexity by pharmacophore triangles. As a whole, VER2008 is slightly more complex than VER2004 by five pharmacophore triangles (NumPh4Triangles in Table 1). The increase in complexity of VER2008 was achieved while maintaining the molecular weight and number of heavy atoms. It is difficult to assess the possible causes behind this as fragments were removed from the library for a variety of reasons (see "Methods"). It is possible that the rather subjective process for adding fragments to the library between VER2004 and VER2008 has unconsciously selected compounds of greater complexity as lessons were learned from fragment screening campaigns. An amalgamated version (VER_ref) has been created as a reference for the analysis presented here by merging fragments from VER2004 and VER2008. After removing the duplicate compounds, VER_ref has 1,605 compounds and its physicochemical characteristics are comparable to either VER2004 or VER2008 versions (Table 1 and Fig. S1).

Characterization by physicochemical properties and pharmacophore space

Maximizing the diversity of a fragment collection not only provides a resource to screen against a variety of targets, but also should increase the number of distinct chemotypes in the screening output. It is equally important that the compounds evolved from the fragments remain in a relevant drug-like space after diversification. The suitability of our fragment library has therefore been assessed by comparison with compounds from the WDI. Fingerprints constructed using 2D 3-point pharmacophore triangles and a number of 2D descriptors were used as a measure of drug likeness.

Table 1 summarises these characteristics for the different fragment libraries in comparison to the WDI. The fragment library contains molecules smaller than the WDI drug reference set. Not surprisingly, the averaged values for the molecular weight (MW), number of heavy atoms (NumHeavy), rotatable bonds (NRot), and number of rings (NumRings) fragments is therefore about half those of the WDI compounds. Vernalis fragments are four times more polar than WDI compounds and their number of detectable pharmacophore triangles (NumPh4Triangles) is also four times less than the averaged number from the WDI set.

Table 2 contains a slightly more detailed comparison of the pharmacophoric complexity of the VER_ref and WDI collections. In total, there are 3,898 and 9,013 unique

Table 1 Average physicochemical properties of various versions of Vernalis (VER) fragment libraries (VER2004, VER2008 and VER_ref) and a drug reference set from World Drug Index (WDI)

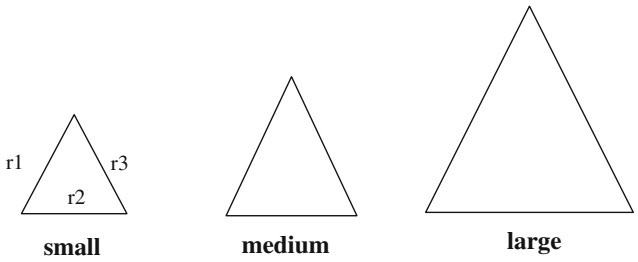
Compound set	Number of compounds	MW	NRot	NumPh4Triangles	NumRings	NumHeavy	SlogP
VER2004	1,275[a]	187.9	2.2	13	1.4	13.3	0.5
VER2008	1,063	193.2	2.2	18	1.6	13.6	0.4
VER_ref	1,605[b]	192.0	2.2	16	1.5	13.6	0.5
WDI	1,141	339.7	4.8	59	2.9	23.5	2.3

MW molecular weight, *NRot* number of rotatable bonds, *NumPh4Triangles* number of pharmacophore triangles, *NumRings* number of rings, *NumHeavy* number of heavy atoms, SlogP [89]

[a] Duplicates have been removed from the published set [15]

[b] Compounds from VER2004 and VER2008 after duplicate removal

Table 2 Comparison of Vernalis fragment library and WDI based on pharmacophore triangles



	VER_ref	WDI	Common	% Of WDI triangles in Vernalis	% Of novel triangles in Vernalis
Total compounds	1,605	1,141	0		
Total pharmacophore triangles	3,898	9,013	2,736	30	30
Small[a]	225	219	144	66	36
Medium[a]	308	389	208	53	32
Large[a]	4	219	4	2	0

[a] Three sets of pharmacophore triangles were examined according to the graph distances (r1–r3 in number of bonds) between vertices (pharmacophore features) of a triangle. Small, medium or large triangles refer all three edges of a triangle to be within 1–3, 4–6 or 7 or more bonds

pharmacophore triangles found in VER_ref and WDI, respectively. Among those, 30% of the WDI triangles can be found in the VER_ref. This, however, does not imply that Vernalis fragments only probe a third of the pharmacophore space compared to WDI. Because of the size difference, some of the pharmacophore triangles in WDI can not exist in the fragments. A comparison of different types of pharmacophore triangles may be more informative [15]. Small, medium and large pharmacophore triangles were constructed for three features separated by 1–3 bonds, between 4 and 6 bonds and between 7 and more bonds, respectively. The numbers of the resulting triangles per compound set are shown in Table 2. For small and medium pharmacophore triangles, 66 and 53% of WDI triangles can be found in VER_ref. There are only four large pharmacophore triangles in the Vernalis library, compared to 219 from WDI. All four large triangles from VER_ref are also part of WDI large pharmacophore collection. The analysis confirms that our fragments do sample more than 50% of drug-like pharmacophore space. In addition, the analysis also revealed that there are novel pharmacophore triangles from the Vernalis library that can not be found in WDI. Both metrics suggest that the fragment library is balanced between relevant chemical space and novelty. This is a property which is needed for the library to be used for a wide range of targets.

Comparison to the published scaffolds

The criteria used to build the Vernalis fragment library are rather simple and largely tailored to identify compounds compatible with our screening methods and synthetic evolution strategies. The distinctive feature of our library design compared to the published protocols elsewhere is the use of pharmacophore fingerprints, rather than the actual chemical substructures, to determine the diversity of the library. It would be of great interest to compare the chemical substructures of the resulting fragments in our SeeDs library with published lists of known preferred scaffolds. To facilitate such analysis, the compounds from the VER_ref library were broken up into distinct scaffolds (molecular frameworks) based on Schuffenhauer's Scaffold Tree method [37]. The Scaffold Tree analysis identified 188 unique rings and 15 types of linkers which constitute 721 different scaffolds in VER_ref. To assess the quality and the coverage of chemical space, benchmark scaffolds were compiled from four studies which described the Vertex SHAPES library [38], Novartis bioactive rings [39], Lilly's analysis of Phase II or later compounds [40] and 72 kinase-targeted scaffolds by Stahura et al. [41].

For scaffold comparison, we have looked at both the coverage of the benchmark scaffolds by VER_ref as well as the breadth or density of coverage (proportion of VER_ref library containing such scaffolds). In Table 3, 23 out of 30 molecular frameworks listed for Vertex SHAPES NMR screening library selection are present in VER_ref and this accounts for 55% of the Vernalis fragments. All Novartis bioactive rings are present in VER_ref and they constitute 65% of the Vernalis library. The coverage of Lilly's most-frequent heterocycles is also excellent (26 out of 30 in VER_ref) but the proportion of Vernalis library containing such rings is less than those in the Vertex library

Table 3 Comparison of scaffolds from Vernalis SeeDs library with published lists

Benchmark scaffolds	VER_ref coverage of benchmark scaffolds	Coverage density in VER_ref library (%)
Vertex SHAPE library [38]	23/30	55
Novartis Bioactive Rings [39]	30/30	65
Lilly heterocycles [40]	26/30	26
Kinase targeted [41]	28/39[a]	23

The coverage density is the proportion of the Vernalis library containing such scaffolds

[a] Among all 72 scaffolds, only 39 passed the SeeDs like filters and 28 of those can be found in VER_ref

and Novartis rings. This is because a number of Lilly scaffolds have multiple entries as the same scaffold with different substitution patterns were counted as different. Finally, of the 72 scaffolds targeted for kinases, 39 pass 'SeeDs like' filters (MW < 250 and logS > -3). Among those scaffolds, 28 are present in 23% of the VER_ref library.

