

Are Target-Family-Privileged Substructures Truly Privileged?

Dora M. Schnur,^{*,†} Mark A. Hermsmeier,[‡] and Andrew J. Tebben[†]

Computer Aided Drug Design and Lead Discovery, Pharmaceutical Research Institute, Bristol-Myers Squibb Company, P.O. Box 5400, Princeton, New Jersey 08543-5400

Received March 31, 2005

One of the early and effective approaches to G-coupled protein receptor target family library design was the analysis of a set of ligands for frequently occurring chemical moieties or substructures. Various methods ranging from frameworks analysis to pharmacophores have been employed to find these so-called target-family-privileged substructures. Although the use of these substructures is common practice in combinatorial library design and has produced leads,¹ the methods used for finding them rarely verified their selectivity for the particular target family from which they were derived. The frequency of occurrence among ligands associated with a target receptor family is not a sufficient criterion for those substructures to receive the label of target-family-privileged substructure. This study explores the question of selectivity of ClassPharmer² generated fragments for a series of target families: GPCRs, nuclear hormone receptors, serine proteases, protein kinases, and ligand-gated ion channels. In addition, a GPCR focused library and a random set of 10k compounds are examined in terms of their target-family-privileged-substructure composition. The results challenge the combinatorial chemistry concept of target-family-privileged substructures and suggest that many of these fragments may simply be drug-like or attractive for various receptors in accordance with the original definition of privileged substructures.^{3,4}

Introduction

The original concept of privileged substructures was put forth by Evans³ and more recently reviewed by Patchett.⁴ They described privileged substructures as those found in ligands across a set of diverse receptors. Further elaboration of the privileged substructure was postulated to lead to selectivity toward a specific target receptor. Although Evans and Patchett evolved this concept within a relatively narrow class of GPCR ligands, subsequent literature methods for finding privileged substructures and common practice in combinatorial chemistry library design not only expanded on their original analysis but also modified the definition of privileged structure to that of commonly occurring fragments within ligands associated with a target-receptor family. Various methods have been employed to find these so-called target-family-privileged substructures that are postulated to be selective for a given target family but promiscuous within that same family of targets. The majority of the commonly used methods are ligand-based and include frameworks analysis,⁵ 4-pt pharmacophores,⁶ and ClassPharmer substructure class generation.^{7,8} These fragments have been used in target family combinatorial library design,^{1,9,10} for virtual screening,¹¹ and for focused screening deck design.¹²

Clearly, the term privileged substructure has taken on a meaning beyond Evans' original intent. It has become identified with those substructures found to be promiscuous within a given target family and carries the implication that these substructures are specific to that target family. The motivation to identify such substructures is derived from the need to avoid off-target affinities early in the discovery process and thereby avoid complications as promising compounds are developed into drugs. If these substructures can be identified, they potentially provide cleaner starting points than the more promiscuous structures do. The question arises then as to how target family

privileged these fragments are. Are we deceiving ourselves in the belief that they are truly selective for the target families from which they are derived? In an era when determining off-target liabilities for potential drugs as early as possible is an important element of drug discovery, this is an important question.

Typically these target-family-privileged structure analyses have attempted to find minimal ligand substructures that have frequent occurrences within the target family. However, this can very easily lead one away from truly privileged substructures and toward those that are merely drug-like and/or promiscuous protein binders. Consider the comparison of the often cited^{5–8} GPCR privileged substructure, biphenyl, and its analogue 2-tetrazolobiphenyl. A substructure search of the 2004 version of MDDR finds that 2-tetrazolobiphenyl appears in 1046 compounds, all of which fall into the activity classes related to the Angiotensin II receptors. The biphenyl substructure is found in 5658 compounds spanning 311 activity classes that include a significant number of GPCRs and also a host of other targets. Although biphenyl may be classified as privileged because of its frequent appearance in GPCRs, it is clear that true privilege does not arise until the tetrazole moiety is included. Biphenyl itself is likely only to be a privileged protein binding element.

Although some of the literature studies involving target-family-privileged substructures compare the fragment-occurrence frequency of GPCR privileged substructures with non-GPCRs as a whole among known drugs,¹³ little analysis has been done on the selectivity of these substructures with respect to other target families.^{7,8} In part, this has been due to the difficulty of collecting or extracting the target family ligand sets from commercial drug databases and corporate collections. Recently however, the publication of an ontology of pharmaceutical ligands by Schuffenhauer et al.¹⁴ removed this roadblock by mapping MDDR activities to published target ontologies. Additionally, the need for target-family knowledge databases to drive target-family-based library design has resulted in a

* To whom correspondence should be addressed. Tel: 609-818-4004. Fax: 609-818-3545. E-mail: dora.schnur@bms.com.

[†] Computer Aided Drug Design.

[‡] Lead Discovery.

number of commercial target family databases from companies, such as Aureus,¹⁵ Jubilant,¹⁶ Sertanty,¹⁷ and Biowisdom.¹⁸

This study examined ligand sets from five target families: G-coupled protein receptors (GPCRs), nuclear hormone receptors (NHRs), serine proteases, protein kinases, and ligand-gated ion channels. Substructure analysis was performed using ClassPharmer² to generate potential privileged substructures for each target family, and then the occurrence of these substructures within each of the other target-family-ligand sets was examined. In addition, a GPCR focused library and a random set of 10k compounds were examined in terms of their target-family-privileged substructure composition. Intra-target family selectivity was also examined qualitatively.

Methods

Ligand Sets. The sets of target-family ligands were extracted from MDDR version 2003.1¹⁹ through a web-based implementation of the target family ontology proposed by Schuffenhauer.¹⁴ A relational data model was constructed, and the tables provided in the Schuffenhauer reference were loaded into an Oracle database. Additionally, a table that mapped the MDDR activity to the MDDR compound identifier was included in the data model. The MDDR structures were stored in a proprietary formatted database indexed on the MDDR compound identifier. This permits rapid retrieval and display of structures in a web browser. A single SQL statement can select all of the structures associated with a node and its child nodes in the ontology. The web interface provides the user with the ability to browse the target family ontology as a hierarchical tree and to display MDDR structures by selection of a node in the ontology.

