Statistical Analysis of the Effects of Common Chemical Substituents on Ligand Potency

Philip J. Hajduk* and Daryl R. Sauer

Pharmaceutical Discovery Division, GPRD, Abbott Laboratories, Abbott Park, Illinois 60064-6098

Received July 10, 2007

The results of a statistical analysis of more than 84000 compounds from lead optimization programs against 30 different protein targets is presented, with a focus on the effects that different chemical substituents have on compound potency. It is observed that the potency changes induced by most chemical groups follows a nearly normal distribution centered near zero (i.e., no effect on potency). However, the widths of the distributions vary significantly between different substituents, and these effects cannot be rationalized by simple physicochemical parameters. In addition, certain substituents consistently bias the distribution toward higher or lower potency, suggesting the existence of preferred and nonpreferred chemical groups for lead optimization. The implications of these results for understanding protein–ligand recognition and for enhancing the efficiency and speed of lead optimization will be discussed.

Introduction

Successful drug discovery typically entails the identification of small molecule leads that have desirable in vitro activity and that can be optimized for potency, selectivity, bioavailability, and safety. This often involves the chemical synthesis of hundreds or even thousands of compounds to find a molecule with the proper balance of properties required for preclinical and clinical development. A host of new technologies have emerged that facilitate the rapid parallel synthesis of large compound libraries, including polymer assisted solution phase (PASP) synthesis[1] and microwave-accelerated PASP synthesis.[2] Unfortunately, it is often the case that a candidate compound cannot be identified that is suitable for clinical development, despite the large number of compounds that have been investigated. This is often due to the fact that the core pharmacophore of a particular lead series contains an insurmountable liability.

In light of this difficulty, pharmaceutical scientists are now employing strategies in which multiple lead series need to be pursued so that attrition due to inherent core liabilities can be managed.[3] Thus, it is often the case that two, three, or even four lead series are simultaneously optimized to increase the probability that at least one will yield a development candidate. While this strategy appears to be sound, the synthesis, purification, and assessment of hundreds of compounds from multiple lead series can create an enormous burden on the discovery process. An additional factor is that chemists are now being required to synthesize multimilligram quantities of each chemical entity so that large numbers of assays can be implemented for compound evaluation, including activity and specificity panels, high-throughput ADME analyses, solubility assessments, and other in vitro tools utilized for lead triage.[4] As a result, fewer compounds can be synthesized around each core. With respect to high-throughput parallel synthesis around a given core, computational analyses[5] are often performed to identify a monomer set that will yield small but chemically diverse libraries that are compatible with the available chemistry. This is based on the reasonable assumption that a library that samples greater diversity space will also sample greater potency space and have the best chance of improving activity. However, given the vast size of the potential chemical universe (estimated to be as large as 10^{60} compounds),[6] small libraries of less than 100 compounds hardly begin to cover the depth and breadth of diversity space. Alternative schemes for identifying optimal sets of chemical substituents have been proposed based on the knowledge of either the three-dimensional structure of the protein target or the chemical structures of existing inhibitors and thus capitalize on known pharmacophore elements or binding energetics to bias the selection process.[7] While such exercises can be successful, the models and force fields typically used in these approaches often fail to accurately capture many critical aspects of ligand binding (including solvation, conformational flexibility, etc.) and thus have only modest predictive ability.[8,9] As a result, there continues to be a need for a greater understanding of substituent effects on ligand binding and tools to aid the chemist in choosing more effective substituents for use in hit-to-lead and lead optimization campaigns.

Several analyses have recently been reported that measure the impact of specific chemical transformations on various measured physicochemical properties (e.g., compound solubility, plasma protein binding, etc.).[10,11] Importantly, these studies contain statistics on the range of effects that each chemical modification yields over relatively large numbers of compound pairs. Such statistics enable the medicinal chemist to derive an expectation (or confidence) that a specific chemical modification will yield the desired result. Several cheminformatic approaches have also been described that attempt to catalog chemical modifications that maintain compound potency (e.g., bioisostere replacements, etc). The work by Sheridan is an early example of this, where drugs and drug-like compounds from the MDDR were analyzed for matched pairs that differed by a single chemical transformation.[12] Drug GURU[13] has recently been described that attempts to apply a set of validated transformations to a given input molecule to generate new ideas for chemical synthesis. Both of these analyses relied on the observation that a given chemical transformation was successfully applied to at least one drug target, but no statistics were obtained as to the frequency with which a given modification was successfully incorporated.

Here we present a statistical analysis of activity data on more than 84000 compounds derived from lead optimization cam-

* To whom correspondence should be addressed. Dr. Philip J. Hajduk, Abbott Laboratories, R46Y, AP-10, 100 Abbott Park Road, Abbott Park, IL 60064-3500. Phone: (847) 937-0368. Fax: (847) 938-2478. E-mail: philip.hajduk@abbott.com.

