Design and Characterization of Libraries of Molecular Fragments for Use in NMR Screening against Protein Targets

Nicolas Baurin,[#] Fareed Aboul-Ela, Xavier Barril, Ben Davis, Martin Drysdale, Brian Dymock, Harry Finch, Christophe Fromont, Christine Richardson, Heather Simmonite, and Roderick E. Hubbard*

Vernalis (R&D) Ltd., Granta Park, Abington, Cambridge CB1 6GB, U.K.

Received June 15, 2004

We have designed four generations of a low molecular weight fragment library for use in NMR-based screening against protein targets. The library initially contained 723 fragments which were selected manually from the Available Chemicals Directory. A series of in silico filters and property calculations were developed to automate the selection process, allowing a larger database of 1.79M available compounds to be searched for a further 357 compounds that were added to the library. A kinase binding pharmacophore was then derived to select 174 kinase-focused fragments. Finally, an additional 61 fragments were selected to increase the number of different pharmacophores represented within the library. All of the fragments added to the library passed quality checks to ensure they were suitable for the screening protocol, with appropriate solubility, purity, chemical stability, and unambiguous NMR spectrum. The successive generations of libraries have been characterized through analysis of structural properties (molecular weight, lipophilicity, polar surface area, number of rotatable bonds, and hydrogen-bonding potential) and by analyzing their pharmacophoric complexity. These calculations have been used to compare the fragment libraries with a drug-like reference set of compounds and a set of molecules that bind to protein active sites. In addition, an analysis of the overall results of screening the library against the ATP binding site of two protein targets (HSP90 and CDK2) reveals different patterns of fragment binding, demonstrating that the approach can find selective compounds that discriminate between related binding sites.

INTRODUCTION

There has been increasing interest over the past 10 years in the use of fragment based methods in drug discovery.[1−7] The central premise is that the majority of compounds that bind to proteins contain a central chemical scaffold that can bind weakly to a particular site on the protein target and that this scaffold is decorated with functional groups that enhance binding potency and selectivity of the compound for that particular target. The aim in fragment based screening is to identify scaffolds that bind at or near to a particular site on the target. This information can then be used to search for and/or design elaborated compounds that will bind with higher affinity and selectivity.

In this paper, we describe the methods used to generate four generations of fragment library to be used in screening against protein targets. The main constraints in the design of the library come from the requirements of the experimental method used to screen the fragments, the aim for the fragments to represent as diverse a set of pharmacophoric functionality as possible, and the need to ensure that the fragments are suitable to be evolved into lead compounds.

There are three main experimental methods that have been developed for identifying which fragments bind to a protein active site. For some targets, there are sufficiently robust enzyme or binding assays that will withstand the high concentrations of fragment (and associated solvent, often DMSO) that are needed to see inhibition, although in many cases this will place an upper limit on affinity of around 200 μM.[3,8] The second alternative is to use protein crystallography, where mixtures of fragments are soaked into crystals of the apo-protein and the identity of which fragment has bound is provided by interpretation of difference electron density maps. This method has the advantage of immediately providing details of the position and orientation of the bound fragment in the active site of the protein. However, there are constraints on the protein system to achieve success. The apo-protein must crystallize with an open active site, not occluded by other molecules from crystal packing. In addition, the binding of small molecular fragments to a protein in a crystal is controlled by the kinetics or mechanism of diffusion as well as by the thermodynamics of binding. However, for appropriate systems, high throughput crystallographic methods have been successful in identifying fragments.[4,5,9,10]

The third method used for fragment screening is NMR spectroscopy.[11−14] Changes in the NMR spectrum of either a putative ligand or of the target protein can be used to identify binding. Both approaches have been widely used and have characteristic advantages and disadvantages.[15] Ligand-detected NMR screening offers particular advantages when dealing with large receptors, since it is not necessary to be able to assign, or even observe, resonances arising from

* Corresponding author: r.hubbard@vernalis.com.
[#] Current address: Dept of Cheminformatics, Aventis Pharma, 13, Quai J. Guesde, F-94403 Vitry/Seine, France.

10.1021/ci049806z CCC: $27.50 © 2004 American Chemical Society
Published on Web 10/09/2004

the receptor. A series of ligand-detected NMR screening experiments such as STD[16] and water-LOGSY[17] have been developed whereby changes in the ligand spectrum result from binding of the ligand to the target protein. Because the protein is essentially silent in these experiments, it is possible to distinguish NMR signals from a number of ligands, allowing mixtures of compounds to be screened at a time. Finally, the experiments can be performed at concentrations to identify binding of fragments with affinities up to 10 mM, a suitable range for scaffolds to be elaborated into high affinity ligands. However, there are some limitations. The change in signal in the NMR experiments is not always associated with binding—a number of other factors can affect the appearance of signals in the absence of binding (such as self-association of the ligand) or result in an absence of signal when a ligand binds (for example, inappropriate exchange rates between free and bound ligand). In addition, the high concentration of ligand used (typically in the region of 2 mM) can lead to signals from nonspecific binding to parts of the protein remote from the active site. In general, this is not an issue for the other two methods. In screening by crystallography, the resulting difference electron density locates the fragment in the active site. For screening by assays, the kinetics of inhibition can be probed to ensure there is competitive binding. In NMR screening, a known ligand can be used as a competitor and a hit validated if the binding signal from the fragment disappears when the competitor ligand is added.

We have developed an in-house strategy for fragment based drug discovery called SeeDs (Selection of Experimentally Exploitable Drug startpoints). In summary, the term covers the selection of fragment libraries described here, the NMR screening protocol to identify fragments that bind competitively to a specific site on the target protein, the determination of the crystal structure of the fragment bound to the protein target active site, and the various approaches for evolving the fragments into hit compounds using substructure searching or the design of small chemical libraries for synthesis. As with any other screening technology, the design of the screening library is a key factor in the success of the overall strategy.

In this paper we describe the methods used to generate four generations of fragment library designed to be used in this SeeDs strategy. NMR based screening places four main constraints on the design of the library. First, the experiments require 2 mM aqueous solubility in the binding experiment and, for practical reasons, 200 mM solubility in DMSO for stock solutions. Second, the fragments should be chemically stable both as stored in library stocks and also in the experimental conditions for screening, which can take a number of days to complete. Third, the size of the library needs to balance the chemical diversity sampled by the fragments with the practical (and cost) considerations of maintaining and screening a very large number of fragments. The aim of the iterative design described here was to develop an overall library of some 1300 diverse fragments with the appropriate chemical stability and solubility properties for use in NMR screening. Finally, the fragments in the library should contain a range of chemical diversity and be suitable for expansion into larger compounds that can be taken forward as potential drug candidates.

METHODS

Compound Libraries. This study selected compounds from the 266 812 listed in the 2001 version of ACD[18] and from an in-house database called rCat of 1 622 763 unique chemical compounds assembled from 23 suppliers.[19]

Reference Set of Small Organic Drug-like Molecules. The World Drug Index[20] was searched for compounds with a United States Adopted Name (USAN) or International Nonproprietary Name (INN). This protocol[21] identifies the subset of molecules that have passed preclinical and Phase I evaluation. This list was further filtered on molecular weight ($250 \leq MW \leq 550$) and to remove compounds containing any atom different to O, N, C, H, Br, I, Cl, F, S, or P to produce a 1141 drug-like compound reference set.

Reference Set of Small Organic Crystallized Molecules. The protein databank[22] (20 025 entries as of March 4, 2003) was analyzed for nonpeptidic ligands that did not contain any atom different to O, N, C, H, Br, I, Cl, F, S, or P. This list was further filtered on molecular weight ($250 \leq MW \leq 550$) to produce a 1094 binding compounds reference set.