Because the most commonly examined GPCR-ligand set is Class A, this subset was used for the analysis. A set of 21 620 ligands was extracted and consisted of 9329 biogenic amines, 7620 peptide binding class A, and 4499 other Class A GPCRs. For nuclear hormone receptors, 2176 ligands were extracted. These consisted of both thyroid and estrogenic receptor ligands. The various ligand-gated ion-channel ligands including glutamate cationic and nicotinic receptor ligands were combined into a set of 3792 compounds. For serine proteases, 3015 ligands for the 8 receptors chymotrypsin, complement inhibitor, elastase, factor Xa, trypsin, factor VIIa, and tryptase were extracted. Similarly, a set of 1079 protein kinase inhibitors was extracted. The structures were stored in 2D MDL sd files²⁰ as input for ClassPharmer.

The test sets of ligands consisted of an ~10k GPCR focused library and an ~10k set of random compounds. The library, R1-core-R2, was loosely designed using privileged substructures derived from the 1999 version of MDDR using frameworks/maximal common substructure analysis. From that analysis, a set of 15 general Markush SLNs represented 90% of the 1999 Class A GPCR set.^{7,8} Reagents (nucleophiles, such as amines and phenols for R1 and R2, respectively) were selected from ACD,²¹ and those with undesirable functional groups (side reactivity, etc.) were filtered out. Using the SLNs, each reagent list was sorted into GPCR-like, nonGPCR-like, and nonGPCR-like but interesting sets. The last set consisted of reagents that provided small R groups such as halogens, small alkyls, etc. Additionally, molecular weight (MW) (<250) and ClogP⁶ (<3.5) cutoffs were used on the reagent lists. All filtering and sorting was done using the Selector module of SYBYL6.7.²² The GPCR-like and nonGPCR-like but interesting lists were further reduced manually on the basis of availability and reactivity considerations. CombiLibMaker²³ was used to generate a library, whose members had molecular weights less than

800. SYBYL Selector was used to further filter the library using a ClogP cutoff (<5) and an MW cutoff <500. The enumerated library was output as SMILES. DiverseSolutions²⁴ cell-based selection was performed from the chemistry space automatically defined by the program from 3D hydrogen-suppressed BCUT descriptors for the filtered enumerated virtual library. The selected virtual products were used to suggest the diverse reagent sets for the library. These reagent lists were modified slightly on the basis of experimentally determined reactivity during reaction trials, that is, some reagents selected computationally were deleted from the set, and a few reagents containing substructures believed specific for a particular GPCR target were also included by the synthetic chemist. The ultimate dimensions of the library design were 4 cores, 64 R1's and 45 R2's. Only compounds that were actually synthesized, isolated, and characterized were used for the analysis. These were obtained from the corporate database in 2D sd format for the analysis.

The 10k random set was extracted from the corporate collection as SMILES using Daylight²⁵ software and converted to 2D sd format using SYBYL UNITY dbtranslate.²⁶

Substructure Generation. ClassPharmer 3.0² was used for the analysis. This software tool uses graph-based analysis to derive keys that capture substructure common features in the ligand training set. The resultant classes or clusters of compounds represented by common substructures can be further analyzed using the R-table generation module and through the importation of activity/selectivity data as property attributes of the classes. The substructures, which are displayed by the viewer module with attached R-group attachment positions indicated, potentially provide a rich source of scaffolds for combinatorial library elaboration, or in the case of this analysis are the putative privileged substructures. Because a redundancy setting controls the number of classes in which a compound may appear, ligands can be broken into a variety of substructure fragments that may not be identified with methods that allow a compound to be assigned only to one cluster. Additionally, compounds that are singletons appear as separate classes. Because test lists of compounds may be filtered through the classes, retention of singletons is an important feature for subsequent virtual screening of compound sets or libraries.

For this analysis, the default settings for homogeneity (medium) and redundancy (medium) were used. Homogeneity controls the size of the substructures generated by adjusting bond topology equivalency, and redundancy controls the frequency with which a compound will appear in multiple classes, that is, be represented by multiple fragments. The choices for both of these parameters are high, medium, and low. For ease in viewing, the resultant substructures and R-tables were generated. The default structure view without R-table generation is the smallest compound in the class with the parent substructure highlighted. No attempt was made to merge classes into supersets because this feature was not yet available.

Target Family Comparisons. ClassPharmer 3.0 classes (or putative privileged substructures) were generated for each target family. Then each target-family-ligand set was filtered through each of the other target family classification sets to find out which classes were occupied and what percentage of each ligand set fit each target family class set. In addition, the GPCR library and the random set were filtered through each classification.

Intra-Target Family Selectivity. Serine protease and class A GPCR classifications were also utilized for a proof of principle for target selectivity analysis. The eight receptors for the serine proteases were assigned a number from 1 to 8 and imported as an attribute for each compound into ClassPharmer.

Table 1. Results of Filtering the Other Target Families through the G-Coupled-Protein-Receptor Class A Substructure Classes: Total Compounds for Each Family that Matched Any Substructure

GPCR class A results		1190 classes		43 singletons	
ontology target family	total cpds	not processed	selected	failed	
GPCRs class A	21 620	354	21 266	not applicable	
biogenic amines			9329		
peptide binding class A			7620		
other class A			4499		
ion channels	3792	21	2236	1535	59%
nuclear hormone receptors	2176	34	648	1494	30%
protein kinases	1079	2	668	409	62%
serine proteases	3015	76	883	2056	29%
GPCR library	10 046	97	9809	140	98%
random set	9911	64	4439	5408	45%

Table 2. Results of Filtering the Other Target Families through the Nuclear Hormone-Receptor Substructure Classes: Total Compounds for Each Family that Matched Any Substructure

nuclear hormone-receptor results		121 classes		19 singletons	
ontology target family	total cpds	not processed	selected	failed	
nuclear hormone receptors	2176	34	2142	not applicable	
GPCRs Class A	21 620	354	2967	18 299	14%
ion channels	3792	21	448	3283	12%
protein kinases	1079	2	116	961	11%
serine proteases	3015	76	234	2705	8%
GPCR library	10 046	97	2230	7719	22%
random set	9911	64	1291	8556	13%