The analysis of scaffolds also revealed that 497 Vernalis fragments (28% of the library) do not contain any of the mentioned published scaffolds. As all the Vernalis fragments satisfy the same lead-like filters, this demonstrates that our approach is able to sample the less explored chemical space as there are additional, novel scaffolds identified. The combination of physicochemical properties and diversity by pharmacophore space not only identifies the majority of the 'privileged' scaffolds reported by others but also a good number of novel, lead-like fragments.

Understanding fragment hit rates

Since the concept of the 'druggable genome' was introduced in 2002 [42], this notion combined with Lipinski's Rule of 5 [43] for drug-like compounds has been regarded as one of the measures to tackle the high attrition rate faced by the pharmaceutical industry. As a step towards describing 'Target Druggability', databases of experimental protein-ligand structures [44–50] have been interrogated by computational chemists to understand the underlying principles influencing ligand binding [51, 52]. The first stage of predicting protein druggability is to predict binding sites for drug-like compounds. Many algorithms [53–66] based on either geometric or energy methods are available for this purpose. For clarity, target druggability in this study is defined as the probability of a protein to bind to small, drug-like compounds with high affinity and specificity [67, 68]. After site identification, quantitatively ranking the druggability of the identified pockets is a more challenging matter. Experimental techniques are available to provide such clues at the gene [69, 70] or protein level [71, 72]. Unfortunately, they are usually slow or costly. In 2005, scientists at Abbott presented a strategy to quickly evaluate protein druggability by screening chemical libraries with 2D heteronuclear-NMR [68]. The observed NMR hit rates were shown to be correlated with a number of surface properties calculated from the binding site.

Encouraged by the study conducted by the Abbott scientists, we have also analyzed our NMR screening data with respect to target druggability. Table 4 summarises the results from fragment screening campaigns against 12 targets (experimental protocols are summarised in Table S1). Many of the targets assessed by our SeeDs approach are of direct pharmaceutical interest such as kinases (PDPK1, CDK2 or JNK3) and heat shock proteins (HSP90 and HSP70). The list also covers targets which some may perceive as difficult or high-risk [73], like protein–protein interactions (PIN-1, PPI-1, PPI-2 and PPI-3). The SeeDs Class 1 hit rate varies from 0.4 to 7.3%. Our fragment screening hit rates are roughly an order of magnitude higher than those reported by Abbott [68]. This could be due to a variety of factors including different sensitivity of chosen NMR techniques, screening buffer conditions, libraries of different content and size. However, like Abbott, the great majority of our fragment hits when crystallized with their respective protein targets bind to the sites which are known to bind small molecules as inferred from existing structural data. Among all 12 targets, high affinity (<300 nM) small-molecular ligands have been reported in the literature [14, 74–79] except for HSP70, PIN-1 and PPI-3 which correspond to the lowest hit rates observed (between 0.4 and 0.7%). For targets which yielded high hit rates (>2%), all nine proteins have potent small molecule binders known to date. So our finding is consistent with Abbott's, even though different techniques were used to detect binding. Our data suggest that fragment binding hit rate by 1D NMR could also be seen as an indication that potent ligands could be developed for a particular binding site (i.e., druggability).

To appreciate the observed hit rates at the atomic level, the protein surface of all 12 targets were characterized by SiteMap [34]. SiteMap reads in a 'bare' protein structure with all non-protein atoms removed and returns the location, volume, surface and shape of potential pockets to bind small molecules. All the identified pockets are also ranked based on their calculated druggability scores. As seen in Table 4, all but two fragment binding sites were correctly predicted and ranked by SiteMap. For PPI-3, the internal crystallographic data could not confidently pinpoint the exact location or the binding modes of the fragment hits. For PIN-1, the fragment binding site was correctly predicted by SiteMap but ranked as the second best druggable site. Examining the SiteMap results on PIN-1 revealed that the most

Table 4 SeeDs screening hit rates for 12 protein targets

Protein	Number of hits			Library size	Class 1 hit rate		High affinity ligands[d]	Fragment binding site ranking[e]
	Total[a]	Class 1[b]	Class 1 series[c]		%	Category		
AK	15	11	10	308	3.6	High	Yes	1
CDK2	109	40	35	1,250	3.1	High	Yes	1
DNA gyrase	54	44	39	855	4.9	High	Yes	1
FAAH	81	63	51	868	7.3	High	Yes	1
HSP70	38	6	5	1,351	0.4	Low	No	1
HSP90	82	60	42	1,351	4.4	High	Yes	1
JNK3	101	55	53	1,351	4.0	High	Yes	1
PDPK1	119	58	54	1,260	4.5	High	Yes	1
PIN-1	13	5	4	1,351	0.4	Low	No	2
PPI-1	40	34	23	1,064	3.2	High	Yes	1
PPI-2	52	24	20	1,068	2.2	High	Yes	1
PPI-3	39	10	9	1,351	0.7	Low	No	N/D[f]
Average	62	34	29		3.2			
Total unique fragments	462	288						

[a] Total number of fragments identified by at least one NMR experiment to interfere with the binding of known competitor compound

[b] Number of fragments identified by all three NMR experiments (STD, Water-LOGSY and CPMG) to interfere with the binding of known competitor compound

[c] Total number of unique chemical series suggested by the clustering results of Class 1 fragment hits with a Tanimoto coefficient of 0.70 and MACCS keys

[d] Reported affinities <300 nM. Please refer to the main text for references

[e] SiteMap [34] ranking of the fragment binding site compared to the rest of the putative sites from the same target identified by SiteMap

[f] Not detected by SiteMap

druggable site was predicted to be the packing site for the N-terminus WW domain of PIN-1 on the catalytic domain (peptidyl-prolyl cis/trans isomerase, PPIase) [80, 81]. The PIN-1 crystal structure (2ITK [82]) used for SiteMap calculation contains the C-terminus catalytic domain only, leaving out the N-terminus WW domain. Since the WW packing site was left bare, SiteMap did recognize the site and prioritized it to be more favorable than the catalytic site from the PPIase domain.

SiteMap also calculates a score (Dscore) for each pocket to indicate its druggability potential. Dscore considers three key aspects of a binding site which are the size, degree of enclosure and hydrophilicity of a binding site [35]. Using Class 1 hit rate of 2% as a cut-off, all targets which yielded high hit rates (>2%) have Dscore greater than 0.8 (Fig. 4; Table 4). For three targets which returned <2% hit rates, two of them have Dscore <0.6. HSP70 is the anomaly from the list which has Dscore greater than 0.8 but resulted in a low hit rate. Examining the HSP70 ATP binding site revealed that it is lined by side-chain atoms of flexible residues and has shown to be very malleable depending on the ligand (internal unpublished data). This dynamic feature observed for the HSP70 ATP binding site is currently unable to be captured during SiteMap calculations. In addition, the

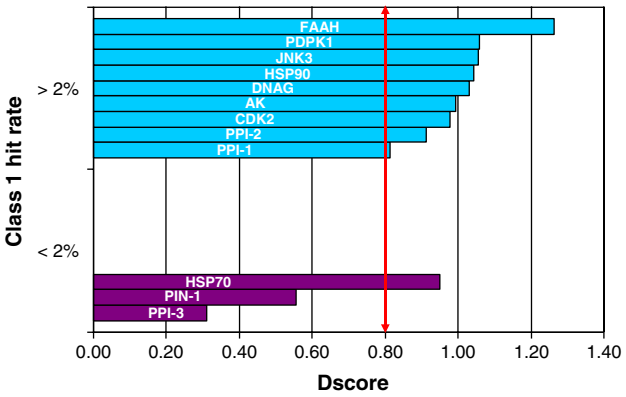


Fig. 4 Targets with observed high (>2%, *blue bars*) and low (<2%, *purple bars*) Class 1 hit rates compared to the druggability score (*Dscore*) calculated by SiteMap [34]. The *red arrow* indicates the minimum Dscore for targets yielding high hit rates for the current data set

HSP70 NMR screen revealed a large amount of non-competitive fragment hits which may also be related to the dynamic properties of the HSP70 binding site. Taking out HSP70, it is encouraging to see that the SiteMap Dscore appears to be a good indicator for the level of NMR hit rate one should expect for a target from our limited list. There is a

'gray' zone of Dscore between 0.6 and 0.8 which is not represented by our data set. With more targets coming through in the future, a better assessment may be possible with respect to Dscore and fragment screening hit rates.