Predicting Aqueous Solubility. We have previously described the development of a linear model for predicting aqueous solubility of small, drug-like compounds.[19] In summary, the model was obtained by a cross-validated PLS regression algorithm that fitted 49 2D descriptors to predict values of solubility for a 3041 molecule training set[23] with a correlation coefficient of 0.82 and a mean absolute error of 0.70. A 21 molecule reference test set confirmed that the model is predictive to within 1 log unit of solubility, the usual error associated with experimental measurement.[24] The results section includes an analysis of solubility prediction for one of the library of fragments.

Filter A—Unwanted Functionality. 2D SMILES representations of molecules were filtered to remove compounds that contained any of the following functionality:[25−27]

- 4 aliphatic carbons except if also contains X−C−C−C−X, X−C−C−X, X−C−X with X=O or N.
- Any atom different from H, C, N, O, F, Cl, S
- -SH, S−S, O−O, S−Cl, N-halogen
- Sugars
- Conjugated system: R=C−C=O, with R different from O, N, or S or aromatic cycle
- (C=O)-halogen, O−(C=O)-halogen, SO_2-halogen, N=C=O, N=C=S, N−C(=S)-N
- Acyclic C(=O)-S, Acyclic C(=S)-O, Acyclic N=C=N
- Anhydride, aziridine, epoxide, ortho ester, nitroso
- Quaternary amines, methylene, isonitrile
- Acetals, thioacetal, N−C−O acetals
- Nitro group, >1 chlorine atom

Filter B—Wanted Functionality. 2D SMILES representations of molecules were filtered to remove compounds that did NOT have at least one ring of at least 5-atoms and did NOT have at least one of the functional groups listed in Table 1. This collection of functionalities was chosen as appropriate groups to allow rapid chemistry to be applied to the fragments for evolution of the compounds. In addition, the groups chosen will add to the polarity of the fragments and aid solubility.

2D 3-Points Pharmacophoric Fingerprinting. The molecular recognition process can be described in terms of a

Table 1: List of Wanted Functional Groups Used by Filter B

R−COOMe	R−COOH
R−NHMe, R−N(Me)$_2$	R−NH$_2$
R−CONHMe, R−CON(Me)$_2$	R−CONH$_2$
R−SO2NHMe, R−SO2N(Me)$_2$	R−SO$_2$NH$_2$
R−OMe	R−OH
R−SMe	

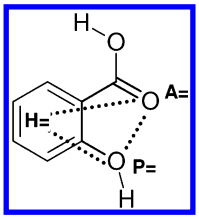

Figure 1. Pharmacophoric triangle detection. The dotted lines define a triangle in salicyclic acid that is composed of 3 pharmacophoric features: piHydrophobic (H=, centroid of the benzene ring), piAcceptor (A=, oxygen of carboxylic acid), and piPolar (P=, oxygen of hydroxyl), where the shortest path between each pair of features is 2 (A= to H=), 1 (P= to H=), and 4 (A= to P=) bonds.

simplified model of interaction where there is a match of localized recognition elements.[28] In this paper, these recognition elements are identified as pharmacophoric triangles and complexity is measured as the number of all possible pharmacophoric triangles in a molecule. The more complex a molecule, the more pharmacophoric triangles it contains.

The Molecular Operating Environment (MOE)[29] software includes an algorithm for identifying the presence of pharmacophoric triangles within compounds, as shown in Figure 1. We defined eight pharmacophoric features: H-bond donor (e.g. NH3), H-bond acceptor (e.g. O-), polar (e.g. OH), hydrophobic (e.g. CH3), pi donor (e.g. NH in pyridine), pi acceptor (e.g. C═O), pi polar (e.g. ═NH), and pi hydrophobic (e.g. aromatic carbon). The algorithm constructs an integer that encodes which features are at the apex of each triangle and also encodes bin numbers representing the number of bonds that make up the shortest path between pairs of features. The set of integers representing the pharmacophoric triangles can be assembled into a vector or modal fingerprint that encodes which features are present in a particular collection of molecules, where the fingerprint contains a particular triangle only once. For a single fragment, the length of this fingerprint gives an indication of the chemical complexity of the fragment. For collections of fragments, this fingerprint can be used to identify which features are present in particular libraries, and the length of the modal fingerprint is a measure of the diversity of the set of molecules.

The binning scheme was derived (Table 2) to optimize the performance of the fingerprinting algorithm in discriminating between sets of close and nonclose analogues derived by Makara.[30]

Clustering. We applied a modified version of the Stochastic Cluster Algorithm (SCA) developed by Reynolds et al.[31] This is a nonhierarchical method, relying on nearest neighbor lists, which speed scales linearly with the database size to cluster. When compared to other nonhierarchical clustering algorithms such as Jarvis-Patrick, the modified SCA algorithm is much faster, speed scaling N instead of N^2 (N is the number of molecules to cluster), and it shows a constant decrease in the number of clusters generated when the similarity threshold applied in the clustering decreases.

Table 2: Bond Binning Scheme Used To Encode the Pharmacophoric Triangles

bin	1	2	3	4	5	6	7	8
no. of bonds	1,2	3	4	5,6	7,8	9,10	11,12	>12

The modified SCA algorithm first applies a sorting by molecular weight of the molecules, which makes it possible to reproduce clustering results whatever the initial sorting of the database.

Quality Control (QC) Protocol. A stock solution of compound was made up to a concentration of 200 mM in d_6-DMSO. Aqueous NMR samples containing 2 mM compound (1% d_6-DMSO) were then made up in a buffer containing 50 mM potassium phosphate pH 7.5, 10% D_2O/90% H_2O. 1D NMR spectra (using excitation sculpting water suppression)[32] and waterLOGSY spectra[17] were acquired on each sample at 298 K on a Bruker DRX600 spectrometer. 1D NMR spectra were also acquired after 24 h in aqueous solution. Compounds failed the QC criteria if

- Not soluble to 2 mM in buffer solution
- NMR spectrum was inconsistent with structure (checked in d_6-DMSO if required)
- Compound was less than 95% pure
- Compound was not stable and soluble for 24 h in buffer solution
- Compound had a positive water LOGSY spectrum (taken to be indicative of self-association and which gives a false positive signal in the NMR screening experiments)
- Compound had no observable nonexchangeable protons
- The relaxation edited spectra showed unusual relaxation properties which could indicate self-association

Software. The Catalyst software[33] provided the pharmacophore searches used to select kinase focused fragments (SeeDs-3). Otherwise, MOE[29] along with Scripting Vector Language (SVL) scripts were used to develop and then apply all the in silico techniques presented in this article.

Screening of Fragment Libraries. The fragments were randomly combined into mixtures of 12 fragments for screening by NMR. For most mixtures, it was possible to deconvolute the spectra for the individual fragments. For the few mixtures where spectral overlap was an issue, fragments were moved to different mixtures. Also, fragments were removed which showed perturbation of spectra in the mixture as this indicates interaction between mixture components. Screening used STD and LOGSY experiments to identify binding of 500 μM mixtures of 12 fragments to samples of each of two proteins. Competitive binding was confirmed by loss of signal when a known inhibitor was added to the mixture. For HSP90, a 10 μM solution in 50 mM KPO_4 buffer, pH 7.5 of the N terminal domain of human HSP90α was used, prepared as described in Wright et al.[34] with PU3[35] as a competitor added at 100 μM. For CDK2, a 10 μM sample of the inactive kinase domain of human CDK2 expressed as a GST fusion (produced essentially as described as in Brown et al.[36]) was used with roscovitine[37] as a competitor added at 100 μM.