Table 3. Results of Filtering the Other Target Families through Ligand-Gated Ion Channel Substructure Classes: Total Compounds for Each Family that Matched Any Substructure

ion Channels		297 classes		12 singletons	
ontology target family	total cpds	not processed	selected	failed	
ion channels	3792	21	3771	not applicable	
glutamate type			2039		
nicotinate type			1732		
nuclear hormone receptors	2176	34	364	1778	17%
GPCRs Class A	21 620	354	5619	15 647	26%
protein kinases	1079	2	309	768	29%
serine proteases	3015	76	239	2700	8%
GPCR library	10 046	97	7069	2880	70%
random set	9911	64	2394	7453	24%

Table 4. Results of Filtering the Other Target Families through the Protein Kinase Substructure Classes: Total Compounds for Each Family that Matched Any Substructure

protein kinases		101 classes		23 singletons	
ontology target family	total cpds	not processed	selected	failed	
protein kinases	1079	2	1077	not applicable	
ion channels	3792	21	566	3205	15%
nuclear hormone receptors	2176	34	92	2050	4%
GPCRs Class A	21 620	354	2026	19 240	9%
serine proteases	3015	76	121	2818	4%
GPCR library	10 046	97	2295	7654	23%
random set	9911	64	1324	8523	13%

The visualization as a distribution histogram for the compounds of each class allowed a crude gauge of the specificity of the classes with regard to serine protease targets. For GPCRs, the ligand set was split into biogenic amines, peptide-binding-protein ligands, and other class A GPCRs. Again, for crude visualization purposes, the biogenic amine target keys were assigned numbers from 1 to 42 and then grouped according to the ontology with the groups assigned numbers from 1 to 15. Similarly, for peptide binding proteins, 41 target keys were numbered and grouped by ontology into 18 supersets to examine their selectivity.

Results and Discussion

Target Family Classes and Filtering. The numbers of classes and compounds per class found for each target family and the class occupancy by other target-family ligands are

summarized in Tables 1–5. Because ClassPharmer normalizes input structures, compounds that contain metals, silicon, salts, and improper valences, or have other structural errors were excluded. Additionally the program is not intended for peptide analysis. As a result of these exclusions, some compounds in the ligand sets were not processed. The percentage of each ligand set found in every other target family classification is summarized in Table 6.

An examination of the results for the GPCR set of substructure classes revealed that significant percentages of the other target-family-ligand sets could be represented by GPCR substructures: 59% of ligand-gated ion-channel ligands, 30% of nuclear hormone-receptor ligands, 62% of protein-kinase ligands, and 29% of serine-protease ligands. Because GPCRs dominate among the pharmaceutical industry drug targets,²⁷ it was no

Table 5. Results of Filtering the Other Target Families through the Serine-Protease Substructure Classes: Total Compounds for Each Family that Matched Any Substructure

ontology target family	serine proteases	not processed	74 singletons	failed	
	323 classes		selected		
	total cpds				
serine proteases	3015	76	2939		not applicable
protein kinases	1079	2	96		9%
ion channels	3792	21	234		6%
nuclear hormone receptors	2176	34	204		9%
GPCRs Class A	21 620	354	3679		17%
GPCR library	10 046	97	5081		48%
random set	9911	64	1718		17%

Table 6. Percentages of Target-Family-Ligand Sets that Occurred in Other Family Substructure Classes^a

target family set	compound sets						
	GPCRs	ion channel	NHRs	protein kinases	serine proteases	GPCR library	random
GPCRs		59%	30%	62%	29%	98%	45%
ion channel	26%		17%	29%	8%	70%	24%
NHRs	14%	12%		11%	8%	22%	13%
protein kinases	9%	4%	15%		4%	23%	13%
serine proteases	17%	6%	9%	9%		48%	17%

^a Columns correspond to compound sets. Rows correspond to target family substructure class sets.

Table 7. Number of Substructure Classes Found for Each Target Family and the Number Occupied by Other Target-Family-Ligand Sets^a

target family set	compound set						
	GPCRs	ion channel	NHRs	protein kinases	serine proteases	GPCR library	random
GPCRs	1190/S43	307	116	130	202	67	549
ion channel	140	297/S12	44	55	274	19	295/S
NHRs	48/S	36	121/S19	20	18	3	55
protein kinases	48/S	34	16	101/S23	20	3	58/S
serine proteases	82	34	24	293	323/S74	8	118

^a Columns correspond to compound sets. Rows correspond to the target family from which the substructure classes were generated. The diagonal corresponds to the number of classes and singletons (S) found for a given target family for its own ligand set.

surprise that 45% of a random set of compounds drawn from a corporate database would fit GPCR substructures. What was somewhat perplexing, if one believes in the concept of target-family-privileged substructures, was the high percentages for the kinases and ion channels. The overlap between the GPCR and ion channels may be rationalized by the similarity in the overall topology of the two target classes. Both proteins form helical bundles within a membrane, and their ligand binding sites are largely aromatic/hydrophobic cavities within this bundle.^{28,29} Underscoring the crossover between the two classes, it is often observed in the course of drug discovery that GPCR ligands are also potent hERG and sodium channel blockers.²⁸ In the case of kinases, a possible partial explanation is that a number of GPCR targets also have ATP binding sites. However, the percentages for the nuclear hormone-receptor-ligand set and the serine-protease-ligand set were also greater than that expected if substructures were privileged.

