On Sampling of Fragment Space

Gergely M. Makara[†]

Merck & Co., Merck Research Laboratories, RY80Y-325, 126 East Lincoln Avenue, Rahway, New Jersey 07065

Received January 9, 2007

Fragment-based lead discovery has over the years matured into an attractive alternative to high-throughput screening (HTS) for lead generation. Several techniques for screening libraries of typically 10^3-10^4 fragments have been reported. In this work, the practical success rates that can be expected from the screening of fragment-like libraries was investigated via interrogating medicinal chemistry databases for several programs with virtual libraries created from commercially available reagents or with libraries of commercially available fragments. The results suggest that hits more potent than typically discovered in today's fragment-based screens can consistently be identified from realistically accessible compound sets under screening conditions similar to commonly used HTS protocols.

Introduction

High-throughput screening (HTS) has been the primary lead discovery paradigm in the past decade. In a typical process, 0.5−2M druglike molecules are used to interrogate targets. The screening compound deck is often prefiltered to eliminate metals and other offenders, and some HTS decks are tailored to conform to user-selected properties such as the rule-of-five. Alternative lead generation techniques include focused screening of libraries made for related target families, literature mining, or recently fragment-based lead discovery. The latter encompasses the screen of a few thousand fragment-like (MW < 300 Da) molecules that may have a higher likelihood of fitting in binding sites as shown by the elegant study of Hann et al.[1] Since small fragments often exhibit binding in affinity ranges that challenge the limits of traditional biochemical assays, NMR and X-ray crystallography are by far the most often applied detection methods in fragment-based lead discovery. As the field matured, the rule-of-three for screening decks[2] and ligand efficiency[3] for weakly active hits have been proposed as guiding principles. The former helps weed out undesirable inputs, while the latter provides a simple ranking mechanism of hits based on the relative free energy of binding per heavy atom.

It has been shown that marketed drugs are more often than not very similar to the leads they were derived from.[4] Thus, both the quality and the quantity of the lead classes available to medicinal chemists are primary drivers for discovering best-in-class medicines. HTS has been quite effective to deliver a few major lead classes for many programs, but its limitations in sampling of chemical space and scalability have been discussed.[5,6] Fragment-based methods have demonstrated value for a number of programs,[7,8] but method development and lead maturation can often become bottlenecks. In addition, known limitations including protein size, false positives, false negatives,[9] and ease of crystallization can hinder its use in general applications against a broad range of therapeutically relevant targets, including G-protein coupled receptors (GPCRs).

As blind screening remains a paramount method for orphan or hard validated targets, a question can be raised: what can be expected from general screening of fragment-like molecules? Can better sampling of chemical space (that is, finding better matches) compensate for the smaller molecular surface area as measured by the potency of primary screening hits? The chemical space of less complex molecules is exponentially smaller than that of druglike molecules.[6] Thus, good sampling in fragment space may be accomplished with a relatively small number of molecules. However, typical fragment libraries used to date consist of only a few thousand molecules[8] for two main reasons. First, enough weak hits are identified even from 1000 fragments, and second, the evolution of these hits to high-quality leads is costly and time-consuming. Typical primary hits in fragment screens possess biochemical activity of millimolar to high micromolar range with rare exceptions in the 50−200 μM range.[10] To our knowledge, no study has been published on linking the potency of primary fragment hits to sampling of fragment space. It can be argued that fragments that capture interactions with the critical residues in a binding pocket may possess potency profiles more similar to HTS hits, which could greatly enhance our ability to rank different fragment-like lead series and focus hit maturation efforts on the most promising classes. In addition, these advanced fragment hits are likely to be more easily progressible than traditional fragment hits, as the key pharmacophore is already incorporated into the hit molecule.

This study was carried out to learn about the usability of fragments in traditional screens and to establish the relationship between the sampling rate and the potency range of hits for fragment libraries. In addition, the performance of libraries "synthesized" with different reagent classes was also compared to shed light on the value of reagent diversity in improving the success rate of blind screening against diverse target families.

Methods

The wealth of available SAR in medicinal chemistry databases can be used to recapitulate lead discovery for these drug discovery targets using virtual screening techniques. (Recapitulation is a process that attempts to find known hits or molecules highly similar to known hits from independently assembled compound sources.) Actives synthesized as part of an *advanced* development program of a major lead series to clinical candidate(s) are typically selected for synthesis on the basis of some medicinal chemistry rationale and tend to be largely independent of what reagents are commercially available. This important principle makes it possible to use the collection of actives in such databases as an unbiased bait set (a set of active structures, where each structure is used as a query, respectively) to query unrelated databases of molecules that are created solely

[†] Phone: 732-594-3053. Fax: 732-594-2130. E-mail: gergely_makara@merck.com.

10.1021/jm0700316 CCC: $37.00
Published on Web 06/14/2007

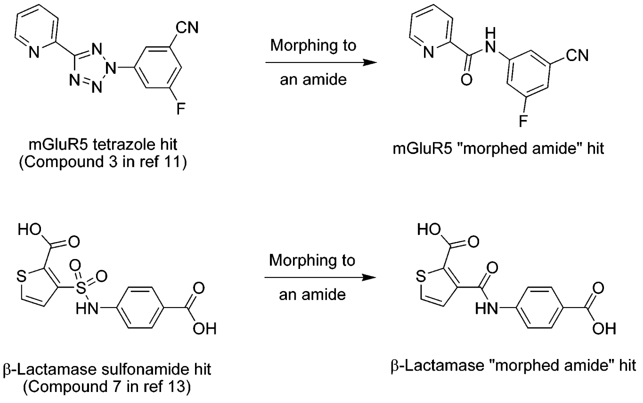


Figure 1. Illustration of the “morphing” process.

on the premise of availability. These virtual fragment libraries accessible by virtual synthesis using commercial building blocks, or commercially available fragments (available chemical compounds, ACC) comprised the virtual primary screening decks in this study. These screening decks were queried with fragment-like actives belonging to major lead classes from development programs, respectively. A screening deck or a subset thereof that yields hits similar to actives for a number of different discovery programs may form a useful basis set for the interrogation of other unrelated targets as well. To this extent, we have investigated the likelihood of success of finding small molecule hits that would show the way to major lead series for six unrelated therapeutic targets, including both enzymes and GPCRs.

An active subset was also created for two of the six therapeutic targets (DPP-4 and mGluR5), where a large number of fragment-like actives was available, to establish the relationship between the number of baits and the number of similar molecules (hits) found in various database subsets.