In order to deduce what a good binding site should be, Dscore and its three components were plotted (Fig. 5). In addition, an additional parameter called site compactness [68] was introduced to assess the binding sites. Site compactness is a ratio of the site volume to its total surface area. The relevant parameters were plotted for three groups of binding sites, experimentally observed binding sites which yielded high or low hit rates (see Table 4) and other binding sites detected by SiteMap which are not experimentally observed (non-binding sites). The averaged values have been calculated for each group and normalized against the ones from the high hit rate group. Consistent with Fig. 4, Dscore clearly discriminates binding sites of high hit rates from the rest. The experimentally observed binding sites also tend to be significantly larger than the sites where no known molecule binds (Size columns in Fig. 5). A 'good' binding site is much better enclosed by the surrounding protein atoms (Enclosure in Fig. 5). The surface property should also be more hydrophobic in nature. The site compactness value also suggests that a good binding site tends to be compact and could be an additional useful indicator to separate more druggable sites from less ones based on our data set. The surface property trend observed here regarding a good binding site agrees entirely with previous findings [54, 61, 68, 83–85]. The combination of size, enclosure and hydrophilicity (Dscore) correlates with observed Class 1 hit rates with an R^2 value of 0.51. Substituting enclosure with compactness while keeping the other two parameters untouched (modified Dscore) improves the correlation with hit rates with an R^2 value of 0.61. Of note is that the equation to calculate original or modified Dscore has not been refit with the current data set. The weighting factors were taken as they were when fitting to the original training set [35]. Refitting the Dscore equation to the current data set would give weighting factors more sensitive to the current dataset leading to higher R^2 values. However, given the size of the data sets and the nature of data acquired are different between our study and the original SiteMap study, it may not be appropriate to do so. Nonetheless, both fragment hits and high-affinity ligands do share the same binding sites and are dictated by the same principle of molecular recognition. This is why assessing NMR screening hit rates with calculated Dscore led to a very reasonable classification.

Analysis of fragment hits

The Vernalis fragment library has been screened against a variety of targets including kinases, ATPases, protein–protein interaction targets and others. In our experience, an average hit rate for the most robust Class 1 hits is 3.2% and this corresponds to 34 unique competitive fragment hits constituting 29 chemical series (Table 4). As a reference, the whole VER_ref library has 1,483 distinct clusters (out of 1,605 fragments) when generated with a Tanimoto coefficient of 0.70. The averaged 3.2% hit rate has given us confidence in this library to find sensible starting points for most of our programs. We are also reassured that the pharmacophore diversity applied during library enumeration has indeed translated into diverse chemical series in the screening output. On the other hand, we also recognized that each screen will have different sensitivity introduced by optimizing the screening condition for each protein target such as varying the buffer conditions and the protein concentration. The identification of hits relies on changes in NMR signals during STD and LOGSY experiments, and it is well-known that changes in signal are not necessarily related to binding, and vice versa. The thermodynamics and kinetics of binding of the competitor ligand will also have an effect.

It is difficult to conduct a detailed analysis of the hits obtained from screening campaigns against different targets run at different times, with different buffer and instrument conditions and interpreted by different operators. However,

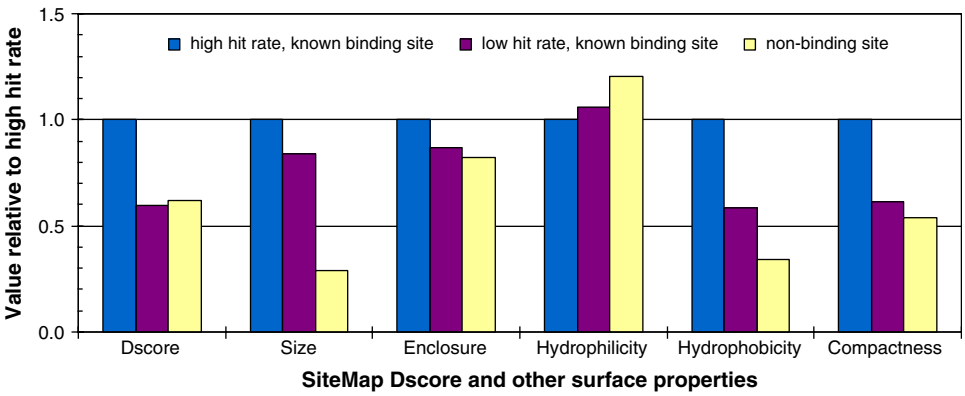


Fig. 5 Comparison of Dscore, its components (size, enclosure and hydrophilicity), hydrophobicity and binding site compactness for experimentally observed (known) binding sites which produced high and low fragment screening hit rates and non-binding sites. The impact of the listed parameters has been normalized against the group of high hit rates

some lessons can be learnt to inform future generations of library design by inspecting the number and the type of hits collected throughout the years based on some common physicochemical properties.

Comparison of Hits, Non-Hits and overall library

We first separated the VER_ref library into two subsets, one consisting of fragments which have been identified as competitive binders (Hits) by at least one NMR experiment (Class 1, 2 or 3, see "Methods") for a given target and the other set containing fragments which have not been recognized as hits by any of our 12 internal projects (Non-Hits). The Hits were made up of 29% of the compounds in the VER_ref library and 18% of the VER_ref library were Class 1 hits. This leaves 71% of the VER_ref library which has not been found to bind competitively with any of the 12 targets.

A number of physicochemical properties have been examined for different sets of fragments to see if any particular compound property could explain why only a third of the library were competitive binders. Figure 6 shows the distribution plots of molecular weight (MW), rotatable bonds (NRot), SlogP, total number of pharmacophore triangles and the number of rings for the VER_ref library, Hits, Class 1 hits and Non-Hits. In terms of averaged molecular weight and rotatable bonds (Table 5), there is little difference between Hits, Non-Hits and the VER_ref library. The histograms in Fig. 6a, b indicate that the distribution of MW and NRot for Hits and Non-Hits are similar to each other and follow the curves of the VER_ref library. However, there is a small tendency for the Hits to have slightly lower MW and NRot. For example, 64% of the Hits are below MW of 200, compared to 59% of Non-Hits and 72% of Hits have two or less rotatable bonds, compared to 64% of Non-Hits. Although only a small effect, it may be that the slightly more rigid fragments (lower NRot) that make up the Hits have a higher entropic gain on binding than the Non-Hits.

There is a clear separation between Hits and Non-Hits when considering SlogP. As shown in Fig. 6c, the Hits appear to be generally more hydrophobic than Non-Hits, with on average twice the hydrophobicity as measured by SlogP (Table 5). This observation agrees with the general observation that hydrophobicity promotes binding [68, 83, 84] and echoes the SiteMap analysis of binding surface properties (Hydrophobicity in Fig. 5). The importance of enhanced hydrophobicity was also highlighted by two other groups which performed fragment screening with various techniques [18, 19]. AstraZeneca (AZ) also reported a similar level of hydrophobicity enhancement for fragment hits ($\sim$two-fold). However, the averaged ClogP value of AZ fragment hits is 2.1 [19], compared to the averaged SlogP value of 0.8 for our Hits. While different algorithms used for logP calculation will lead to slightly different values [86], the over twofold difference of logP observed for fragment hits is most likely due to the logP difference in the starting library. The VER_ref library has a mean logP of 0.5, compared to the mean logP of 1.4 for the AZ library. VER_ref may be so hydrophilic because of the high solubility requirement for the NMR detection methods used. We found that the use of a water solubility filter in the first two iterations of our fragment library design cycle [15] drove the SlogP values to the lower (more hydrophilic) range.