Similar observations may apply for the other target families. For the ligand-gated ion channels, 26% of GPCRs, 17% of nuclear hormone receptors, 29% of protein kinases, and 8% of serine proteases matched the substructure classes. For this family, 24% of the random set matched the substructures. This suggested that the GPCR and kinase percentages were significantly high. For nuclear hormone receptors, 9% of GPCRs, 12% of ligand-gated ion channels, 11% of protein kinases, and 8% of serine proteases matched the substructure classes. Only 13% of the random set matched these substructure classes. For protein kinases, 9% of GPCRs, 15% of ligand-gated ion channels, 4% of nuclear hormone receptors, and 4% of serine proteases matched the substructure classes. Again 13% of the random set

matched these substructure classes. Interestingly, the percentage of nuclear hormone-receptor ligands may have been significant. For serine proteases, 17% of the GPCRs, 6% of the ligand-gated ion channels, 9% of the nuclear hormone receptors, and 9% of the protein kinases matched the substructure classes. In this case, 17% of the random compounds matched the substructure set. The percentage of GPCRs was regarded to be high for this set.

Numbers of ligands alone are an insufficient measure of selectivity/nonselectivity, particularly in view of the inequity in the sample size of the GPCR ligands relative to the other target family sets. It is also true that the random set used for comparison was not truly random but was, in part, dependent on the historic target focus of the corporate database from which it was selected. Therefore, in addition to examining the total number of compounds from other target families that occupied each set of classes, the number of classes or substructures per family occupied by other families was also tabulated and appears in Table 7. The first observation from this table was that singletons were rarely filled by alternate target-family ligands. Because these tended to be target-specific structures with no analogues in the database, this is not surprising, and these singleton substructures can be omitted from consideration as privileged substructures.

Table 8 analyzes these data by the percentage of occupied classes. As can be seen from the Table, the GPCR ligand set occupied 40 to 48% of the substructure classes generated by all but the serine-protease target family for which it occupied only 25%. The ion-channel ligand set occupied 26 to 34% of the classes generated by all but that by the serine-protease target

Table 8. Percentages of Substructure Classes (Singletons Excluded) Occupied by Each Target Family^a

target family set	compound sets					
	GPCRs	ion channel	NHRs	protein kinases	serine proteases	random
GPCRs	na	26	10	11	17	46
ion channel	47	na	15	19	92	99
NHRs	40	30	na	17	15	45
protein kinases	48	34	16	na	20	57
serine proteases	25	11	7	91	na	37

^a Columns correspond to compound sets. Rows correspond to target family substructure class sets; na: not applicable.

family for which it occupied only 11%. The nuclear hormone-receptor ligand set seemed to contain the fewest out of the family substructures with occupied structure classes ranging from 7 to 16%. Interestingly, the kinase ligand set occupied 91% of the serine-protease substructure classes, whereas the serine-protease ligands occupied only 20% of the kinase substructure classes.

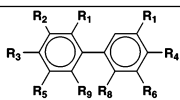
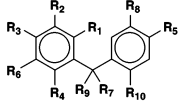
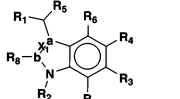
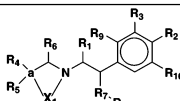
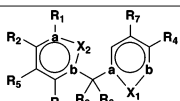
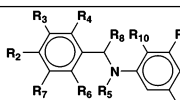
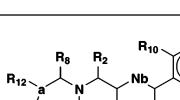
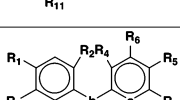
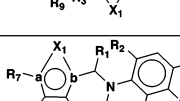
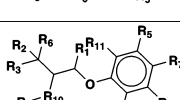
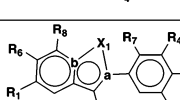
Table 9. MDDR GPCR Activity Key Codes Associated with Chart 3

MDDR activity key	biogenic amine group key	biogenic amine key	MDDR activity key	biogenic amine group key	biogenic amine key
5-HT1A_agonist	1	1	adrenergic_beta1_agonist	7	22
5-HT1A_antagonist	2	2	adrenergic_beta_agonist	8	23
5-HT1B_agonist	3	3	adrenergic_beta_blocker	9	24
5-HT1C_agonist	4	4	adrenoceptor_(beta3)_agonist	1	25
5-HT1D_agonist	5	5	adrenoceptor_alpha1_agonist	2	26
5-HT1D_antagonist	6	6	adrenoceptor_alpha2_antagonist	3	27
5-HT1F_agonist	7	7	anticholinergic	1	28
5-HT1_agonist	8	8	anticholinergic_ophthalmic	2	29
5-HT2A_antagonist	9	9	antihistaminic	1	30
5-HT2B_antagonist	10	10	antimuscarinic	1	31
5-HT2C_antagonist	11	11	dopamine_(D1)_agonist	1	32
5-HT2_antagonist	12	12	dopamine_(D1)_antagonist	2	33
5-HT4_agonist	13	13	dopamine_(D2)_agonist	3	34
5-HT4_antagonist	14	14	dopamine_(D2)_antagonist	4	35
5-HT_antagonist	15	15	dopamine_(D3)_antagonist	5	36
adrenergic_ophthalmic	1	16	dopamine_(D4)_antagonist	6	37
adrenergic_alpha1_blocker	2	17	dopamine_agonist	7	38
adrenergic_alpha2_agonist	3	18	H2_antagonist	1	39
adrenergic_alpha2_blocker	4	19	muscarinic_(M1)_agonist	1	40
adrenergic_alpha_blocker	5	20	muscarinic_(M2)_Antagonist	2	41
adrenergic_beta1blocker	6	21	muscarinic_M3_antagonist	3	42
	peptide binding group key	peptide binding key		peptide binding group key	peptide binding key
anaphylatoxin_receptor_antagonist	1	1	IL-8_inhibitor	9	22
angiotensin_IIblocker	2	2	neurokinin_agonist	10	23
angiotensin_II_AT1_antagonist	2	3	neurokinin_antagonist	10	24
angiotensin_II_AT2_antagonist	2	4	neurokinin_NK2_antagonist	10	25
bombesin_antagonist	3	5	Rneurokinin_NK3_antagonist	10	26
bradykinin_antagonist	4	6	neuropeptide_Y_antagonist	11	27
bradykinin_BK1_antagonist	4	7	neurotensin_receptor_antagonist	12	28
bradykinin_BK2_antagonist	4	8	opioid_agonist	12	29
CCK_A_agonist	5	9	opioid_mixed_agonistnantagonist	13	30
CCK_A_antagonist	5	10	oxytocin	14	31
CCK_agonist	5	11	oxytocin_antagonist	14	32
CCK_antagonist	5	12	somatostatin_analog	15	33
CCK_B_agonist	5	13	somatostatin_antagonist	15	34
CCK_B_antagonist	5	14	substance_P_antagonist	16	35
endothelin_agonist	6	15	vasopressin_antagonist	17	36
endothelin_antagonist	6	16	vasopressin_V1_antagonist	17	37
endothelin_ETA_antagonist	6	17	vasopressin_V2_antagonist	17	38
endothelin_ETB_antagonist	6	18	delta_agonist	18	39
galanin_antagonist	7	19	kappa_agonist	18	40
gastrin-releasing_peptide_antagonist	8	20	mu_agonist	18	41
gastrin_agonist	8	21			
	other GPCR group key	other GPCR key		other GPCR group key	other GPCR key
adenosine_(A1)_agonist	5	1	LHRH_antagonist	2	11
adenosine_(A1)_antagonist	5	2	melatonin_agonist	4	12
adenosine_(A2)_agonist	5	3	melatonin_antagonist	4	13
adenosine_(A2)_antagonist	5	4	P2T_purinoreceptor_pntagonist	5	14
adenosine_A3_antagonist	5	5	PAF_analog	6	15
cannabinoid_agonist	1	6	PAF_antagonist	6	16
FSH	3	7	PGE2_antagonist	7	17
GHR_promoting_agent	8	8	prostaglandin	7	18
gonadotropin	3	9	TRH_analog	8	19
LHRH_agonist	2	10	thromboxane_antagonist	7	20