The diversity and size of off-the-shelf reagents for different chemical classes vary greatly. Inarguably amines and acids constitute the most accessible classes with thousands of chemicals commercially obtainable from the public domain. To study the effect of reagent diversity, hit rates obtained by the virtual screens of libraries enumerated around amides and amide isostere motifs, respectively, were also compared. Examples for such bioisostere pairs include amides vs triazoles, amides vs tetrazoles, and less frequently amides vs sulfonamides. Thus, a comparative study for three targets was carried out: mGLuR5 antagonists,[11,12] inhibitors for an in-house dehydrogenase program (dehydrogenase-1), and recently published reversible β-lactamase inhibitors.[13] To achieve that, for the first two targets both fragment-like heterocyclic actives and amide actives were extracted from our corporate medicinal chemistry database, respectively. For β-lactamase, both sulfonamides and amides were taken from the literature.[13,14] To gain further insight into the role of available reagents, for all three cases a third “active set” was created by artificially altering the central heterocyclic moieties to the corresponding secondary amides (“morphed amides”, Figure 1) in order to look at more specifically *only* the effect of reagent size and diversity on likelihood of success. The hit rate in a virtual amide library derived from using these morphed amide actives as baits was compared to that in the corresponding virtual heterocyclic library using the actual heterocyclic actives as baits.

There are many medicinal chemistry programs that are initiated with druglike leads and no fragment-like molecules as actives are synthesized for the major lead chemotype. It is, therefore, important whether an extrapolation can be made to programs that do not have fragment-like actives for the major lead series (that is, programs that cannot be studied by the virtual recapitulation technique used herein). To probe this question, two major chemical classes currently being pursued in house for a GPCR agonist program (GPCR3 agonist) were subjected to recapitulation via virtual screening. For both classes a key pharmacophore element (~140 Da size) was chosen and a substructure search in the commercial fragment database (ACC) using the key pharmacophore was carried out to select compounds for biological evaluation, respectively. Fragments for one of the two classes showed full agonist behavior in a cAMP assay. These fragments contained an amide motif and served as baits for virtual recapitulation for this program.

Last, a few fragment-like molecules that encompass the key β-amino acid amide motif in sitagliptin (**1**) were synthesized for dipeptidyl peptidase IV (DPP-4).[15] Surprisingly, no small molecule *cluster* leads containing this simple but critical motif have been discovered in HTS campaigns at Merck or reported by others. The question can be raised whether the lack of straightforward leads for this class of DPP-4 inhibitors is due to the sampling problem in druglike chemical space. To this extent, a simple SAR series (**2**−**6**, Scheme 1, Table 1) was purchased and synthesized to get an idea on the range of potency

Scheme 1. DPP-4 Inhibitors

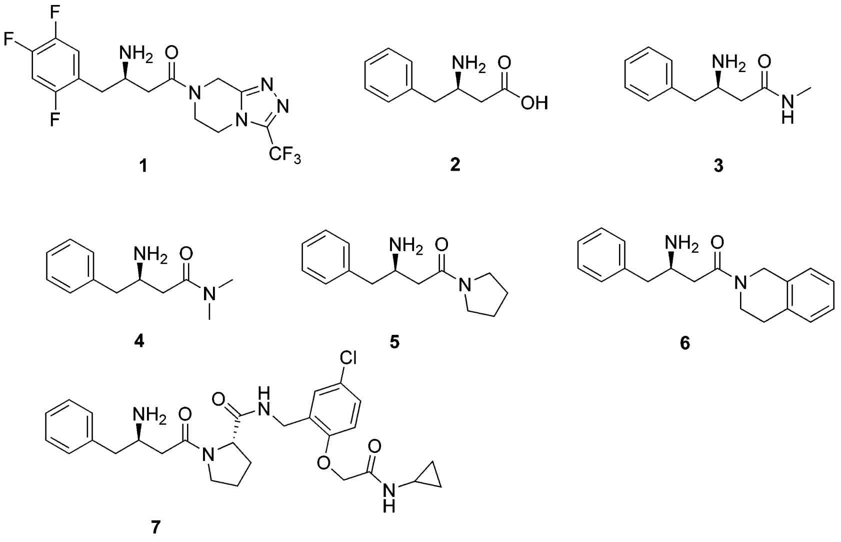


Table 1. Inhibitory Activities (μM) and Properties of Fragments Evaluated for DPP-4 As Compared to HTS Hit **7**

entry	MW	LE[a]	DPP-4	DPP-8	DPP-9	QPP	PEP	FAP
2	179	0.35	450	ND	ND	ND	ND	ND
3	192	0.48	13	460	>1000	680	>1000	>1000
4	206	0.50	3	ND	ND	ND	ND	ND
5	232	0.48	1	ND	ND	ND	ND	ND
6	294	0.40	0.4	ND	ND	ND	ND	ND
7	513	0.22	1.9	>10	>100	66	ND	ND

[a] LE: ligand efficiency.

for the minimum pharmacophore. This pharmacophore is known to possess a reverse binding mode as compared to α-amino acid substrate analogs.[15] Actives from this limited set (**3**−**6**) were used in queries of commercial and virtual databases. Compound **3**, which appeared to comprise the combination of minimum pharmacophore elements, was also subjected to counterscreening against a panel of related enzymes, which has previously been used to quantify the off-target liability for DPP-4 inhibitors.[15]

In general, the fragment-like baits from the in-house databases had potency <50 μM, and for most sets it was <5 μM (Tables 2 and 3). Ligand efficiencies for all but six bait actives across the six therapeutic targets were higher than 0.3 with an average and median of ∼0.45. Similarity searches were carried out using a cutoff of 0.75 Tanimoto (Tc) to distinguish prospective actives from inactives using FCFP_4 fingerprints.[16] This threshold is in accord with the recent recommendation of Tc values of 0.7 and 0.8 being optimal in similarity searches for different fingerprint types, respectively.[17] It has been shown that 2D fingerprints are unreliable descriptors of molecular shape and the three-dimensional distribution of pharmacophore points and are prone to select for similar molecular frameworks and connectivity.[18,19] This artifact, however, was thought to be advantageous for the purpose of finding compounds that both *look and act* in biological settings like the bait molecules.

Results

Fragment Libraries for Lead Series Containing Amide Bonds. Results for the virtual recapitulation of hit discovery from de novo amide libraries for major lead series developed for six therapeutic targets are depicted in Table 2. A total of 2.1 million of the 6.2 million amides in the molecular weight range of 160−300 Da passed the filters for fragment-like properties. A 25−35% pass rate using these filters, which include calculated physicochemical properties as well as unwanted functionalities, in our experience is fairly typical once the molecular weight boundaries are met. The results of the virtual screenings suggest that a sufficient number of similar compounds (hits) can be identified in the full library through the 50K subset to give confidence in our ability to find starting hits for all programs except for HIV integrase. Importantly, the hit rate distribution obtained from the 200K and 50K virtual amide libraries is more uniform than that given by ACC.