GENERATION OF THE LIBRARIES

SeeDs-1—The Initial Library. The ACD was searched for compounds supplied by the Aldrich (21 595) and Maybridge (65 538 from their general screening file) companies

who at that time were recognized as being able to rapidly supply compounds in the appropriate format and then filtered on molecular weight ($110 \leq MW \leq 250$, with an upper limit of 350 for compounds containing sulfonamide). The resulting compounds were filtered to remove molecules that contained metals, linear carbon chains longer than 5C or reactive groups to give 7545 molecules. These were visually inspected by a medicinal chemist to identify potentially insoluble fragments and to ensure that most fragments contained suitable functional groups to allow rapid chemical evolution of the fragment scaffold. 1078 fragments were ordered and 723 passed QC. The majority of the QC failures were due to instability, either in the aqueous conditions used in screening or at the high concentrations in DMSO used for storage.

This initial SeeDs library also allowed an assessment of the performance of the solubility prediction model. Of the 723 fragments that passed QC, 636 (88%) were correctly predicted to have solubility $\geq$ 2 mM. Of the 87 false negatives, 32 were predicted to have solubility between 500 and 2000 μM, 48 between 100 and 500 μM, and 7 to have solubility between 30 and 100 μM. Of the 95 molecules that failed QC on solubility, 84 (88%) were correctly predicted as insoluble at 2 mM. Of the remaining 11 false positives, 6 were predicted to have solubility between 4 and 10 mM and the remaining 5 to have solubility between 16 and 38 mM.

These results demonstrate that the solubility prediction model performs reasonably well on the fragments and provide confidence that the method allows very large numbers of compounds to be rapidly assessed with an acceptable error rate.

SeeDs-2—The Comprehensive Library. The process for generating the SeeDs-1 library required intense input from medicinal chemists. The criteria on which fragments were discarded was systematically analyzed, challenged, and discussed. From this, a series of functionality filters were derived that are described in the methods section. Filter A is a collection of criteria to identify unwanted features in compounds. Filter B identifies wanted chemical functionality that can be used for rapid chemical elaboration or evolution of the fragment. The major advantage of these filters is that they can be applied in silico to select appropriate fragments from very large libraries of compounds.

The aim of the calculations for the SeeDs-2 library was to use these in silico filters to identify additional fragments from a combination of the compounds in ACD and the in-house rCat database.[19] 93 416 of the 266 812 ACD compounds were already in rCat, so the total number of compounds considered was 1.79M. The compounds were first filtered on molecular weight ($100 \leq MW \leq 250$, with an upper limit of 350 for compounds containing sulfonamide) and then analyzed with both filter A and filter B. After removing any duplicates of the SeeDs-1 library, this resulted in 107 000 compounds, of which 43 458 were predicted to have solubility $\geq$ 2 mM. These were then clustered using 2D, 3-point pharmacophoric features, with the similarity threshold adjusted to provide about 3000 clusters. 2918 compounds (the centroid of each cluster) were selected for visual inspection by a panel of medicinal chemists to review the compounds for their chemical tractability, cost, and availability. From this analysis, 395 compounds were purchased, of which 357 remained after QC.

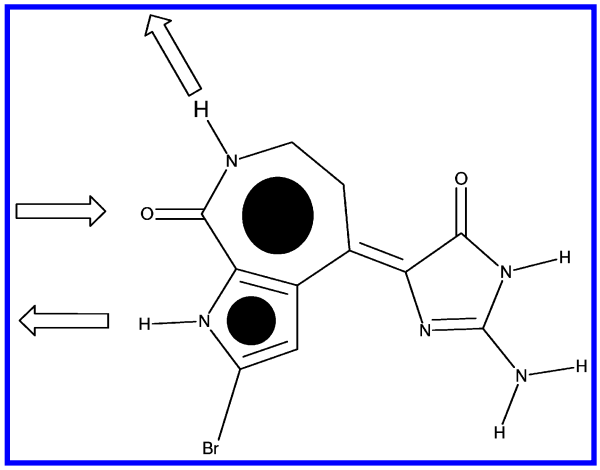

Figure 2. The structure of hymenialdisine. Arrows mark the location and direction of the polar pharmacophore points, while the ellipses mark the hydrophobic features.

SeeDs-3—The Kinase Focused Library. The aim of this library was to identify fragments in the 1.62M rCat database that may have a higher probability of binding to protein kinases. The design of a target-specific library of molecular fragments has been successfully applied to DNA gyrase[3] and, although our existing library (SeeDs-1 and -2) already covered the "kinase pharmacophoric space", we deemed important to increase the density of population of molecules satisfying the distinctive pattern of hydrogen bond donors and acceptors provided in the ATP recognition site of protein kinases.[38] The majority of known kinase inhibitors interact with this motif that consists of a backbone carbonyl of one residue and the backbone NH and carbonyl of the residue at the +2 position. Different inhibitors satisfy different combinations of this hydrogen bonding potential.

One aim was to select molecules with the potential for binding to a range of kinases, and so the principle constraint was to target these well conserved, backbone mediated interactions. The main filter used were a combination of pharmacophores built by considering the conformation of hymenialdisine bound to CDK2, as shown in Figure 2 (PDB code 1DM2[39]). Hymenialdisine was chosen as a template for a number of reasons. Crystallographic data are available to show that it binds to the region of interest and satisfies the donor and acceptor binding capacity of the binding site in a similar way to most kinase ligands. Also, hymenialdisine is relatively planar, like many of the known kinase ligands, allowing representative selection of hydrophobic query points.

In addition to the requirement for hydrogen bonding groups, the adenine binding cleft of the ATP site tends to favor moderately planar substrates. Each of the pharmacophore queries consisted of one donor, one acceptor, and one of the 2 possible hydrophobic points. Hydrogen bond donor and acceptor directionality was inferred from the crystal structure. Constraint tolerances of 1.5 Å were used for donor—acceptor distance and 2—2.5 Å to the hydrophobe. These loose tolerances were chosen to represent the flexibility seen across kinase structures and also to allow the pharmacophore queries to maximize the diversity of fragments found.

The results of all four pharmacophore queries were combined and filtered on physicochemical parameters (mo-

lecular weight 150−250, SlogP between 2.5 and −2.5) and the presence of suitable chemical handles for elaboration (filter B). This was followed by removal of the 37 compounds that were duplicates of SeeDs-1 or SeeDs-2 fragments to give 3529 compounds.

This set of compounds was then analyzed for unwanted functionality. The general filter, filter A was applied, with the additional requirement that compounds were discarded if they contained

- Any halogen
- Any carbocycle
- Any sulfur atom unless oxidized or in an aromatic ring
- More than 2 rings
- Any bridged systems
- Chains longer than two carbons
- 2-Hydroxypyridines/pyridones.

These additional filters came from analysis of the compounds discarded during the visual inspection stage of generating the SeeDs-2 library.

The surviving data set of 2042 compounds was analyzed for diversity using two methods. In the first, a Tanimoto similarity metric was applied to MACCS structural keys, resulting in a data set of 500 compounds, which was reduced to 383 after applying a 2 mM predicted solubility cutoff. In the second approach, diversity was analyzed using the 2D, 3-point pharmacophoric fingerprint method, resulting in 170 additional fragments. This combination of diversity analyses was used to increase the number of compounds being considered. These two sets of compounds were pooled and inspected visually for chemical tractability, availability, and cost. From this analysis, 204 compounds were purchased of which 174 passed QC tests.