Similarly, the serine protease compound set filled 92% of the target gated ion channel substructure classes while the ion channel compounds filled only 11% of the serine protease substructure classes. When these percentage occupancies are compared with those from the 10k random compound set, it becomes apparent that the higher percentages are an indication of nonprivileged substructure classes. Although there are hints that some target-family-privileged substructures may in fact exist in these sets, it is clear that many of the substructures found are likely to have the potential to be promiscuous across target families.

Digging deeper into the analysis, it is informative to look at specific substructures from a target family and determine how many compounds from other target families also had the same structure. Substructures for which the other target-family occupancies are high cannot be considered to be target-family-privileged. Some examples of GPCR fragments and their nonGPCR occupancies are shown in Chart 1. As discussed above, the biphenyl substructure (compound 1) is found quite

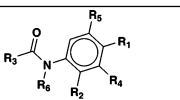
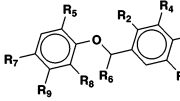
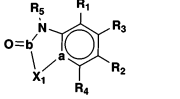
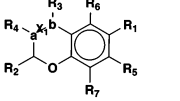
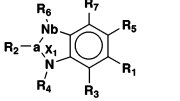
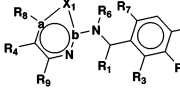
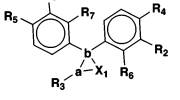
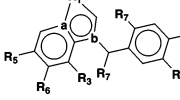
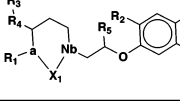
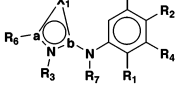
Chart 1. Examples of Substructures Generated from the Class A G-coupled Protein Receptor Ligand Set and Their Occurrences in Other Target Families^a

	ID	GPCR	Kinase	Protease	Ion Channel	NHR
	1	778	1	92	22	18
	2	607	2	33	12	60
	3	391	44		138	5
	4	299		13	40	1
	5	270	2	5	99	39
	6	266	20	70	3	11
	7	248	6	36	33	3
	8	234	1	1	91	61
	9	170	15	37	13	6
	10	166	1	38	27	6
	11	155	2	16	1	88

^a Attachment points are indicated by R groups. In the rings, variations are indicated by a–b connection points for X groups.

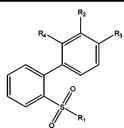
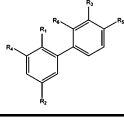
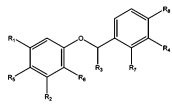
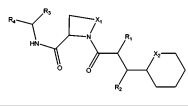
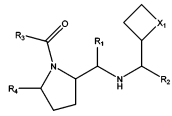
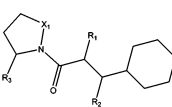
frequently in the GPCR ligands. However, it is not exclusive to the GPCRs, for it is quite common to the other target classes. Further inspection of the data in Chart 1 reveals that those substructures defined as privileged on the basis of their frequency within the GPCR target class are in fact common elements of the kinase, protease, ion channel, and NHR ligands. One must consider substructures with somewhat more functionality before privilege is observed as shown in the serine protease examples discussed below.

Intra-Target Family Selectivity. Some sample results from the serine protease analysis are visualized in Chart 2. For clarity, the eight targets or activity keys were numbered: 1. Chymotrypsin_inhibitor, 2. complement_inhibitor, 3. elastase_inhibitor,

	ID	GPCR	Kinase	Protease	Ion Channel	NHR
	12	153	25	7	73	6
	13	131	20	62		33
	14	116	45		307	10
	15	114	4	7	121	36
	16	77	16	5	87	6
	17	53	30	30	40	2
	18	48		8	87	2
	19	40	2	6	25	58
	20	28		6	1	134
	21	20	187	1	1	

4. factor_VIIa_inhibitor, 5. factor_Xa_inhibitor, 6. thrombin_inhibitor, 7. trypsin_inhibitor, and 8. tryptase_inhibitor. In the examples shown, ClassPharmer class (or substructure) 89, a phenyl benzyl ether, is found to be nonselective across the targets. It is also found in GPCR literature as a privileged substructure.⁶ Similarly, class 12 biphenyl is nonselective and also a literature-GPCR-privileged substructure.⁶ In fact, this substructure was found experimentally to bind to a wide range of proteins.³⁰ In contrast, ClassPharmer classes 16 and 162 are selective for factor Xa. In fact, if R1 = O in structure 162, then the resulting keto does indeed yield one of the classic warheads³¹ for the P1 pocket and may be considered a privileged substructure. Similarly, for class 16, if R2 = acyl, then this somewhat