The consequences of having such a hydrophilic library are important and help a general strategy for evolving hits from the fragments. Arguably, the central scaffold is making the key directional interactions (polar, hydrogen bonding etc.) in the target binding site and should be optimal before adding the usually more hydrophobic bulk that is used to increase potency. However, this analysis has highlighted that the fragment library could stand further enrichment with more hydrophobic fragments. As the in silico water solubility model we use has been reasonably predictive (88% of the fragments correctly predicted to be soluble at 2 mM [15]), we should be able to bias subsequent library curation with fragments with relatively higher SlogP. As shown in Fig. 6c, applying a cut-off of 2 mM for water solubility still allows us to sample an ample range of SlogP.

The pharmacophore triangles present in Hits are little different from those seen for Non-Hits in terms of averaged number of triangles encoded by each fragment (Table 5; Fig. 6d). The distribution curves of pharmacophore triangles for the four fragment sets (VER_ref, Hits, Class 1 hits and Non-Hits) are also very similar and Hits do not enrich a particular range of pharmacophore triangles (Fig. 6d). Plotting the same histogram with a finer bin width revealed that an area populated more frequently by Hits compared to Non-Hits is when the number of pharmacophore triangles is greater than 4 (data not shown). This could suggest a minimum requirement of pharmacophore complexity for a fragment for specific binding. Curiously, Class 1 hits appear to be more complex than Class 2–3 hits having, on average, an additional three pharmacophore triangles (Table 5). The enhanced pharmacophore complexity observed for Class 1 hits is likely to improve the chance of a small fragment for specific protein binding by having an appropriate combination of pharmacophore features. Moreover, the binding is likely to be achieved with sufficient affinity which allows such fragments to be detected consistently in a series of NMR screening experiments. A preliminary analysis of the experimental results from NMR screening suggests that STD registers the most hits for more targets than the other methods, but that LOGSY and CPMG can detect more hits for some targets. There appears to be no general rule of which NMR experiment is most

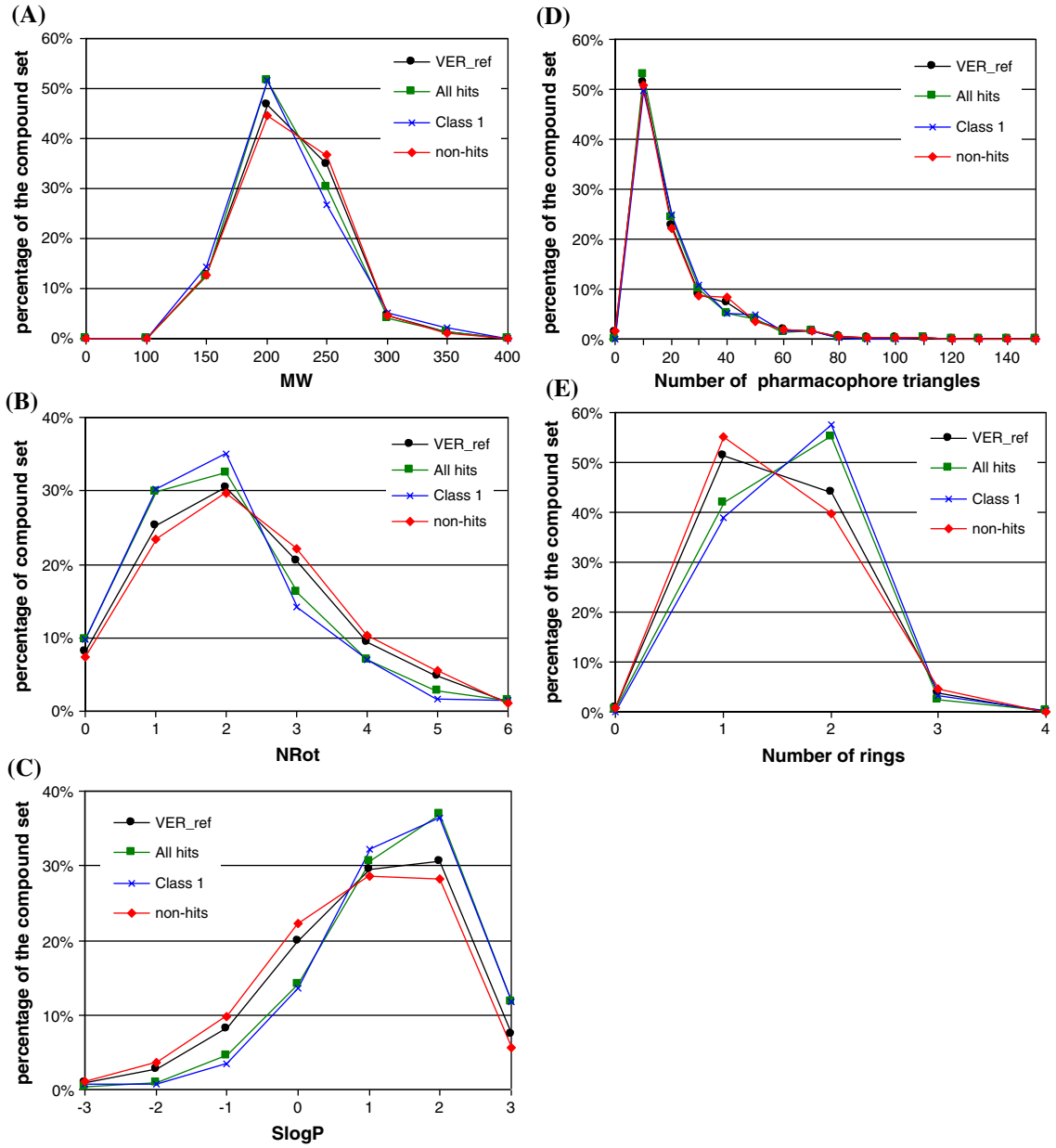


Fig. 6 Distribution plots of **a** molecular weight (*MW*), **b** number of rotatable bonds (*NRot*), **c** SlogP, **d** number of pharmacophore (*ph4*) triangles and **e** number of rings for the whole library (VER_ref), all hits (Class 1–3), Class 1 hits and non-hits

robust for which type of target, though the tendency is that STD works reliably for kinases and LOGSY where water molecules are important in ligand binding, as in HSP90.

Figure 6e summarises the number and types of ring structures present in the different categories. Although the number of rings is the same, the Hits contain more two-ring compounds than the Non-Hits. An increased number of rings is an efficient way to improve complexity without being penalized by entropy. In addition, a ring allows more pharmacophore features to be incorporated without significantly increasing the number of rotatable bonds. This may make some fragments more efficient binders than the others and more easily seen to bind by NMR. Two simple measures were devised to look at the degree of overall cyclization and of aromatization. The degree of cyclization (cyclicity) is determined by the ratio between the number of ring bonds and total number of bonds between heavy atoms and the degree of aromatization (aromaticity) is assessed by the ratio between the number of aromatic bonds and total number of bonds between heavy atoms. As shown in Table 6, while the overall cyclicity has remained fairly constant among all fragment sets, the Hits appear to be more aromatic than Non-Hits by 13%. The Class 1 hits were shown to have one of the highest aromaticity.