Relative Hit Rates. The number of similar molecules found is nearly proportional to the number of baits for DPP-4 and for the two largest library sets for mGluR5 (Figure 2; in the latter case, the variance for the small library sets was very high due to the low numerical values). This finding is a very important point considering that surely not all active fragment-like molecules present in the fragment universe have been synthesized and assayed for any of the six programs studied. Thus, our bait sets must be incomplete, meaning that the results depicted in Tables 2 and 3 are indicative of the upper limit of required library size but not the actual number necessary. Extrapolating from the values for the full DPP-4 bait column, one would conclude that this upper limit of the required virtual amide library could be 20 000 or even less. This means that ∼1% of the filtered fragment library (or 0.3% of the 6.2M small molecule library) can represent the full commercially available amide fragment space (as made by a single amide bond formation step) for the purpose of finding actives with the biochemical activity cutoffs shown in Table 2. In addition, applying a universal 50 μM cutoff for actives would surely increase the number of true fragment-like actives and thereby decrease the required library size even further. Figure 2 also reveals a saturation phenomenon in relative hit rates when the number of baits gets high compared to the size of the queried database: for mGLuR5 the relative hit rate obtained using 250 bait molecules drops sharply at the full library level (Table 4, Figure 2). The latter phenomenon makes a lot of sense, and a similar but less profound trend can be seen for DPP-4.

Amides versus Other Lead Motifs. It is evident from Table 3 that reagent availability and diversity play an important role and can be considered determining factors in the success rate of discovery fragment libraries. The virtual amide library in general enables hit identification for both the “morphed” amide and “real” amide bait sets as well as for the one amide derivative of β-lactamase sulfonamides. To exemplify the nature of similar hits in our virtual screen, the best matching compounds for the β-lactamase series are shown (Scheme 2). Evidently, based on the known SAR,[13] reasonable starting points would be identified from a sulfonamide library of at least 10 000 compounds but not from a library of 4000 compounds. Weak but closely related amides would be found in a general amide library of 50 000 as well (Scheme 2). On the other hand, only amide but not

Table 2. Hit Rates Obtained for Amide Baits in Fragment-like Acyclic Virtual Amide Libraries, and in Available Chemical Compound (ACC) Sets

	library size	DPP-4 inhibitors		integrase inhibitors	DH-1 inhibitors	mGluR5 antagonists	GPCR1 agonists	GPCR2 agonists	GPCR3 agonists[a]
no. of baits		309[b]	4[c]	39	70	71	52	13	8
potency (μM)		0.0006−49	0.4−13	0.01−30	0.73−4.8	0.007−28	0.004−4.8	1.6−26	17−82
full amides	2167K	6571	559	190	738	1121	670	306	268
subset amides[d]	200K	355	30	8.5	56	86.5	44	18	19.5
subset amides[d]	50K	71.5	5.5	0.5	16.5	20	7.5	6.5	2
subset amides[d]	20K	16	0	1	6.5	11	4	4	1.5
subset amides[d]	10K	5	0	0.5	4.5	4.5	2	2	1
subset amides[d]	5K	3	0	0	2	2.5	1.5	0.5	0
full ACC	66K	2	3	3	31	164	24	65	NA
subset ACC[d]	6K	0	0	0	1	2.5	1	4.5	NA
subset ACC[d]	2K	0	0	0	1	1.5	1	0.5	NA
subset ACC[d]	0.5K	0	0	0	0	0.5	0	0.5	NA

[a] Baits were derived from a set of fragments purchased and screened against GPCR3. Fragments for purchasing were identified via the use of a substructure query using a known pharmacophore motif. [b] All fragment-like hits. [c] Model amide fragments shown in Table 1. [d] Each library subset was created by both method A and B, respectively; values are the average of the two methods.

Table 3. Hit Rates Obtained for Different Chemotypes for Three Targets

	% of library	DH-1 inhibitors			mGluR5 antagonists			β-lactamase inhibitors		
bait class		triazoles[a]	morphed amides[b]	amides[c]	tetrazoles[d]	morphed amides[b]	amides[c]	sulfonamides[e]	morphed amides[b]	amides[c]
no. of baits		279	54	70	250	186	71	8	8	1
potency (μM)		0.00003−4.8	NA	0.73−4.8	0.00025−39	NA	0.007−28	1.1−310	NA	340
full library	100	2	434	738	435	4360	1121	30	111	71
subset library[f]	10	0	43.5	56	36	317	86.5	2.5	4.5	3.5
subset library[f]	2.5	0	11	16.5	3	79	20	0.5	3	1
full ACC[g]	100	3	65	31	1	417	164	15	48	24
subset ACC[g,f]	9	0.5	4.5	1	0	12.5	2.5	1.5	3.5	1
subset ACC[g,f]	3	0	1	1	0	3.5	1.5	0	1.5	0.5

[a] Full fragment-like virtual triazole library had 3961 members. [b] Connecting heterocyclic/sulfonamides motif was replaced by an amide bond (Figure 1); full fragment-like acyclic virtual amide library had 2.1M members. [c] Full fragment-like acyclic virtual amide library had 2.1M members. [d] Full fragment-like virtual tetrazole library had 33 756 members. [e] Full fragment-like virtual sulfonamide library had 79 097 members. [f] Each library subset was created by both method A and B, respectively; values are the average of the two methods. [g] Available chemical compounds (ACC).

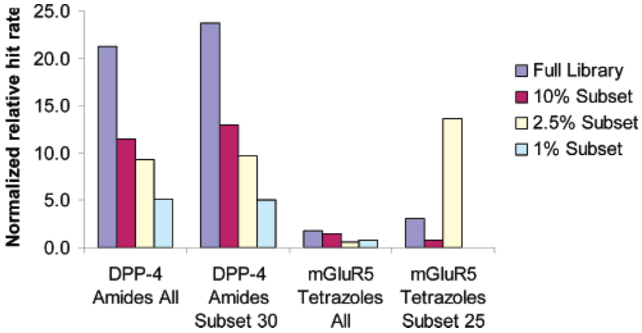


Figure 2. Normalized relative hit rates [(number of hits × 100)/(number of baits × percent library size)] using full and partial bait sets.