Chart 2. Serine-Protease-Substructure Selectivity Examples^a

Class substructure	Class ID	Compounds in class	Class Compounds per Activity Key*							
			1	2	3	4	5	6	7	8
	2	103				9	94			
	12	92		1		8	78	5		
	89	62	5		24	1	7	3	21	1
	16	61							61	
	25	51					1	50		
	162	51							51	
			1	2	3	4	5	6	7	8

^a The activity keys are: 1. chymotrypsin_inhibitor, 2. complement_inhibitor, 3. elastase_inhibitor, 4. factor_VIIa_inhibitor, 5. factor_Xa_inhibitor, 6. thrombin_inhibitor, 7. trypsin_inhibitor, and 8. tryptase_inhibitor

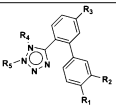
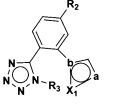
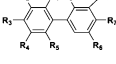
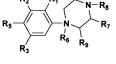
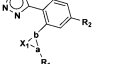
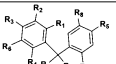
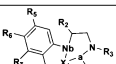
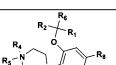
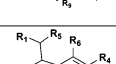
more functionalized structure may in fact be a privileged substructure. Class 2 is not selective for a single target but is selective for both factor targets. Because these targets have very high homology,³² this is not surprising. Class 25 appears selective for factor Xa and thrombin. Whether this selectivity is real or an artifact of target-family-based design is unclear. Rational design has been extensively employed for this target family because numerous X-ray crystal structures are available.³³ Where they are not, homology modeling of the target is a reasonable design method, as is the modification of known serine-protease-ligand chemotypes for selectivity for the desired target.

The examples from the GPCR analysis are visualized in Chart 3. The classes found ranged from selective (a few keys hit) to nonselective (a wide range of keys hit). The ligands were divided according to target ontology into biogenic amines, peptide binding groups, and others for this analysis. There were 42 activity keys found in MDDR for biogenic amines (bioamine key) and 41 activity keys for peptide-binding (pepbindingkey) subfamilies, respectively. To simplify the charts, these keys were grouped according to ontology into sets resulting in 15 superkeys for the biogenic amines and 18 superkeys for the peptide-binding family. In the examples shown, the ligand classes associated with the peptide-binding protein targets appeared more selective than those of the biogenic amines. This can be rationalized by the differing regions that give rise to ligand affinity within the GPCRs. The biogenic amines derive their binding affinity from specific interactions between conserved

residues within the trans-membrane bundle and functionality on the small ligands.²⁹ However, in the peptide-binding family, very little binding affinity can be ascribed to interactions between residues within the bundle and the much larger endogenous peptide ligands.²⁹ Instead, a majority of the binding affinity arises from a large number of interactions between the loops and the residues on the surface of the peptide. Although there are specific interactions between the termini of the ligand and residues within the bundle, they do not generally contribute to affinity. Instead, they are critical for the activation of the receptor and would be predicted to be specific for a given receptor family. The small-molecule ligands that have been discovered for the peptide-binding GPCRs bind within this variable trans-membrane region and tend to be more selective toward a particular family.

GPCR Library Selectivity. Roughly 98% of library A compounds fell into structure classes defined by MDDR class A GPCRs compared to 45% of the random compound set selected from the corporate database. Sixty-seven GPCR substructure classes were represented but four classes were found in 99% of the library compounds. Because the library was designed using frameworks-analysis-derived privileged substructures to filter/bias the reagent sets defining the R groups, these results are unsurprising if the synthetic chemist used the reagents specified as GPCR-like in the actual synthesis. The four substructure classes that represented most of the library contained phenyl rings; again an unsurprising result for a synthesis that used reagents such as anilines and phenols.

Chart 3. G-Coupled-Protein-Receptor-Substructure Selectivity Examples^a

structure	No of cpds	Biogenic amine group keys:															Peptide binding group keys:																		other	
		number of compounds per structure															number of compounds per structure																			
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18		
	994																196																			28
	799																782																			17
	778	9	3	1	1	9	36			19	1			9	3	604					50			3	8						2	20	1	10		
	706	110	221	13	51	61	165	2		44	1	1	18		110	221														6		2	1	1	1	8
	676																667																			9
	607	191	19	22	6			3	2	25			3		191	13			4	61			1	3	16					94	48					
	519	92	151	6	39	37	136	2		40	1	1	6		92										1	5				1	1			100		
	418	64	219	2	31	14	78	2		2					64											1			1	3	1					
	391	27	21	13	2	159	23	3	57	13	2	2		13	27	4			1	1				7	10						2			31		

^a Classes found range from selective (a few keys hit) to nonselective (a wide range of keys hit). The ligands were divided according to target ontology into biogenic amines, peptide-binding groups, and others for this analysis. There were 42 activity keys found in MMDR for biogenic amines and 41 activity keys for peptide-binding subfamilies, respectively. To simplify the histograms, these keys were merged according to ontology into sets resulting in 15 group keys for the biogenic amines and 18 group keys for the peptide-binding family. See Table 9 for details on the keys.