Table 5 Average properties for various compound sets as plotted in Figs. 6 and 7

Compound set	MW	NRot	SlogP	NumPh4Triangles	NumRings
VER_ref	192	2.2	0.5	16	1.5
Non-hits[a]	194	2.3	0.4	16	1.5
All hits	189	2.0	0.8	15	1.6
Class 1[b]	190	1.9	0.8	17	1.6
Class 2–3[b]	188	2.1	0.8	14	1.5
Kinase hits	186	1.8	0.9	16	1.6
PPI hits	202	2.1	1.2	16	1.8

[a] Fragments which have not been identified as competitive binders for the 12 targets

[b] Please refer to "Methods" for the definition of Class 1, 2 and 3

Table 6 Cyclicity and aromaticity of the fragments

Compound set	Cyclicity[a]	Aromaticity[b]
VER_ref	0.58	0.45
All hits	0.61	0.55
Class_1[c]	0.63	0.57
Class_2_3[c]	0.59	0.50
Non-Hits[d]	0.57	0.42
Kinase hits	0.64	0.58
PPI hits	0.64	0.57

[a] Cyclicity is defined as the ratio between the number of ring bonds and total number of bonds between heavy atoms

[b] Aromaticity is the ratio between the number of aromatic bonds and total number of bonds between heavy atoms

[c] Please refer to "Methods" for the definition of Class 1, 2 and 3

[d] Fragments which have not been identified as competitive binders for the 12 targets

These analyses show that there are subtle differences in the distribution of the various properties between Hit and Non-Hit fragments which reflect the properties to be expected for increased affinity of binding. However, these differences are small and the average values for most of the properties are quite similar between Hits and Non-Hits. This confirms that the library design procedure has generated a suitable library for screening. The next section considers how well this library performs across a range of different targets.

Analysis of hits across different targets

An early concern about fragments was whether specific binding could be achieved with such small compounds. We have analysed the output from screening the library against 12 targets to assess the degree of target discrimination by fragments. Among all fragment hits, 62% of the fragments were competitive binders with just one target and another 24% were hits for two targets. Considering how small fragments are, this reflects quite a high target specificity. There are nine fragments which were identified as competitive binders 42–50% of the time (this corresponds to 5–6 of the 12 screens shown in Table 4). This pool of nine fragments, representing 0.6% of the library (VER_ref), appears to be rather versatile binders. Interestingly, among the nine versatile fragments, they showed a preference for ATP binding proteins as their average probability (80%) of hitting an ATP binding protein is higher than the expected average (58%). This may suggest these 9 fragments to be 'generic' scaffolds for ATP binding proteins. All but two of the nine fragments have a bicyclic ring with nitrogen and oxygen atoms decorated around the ring. Flexible alignment of those fragments with adenine suggests multiple solutions for the overlays and some of the overlays have been verified by X-ray crystallography (internal unpublished data).

The nature of the chemical space covered by the fragment hits has been assessed by clustering on Tanimoto similarity with MACCS fingerprints and the results are summarized in Table 4. When clustering at a similarity threshold of 0.70, more than 84% of the fragment hits are unique clusters and the mean Tanimoto similarity score is 0.40 (±0.15) between all cluster centroid fragments for a given target. This suggests that the redundancy of chemical scaffolds in the fragment hits is very small. The diversity in fragment hits seems to be preserved across all 12 targets. This has shown that the pharmacophore diversity driven selection has been productive in adding fragments with distinctive features.

Figure 7 plots the average properties of all the fragment hits and the properties of the hits for two protein families, kinases and protein–protein interaction (PPI). Overall, the properties are quite similar across the different targets though with some interesting differences in detail. Fragment hits for PPI are slightly heavier on average by a MW of about a carbon (Fig. 7a) and this is also reflected in the number of heavy atoms (14 for PPI hits and 13 non-PPI hits). The size increase observed for PPI hits was achieved not by increasing the number of rotatable bonds (Fig. 7b) but mainly by having a higher number of rings (Fig. 7e). The fragment hits for PPI also appear to be more hydrophobic than hits for other targets (Fig. 7c). These trends for the PPI hits are consistent with the nature of their binding sites, as most known PPI binding sites are a collective set of small and shallow hydrophobic pockets.

Although no differences have been observed for the average number of pharmacophore triangles among different sets of hits (Fig. 7d), the level of complexity required for a fragment to be detected in binding varies

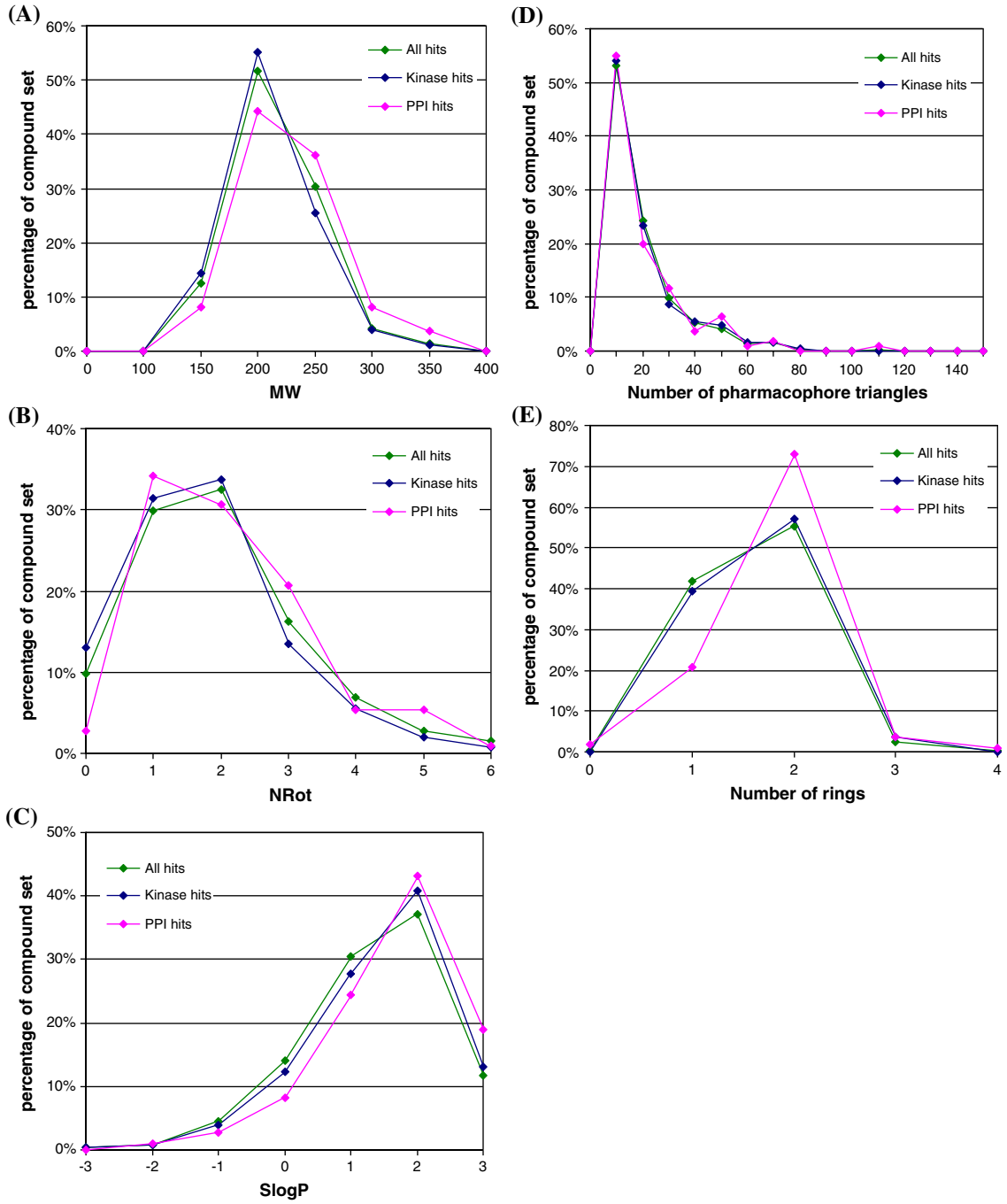


Fig. 7 Distribution plots of **a** molecular weight (*MW*), **b** number of rotatable bonds (*NRot*), **c** SlogP, **d** number of pharmacophore triangles and **e** number of rings for all hits, kinase hits and Protein–protein interaction (*PPI*) hits

from one target to another. Figure 8 plots the averaged pharmacophore complexity of both the Hits and the Class 1 hits for each target. HSP70 appears to be the most demanding target as it requires the most complex fragments (20 and 27 triangles for all and Class 1 hits) among all targets studied. This could in part explain why its hit rate was among the lowest as fewer fragments have the complexity required for HSP70 binding. The suggested complexity from three kinase screens is roughly the same (16–18 triangles), except for CDK2 Class 1 hits which appear to be most complex among all three kinases. Two of the PPIs are also on the high end of molecular complexity which correlates with the bigger size observed for the hits.