SeeDs-4—The Incrementally Diverse Library. This final library adopted a modified approach to identify additional, diverse fragments. The 2D, 3-point pharmacophores from all the members of the SeeDs-1, SeeDs-2, and SeeDs-3 libraries were concatenated and used as a novelty screen for the 43 358 fragments that remained after the solubility filter selection described for the SeeDs-2 library. 3794 fragments were found to have at least one novel pharmacophoric triangle from which 717 distinctive, and unique fragments were identified after clustering at a 50% similarity threshold. This particular library was subjected to a more rigorous visual inspection where a panel of 7 chemists all had to agree on medicinal chemistry tractability of each compound before acceptance. From this, only 65 compounds were judged acceptable after visual inspection and 61 passed QC.

SUMMARY ANALYSES OF THE LIBRARIES

Number of Fragments Found. The overall aim was to generate a total fragment library of between 1200 and 1500 fragments to achieve a balance between achieving sufficient diversity of fragments, the tractability of the fragments for medicinal chemistry, the cost of purchase, the practicalities of library maintenance, and using the library in screening by NMR. The selection process evolved both between and while the different libraries were being designed, resulting in a number of inconsistencies. For example there was a varying strategy on fragments containing halogen atoms which add molecular weight and in general decrease solubility, both unfavorable properties for a fragment. Such fragments were explicitly excluded for SeeDs-1 and SeeDs-3, and only 8 fragments containing Cl and 3 containing F were selected for the other libraries. However, as with all the criteria applied, the aim is to identify small, core scaffolds which can then be evolved by library design or conventional medicinal chemistry.

Table 3: Summary of Number of Compounds Considered and Selected in Generating the Four SeeDs Libraries

	compounds initially considered	number visually inspected	number purchased (%)[a]	number passing QC (%)[b]
SeeDs-1	87133	7545	1078 (14%)	723 (67%)
SeeDs-2	1.79M	2918	395 (14%)	357 (90%)
SeeDs-3	1.62M	383	204 (53%)	174 (85%)
SeeDs-4	1.79M	717	65 (9%)	61 (94%)

[a] Number purchased/number inspected. [b] Number passing QC/number purchased.

Table 3 summarizes the number of compounds considered at the main stages in the selection process for each sublibrary and the number of fragments finally selected. The number of compounds considered for the SeeDs-1 library was limited as most of the analyses were performed manually. The rules developed for SeeDs-1 were encapsulated in series of in silico filters which allowed a much larger number of compounds to be considered for subsequent libraries. The improvement in the success rate for compounds in meeting QC criteria between SeeDs-1 and the other libraries is primarily due to the success of the solubility filter.

The attrition rate at the visual inspection stage was fairly constant for SeeDs-1, -2, and -4. Although a rather subjective selection process, the primary consideration was an assessment of chemical tractability and suitability to be used as the core skeleton for building into a lead or drug-like molecule. The higher percentage of compounds selected for SeeDs-3 mostly reflects that the compounds were selected to satisfy a kinase binding pharmacophore, which will focus the selection to compounds that will have the combination of functionality required to survive the visual inspection process.

A decreasing number of fragments were found that satisfied the in silico filters at each stage, such that for SeeDs-4, only 717 compounds out of 1.79M passed the property and diversity criteria. This indicates that we are close to exhausting the diversity space within the available compound catalog used. Further additions to the fragment library will only come from screening of other compound collections. Alternatively, the current library could be searched for missing chemical diversity (such as identifying regions of 3 point pharmacophore space that are particularly barren) and designing novel fragments for bespoke synthesis.

Physicochemical Characteristics. Table 4 summarizes the physicochemical properties of the individual and cumulative libraries. A number of trends can be identified.

The fragments in the later libraries (SeeDs-2, -3, and -4) show slightly higher molecular weight. This probably reflects that novel compounds could only be identified from the larger catalog of available compounds by selecting slightly larger molecules, and also that the compounds in the first library were not required to have wanted functionalities (filter B), allowing simpler compounds to be included. More marked is the lowering of SlogP, between SeeDs-1 and the other

Table 4: Physicochemical Properties of the Individual and Cumulative Libraries, Showing the First Quartile/Average/Last Quartile Values

libraries	no. of molecules	MWeight	SlogP	rotatable bonds	h-bond acceptors	h-bond donors	length of 2D 3-points fingerprint
SeeDs-1	723	162/187/210	0.25/0.91/1.63	1/2.2/3	2/3.0/4	0/0.9/1	7/17.2/20
SeeDs-2	357	152/186/218	−0.87/−0.08/0.87	1/2.2/3	2/3.2/4	1/1.6/2	10/21.3/28
SeeDs-3	174	177/199/223	−0.60/0.16/1.03	1/2.7/4	3/3.5/4	1/1.5/2	10/21.7/28
SeeDs-4	61	178/211/236	−0.46/0.31/1.26	1/2.6/4	2/2.8/4	1/1.6/2	10/23.3/34
SeeDs-1+2	1080	158/187/215	−0.13/0.59/1.47	1/2.2/3	2/3.0/4	1/1.2/2	8/18.6/24
SeeDs-1+2+3	1254	162/188/217	−0.23/0.53/1.40	1/2.3/3	2/3.1/4	1/1.2/2	9/19.0/25
SeeDs-1+2+3+4	1315	163/189/218	−0.25/0.52/1.39	1/2.3/3	2/3.1/4	1/1.2/2	9/19.2/25

Table 5: Comparison of the 2D 3-Point Pharmacophoric Fingerprint of the SeeDs Library Compared to that of the Drug-like and the Binding Reference Set, Considering Fingerprints that Are Found Containing Different Numbers of Bonds between Features[a]

A − length of the modal fingerprint
B − length of the drug-like reference set-shared modal fingerprint [absolute and (percentage)]
C − length of the binding reference set shared modal fingerprint [absolute and (percentage)]

		bond limit 1−3			bond limit 4−8			bond limit 9-			no bond limit		
libraries	no. of molecules	A	B	C	A	B	C	A	B	C	A	B	C
SeeDs-1	723	117	95 (53%)	106 (44%)	497	388 (36%)	447 (29%)	0	0 (0%)	0 (0%)	2286	1915 (23%)	2038 (18%)
SeeDs-1+2	1080	169	119 (67%)	139 (58%)	659	488 (46%)	557 (36%)	0	0 (0%)	0 (0%)	3214	2525 (31%)	2742 (24%)
SeeDs-1+2+3	1254	170	119 (67%)	139 (58%)	711	518 (48%)	596 (39%)	0	0 (0%)	0 (0%)	3468	2690 (33%)	2941 (25%)
SeeDs-1+2+3+4	1315	175	123 (69%)	142 (59%)	747	539 (50%)	623 (40%)	0	0 (0%)	0 (0%)	3693	2848 (35%)	3102 (27%)
drug-like reference	1141	178	178 (100%)	139 (58%)	1072	1072 (100%)	888 (58%)	220	220 (100%)	108 (36%)	8260	8260 (100%)	6679 (57%)
binding reference	1094	240	139 (79%)	240 (100%)	1544	888 (83%)	1544 (100%)	303	108 (49%)	303 (100%)	11672	6679 (81%)	11672 (100%)

[a] For a bond limit of 1−3, only pharmacophoric triangles where each of the features is separated by 3 or less bonds is considered. For a bond limit of 4−8 bonds, only features separated by between 4 and 8 bonds are considered. Three numbers A, B, and C are provided for each bond limit, reporting the length of the modal fingerprint for each library (A), and the length (and percentage) of the fingerprint or number of features shared between the library and the drug-like (B) and binding (C) reference set. For example, for a bond limit of 1−3 bonds, the 723 fragments in SeeDs-1 share 95 features with the drug-like reference set, which is 53% of the total of 178 features in the reference set.

libraries. This is presumably the impact of the solubility filter that was applied.