Table 4. Hit Rates Obtained for Full and Subset Bait Sets for Two Targets

	% of library	DPP-4 inhibitors		mGluR5 antagonists	
bait class		all amides	subset amides	all tetrazoles	subset tetrazoles
no. of baits		309	30[a]	250	25[a]
full library[b]	100	6571	712	435	77.5
subset library[c]	10	355	38.75	36	5.25
subset library[c]	2.5	71.5	7.25	3	0.25
subset library[c]	1	16	1.5	2	0.75

[a] A combination of subsets selected by methods A and B. [b] Full fragment-like acyclic virtual amide library had 2.1M members; full fragment-like virtual tetrazole library had 33 756 members. [c] Each library subset was created by both method A and B, respectively; values are the average of the two methods.

sulfonamide fragment hits similar to the β-lactamase series are found in ACC. The latter is partially due to the lower representation of sulfonamides in the fragment-like ACC. The increased molecular weight and lower solubility of sulfonamides as compared to amides disqualify many sulfonamides from our ACC deck. Triazole library subsets fail for dehydrogenase-1, while our tetrazole libraries highly efficiently deliver starting points for the mGluR5 chemotype. It is noteworthy that even for mGluR5 the relative hit rates are more than an order or magnitude larger for the amide libraries than that from tetrazole libraries (Figure 3).

Fragment-like Hits for Medicinal Chemistry Programs Devoid of Known Fragment-like Actives. The two primary pharmacophore motifs in our GPCR3 agonist database contained an amide and a heterocyclic motif, respectively. These motifs were the basis of a generic substructure query in our fragment-like ACC database. After visual inspection of the initial hits obtained with this substructure query, a total of 80 compounds were purchased and screened in our GTP binding assay. Eight compounds that showed good single-point inhibition were titrated in a cAMP assay to establish agonistic behavior and potency. These agonists when applied as baits in our virtual screening protocol resulted in hit rates equivalent to the other targets in Table 2. Thus, it is possible to re-engineer fragment-like starting points for this development program and to show that such hits could be identified from a random lead discovery screening event as well. Hit rates in the ACC database for this target are likely to be biased and therefore be of low significance, as the actual baits were selected and purchased from the ACC set.

Dipeptidyl Peptidase IV Inhibitors. DPP-4 inhibitors have been demonstrated to possess therapeutic potential in the treatment of type 2 diabetes.[20] Most DPP-4 inhibitors reported to date incorporate an α-amino acid moiety[15] or aminomethyl-heterocycle[21] or a β-amino acid amide.[15] The latter is known to be ordered in an opposite binding orientation of that of α-amino acids.[15,22] The HTS hit rate for DPP-4 appears to have been low (∼0.01% range) according to publicly available reports[23] and our internal observation. It is also likely that no straightforward lead toward the β-amino acid class has been found in lead generation efforts throughout the industry other than peptidomimetic **7** (Scheme 1) reported earlier.[24] That is rather surprising considering the simplicity of this chemical class and that a wide range of amide substitutions are tolerated as long as the β-amino acid motif is retained (Table 1). All fragments tested had good ligand efficiency (Table 1), and the minimum critical pharmacophore (compound **3**) depicted fairly good off-target selectivity against a wide range of serine proteases. Using **3**−**6** as baits, good relative hit rates were obtained in the virtual amide library but not in the ACC. In the latter case, the three hits contained no β-amino acid amides as opposed to the virtual amide hit sets, where several molecules contained the key pharmacophore of our interest (**3**). In addition, the hit rate in ACC for using the full 309 member DPP-4 bait set, which included both α- and β-amino acids, was very low: only two α-amino acid amides were within 0.75 Tanimoto similarity of any of the 309 bait compounds. With SAR available for thousands of compounds in our in-house database, it can be concluded that many members (Scheme 3) of the "similar" set at 50K virtual amide level are expected to have good potency in the biochemical assay. In fact, some molecules in the "dissimilar" sets could be actives as well (Scheme 3). As expected, some actives are missed by our selection method of Tanimoto coefficient using 2D fingerprints, but on the other hand, many members of our similar hits would be weak inhibitors. Thus, statistically the data presented in Tables 2 and

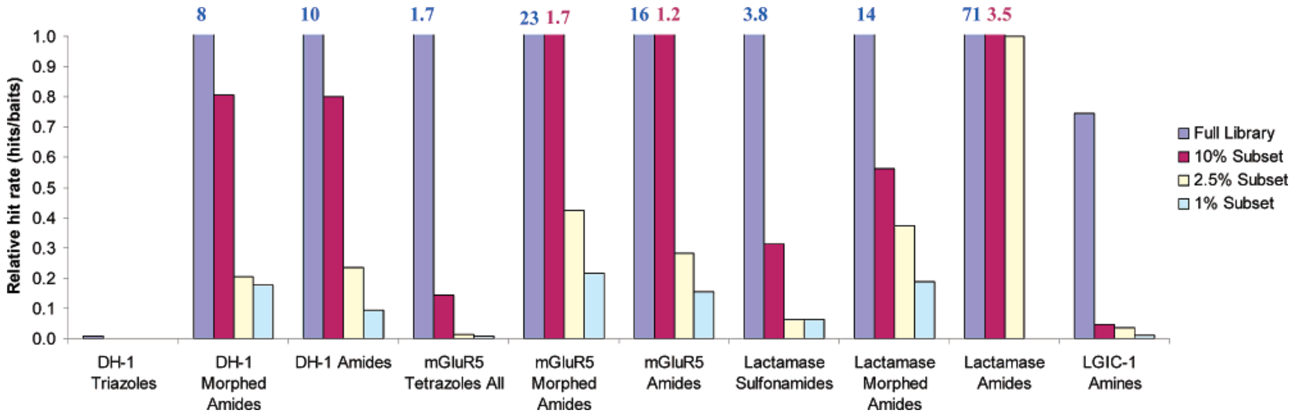
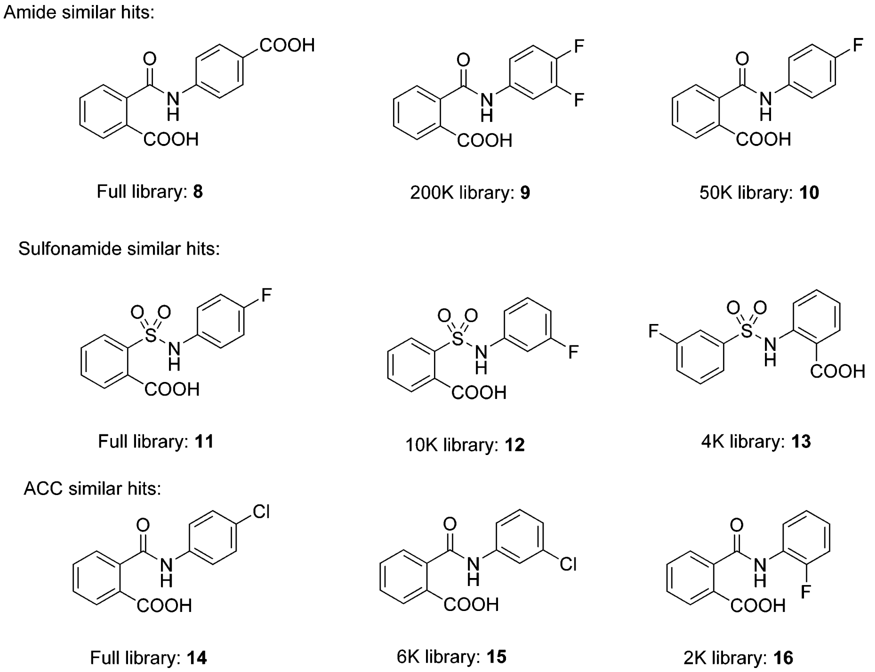