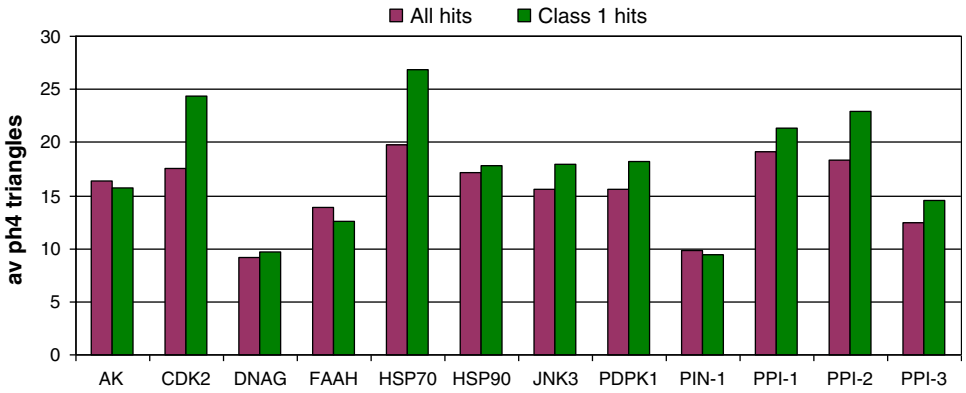


Fig. 8 Pharmacophore complexity observed for all fragments hits and Class 1 hits for 12 protein targets

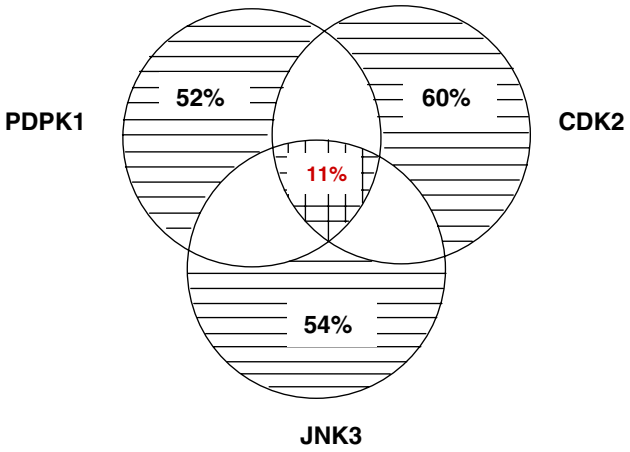


Fig. 9 Overlap of kinase fragment hits. The *horizontal lines* indicate the portion of unique fragment hits to each kinase. The *crossed area* (11%) is the portion of common hits to all three kinases

Among all 12 targets, there are three kinases and their hits were identified by different competitor ligands all binding to the ATP binding site. Kinases are one of the protein families which yielded the highest hit rate. Although all three kinases share 'high' sequence identity and close similarity of protein structure, there is surprising selectivity in the fragment hits observed. About 11% of the fragment hits are common to all three kinases (Fig. 9). Between any pair of two kinases, the overlap of hits is between 20 and 30%. This means at least 52% of the kinase fragment hits are unique to one of these three kinases. The number is 12% less than the overall uniqueness (64%) observed for a fragment hit to any target. A good number of the hits were crystallized with their respective targets and all of them were found to bind to the ATP binding site exhibiting the conventional hinge binding pharmacophore, although sometimes with different binding modes [8]. It is suspected that different steric and electrostatic properties are exhibited around the hinge region for PDPK1, CDK2 and JNK3 even with subtle and conservative substitution of amino acids. Some fragment hits seem to be able to respond to subtle differences.

Conclusions

One of the biggest problems faced in designing drug molecules is the vast chemical space [87, 88] to be explored. Fragment-based approaches identify a subset of chemical moieties responsible for key molecular recognitions early on and this allows scientists to devote their time in a much reduced and relevant chemical space. Part of the efficiency of fragment-based screening depends on the diversity and hit-likeness of the compounds in the screening library. In this study, we have described the design and the evolution of Vernalis fragment library. The assessment of the Vernalis library with respect to known drug molecules suggests that it is possible to capture drug-like chemical space by small fragments. Comparison of the Vernalis library and benchmark scaffolds provides another way to assess the coverage of chemical space by our library. The results suggest that the pharmacophore diversity approach applied during library enumeration has yielded a combination of known and novel scaffolds in the library.

The true test of the Vernalis library is the results from screening against 12 diverse protein targets by NMR. The averaged hit rate for Class 1 hits is 3.2% and this corresponds to 34 unique competitive fragment hits constituting 29 chemical series. Not only have we demonstrated a sufficiently high hit rate for a variety of targets, but also truly diverse hits were obtained for each target without tailoring the library specifically for a target. The analysis of fragment hits has highlighted the significance of hydrophobicity in binding, as fragment hits were found to be twice as lipophilic than non-hits. The enhanced SlogP values revealed in the fragment hits also allowed us to improve our library enumeration process by incorporating a specific SlogP filter after the water solubility calculation. Through the examination of chemical complexity and the composition of rings, the effectiveness of fragment hits seems to be a result of the right combination of pharmacophore features presented in an entropically favoured way with enhanced aromaticity. Although all fragment hits

share similar physicochemical properties, different patterns were revealed for different protein families. For example, the fragment hits for protein–protein interaction targets appear to be heavier and more hydrophobic than other fragment hits. Both features nicely complement the characteristic of the binding sites from protein–protein interaction targets. This also implies that fragment hits are discriminative and genuine binders. We have also noted that fragment screening hit rates can indicate the level of difficulty in progressing a target in terms of deriving high-affinity compounds (druggability). The druggability score (Dscore) calculated by SiteMap has allowed us to differentiate targets of high and low hit rates. The characterization of the binding pockets also revealed the key contributions of surface properties to distinguish druggable from non-druggable binding sites.

A well-designed library needs to consider diversity, drug/lead-like properties, solubility, synthetic accessibility, maintenance and efficient data analysis. To make it truly useful, it also has to incorporate the demands of the chosen screening technique, medicinal chemists and compound management teams. A library also has to be continuously evolving to keep up to date with the demands of the different discovery projects and the changing nature of the targets under study. Our retrospective analysis of the Vernalis library and screening experiences has not only confirmed that the library is fit for purpose. It has also provided knowledge and directions for future library evolution while maintaining its quality as a resource for initiating drug discovery against very different classes of target. Finally, fragment library design is only a component of the whole fragment based drug discovery process. To capitalize on a sound design library requires a seamless integration of structures, computational and medicinal chemistry. Tools which make it easier for bench chemists to assimilate information about fragment hits into compound design are one of the next steps for the methods. We also hope that the analysis presented in this study will contribute to the development of new tools and future improvement of the fragment based approach.