However, as can be seen in Tables 2–6, significant percentages of the library also fell into structure classes defined by other MDDR target families. For ligand-gated ion channels, 70% of the compounds (vs 24% of the random) matched substructure classes. An examination of the nineteen substructures involved revealed that this high percentage, compared to the GPCR dataset, was due to the incorporation of part of the proprietary core in some of the substructures. For NHRs, 22% of the library compounds versus 13% of the random compounds were found. This percentage is higher than that for any of the target families filtered through this set of substructures and was represented by only three substructures. Again, an examination of the substructures revealed that two were similar to those found for the ion channels and were a result of the proprietary core. Correspondingly, 23% of the compounds versus 13% of the random compounds were found to match the protein-kinase substructures, and 48 versus 17% of the random compounds matched the serine-protease substructures. The protein kinases were represented by three substructures, and the serine proteases covered eight substructures. In all cases, the percentage of compounds matching target-family substructures was both higher than that of random compounds and higher than that of any other target family. Clearly, the substructures represented in the library were not truly G-coupled-protein-receptor selective, although they were found to be GPCR-privileged structures

by a standard methodology. This was due to both the proprietary core and the promiscuity of the GPCR-privileged structure containing reagents actually used for the library. Most of the GPCR substructures with significant library populations had a wide range of associated MDDR activity keys, and the library was found to hit numerous HTS screens (unpublished results). Although the library may have been useful for general screening and did tend to yield actives for GPCR targets, it did not meet the intended design criteria of a GPCR-focused library.

Conclusions

It is clear from this analysis that care must be taken to validate the selectivity of a potential target-family-privileged substructure across target families. Generating maximal common substructures and tabulating intra-target family occurrences within a drug database is insufficient. So-called target-family-privileged substructures may occur with high frequency among the ligands of a particular target family but may not in reality be selective for that family.

Substructures that are not selective for a particular target family pose potential risks for off-target liabilities outside the desired target family. On the positive side, this lack of selectivity may be an asset in combinatorial library design if the resultant libraries are screened against a wide variety of targets in search

of leads and if selectivity may be introduced by appropriate R-group elaboration.

Obviously, certain chemical fragments occur with high frequency in commercial drug databases, and they provide a useful tool for designing drug-like compounds. Why they occur in such high frequencies is clearly a topic for debate. It is entirely possible that their frequency is, at least in part, merely an artifact of the fragments found in commonly available commercial reagent sets. The observed frequencies may also derive from analogue design methods such as those of Topliss³⁴ or Hansch and Free-Wilson³⁵ that medicinal chemists use for the follow-up of structure–activity relationships and the available synthetic reactions for core and/or substituent variation. Analyses of these possibilities are beyond the scope of this work but provide areas for further study.

The significance of putative target-family-privileged substructures should be examined with regard to the actual receptors where possible. Unfortunately, correlation of the promiscuity and selectivity of these fragments to specific interactions within the GPCRs has been hindered by a lack of the crystal structures of integral membrane proteins. However, the crystal structure of bovine rhodopsin³⁶ has provided a template from which the homology models of class A GPCRs can be built. One article has appeared that deduces the conserved set of mostly aromatic amino acids within the 5HT₆, MC₄, GHS, and AG₂ receptors and relates them to the binding of a small set of GPCR-privileged structures.³⁷ This analysis will likely be expanded to include many more receptors soon.

It has been well established within several enzyme classes that privilege can exist through warheads that participate in specific interactions within the protein or ions and cofactors such as the hydroxamates in MMPs³⁸ and the benzamidines for serine proteases.³⁹ However, our results demonstrate that these fragments are indeed rare and may be difficult to identify utilizing techniques that merely analyze fragment frequency. In general, the high frequency fragments tend to be fairly rigid substructures and are often aromatic. This makes sense when one considers the nature of drug receptors. Hydrophobic pockets are commonplace, and π -stacking with phenylalanines and tyrosines is commonly observed. Many so-called privileged substructures from target-family-ligand-fragment analysis might be better described as drug-like or receptor privileged rather than target-family-privileged substructures.

In conclusion, our analysis of target family ligands in MDDR supports the original definition of privileged substructures by Evans³ and as reviewed by Patchett⁴ but contradicts the common assumption that privileged substructures are target-family selective.

Acknowledgment. We thank the members of the BMS CADD group, Patrick Lam and Ruth Wexler (BMS chemistry), Brad Pearce (BMS New Leads), Professor Robert Pearlman of the University of Texas at Austin, and Roy Vaz, (formerly with BMS, currently, with Sanofi-Aventis) for helpful discussions, and Patricia Bacha of Bioreason, Inc. for technical assistance with ClassPharmer.