Figure 3. Relative hit rates (number of hits/number of baits) for programs with non-amide leads: dehydrogenase-1, mGluR5 antagonists, β-lactamase, and ligand gated ion channel-1 (LGIC-1: for values, see Experimental Section) with numerical values shown for bars that are out of scale.

Scheme 2. Representative Examples of Virtual Hits Found for β-Lactamase



Amide similar hits:

Full library: **8**

200K library: **9**

50K library: **10**

Sulfonamide similar hits:

Full library: **11**

10K library: **12**

4K library: **13**

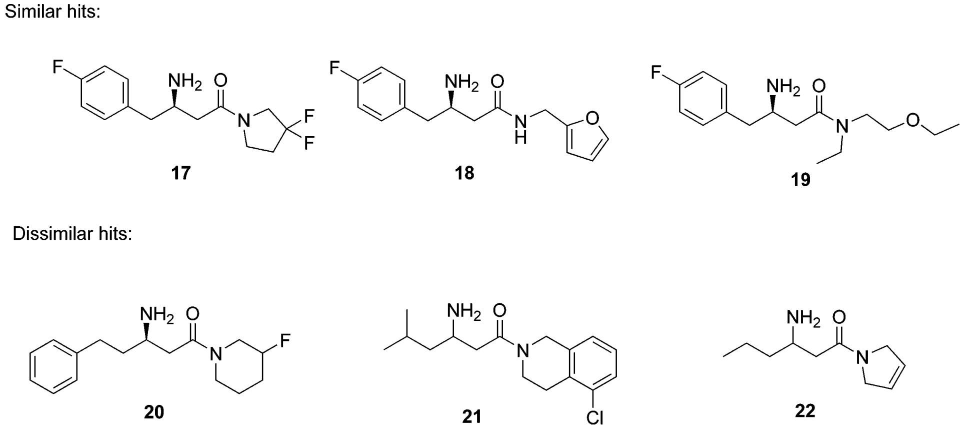
ACC similar hits:

Full library: **14**

6K library: **15**

2K library: **16**

Scheme 3. Representative Examples of Virtual Hits Found in the "Similar" and the "Dissimilar" Clusters for DPP-4



Similar hits:

17

18

19

Dissimilar hits:

20

21

22

3 may be a good approximation of the number of actual hits in the virtual and commercial datasets, although the actual selected members may contain false positives.

Hits in the Available Chemical Compound Set. The hit rate for the various chemotypes in the ACC set shows a large variance (Tables 2 and 3). This collection is not limited to amides and thus more diverse from the chemical connectivity point of view than our virtual amide library. From Table 3 it is also evident that screening the entire collection of 66 000 compounds does little to enhance our ability to find hits like the triazole and tetrazole bait molecules. The hit rate for amide chemotypes in the ACC set varies greatly as opposed to that in the comparable 50 000 member virtual set. Specifically looking at the DPP-4 case, the two similars in the ACC group did not include a β-amino acid amide motif.

Discussion

Fragment-based lead discovery techniques using screening decks of 10^3 compounds have been consistently successful to deliver weak hits against diverse sets of targets. Finding more of these weak fragment hits has not been desirable, however, due to cost considerations and throughput limitations in primary screening and follow-up. Nevertheless, the effect of the fragment library size on the expected output of primary fragment screens is of significant interest. Meaningful advantages in the lead discovery process could potentially be realized if primary fragment hits reach a potency level where they can directly be evaluated in biological settings including off-target assays. For that to happen, the ligand efficiency index of the fragment hits does not have to be extremely high: molecules of 200−300 Da with potency range of 0.5−20 μM require a ligand efficiency (LE) range of 0.28−0.58. That appears to be an achievable range because it is in accord with both the average LE value of ~0.45 for the bait molecules in this study and other literature values reported for fragments.[25] Thus, it can be hypothesized that the likelihood of consistently delivering potent fragment hits appears to be a matter of sampling of fragment space.

A few important points are worth highlighting regarding the premise of this study. (A) Molecular frameworks of actives in medicinal chemistry databases at Merck for advanced programs are not significantly biased by available reagents, and therefore, these datasets can provide useful bait molecules (or training sets) for querying accessible compound collections. (B) In development programs, typically only a small number of fragment-like actives are made once the primary SAR is established. (C) The hits obtained via interrogation of large virtual databases with relatively few bait molecules appears to be proportional to the number of baits. (D) The study design parameters (MW < 300 Da, hit potency < 50 μM) employed in this work are in good accord with recently published guidance[26] for obtaining rule-of-five compliant optimized candidates. The lower molecular weight limit (MW > 160 Da) eliminates most of the very small fragments that are expected to show weak and promiscuous binding modes.[1,27,28]

The virtual recapitulation of major development series in this work unveiled several trends. Some of these candidate series belonged to chemotypes synthetically accessible with vast and readily available chemical classes, while the practically available chemical space around other series was limited by the reagent pool and the synthetic schemes. The diversity in chemical functionality of the reagents for fragment libraries plays an important role in the uniform success in Table 2 of our amide library, as evidenced by hit rate comparisons among different chemical classes in Table 3. Artificially created "morphed" amide baits, which may not actually exist or been tested for activity, provide an opportunity to directly evaluate the effect chemical inputs may have on library performance. It was found that the latter is highly influenced by the available reagents that can be used to construct the screening deck. Side chains available in amine and acid reagent classes span much larger functional and connectivity diversity than the reagent pools used to build the non-amide sets. Poorly performing (a) triazole, (b) tetrazole, (c) sulfonamide, and (d) amine virtual libraries (Figure 3) are plagued by detrimental lack of functional diversity among (a) both hydrazines and amides, (b) carbonyls, (c) sulfonyl chlorides, and (d) carbonyls, respectively. The need for reagent diversity to sample chemical space is manifested in an obvious trend in Table 3: "morphed" baits are much more effective than the corresponding non-amide baits in finding similar "hits". This finding was further substantiated by using actual amide baits, which are close analogs of the non-amide bait compounds (Table 3) and have been tested in assays. Thus, it can be argued that a library development strategy for fragment screening decks should consider reagent diversity as one of the primary drivers. Consequently, libraries of unique scaffolds with highly limited side chain diversity may initially be deprioritized, as their contribution to the *systematic* success rate of a general fragment screening deck will be minor.