Acknowledgments The authors thank many scientists at Vernalis to make the analysis possible. In particular, this analysis has benefit greatly from fruitful discussion and data mining by Heather Simmonite, Ben Davis and James Murray. We also express our gratitude to Chemical Computing Group and Schrodinger Inc., for answering our questions and helping us with the programs MOE and SiteMap to extract information needed for the analysis.

References

1. Hajduk PJ, Greer J (2007) Nat Rev Drug Discov 6:211. doi:10.1038/nrd2220
2. Jhoti H, Cleasby A, Verdonk M, Williams G (2007) Curr Opin Chem Biol 11:485. doi:10.1016/j.cbpa.2007.07.010
3. Congreve M, Chessari G, Tisi D, Woodhead AJ (2008) J Med Chem 51:3661. doi:10.1021/jm8000373
4. Shuker SB, Hajduk PJ, Meadows RP, Fesik SW (1996) Science 274:1531. doi:10.1126/science.274.5292.1531
5. Nienaber VL, Richardson PL, Klighofer V, Bouska JJ, Giranda VL, Greer J (2000) Nat Biotechnol 18:1105. doi:10.1038/80319
6. Lepre CA, Moore JM, Peng JW (2004) Chem Rev 104:3641. doi:10.1021/cr030409h
7. Antonysamy SS, Aubol B, Blaney J, Browner MF, Giannetti AM, Harris SF, Hébert N, Hendle J, Hopkins S, Jefferson E, Kissinger C, Leveque V, Marciano D, McGee E, Nájera I, Nolan B, Tomimoto M, Torres E, Wright T (2008) Bioorg Med Chem Lett 18:2990. doi:10.1016/j.bmcl.2008.03.056
8. Hubbard RE, Davis B, Chen I, Drysdale MJ (2007) Curr Top Med Chem 7:1568. doi:10.2174/156802607782341109
9. Mayer M, Meyer B (1999) Angew Chem Int Ed Engl 38:1784. doi:10.1002/(SICI)1521-3773(19990614)38:12<1784::AID-ANIE1784>3.0.CO;2-Q
10. Meiboom S, Gill D (1958) Rev Sci Instrum 29:688. doi:10.1063/1.1716296
11. Dalvit P, Pevarello DP, Tato M, Veronesi M, Vulpetti A, Sundstrom M, Bio J (2000) NMR 18:65. doi:10.1023/A:1008354229396
12. Petros AM, Dinges J, Augeri DJ, Baumeister SA, Betebenner DA, Bures MG, Elmore SW, Hajduk PJ, Joseph MK, Landis SK, Nettesheim DG, Rosenberg SH, Shen W, Thomas S, Wang X, Zanze I, Zhang H, Fesik SW (2006) J Med Chem 49:656. doi:10.1021/jm0507532
13. Howard N, Abell C, Blakemore W, Chessari G, Congreve M, Howard S, Jhoti H, Murray CW, Seavers LCA, van Montfort RLM (2006) J Med Chem 49:1346. doi:10.1021/jm050850v
14. Brough PA, Aherne W, Barril X, Borgognoni J, Boxall K, Cansfield JE, Cheung K-MJ, Collins I, Davies NGM, Drysdale MJ, Dymock B, Eccles SA, Finch H, Fink A, Hayes A, Howes R, Hubbard RE, James K, Jordan AM, Lockie A, Martins V, Massey A, Matthews TP, McDonald E, Northfield CJ, Pearl LH, Prodromou C, Ray S, Raynaud FI, Roughley SD, Sharp SY, Surgenor A, Walmsley DL, Webb P, Wood M, Workman P, Wright L (2008) J Med Chem 51:196. doi:10.1021/jm701018h
15. 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 Comput Sci 44:2157. doi:10.1021/ci049806z
16. Jacoby E, Davies J, Blommers MJJ (2003) Curr Top Med Chem 3:11. doi:10.2174/1568026033392606
17. Rees DC, Congreve M, Murray CW, Carr R (2004) Nat Rev Drug Discov 3:660. doi:10.1038/nrd1467
18. Schuffenhauer A, Ruedisser S, Marzinzik AL, Jahnke W, Blommers M, Selzer P, Jacoby E (2005) Curr Top Med Chem 5:751. doi:10.2174/1568026054637700
19. Albert JS, Blomberg N, Breeze AL, Brown AJ, Burrows JN, Edwards PD, Folmer RH, Geschwindner S, Griffen EJ, Kenny PW, Nowak T, Olsson LL, Sanganee H, Shapiro AB (2007) Curr Top Med Chem 7:1600. doi:10.2174/156802607782341091
20. Hubbard RE, Chen I, Davis B (2007) Curr Opin Drug Discov Devel 10:289
21. Hajduk PJ, Bures M, Praestgaard J, Fesik SW (2000) J Med Chem 43:3443. doi:10.1021/jm000164q
22. Muegge I, Heald SL, Brittelli D (2001) J Med Chem 44:1841. doi:10.1021/jm015507e
23. Bemis GW, Murcko MA (1999) J Med Chem 42:5095. doi:10.1021/jm9903996
24. Bemis JW, Murcko MA (1996) J Med Chem 39:2887. doi:10.1021/jm9602928