The number of rotatable bonds increases slightly with the later libraries. However, the overall flexibility of the fragments is relatively conserved across the different libraries. Most fragments have a cyclic, rigid core as expected, but the requirements of the functionality filters ensure most have one or two functional groups attached, which provide the flexibility.

There are small variations in the hydrogen bond donor and acceptor profiles for the different libraries. The increased number of acceptors in the SeeDs-3 library is understandable as the initial selection was to satisfy a kinase binding pharmacophore. The smaller number of donors in SeeDs-1 probably reflects the subjectivity of the visual inspection process—for subsequent libraries, the in silico functionality filters increased the functionality present. This improvement in functionality is also reflected in the increase in the length of 2D 3-point fingerprints, a reflection of the increase of pharmacophoric complexity of the fragments.

Pharmacophoric Characterization. The central purpose of the library is to provide fragments that will bind to the active site of any target and will also provide suitable starting points for drug discovery. In the following analysis, we compare the 2D 3-point pharmacophore properties of the SeeDs fragment libraries to two sets of compounds. The first is a drug-like reference set, containing compounds taken from the World Drug Index (see Methods). The second is a set of compounds that are present in the structures of proteins deposited in the protein databank, the binding compounds reference set.

We have characterized the SeeDs and reference libraries by analysis of the modal fingerprints derived from the 2D 3-point pharmacophoric triangles present in the sets of compounds (see Methods). The fragment library contains smaller molecules than either of the two reference libraries. Therefore, the modal fingerprints for the different libraries were calculated for various bond limit ranges, considering the pharmacophoric triangles that are from features separated by less than 3 bonds, between 4 and 8 bonds, and greater than 9 bonds. The results are shown in Table 5. The analysis focuses on how many of the pharmacophoric triangles present within the fragment library are the same as those present in the two reference libraries, which is a measure of the similarity of chemical diversity between the compound collections. For both analyses, the fragment library has a higher percentage of features in the 1−3 bond limit that are similar to the reference compounds than in the 4−8 bond limit and no 3 point pharmacophores that are made up solely of features separated by 9 or more bonds. This is expected because of the smaller molecular weight of the fragments. It can be seen that each subsequent SeeDs library has added fragments that increase the coverage of the pharmacophore space present in the reference sets of compounds. However, the analyses show that although the binding set of compounds contains more pharmacophoric features than the drug-like

Table 6: Pharmacophoric Characterization of the Four SeeDs Libraries

library	no. of molecules	average length of the 2D 3-points pharmacophoric fingerprint	no. of clusters at 80% similarity	mean occupancy (molecule per cluster)
SeeDs-1	723	17.2	617	1.18
SeeDs-2	357	21.3	345	1.04
SeeDs-3	174	21.7	167	1.04
SeeDs-4	61	23.3	61	1

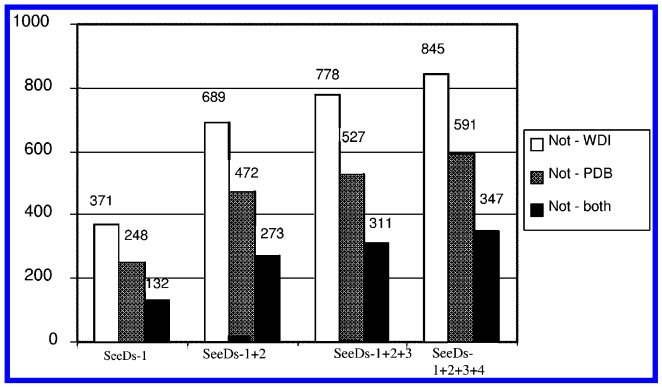

Figure 3. Number of pharmacophoric triangles for the different SeeDs libraries that are novel with respect to the drug-like reference set (Not-WDI), novel with respect to the binding reference set (Not-PDB), or novel with respect to both libaries (Not-both). The number of novel triangles was calculated from Table 5. For example, the SeeDs-1 library contains 2286 different pharmacophoric triangles, of which 1915 are present in the drug-like reference set, which gives 371 triangles that are novel to the SeeDs-1 library.

compounds, the fragment library has a higher percentage of the features present in drug-like compounds (35%) than present in the binding set of compounds (27%). However, the analyses show that the combined SeeDs1+2+3+4 library is still missing features that are found in known drugs or binding compounds.

An interesting side note from this analysis is that the binding reference set of compounds contains a large percentage (81%) of the features present in the drug-like reference set, whereas the drug-like reference set only covers 57% of the features present in the binding reference set. This disparity is seen across all lengths of fingerprints, so is not due to any difference in size of the compounds in the libraries. The difference probably reflects the increased number of nondrug-like features present in the binding set, which includes many carbohydrates and enzyme cofactors.

The diversity of the individual libraries was analyzed by clustering the fingerprints for individual molecules. As summarized in Table 6, the number of fragments per cluster is very similar for all the libraries, showing that the selection procedures are successful in adding fragments with distinctive features. However, the average length of the fingerprint increases with each successive library, reflecting an increasing complexity in the fragments as the criteria for selection were developed.

Fragment Novelty. Figure 3 analyzes how many of the pharmacophoric features in the various fragment libraries are novel compared to those found in the drug-like, the binding, and in both reference sets of compounds. This shows that the number of novel features found in the libraries increases with each generation of library. In addition, it shows that many of the features in the fragments that are novel compared to the drug-like reference set are found in the binding reference set and vice versa.

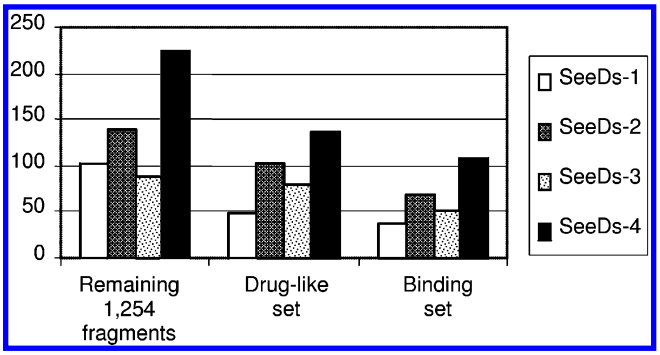

Figure 4. Average number of novel pharmacophoric triangles for 100 subsets of 61 fragments randomly selected from the four SeeDs libraries, compared to the triangles present in the remaining SeeDs or the two reference sets of compounds.

The SeeDs-4 library added 61 fragments designed specifically to maximize the number of novel pharmacophoric triangles in the SeeDs library. The success of this strategy was analyzed by randomly selecting 61 fragments from each of the SeeDs libraries (1, 2, and 3) and identifying how many novel triangles the selection contained compared to the triangles present in either the remaining 1254 fragments, in the drug-like reference set or in the binding reference set of compounds. The calculation was repeated for 100 different random selections of fragments to provide the average figures shown in Figure 4.

There is a common pattern to the novelty of the fragments from each of the four SeeDs libraries for each of the three comparisons made. With the exception of SeeDs-3, there is a general increase in novel content in progressing from SeeDs-1 to SeeDs-4, which is not surprising as new chemical diversity was an important design constraint for both SeeDs-2 and SeeDs-4. SeeDs-3 is the kinase targeted library, where the design focus was on close elaboration around a kinase binding motif. It is therefore not surprising that there is relatively less novelty in this library when compared to the other libraries. A striking feature of this analysis is that, on average, fragments from each of the four libraries contain novel pharmacophoric features compared to the remaining fragments.