Considering the relatively small number of baits used in our study, it appears that a few thousand compounds are sufficient to represent the commercially accessible amide fragment space for the purpose of randomly generating hit classes with potency <50 μM. From the results, it is also evident that a subset of a few hundred amides would be an insufficient screening deck to rely on with the expectation of obtaining potent fragment hits against diverse targets. This notion is consistent with the reported potency range of unoptimized fragment hits.[10,25] On the basis of the data in Table 3 and Figure 3 it can also be proposed that molecules in these amide libraries are useful surrogates for several neutral, heterocyclic systems for hit discovery purposes. The structure of these potent amide fragment hits can later be modified into various heterocyclic bioisosteres during lead optimization, an analoging process routinely explored in medicinal chemistry development work.

Average virtual hit rate of 0.035% in the 50K amide library across the seven targets is lower than that found in NMR-based fragment screenings.[10] The difference could partially be due to the low number of baits for most targets as discussed earlier and the difference in activity range of interest (<85 μM herein vs 10−5000 μM in the NMR studies). Most importantly, this hit rate is only slightly lower than the confirmed hit rate of typical HTS campaigns. On the other hand, it is not to suggest that the hit discovery challenges facing the pharmaceutical industry can be solved by merely screening amide libraries. A carefully designed general fragment library spanning various key medicinally relevant connectivity classes would have to be built using diverse reagent classes to expect good performance against multiple protein families. In addition, there is also a significant value in screening the highly diverse commercially available fragment collection (ACC) to achieve better sampling rates, especially for chemical classes that are difficult to access via parallel synthesis. The screening results from this collection, however, are expected to be more inconsistent (Tables 2 and 3).

Our findings for DPP-4 and other targets in this study point to the importance of sampling in fragment-based lead discovery and suggest that carefully designed and more densely populated fragment sets may deliver hits superior to those obtained with today's smaller fragment libraries. Such screening decks may eliminate the need for costly follow-up of many extremely weak fragment leads that, due to their weak potency, would otherwise not capture the attention of medicinal chemists. Alternatively, one can elect not to carry out follow-up on more than a few fragment hits. However, with no compelling biological data at hand, it is difficult if not impossible to tell a priori which molecular framework would lead to more progressible leads or ultimately better drugs. The HTS hit **7** from the Merck sample collection is equipotent to fragment **4** and is only slightly more potent than the minimum pharmacophore **3**. It has been shown that the right-hand side of the 513 Da **7** does not contribute to binding to DPP-4.[24] Thus, its potency, selectivity profile, and high ligand efficiency would make **3** a highly attractive starting point for lead discovery. The virtual screening data (Table 2) presented herein also reveal that even better starting points (Scheme 3) as well as some limited SAR would be realized from screening a systematically built general fragment-like amide library.

Conclusion

The results presented herein support the hypothesis that a relatively small number of readily available fragment-like compounds can yield potent actives against a large variety of targets. The data also imply that reagent diversity for chemical transformations selected for fragment library synthesis has a great impact on the hit rate in primary screening of fragments. Extrapolation of our results to other types of libraries may form the basis for the construction of a general fragment library for lead discovery. These studies are currently underway in our group and will be reported at a later date.

Experimental Section

The virtual amide libraries were constructed as follows. Amines and acids were extracted using a Pipeline Pilot (version 4.5) protocol, from MDL's ACD (2004.4 ed.), respectively. Both reagent sets were filtered to remove duplicates and incompatible and reactive functional groups (amines, acids etc.), followed by an availability filter (≤$500/g listed in ACD from at least one vendor). The reagents then were adjusted to their neutral form, and counterions or salts were removed. The full library then was enumerated using Pipeline Pilot, followed by deprotection of Boc and O-tBu functional groups. Duplicates were removed after normalization of the tautomer forms with Pipeline Pilot. The remaining molecules were subjected to the following filters: S/N/O count < 8, N/O count > 1, H-bond donor count < 4, $-3 \le \log P \le 3$, unknown stereocount < 2, rotatable bond count < 10, 50 μM $<$ calculated solubility $<$ 50 mM, and a list of unwanted functional groups due to reactivity, toxicity, or PK considerations. Solubility was calculated as implemented in Pipeline Pilot.[31]

The virtual triazole library was created as above but via the reaction of hydrazides with amides as published.[29,30] The virtual tetrazole library was created as above but via the reaction of aryl amines with aldehydes as published.[12] The virtual sulfonamide library was created as above but via the sulfonylation of amines with sulfonyl chlorides.

The virtual amine library was created as above but via the reductive alkylation of aryl amines with aldehydes. The total amine library was composed of 126 000 members. Ligand-gated ion channel-1 bait set had 43 actives with potency ranging 47−9300 nM. The averaged hit rate for 100%, 10%, 2.5%, and 1% library subsets were 32, 2, 1.5, 0.5, respectively.

The available chemical compound set (ACC) was created by combining entries in ACD (MDL, 2004.4 ed.) with the small molecule offerings of 18 vendors that regularly update their portfolio to Merck Research Laboratories. These vendors in our experience successfully deliver on >95% of the compounds ordered. The combined database was subjected to filters as described for amides to yield 66 469 independent entries.

Subsets were created with Pipeline Pilot as follows. First the parent set was randomized and then the desired subset size was selected by either FCFP_4 fingerprints (method A) and by molecular weight (method B).

Biochemical assay conditions for DPP-4 and related proteases were carried as published.[15]

Compounds **3**−**7** were synthesized using the traditional amide coupling protocol with the Boc-β-amino acid (1 equiv), the amine (1 equiv), immobilized PS-carbodiimide (3 equiv), HOBt (1 equiv), and diisopropylethylamine (6 equiv) in THF at room temperature. After overnight reaction, PS−NCO (5 equiv) and PS trisamine (5 equiv) were added, and the mixture was shaken for 5 h at room temperature. The resins were filtered off and the volatiles were evaporated in a SpeedVac. To the residue was added TFA:water = 95:5 for 1 h and the product was dried in a SpeedVac before preparative HPLC purification.