25. http://www.daylight.com/dayhtml_tutorials/languages/smarts/index.html
26. Furet P, Meyer T, Strauss A, Raccuglia S, Rondeau J-M (2002) Bioorg Med Chem Lett 12:221. doi:10.1016/S0960-894X(01)00715-6
27. Thomson Scientific 3501 Market Street, Philadelphia, PA 19104, U.S.A., http://thomsonderwent.com/products/lr/wdi/
28. MOE (The Molecular Operating Environment) Version 2008.10, Chemical Computing Group Inc., 1010 Sherbrooke Street West, Suite 910, Montreal, Canada H3A 2R7. http://www.chemcomp.com
29. Kirsten G (2008) GpiDAPH3, Chemical computing group, personal communication
30. http://accelrys.com/products/scitegic/
31. Jarvis RA, Patrick EA (1973) Trans IEEE Comput C-22:1025
32. Durant JL, Leland BA, Henry DR, Nourse JG (2002) J Chem Inf Comput Sci 42:1273. doi:10.1021/ci010132r
33. MACCS (Molecular ACCess System) Symyx Technologies, Inc, 415 Oakmead Parkway, Sunnyvale, CA 94085
34. SiteMap 2.2 (2008) Schrodinger LLC, 120 West 45th Street, New York, NY 10036
35. Halgren TA (2009) J Chem Inf Model 49:377. doi:10.1021/ci800324m
36. Protein Preparation Wizard (2008) Schrodinger LLC, 120 West 45th Street, New York, NY 10036
37. Schuffenhauer A, Ertl P, Roggo S, Wetzel S, Koch MA, Waldmann H (2007) J Chem Inf Model 47:47. doi:10.1021/ci600338x
38. Fejzo J, Lepre CA, Peng JW, Bemis GW, Ajay, Murcko MA, Moore JM (1999) Chem Biol 6:755. doi:10.1016/S1074-5521(00)80022-8
39. Ertl P, Jelfs S, Muhlbacher J, Schuffenhauer A, Selzer P (2006) J Med Chem 49:4568. doi:10.1021/jm060217p
40. Broughton HB, Watson IA (2005) J Mol Graph Model 23:51. doi:10.1016/j.jmgm.2004.03.016
41. Stahura FL, Xue L, Godden JW, Bajorath J (1999) J Mol Graph Model 17:1. doi:10.1016/S1093-3263(99)00015-7
42. Hopkins AL, Groom CR (2002) Nat Rev Drug Discov 1:727. doi:10.1038/nrd892
43. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (1997) Adv Drug Deliv Rev 23:3. doi:10.1016/S0169-409X(96)00423-1
44. Roche O, Kiyama R, Brooks CL (2001) J Med Chem 44:3592. doi:10.1021/jm000467k
45. Chen X, Lin Y, Liu M, Gilson MK (2002) Bioinformatics 18:130. doi:10.1093/bioinformatics/18.1.130
46. Wang R, Fang X, Lu Y, Wang S (2004) J Med Chem 47:2977. doi:10.1021/jm030580l
47. Wang R, Fang X, Lu Y, Yang C-Y, Wang S (2005) J Med Chem 48:4111. doi:10.1021/jm048957q
48. Hu L, Benson ML, Smith RD, Lerner MG, Carlson HA (2005) Proteins 60:333. doi:10.1002/prot.20512
49. Kellenberger E, Muller P, Schalon C, Bret G, Foata N, Rognan D (2006) J Chem Inf Model 46:717. doi:10.1021/ci050372x
50. Block P, Sottrifer CA, Dramburg I, Klebe G (2006) Nucleic Acids Res 34:D522. doi:10.1093/nar/gkj039
51. Smith RD, Hu L, Falkner JA, Benson ML, Nerothin JP, Carlson HA (2006) J Mol Graph Model 24:414. doi:10.1016/j.jmgm.2005.08.002
52. Carlson HA, Smith RD, Khazanov NA, Kirchhoff PD, Dunbar JB Jr, Benson ML (2008) J Med Chem 51:6432. doi:10.1021/jm8006504
53. Miranker A, Karplus M (1991) Proteins 11:29. doi:10.1002/prot.340110104
54. Laskowski RA (1995) J Mol Graph 13:323. doi:10.1016/0263-7855(95)00073-9
55. Hendlich M, Rippmann F, Barnickel G (1997) J Mol Graph Model 15:359. doi:10.1016/S1093-3263(98)00002-3
56. Liang J, Edelsbrunner H, Fu P, Sudhakar PV, Subramaniam S (1998) Proteins 33:1. doi:10.1002/(SICI)1097-0134(19981001)33:1<1::AID-PROT1>3.0.CO;2-O
57. Brady GP Jr, Stouten PF (2000) J Comput Aided Mol Des 14:383. doi:10.1023/A:1008124202956
58. An J, Totrov M, Abagyan R (2004) Genome Inf 15:31
59. Laurie AT, Jackson RM (2005) Bioinformatics 21:1908. doi:10.1093/bioinformatics/bti315
60. Coleman RG, Salzberg AC, Cheng AC (2006) J Chem Inf Model 46:2631. doi:10.1021/ci600229z
61. Nayal M, Honig B (2006) Proteins 63:892. doi:10.1002/prot.20897
62. Cheng AC, Coleman RG, Smyth KT, Cao Q, Soulard P, Caffrey DR, Salzberg AC, Huang ES (2007) Nat Biotechnol 25:71. doi:10.1038/nbt1273
63. Halgren T (2007) Chem Biol Drug Des 69:146. doi:10.1111/j.1747-0285.2007.00483.x
64. Landon MR, Lancia DR Jr, Yu J, Thiel SC, Vajda S (2007) J Med Chem 50:1231. doi:10.1021/jm061134b
65. Harris R, Olson AJ, Goodsell DS (2007) Proteins 70:1506. doi:10.1002/prot.21645
66. Schalon C, Surgand JS, Kellenberger E, Rognan D (2008) Proteins 71:1755. doi:10.1002/prot.21858
67. Hajduk PJ, Huth JR, Tse C (2005) Drug Discov Today 10:1675. doi:10.1016/S1359-6446(05)03624-X
68. Hajduk PJ, Huth JR, Fesik S (2005) J Med Chem 48:2518. doi:10.1021/jm049131r
69. Hardy LW, Peet NP (2004) Drug Discov Today 9:117. doi:10.1016/S1359-6446(03)02969-6
70. Xin H, Bernal A, Amato FA, Pinhasov A, Kauffman J, Brenneman DE, Derian CK, Andrade-Gordon P, Plata-Saalaman CR, Ilyin SE (2004) J Biomol Screen 9:286. doi:10.1177/1087057104263533
71. Savchuk NP, Balakin KV, Tkachenko SE (2004) Curr Opin Chem Biol 8:412. doi:10.1016/j.cbpa.2004.06.003
72. Darvas F, Dorman G, Puskas LG, Bucsai A, Urge L (2004) Med Chem Res 13:643. doi:10.1007/s00044-004-0108-5
73. Whitty A, Kumaravel G (2006) Nat Chem Biol 2:112. doi:10.1038/nchembio0306-112
74. Hajduk PJ, Gomtsyan A, Didomenico S, Cowart M, Bayburt EK, Solomon L, Severin J, Smith R, Walter K, Holzman TF, Stewart A, McGaraughty S, Jarvis MF, Kowaluk EA, Fesik SW (2000) J Med Chem 43:4781. doi:10.1021/jm000373a
75. Richardson CM, Nunns CL, Williamson DS, Parratt MJ, Dokurno P, Howes R, Borgognoni J, Drysdale MJ, Finch H, Hubbard RE, Jackson PS, Kierstan P, Lentzen G, Moore JD, Murray JB, Simmonite H, Surgenor AE, Torrance CJ (2007) Bioorg Med Chem Lett 17:3880. doi:10.1016/j.bmcl.2007.04.110
76. Angehrn P, Buchmann S, Funk C, Goetschi E, Gmuender H, Hebeisen P, Kostrewa D, Link H, Luebbers T, Masciadri R, Nielsen JE, Reindl P, Ricklin F, Schmitt-Hoffmann A, Theil F-P (2004) J Med Chem 47:1487. doi:10.1021/jm0310232
77. Seierstad M, Breitenbucher JG (2008) J Med Chem 51:7327. doi:10.1021/jm800311k
78. Zhao H, Serby MD, Xin Z, Szczepankiewicz BG, Liu M, Kosogof C, Liu B, Nelson LTJ, Johnson EF, Wang S, Pederson T, Gum RJ, Clampit JE, Haasch DL, Abad-Zapatero C, Fry EH, Rondinone C, Trevillyan JM, Sham HL, Liu G (2006) J Med Chem 49:4455. doi:10.1021/jm060465l
79. Gopalsamy A, Shi M, Boschelli DH, Williamson R, Olland A, Hu Y, Krishnamurthy G, Han X, Arndt K, Guo B (2007) J Med Chem 50:5547. doi:10.1021/jm070851i

80. Ranganathan R, Lu K, Hunter T, Noel J (1997) Cell 89:875. doi:10.1016/S0092-8674(00)80273-1
81. Bayer E, Goettsch S, Mueller J, Griewel B, Guiberman E, Mayr L, Bayer P (2003) J Biol Chem 278:26183. doi:10.1074/jbc.M300721200
82. Zhang Y, Daum S, Wildemann D, Zhou XZ, Verdecia MA, Bowman ME, Lucke C, Hunter T, Lu K-P, Fischer G, Noel JP (2007) ACS Chem Biol 2:320. doi:10.1021/cb7000044
83. Ruppert J, Welch W, Jain AN (1997) Protein Sci 6:524
84. Sotriffer C, Klebe G (2002) Farmaco 57:243. doi:10.1016/S0014-827X(02)01211-9
85. Campbell SJ, Gold ND, Jackson RM, Westhead DR (2003) Curr Opin Struct Biol 13:389. doi:10.1016/S0959-440X(03)00075-7
86. Wang R, Fu Y, Lai L (1997) J Chem Inf Comput Sci 37:615. doi:10.1021/ci960169p
87. Oprea TI (2000) J Comput Aided Mol Des 14:251. doi:10.1023/A:1008130001697
88. Ertl P (2003) J Chem Inf Comput Sci 43:374. doi:10.1021/ci0255782
89. Wildman SA, Crippen GM (1999) J Chem Inf Comput Sci 39:868. doi:10.1021/ci9903071