The analysis also shows that the fragments provide novel features relative to the drug-like and binding reference sets of compounds. As shown in Table 5 above, the fragments have more similarity in pharmacophoric features to the drug-like (69% pharmacophore coverage) than to the binding (59%) reference set of compounds. However, on average, there are more novel triangles in the fragments compared to the drug-like reference set than compared to the binding set. This apparent discrepancy reflects the difference of diversity of the reference libraries—the World Drug Index contains less complex molecules (as shown by the length of the modal fingerprint), so it is easier to find novel triangles in the fragments when compared to the drug-like reference set.

Comparison to Other Published Fragment Libraries. We have compared the chemical substructures in our SeeDs libraries with the publicly available scaffolds from two groups. The Vertex group proposed 41 scaffolds for use in NMR screening,[2] and the Bajorath group identified 72 scaffolds[40] for construction of combinatorial libraries to target the ATP-binding site of proteins.

Out of the 1315 fragments in the four SeeDs libraries, 799 have at least one of the Vertex scaffolds and 240 at least one of Bajorath's scaffolds and together 906 (69%) of the SeeDs fragments have at least one scaffold in common with the Vertex or Bajorath fragments. The remaining 409 (31%) SeeDs do not contain any of the published scaffolds. As all the SeeDs scaffolds satisfy the same drug-like criteria, this demonstrates that there are additional, novel scaffolds identified through our approach.

Of the 41 Vertex scaffolds, 16 are present in the SeeDs libraries, and 12 are not present and pass the filter A and molecular weight criteria. The remaining 13 are not found in the SeeDs libraries but do not pass the filters, mainly because of the presence of fused rings of 3 or more rings and molecular weight greater than 250. Of the 72 Bajorath scaffolds, 25 are present in the SeeDs libraries, 13 are not present and pass all the filters, and 34 are not present but do not pass the filters. Again, most of these 34 scaffolds contain 3+ fused rings and are of too high molecular weight.

This analysis indicates that the methods identify the majority of the acceptable scaffolds identified by other groups and in addition a substantial number of novel, drug-like fragments.

There has also recently been a proposal[41] that the hits found in fragment screening satisfy a 'rule of three' where, on average the molecular weight is <300, the number of hydrogen bond donors is ≤ 3, the number of hydrogen bond acceptors is ≤ 3, and ClogP is ≤ 3. In addition, it was suggested that fragments should also have less than 3 rotatable bonds (NROT) and a polar surface area (PSA) of less than 60 Å^2. The combined SeeDs library of 1315 fragments has the following properties:

- 1303 (99%) have MW ≤ 300
- 1302 (99%) have ≤ 3 hydrogen bond donors
- 1184 (90%) have less than 3 hydrogen bond acceptors
- 1302 (99%) have SlogP (calculated with MOE[26]) ≤ 3
- 1062 (81%) have NROT ≤ 3
- 1046 (80%) have PSA ≤ 60 Å^2

In an ideal screening library, the properties of the hits found across a number of targets will be similar to the average of the whole library, otherwise the library would be covering nonbinding chemical space. The above analysis suggests that our library design selects "hit-like" fragments, which makes it more efficient. This is not surprising, as the molecular weight, functionality, and solubility filters will discriminate against fragments that do not meet the criteria. However, the 'rule of three' is a simple measure of the upper limit of some of the desirable properties but says nothing about the lower limit of these same properties or how they should be distributed. The additional criteria adopted in our selection process are the inclusion of more subjective, medicinal chemistry criteria so that the fragments provide a tractable basis for evolution into lead compounds.

Hit Rates in Fragment Screening. The first three SeeDs libraries have been screened for binding to two different proteins, HSP90 and CDK2, using a competitive NMR experiment. This typically identified fragments that bind to a protein active site with better than 5 mM affinity. Figure 5 shows the percentage of fragments from each of the libraries that has been classified as a hit. There will be differences in the sensitivity of the screening between the two proteins, due to changes in the buffer conditions, the thermodynamics and kinetics of binding of the competitor ligand, and the protein concentration. In addition, 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 binding does not necessarily give rise to a change in signal. It is therefore difficult to make detailed comparisons of the results obtained for different proteins. However, comparisons can be made of the relative hit rates achieved for the different libraries for a particular target.

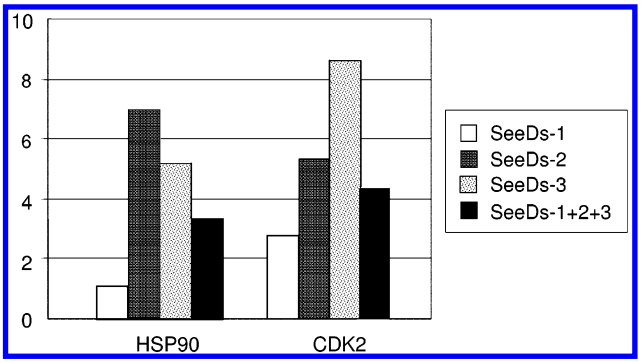

Figure 5. Percentage hit rate in NMR based screening of the different fragment libraries against two protein targets.

The results for CDK2 are the most straightforward to explain. The most successful library is SeeDs-3 which is the kinase library that was designed using a pharmacophore derived from a CDK2 inhibitor. There is a marked increase in the success rates for SeeDs-2 over SeeDs-1 that probably reflects the increase in quality and some aspect of the diversity of the library (see later). The number of hits for each library is quite different for HSP90, with much fewer hits for SeeDs-1 and a great increase in hits from SeeDs-2. This difference in hit rates between CDK2 and HSP90 reflects how, even though the fragments are quite small and they bind with quite low affinity, the pattern of binding is different between active sites. Although both proteins bind ATP, the HSP90 active site contains a number of conserved water molecules that mediate the interactions between the ligand and protein, a configuration that is not seen in any kinase. This is probably why the hit rate for the kinase library, SeeDs-3 is relatively low for HSP90 compared to CDK2.

The observed increase in success between SeeDs-1 and SeeDs-2 libraries for both proteins is not due to the chemical diversity constraints used in selecting fragments for SeeDs-2. As discussed above (see Table 6), the complexity of the fragment libraries is very similar. However, as shown in Figure 6, the fragments that are found as hits in the NMR screening are more complex than the average complexity found for the set of SeeDs libraries. This, presumably, increases the likelihood that such a small fragment will have an appropriate mix of functionality that will not only interact with the protein target but will also bind with sufficient affinity to be detected in the NMR screening experiment. As demonstrated in a recent paper by Hann et al.[28] there is a tradeoff between the probability of a molecule binding to a receptor (the simpler the molecule the better) and the probability that binding will be measurable (the more complex the ligand, the greater the affinity and, hence, the more likely to be measurable). By using NMR as the screening tool, we have shifted the limit of detection from

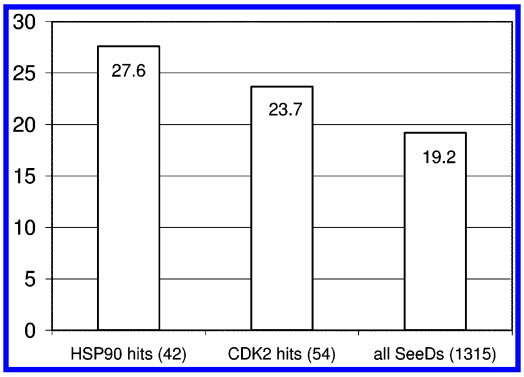

Figure 6. plot of the average length of the 2D 3-point fingerprint for the fragments that are identified as hits against HSP90, as hits against CDK2 and for all the fragments in the combined SeeDs libraries.

the micromolar to the millimolar range and, in consequence, the probability of finding molecules that bind has increased very significantly. The fact that the average hit is more complex than the average compound in the library seems to indicate that, for the two targets reported here, some compounds bind too weakly to be detected and that the library could yield slightly better hit rates by increasing its complexity. As seen in Figure 6, the level of complexity of the ligands necessary to detect binding is greater for HSP90 than for CDK2, explaining the lower hit rate. This suggests that the balance between complexity and detectable binding varies between these two targets. It is worth noting that the sublibraries designed using our in silico methods (SeeDs-2, -3, -4) appear to have an appropriate level of complexity (particularly for CDK2), whereas the manually selected SeeDs-1 library was perhaps too simple to yield the optimal hit rates. There are insufficient data for a robust analysis, but the results and chemical intuition suggest there will be a lower limit on the degree of complexity of useful fragments. For example, English et al.[43] showed that although discrete binding sites can be observed for high concentrations of simple fragments or organic solvents, the active site on a protein will accommodate a great diversity of different low affinity compounds at the concentrations required to see binding.