Compound **3**: ^{1}H NMR (DMSO-d_6, 400 MHz) δ 7.79 (1H, br d), 7.23 (2H, t), 7.14 (3H, t), 3.20 (1H, m), 2.60-2.40 (5H, m), 2.04 (1H, dd), 1.93 (1H, dd); ^{13}C NMR (DMSO-d_6, 400 MHz) δ 172.16, 140.11, 129.90, 128.84, 126.60, 50.79, 44.41, 43.43, 40.80, 25.99; HRMS calcd for $C_{11}H_{16}N_2O$ 193.1341, found 193.1350.

Compound **4**: ^{1}H NMR (DMSO-d_6, 400 MHz) δ 7.23 (2H, t), 7.15 (3H, t), 3.20 (1H, br), 2.85 (3H, s), 2.75 (3H, s), 2.60 (1H, dd), 2.52 (1H, dd), 2.28−2.16 (2H, m); ^{13}C NMR (DMSO-d_6, 400 MHz) δ 171.76, 140.33, 129.87, 128.84, 126.59, 50.62, 43.97, 40.81, 37.37, 35.33; HRMS calcd for $C_{12}H_{18}N_2O$ 207.1497, found 207.1506.

Compound **5**: ^{1}H NMR (DMSO-d_6, 400 MHz) δ 7.23 (2H, t), 7.14 (3H, d), 3.35-3.18 (5H, br m), 2.60 (1H, dd), 2.49 (1H, dd), 2.21 (1H, dd), 2.12 (1H, dd), 1.78 (2H, m), 1.68 (2H, m); ^{13}C NMR (DMSO-d_6, 400 MHz) δ 170.18, 140.32, 129.86, 128.83, 126.58, 50.47, 46.60, 45.77, 44.06, 42.08, 26.21, 24.57; HRMS calcd for $C_{14}H_{20}N_2O$ 233.1654, found 233.1662.

Compound **6**: ^{1}H NMR (DMSO-d_6, 400 MHz) δ 7.22 (2H, m), 7.15 (7H, m), 4.54 (2H, m), 3.63-3.53 (2H, m), 3.25 (1H, m), 2.76 (1H, br t), 2.72-2.60 (2H, m), 2.53 (1H, m), 2.35 (2H, m); ^{13}C NMR (DMSO-d_6, 400 MHz) δ 170.87, 170.78, 140.29, 135.47, 135.20, 134.37, 134.01, 129.89, 129.18, 129.00, 128.85, 127.11, 127.07, 126.96, 126.86, 126.79, 126.72, 126.82, 50.75, 50.71, 47.16, 44.11, 44.03, 43.98, 43.26, 29.40, 28.62; HRMS calcd for $C_{19}H_{22}N_2O$ 295.1810, found 295.1818.

Acknowledgment. The author is indebted to Xiaoping Zhang and Ning Ren for the biochemical data and to Ann Weber, Kevin Chapman, and Steven Colletti for the insightful discussions. The contribution of John Athanasopoulos in collecting the high-resolution mass spectra is greatly appreciated as well.

References

(1) 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.

(2) Congreve, M.; Carr, R.; Murray, C.; Jhoti, J. A 'Rule of Three' for Fragment-Based Lead Discovery? *Drug Discovery Today* **2003**, *8*, 876−877.

(3) Hopkins, A. L.; Groom, C. R.; Alex, A. Ligand Efficiency: A Useful Metric for Lead Selection. *Drug Discovery Today* **2004**, *9*, 430−431.

(4) Proudfoot, J. R. Drugs, Leads, and Drug-Likeness: An Analysis of Some Recently Launched Drugs *Bioorg. Med. Chem. Lett.* **2002**, *12*, 1647−1650.

(5) Harper, G.; Pickett, S. D.; Green, D. V. S. Design of a Compound Screening Collection for Use in High Throughput Screening. *Comb. Chem. High Throughput Screening* **2004**, *7*, 63−70.

(6) Hann, M. M.; Oprea, T. I. Pursuing the Leadlikeness Concept in Pharmaceutical Research. *Curr. Opin. Chem. Biol.* **2004**, *8*, 255−263.

(7) Makara, G. M.; Athanasopoulos, J. Improving Success Rates for Lead Generation Using Affinity Binding Technologies. *Curr. Opin. Biotechnol.* **2005**, *16*, 666−673.

(8) Carr, R. A. E.; Congreve, M.; Murray, C. W.; Rees, D. C. Fragment-Based Lead Discovery: Leads by Design. *Drug Discovery Today* **2005**, *10*, 987−992.

(9) 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 Dihydroneopterin Aldolase Using CrystaLEAD High-Throughput X-ray Crystallographic Screening and Structure-Directed Lead Optimization. *J. Med. Chem.* **2004**, *47*, 1709−1718.

(10) Hajduk, P.; Huth, J. R.; Fesik, S. W. Druggability Indices for Protein Targets Derived from NMR-Based Screening Data. *J. Med. Chem.* **2005**, *48*, 2518−2525.

(11) Poon, S. F.; Eastman, B. W.; Chapman, D. W.; Chung, J.; Cramer, M.; Holtz, G.; Cosforda, N. D. P.; Smith, N. D. 3-[3-Fluoro-5-(5-pyridin-2-yl-2*H*-tetrazol-2-yl)phenyl]-4-methylpyridine: A Highly Potent and Orally Bioavailable Metabotropic Glutamate Subtype 5 (mGlu5) Receptor Antagonist. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 5477−5480.

(12) Smith, N. D.; Poon, S. F.; Huang, D.; Green, M.; King, C.; Tehrani, L.; Roppe, J. R.; Chung, J.; Chapman, D. P.; Cramera, M.; Cosforda, N. D. P. Discovery of Highly Potent, Selective, Orally Bioavailable, Metabotropic Glutamate Subtype 5 (mGlu5) Receptor Antagonists Devoid of Cytochrome P450 1A2 Inhibitory Activity. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 5481−5484.

(13) Tondi, D.; Morandi, F.; Bonnet, R.; Costi, M. P.; Shoichet, B. K. Structure-Based Optimization of a Non-β-lactam Lead Results in Inhibitors That Do Not Up-Regulate β-Lactamase Expression in Cell Culture. *J. Am. Chem. Soc.* **2005**, *127*, 4632−4639.