References

- Horton, D. A.; Bourne, G. T.; Smythe, M. L. The combinatorial synthesis of bicyclic privileged structure or privileged substructures. *Chem. Rev.* **2003**, *103*, 893–930.
- Classpharmer is available from Bioreason, Inc., 3900 Paseo del Sol Santa Fe, NM 87507, <http://bioreason.com/>
- Evans, B. E.; Rittle, K. E.; Bock, M. G.; Dipardo, R. M.; Freidinger, R. M.; Whiter, W. L.; Lundell, G. F.; Veber, D. F.; Anderson, P. S.; Chang, R. S. L.; Lotti, V. J.; Cerino, D. J.; Chen, T. B.; Kling, P. J.; Kunkel, K. A.; Springer, J. P.; Hirshfield, J. Methods for drug discovery-development of potent, selective, orally effective cholecystokinin antagonists. *J. Med. Chem.* **1988**, *31*, 2235–2246.
- Patchett, A.; Nargund, R. P. Privileged structures – an update. *Annu. Rep. Med. Chem.* **2000**, *35*, 289–298.
- Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.
- Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *J. Med. Chem.* **1999**, *42*, 3251–3264.
- Schnur, D. M. and Hermsmeier, M. Recent approaches to target class design. Presented at the 36th Mid Atlantic Regional Meeting of the American Chemical Society, Princeton, NJ, 2003.
- Schnur, D. S.; Beno, B. R.; Good A.; Tebben, A.; Approaches to target class library design. In *Chemoinformatics: Concepts, Methods and Tools for Drug Discovery, Methods in Molecular Biology*; Bajorath, J., Ed.; Humana Press: Totowa, NJ, **2004**; pp 355–377.
- Lamb, M. L.; Bradley, E. K.; Beaton, G.; Bondy, S. S.; Castellino, A. J.; Gibbons, P. A.; Suto, M. J.; Grootenhuis, P. D. J. Design of a gene family screening library targeting G-protein coupled receptors. *J. Mol. Graphics Modell.* **2004**, *23*, 15–21.
- Mueller, G. Medicinal chemistry of target family directed masterkeys. *Drug Discovery Today* **2003**, *8*, 681–691.
- Bleicher, K. H.; Green, L. G.; Martin, R. E.; Rogers-Evans, M. Ligand identification for G-protein-coupled receptors: A lead generation perspective. *Curr. Opin. Chem. Biol.* **2004**, *8*, 287–296.
- Merlot, C.; Domine, D.; Cleva, C.; Church, D. J. Chemical structures in drug discovery. *Drug Discovery Today* **2003**, *8*, 594–602.
- Lowrie, J. F.; Delisle, R. K.; Hobbs, D. W.; Diller, D. J. The different strategies for designing GPCR and kinase targeted libraries. *Comb. Chem. High Throughput Screening* **2004**, *7*, 495–510.
- Schuffenhauer, A.; Zimmermann, J.; Stoop, R.; van der Vyver, J.-J.; Lecchini, S.; Jacoby, E. An ontology for pharmaceutical ligands and its application for in silico screening and library design. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 947–955.
- Aureus Pharmaceuticals, 174, Quai de Jemmapes, 75010 Paris, France; <http://www.aureus-pharma.com>.
- Jubilant Biosys, Ltd., 8575 Window Latch Way, Columbia, MD 21045; <http://www.jubilantbiosys.com>.
- Sertanty Inc., 1735 N. First St. #102, San Jose CA, 95112; <http://www.sertanty.com>.
- Biowisdom Ltd., Babraham Hall, Babraham, Cambridge, CB2 4AT, United Kingdom; <http://www.biowisdom.com/>.
- MDL Drug Data Report (MDDR) available from MDL Information Systems Inc., 14600 Catalina Street, San Leandro CA, 94577; <http://www.mdl.com>.
- Compound structure file format, MDL Information Systems Inc., 14600 Catalina Street, San Leandro CA, 94577; <http://www.mdl.com>.
- Chemicals directory available from MDL Information Systems Inc., 14600 Catalina Street, San Leandro CA, 94577; <http://www.mdl.com>.
- SYBYL available from Tripos, Inc., 1699 South Hanley Road, St. Louis, MO 63144-2319; <http://www.tripos.com>.
- CombiLibMaker, developed by Pearlman et al., University of Texas at Austin, available as a module of SYBYL from Tripos, Inc., 1699 South Hanley Road, St. Louis, MO 63144-2319; <http://www.tripos.com>. More recent and greatly improved versions for windows and linux are available as Optive Benchware from Optive Research, Inc., 12331-A Riata Trace Parkway, Suite 110, Austin, TX 78727; <http://www.optive.com>.
- DiverseSolutions developed by Pearlman et al., University of Texas at Austin, available from Tripos, Inc., 1699 South Hanley Road, St. Louis, MO 63144-2319; <http://www.tripos.com> or from Optive Research, Inc., 12331-A Riata Trace Parkway, Suite 110, Austin, TX 78727; <http://www.optive.com>.
- Daylight Chemical Information Systems, Inc., 27401 Los Altos, Suite 360, Mission Viejo, CA 92691; <http://www.daylight.com/>.
- SYBYL UNITY available from Tripos, Inc., 1699 South Hanley Road, St. Louis, MO 63144-2319; <http://www.tripos.com>.
- Ashton, M.; Charlton, M. H.; Schwarz, M. K.; Thomas, R. J.; Whittaker, M.; The Selection and design of GPCR Ligands. *Comb. Chem. High Throughput Screening* **2004**, *7*, 441–452.
- Aronov, A. M. Predictive in silico modeling for hERG channel blockers. *Drug Discovery Today*, **2005**, *10*, 149–155.
- Fanelli, F.; DeBenedetti, P. G. Computational modeling approaches to structure–function analysis of G protein-coupled receptors. *Chem. Rev.* **2005**, *9*, 3297–3351.

- (30) Hajduk, P. J.; Bures, M.; Praestgaard, J.; Fesik, S. W. Privileged molecules for protein binding identified from NMR-based screening. *J. Med. Chem.* **2000**, *43*, 3443–3447.
- (31) Demuth, H. U. Recent developments in inhibiting cysteine and serine proteases. *J. Enzyme Inhib.* **1990**, *3*, 249–278.
- (32) Oshiro C.; Bradley E. K.; Eksterowicz J.; Evensen E.; Lamb M. L.; Lanctot J. K.; Putta S.; Stanton R.; Grootenhuys P. D. J. Performance of 3D-database molecular docking studies into homology models. *J. Med. Chem.* **2004**, *47*, 764–767.
- (33) Mueller, M. M.; Sperl, S.; Sturzebecher, J.; Bode, W.; Moroder, L. (R)-3-amidinophenylalanine-derived inhibitors of factor Xa with a novel active-site binding mode. *Biol. Chem.* **2002**, *383*, 1185–1191.
- (34) Topliss, J. G. Some observations on classical QSAR. *Perspect. Drug Discovery Des.* **1993**, *1*, 253–268.
- (35) Kubinyi, H. 2D QSAR models: Hansch and Free-Wilson analyses. *Comput. Med. Chem. Drug Discovery* **2004**, 539–570.
- (36) Palczewski, K.; Kumasaka, T.; Hori, T.; Behnke, C. A.; et al. Crystal structure of rhodopsin: A G protein-coupled receptor. *Science* **2000**, *289*, 739–745.
- (37) Bondensgaard, K.; Ankensen, M.; Thogersen, H.; Hansen, B. S.; Wulff, B. S.; Bywater, R. P. Recognition of privileged structures by G-protein coupled receptors. *J. Med. Chem.* **2004**, *47*, 888–899.
- (38) Skiles, J. W.; Gonella, N. C.; Jeng, A. Y. The design, structure and therapeutic application of matrix metalloproteinase inhibitors. *Curr. Med. Chem.* **2001**, *8*, 425–474.
- (39) Pauls, H. W.; Ewing, W. R. The design of competitive, small-molecule inhibitors of coagulation factor Xa. *Curr. Top. Med. Chem.* **2001**, *1*, 83–100.

JM0502900