CONCLUDING REMARKS

We have described the design of a library of 1315 fragments suitable for use in NMR screening experiments to identify diverse chemical scaffolds that bind to protein targets. The design of any chemical library requires some measure of the chemical space to be covered and a method to select compounds based on diversity in order to populate the desired chemical space with similar density. To our knowledge, the design of libraries specifically for small molecule screening has been described in detail only twice before. An early example of screening for small molecules using enzymatic methods was described by Boehm et al.,[3] where a library was designed ad-hoc to maximize the probability of binding DNA gyrase. In contrast, with the exception of SeeDs-3, our libraries were designed to be used against a wide range of targets, and all the fragments were chosen to be suitable for use in NMR screening. In this regard, our libraries bear more resemblance to previously documented library design exercises for NMR-based screening.[2,17] There are many similarities between the selection criteria applied by the Novartis group[17] and those described in this paper. There are subtle differences in the physicochemical and diversity filters applied, but the main contrast between the two approaches is our intensive use of medicinal chemistry scrutiny in the final choice of fragments to be purchased. There is more of a contrast between our approach and the so-called SHAPES strategy[2] of the Vertex group that focused on the use of chemical scaffolds frequently found in commercial drugs. This approach is based on the assumption that molecular fragments identified in this way as well as its derivatives will resemble known drugs and are more likely to have desirable physicochemical and biological properties. The drawback of this method, as already reported in the literature,[42] is that novelty is not addressed (i.e. it does not cover the scaffolds of yet undiscovered drugs) and patentability might be an issue because the library is limited to fragments with preexisting intellectual property. By not restricting the library to a limited set of molecular cores our library incorporates much more diversity, and we have placed particular importance on filtering strategies to build in chemical tractability and drug-likeness.

The main requirement on the fragments of the library for use in NMR screening is high solubility and appropriate NMR spectra, and thus the entire library is also suitable for screening using standard binding or enzyme assay methods. Additional selection of appropriate sets of compounds would be required using shape based criteria to configure the library for use in crystal soaking based fragment screening.

The classification of compounds on chemical tractability is necessarily subjective. Most medicinal chemists agree that certain chemical features will confer reactivity or toxicity in any compound they are found in. In addition, it is possible to use rule-based drug-like filters such as the Lipinski[21] and Veber rules[44] in combination with more specific ADMET models (such as aqueous solubility, Caco-2 membrane permeation, toxicity, or metabolism models)[19] to provide fast in silico tools to delimit drug-like property space. However, it is more difficult to systematically analyze molecules for their synthetic tractability and suitability to be taken forward for hit or lead optimization. For these reasons, the visual inspection stage of selecting fragments for inclusion in the libraries is the most subjective part of the library generation process.[45]

ACKNOWLEDGMENT

We would like to thank Dr. Dimitri Bondarev, from the Chemical Computing Group, Montreal, Canada, for in depth discussions and close support in the tailoring of the ChemComp 2D 3-Points pharmacophoric fingerprint algorithm. David Morley, now of Enspiral Discovery, provided extensive support in writing, maintaining, and debugging scripting and databasing applications. We are also very grateful to the many other medicinal chemists at Vernalis who reviewed the selection of fragments, in particular Justin Bower, Martin Parratt, and Andy Potter.

REFERENCES AND NOTES

(1) Shuker, S. B.; Hajduk, P. J.; Meadows, R. P.; Fesik, S. W. Discovering high affinity ligands for proteins: SAR by NMR. *Science* **1996**, *274*, 1531−1534.