(14) Powers, R. A.; Morandi, F.; Shoichet, B. K. Structure-Based Discovery of a Novel, Noncovalent Inhibitor of AmpC β-Lactamase. *Structure* **2002**, *10*, 1013−1023.

(15) Kim, D.; Wang, L.; Beconi, M.; Eiermann, G. J.; Fisher, M. H.; He, H.; Hickey, G. J.; Kowalchick, J. E.; Leiting, B.; Lyons, K.; Marsilio, F.; McCann, M. E.; Patel, R. A.; Petrov, A.; Scapin, G.; Patel, S. B.; Roy, R. S.; Wu, J. K.; Wyvratt, M. J.; Zhang, B. B.; Zhu, L.; Thornberry, N. A.; Weber, A. E. (2*R*)-4-Oxo-4-[3-(Trifluoromethyl)-5,6-dihydro[1,2,4]triazolo[4,3-*a*]pyrazin-7(8*H*)-yl]-1-(2,4,5-trifluorophenyl)butan-2-amine: A Potent, Orally Active Dipeptidyl Peptidase IV Inhibitor for the Treatment of Type 2 Diabetes. *J. Med. Chem.* **2005**; *48*, 141−151.

(16) Jensen, B. F.; Vind, C.; B. S.; Padkjaer, S. B.; Brockhoff, P. B.; Refsgaard, H. H. F. In Silico Prediction of Cytochrome P450 2D6 and 3A4 Inhibition Using Gaussian Kernel Weighted k-Nearest Neighbor and Extended Connectivity Fingerprints, Including Structural Fragment Analysis of Inhibitors versus Noninhibitors. *J. Med. Chem.* **2007**, *50*, 501−511.

(17) Godden, J. W.; Stahura, F. L.; Bajorath, J. Anatomy of Fingerprint Search Calculations on Structurally Diverse Sets of Active Compounds. *J. Chem. Inf. Model.* **2005**, *45*, 1812−1819.

(18) Makara, G. M. Measuring Molecular Similarity and Diversity: Total Pharmacophore Diversity. *J. Med. Chem.* **2001**, *44*, 3563−3571.

(19) Jain, A. N. Morphological Similarity: A 3D Molecular Similarity Method Correlated with Protein-Ligand Recognition. *J. Comput.-Aided Mol. Design* **2000**, *14*, 199−213.

(20) Wiedeman, P. E.; Trevillyan, J. M. Dipeptidyl Peptidase IV Inhibitors for Type 2 Diabetes and Metabolic Syndrome. *Drug Discovery Today*: *Ther. Strategies* **2005**, *2*, 143−149.

(21) Peters, J.; Weber, S.; Kritter, S.; Weiss, P.; Wallier, M. B.; Hennig, M.; Kuhn, B.; Loeffler, B.-M. Aminomethylpyrimidines as Novel DPP-IV Inhibitors: A 105-fold Activity Increase by Optimization of Aromatic Substituents. *Bioorg. Med. Chem Lett.* **2004**, *14*, 1491−1493.

(22) Nordhoff, S.; Cerezo-Galvez, S.; Feurer, A.; Hill, O.; Matassa, V. G.; Metz, G.; Rummey, C.; Thiemann, M.; Edwards, P. J. The Reversed Binding of β-Phenethylamine Inhibitors of DPP-IV: X-ray Structures and Properties of Novel Fragment and Elaborated Inhibitors. *Tetrahedron Lett.* **2006**, *16*, 1744−1748.

(23) Ward, R. A.; Perkins, T. D. J.; Stafford, J. Structure-Based Virtual Screening for Low Molecular Weight Chemical Starting Points for Dipeptidyl Peptidase IV Inhibitors. *J. Med. Chem.* **2005**, *48*, 6991−6996.

(24) Xu, J.; Ok, H. O.; Gonzalez, E. J.; Colwell, L. F.; Habulihaz, B.; He, H.; Leiting, B.; Lyons, K. A.; Marsilio, F.; Patel, R. A.; Wu, J. K.; Thornberry, N. A.; Weber, A. E.; Parmee, E. R. Discovery of Potent and Selective β-Homophenylalanine Based Dipeptidyl Peptidase Inhibitors. *Bioorg. Med. Chem Lett.* **2004**, *14*, 4759−4762.

(25) McClure, K. F.; Abramov, Y. A.; Laird, E. R.; Barberia, J. T.; Cai, W.; Carty, T. J.; Cortina, S. R.; Danley, D. E.; Dipesa, A. J.; Donahue, K. M.; Dombroski, M. A.; Elliott, N. C.; Gabel, C. A.; Han, S.; Hynes, T. R.; LeMotte, P. K.; Mansour, M. N.; Marr, E. S.; Letavic, M. A.; Pandit, J.; Ripin, D. B.; Sweeney, F. J.; Tan, D.; Tao, Y. Theoretical and Experimental Design of Atypical Kinase Inhibitors: Application to p38 MAP Kinase. *J. Med. Chem.* **2005**, *48*, 5728−5737.

(26) Hajduk, P. H. Fragment-based Drug Design: How Big Is Too Big? *J. Med. Chem.* **2006**, *49*, 6972−6976.

(27) Babaoglu, K.; Shoichet, B. K. Deconstructing Fragment-Based Inhibitor Discovery. *Nature Chem. Biol.* **2006**, *2*, 720−723.

(28) Hajduk, P. H. Puzzling through Fragment-Based Drug Design. *Nature Chem. Biol.* **2006**, *2*, 658−659.

(29) Olson, S.; Aster, S. D.; Brown, K.; Carbin, L.; Graham, D. W.; Hermanowski-Vosatka, A.; LeGrand, C. B.; Mundt, S. S.; Robbins, M. A.; Schaeffer, J. M.; Slossberg, L. H.; Szymonifka, M. J.; Thieringer, R.; Wright, S. D.; Balkovec, J. M. Adamantyl Triazoles as Selective Inhibitors of 11β-Hydroxysteroid Dehydrogenase Type 1. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 4357−4362.

(30) Omodei-Sale, A.; Consonni, P.; Galliani, G. A New Class of Nonhormonal Pregnancy-Terminating Agents. Synthesis and Contragestational Activity of 3,5-Diaryl-s-triazoles. *J. Med. Chem.* **1983**, *26*, 1187−1192.

(31) Tetko, I. V.; Tanchuk, V. Y.; Kasheva, T. N.; Villa, A. E. P. Estimation of Aqueous Solubility of Chemical Compounds Using E-State Indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1488−1493.

JM0700316