(2) Fejzo, J.; Lepre, C. A.; Peng, J. W.; Bemis, G. W.; Ajay; Murcko, M. A.; Moore, J. M. The SHAPES strategy: an NMR-based approach for lead generation in drug discovery. *Chem Biol*. **1999**, *10*, 755−69.
(3) Boehm, H. J.; Boehringer, M.; Bur, D.; Gmuender, H.; Huber, W.; Klaus, W.; Kostrewa, D.; Kuehne, H.; Luebbers, T.; Meunier-Keller, N.; Mueller, F. Novel inhibitors of DNA gyrase: 3D structure based biased needle screening, hit validation by biophysical methods, and 3D guided optimization. A promising alternative to random screening. *J. Med. Chem*. **2000**, *14*, 2664−2674.
(4) Nienaber, V. L.; Richardson, P. L.; Klighofer, V.; Bouska, J. J.; Giranda, V. L.; Greer, J. Discovering novel ligands for macromolecules using X-ray crystallographic screening. *Nat. Biotechnol*. **2000**, *10*, 1105−1108.
(5) Sharff, A.; Jhoti, H. High Throughput crystallography to enhance drug discovery. *Curr. Opin. Chem. Biol.* **2003**, *7*, 340−345.
(6) Erlanson, D. A.; McDowell, R. S.; O'Brien, T. Fragment-Based Drug Discovery. *J. Med. Chem.* **2004**, 3463−3482.
(7) Rees, D. C.; Congreve, M.; Murray, C. W.; Carr, R. Fragment-Based Lead Discovery. *Nature Rev. Drug Discovery* **2004**, *3*, 660−672.
(8) Maly, D. J.; Choong, I. C.; Ellman, J. A. Combinatorial target-guided ligand assembly: Identification of potent subtype-selective c-Src inhibitors. *PNAS* **2000**, *97*, 2419−2424.
(9) Lesuisse, D.; Lange, G.; Deprez, P.; Benard, D.; Schoot, B.; Delettre, G.; Marquette, J. P.; Broto, P.; Jean-Baptiste, V.; Bichet, P.; Sarubbi, E.; Mandine, E. SAR and X-ray. A new approach combining fragment-based screening and rational drug design: application to the discovery of nanomolar inhibitors of Src SH2. *J. Med. Chem*. **2002**, *45*, 2379−2387.
(10) Sanders, W. J.; Nienaber, V. L.; Lerner, C. G.; McCall, J. O.; Merrick, S. M.; Swanson, S. J.; Harlan, J. E.; Stoll, V. S.; Stamper, G. F.; Betz, S. F.; Condroski, K. R.; Meadows, R. P.; Severin, J. M.; Walter, K. A.; Magdalinos, P.; Jakob, C. G.; Wagner, R.; Beutel, B. A. Discovery of potent inhibitors of dihydroneptin aldolase using CrystaLEAD high-throughput crystallographic screening and structure-directed lead optimization. *J. Med. Chem*. **47**, 1709−1718.
(11) Pellecchia, M.; Sem, D. S.; Wuthrich, K. NMR in drug discovery. *Nat. Rev. Drug Discovery* **2002** *1*, 211−219.
(12) Coles, M.; Heller, M.; Kessler, H. NMR-based screening technologies. *Drug Discovery Today* **2003**, *8*, 803−810.
(13) Fejzo, J.; Lepre, C.; Xialoling, X. Application of NMR screening in Drug Discovery. *Curr. Top. Med. Chem.* **2003**, *3*, 81−97.
(14) Jacoby, E.; Davies, J.; Blommers, M. J. Design of small molecule libraries for NMR screening and other applications in drug discovery. *Curr. Top. Med. Chem*. **2003**, *3*, 11−23.
(15) Meyer, B.; Peters, T. NMR spectroscopic techniques for screening and identifying ligand binding to protein receptors. *Angew. Chem. Int. Ed.* **2003**, *42*, 864−890.
(16) Mayer, M.; Meyer, B. Characterization of Ligand Binding by Saturation Transfer Difference NMR Spectra. *Angew. Chem. Int. Ed*. **1999**, *35*, 1784−1788.
(17) Dalvit, C.; Pevarello, P.; Tato, M.; Veronesi, M.; Vulpetti, A.; Sundstrom, M. Identification of compounds with binding affinity to proteins via magnetization transfer from bulk water. *J. Biomol. NMR* **2000**, *18*, 65−68.
(18) MDL Information Systems, Inc., 14600 Catalina Street, San Leandro, CA 94577, U.S.A. http://www.mdl.com.
(19) Baurin, N.; Baker, R.; Richardson, C.; Chen, I.; Foloppe, N.; Potter, A.; Jordan, A.; Roughley, S.; Parratt, M.; Greaney, P.; Morley, D.; Hubbard, R. E. Drug-like annotation and duplicate analysis of a 23-supplier chemical database totalling 2.7 million compounds. *J. Chem. Inf. Comput. Sci*. **2004**, *44*, 643−651.
(20) Thomson Scientific, 3501 Market Street, Philadelphia, PA 19104, U.S.A., http://thomsonderwent.com/products/lr/wdi/.
(21) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev*. **1997**, *23*, 3−25.
(22) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(23) Huuskonen, J.; Rantanen, J.; Livingstone, D. Prediction of aqueous solubility for a diverse set of organic compounds based on atom-type electrotopological state indices. *Eur. J. Med. Chem*. **2000**, *35*, 1081−1088.
(24) Myrdal, P. B.; Manka, A. M.; Yalkowsky, S. H. AQUADAC 3: Aqueous Functional Group Activity Coefficients: Application to the estimation of aqueous solubility. *Chemosphere* **1995**, *30*, 1619−1637.
(25) Muegge, I.; Heald, S. L.; Brittelli, D. Simple selection Criteria for Drug like-like Chemical Matter. *J. Med. Chem*. **2001**, *44*, 1841−1846.
(26) Bemis, G. W.; Murcko, M. A. The properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem*. **1996**, *39*, 2887−2893.
(27) Bemis, G. W.; Murcko, M. A. The properties of Known Drugs. 2. Side Chains. *J. Med. Chem*. **1999**, *42*, 5095−5099.
(28) Hann, M. M.; Leach, A. R.; Harper G. Molecular complexity and its impact on the probability of finding leads for drug discovery. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 856−864.
(29) Chemical Computing Group Inc., Montreal, H3A 2R7 Canada, http://www.chemcomp.com.
(30) Makara, G. M. Measuring molecular similarity and diversity: total pharmacophore diversity. *J. Med. Chem*. **2001**, *44*, 3563−3571.
(31) Reynolds, C. H.; Druker, R.; Pfahler, L. B. Lead Discovery Using Stochastic Cluster Analysis (SCA): A New Method for Clustering Structurally Similar Compounds. *J. Chem. Inf. Comput. Sci.* **1998**; *38*, 305−312.
(32) Hwang, T. L.; Shaka, A. J. Multiple-pulse mixing sequences that selectively enhance chemical exchange or cross-relaxation peaks in high-resolution NMR spectra. *J. Magn. Reson*. **1998**, *135*, 280−287.
(33) Accelrys Inc. 9685 Scranton Road. San Diego, CA 92121-3752, U.S.A. http://www.accelrys.com.
(34) Wright, L.; Barril, X.; Dymock, B.; Sheridan, L.; Surgenor, A.; Beswick, M.; Drysdale, M.; Collier, A.; Massey, A.; Davies, N.; Fink, A.; Fromont, C.; Aherne, W.; Boxall, K.; Sharp, S.; Workman, P.; Hubbard, R. E. Structure−Activity Relationships in purine-based inhibitor binding to HSP90 isoforms. *Chem. Biol.* **2004**, in press.
(35) Chiosis, G.; Timaul, M. N.; Lucas, B.; Munster, P. N.; Zheng, F. F.; Sepp-Lorenzino, L.; Rosen, N. A small molecule designed to bind to the adenine nucleotide pocket of HSP90 causes Her2 degradation and the growth arrest and differentiation of breast cancer cells. *Chem. Biol.* **2001**, *8*, 289−299.
(36) Brown, N. R.; Noble, M. E. M.; Lawrie, A. M.; Morris, M. C.; Tunnah, P.; Divita, G.; Johnson, L. N.; Endicott, J. A. Effects of phosphorylation of threonine 160 on cyclin-dependent kinase 2 structure and activity. *J. Biol. Chem*. **1999**, *274*, 8746−8756.
(37) De Azevedo, W. F.; Leclerc, S.; Meijer, L.; Havlicek, L.; Strnad, M.; Kim, S. H. Inhibition of cyclin-dependent kinases by purine analogues: crystal structure of human cdk2 complexed with roscovitine. *Eur. J. Biochem.* **1997**, *243*, 518−526.
(38) Furet, P.; Meyer, T.; Strauss, A.; Raccuglia, S.; Rondeau, J.-M. Structure-Based Design and Protein X-ray Analysis of a Protein Kinase Inhibitor. *Bioorg. Med. Chem. Lett.* **2002**, *12*, 221−224.
(39) Meijer, L.; Thunnissen, A. M.; White, A. W.; Garnier, M.; Nikolic, M.; Tsai, L. H.; Walter, J.; Cleverley, K. E.; Salinas, P. C.; Wu, Y. Z.; Biernat, J.; Mandelkow, E. M.; Kim, S. H.; Pettit, G. R. Inhibition of cyclin-dependent kinases, GSK-3beta and CK1 by hymenialdisine, a marine sponge constituent. *Chem. Biol.* **2000**, *7*, 51−63.
(40) Stahura, F. L.; Xue, L.; Godden, J. W.; Bajorath, J. Molecular scaffold-based design and comparison of combinatorial libraries focused on the ATP-binding site of protein kinases. *J. Mol. Graph Model*. **1999**, *19*, 1−9.
(41) Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A. A 'rule of three' for fragment-based lead discovery? *Drug Discovery Today* **2003**, *8*, 876−877.
(42) Lepre, C. A. Library design for NMR-based screening. *Drug Discovery Today* **2001**, *6*, 133−140.
(43) English, A. C.; Groom, C. R.; Hubbard, R. E. Experimental and computational mapping of the binding surface of a crystalline protein. *Prot. Eng*. **2001**, *14*, 47−59.
(44) Veber, D. F.; Johnson, S. R.; Cheng, H.-Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem*. **2002**, *45*, 2615−2623.
(45) Takaoka, Y.; Endo, Y.; Yamanobe, S.; Kakinuma, H.; Okubo, T.; Shimazaki, Y.; Ota, T.; Sumiya, S.; Yoshikawa, K. Development of a Method for Evaluating Drug-Likeness and Ease of Synthesis Using a Data Set in Which Compounds Are Assigned Scores Based on Chemists' Intuition. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1269−1275.

CI049806Z