10.1021/jm070838y CCC: $40.75 © 2008 American Chemical Society
Published on Web 01/04/2008

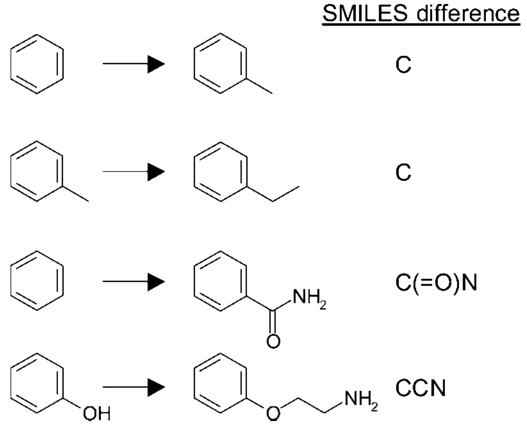

Figure 1. Examples of the types of pairwise comparisons that are utilized in the analysis along with the SMILES string for the substituent that differs.

paigns against 30 different protein targets. Distributions of the potency effects for more than 120 different substituents or transformations allowed for quantitative comparisons of both the average and the global effect of each functionality on ligand binding. It is observed that most modifications, on average, have no significant net effect on compound potency, but large variations in profile widths are obtained. Interestingly, several substituents do appear to bias for or against gains in compound potency, suggesting that certain functionalities might be statistically preferred for protein targets. The results presented here provide insight into the energetics of molecular recognition and suggest alternative approaches for the design and selection of chemical functionalities for use in high-throughput organic synthesis and lead optimization.

Results and Discussion

Data Analysis. Using software tools available from Daylight (see Methods),[14] exhaustive pairwise comparisons of more than 84000 compounds were performed to identify 50127 compound pairs that differed by only a single substituent or could be defined by a simple chemical transformation (see Figure 1). For each modification, potency profiles were then constructed by comparing the potency of the parent (or reference) compound to the modified (or test) compound. The potency change was represented as the base-10 logarithm of the ratio of IC_{50} (or K_I) values of the test compound over the reference compound. As a result, an increase in activity (or decrease in IC_{50} value) imparted by the transformation is indicated by a negative value. Examples of representative potency profiles are shown in Figure 2. While the potency distributions appear to be near-normal and centered near 0.0, the majority fail standard statistical tests for normal distributions (e.g., the Shapiro−Wilk or Anderson−Darling tests). Thus, while averages and standard deviations are listed for each chemical modification, statistical differences between distributions were analyzed using 2 × 2 contingency tests on the cumulative probability of increasing (F(−1.0)) or decreasing (F(1.0)) the potency by 1 log unit relative to the parent compound. Shown in Tables 1−4 are the statistics generated for 127 additions or chemical modifications (a complete listing of each modification with corresponding reference structures is given in the Supporting Information, Table S1, while a file containing all 50,127 transformations used to derive the data in Tables 1−4 is given in Supporting Information, Table S2). Overall, the average effect on potency for these additions was +0.08 log units, and approximately 80% of all 50127 pairwise comparisons exhibited an average value for the potency change within 1 log unit of 0.0.

It is important to note that, unlike metabolic stability or plasma protein binding, the potency profile characteristics listed in Tables 1−4 are the cumulative effects over multiple targets, each of which has its own distinct structure–activity relationship (SAR) and chemotype preference. Thus, measures were implemented to guard against target bias and ensure that the profiles for each substituent are representative of a range of targets. First, a chemical transformation was only included if it was represented in the data for at least 15 of the 30 different protein targets. Second, potency profiles for each modification were iteratively calculated after removing a single target from the data set and re-examining the potency distributions. If removal of data from a single target produced a statistical change (see Methods for details), then the data for that target was removed. Using these criteria, there was surprisingly little bias in the distributions derived from any single target, indicating that the profiles described in this work are likely generally applicable to a wide range of protein targets. Finally, to guard against false discovery, the *p*-values were adjusted using the Benjamini−Hochberg method.[15] Only the corrected *p*-values are listed in Tables 1−4.

As shown in Tables 1−4, many modifications exhibit statistically significant changes in the F(−1.0) and F(1.0) values when compared to those for a methyl group. However, care must be taken to ensure that this result is not simply an anomaly resulting from small sample sizes. Figure 3 shows the distribution of F(−1.0) and F(1.0) values as a function of the number of comparisons (*N*) identified in the data set. As expected, substituents with small *N* exhibit a broader range of distribution values than those with greater representation. Nonetheless, substituents that are statistically differentiable from a methyl group at the 99.9% confidence limit ($p < 0.001$, yellow boxes in Figure 3) are distributed fairly uniformly over the range of sample sizes. This indicates that substituents exhibiting statistically meaningful differences relative to a methyl group are not overly biased toward those with small numbers of comparisons.

Effects of Assay Artifacts. Assay variation and systematic assay errors can potentially confound this analysis and render any observed differences between the various modifications irrelevant. To investigate the dependence of these results on assay error, two analyses were performed. First, random noise was added to the experimental data assuming a Gaussian distribution and a defined standard deviation in the pK_D values. A total of 10 different sets of simulated data were generated for a variety of standard deviations in the pK_D values, from which means and standard errors in each of the distribution parameters were derived. As shown in Figure 4, for F(−1.0) values, only marginal changes are observed when the standard deviation in the pK_D value is less than 0.4 units (approximately a factor of 2.5 in K_D). As expected, there is a trend toward increasing values for this parameter (reflecting a simple broadening of the distribution), but the error ranges on the F(−1.0) values remain small until pK_D errors of 0.5 are simulated. As an assay error of less than 0.4 log units is typically targeted during quality control of in vitro biochemical assays,[16,17] these simulations suggest that normal assay variation will not significantly affect the results of this analysis.

A second analysis attempted to investigate the influence of systematic assay error, which is of course much more difficult to simulate. One potential source of systematic assay error is the influence of compound solubility, such that less soluble compounds can yield inaccurate or widely variable assay responses. The fact that there was little dependence of the

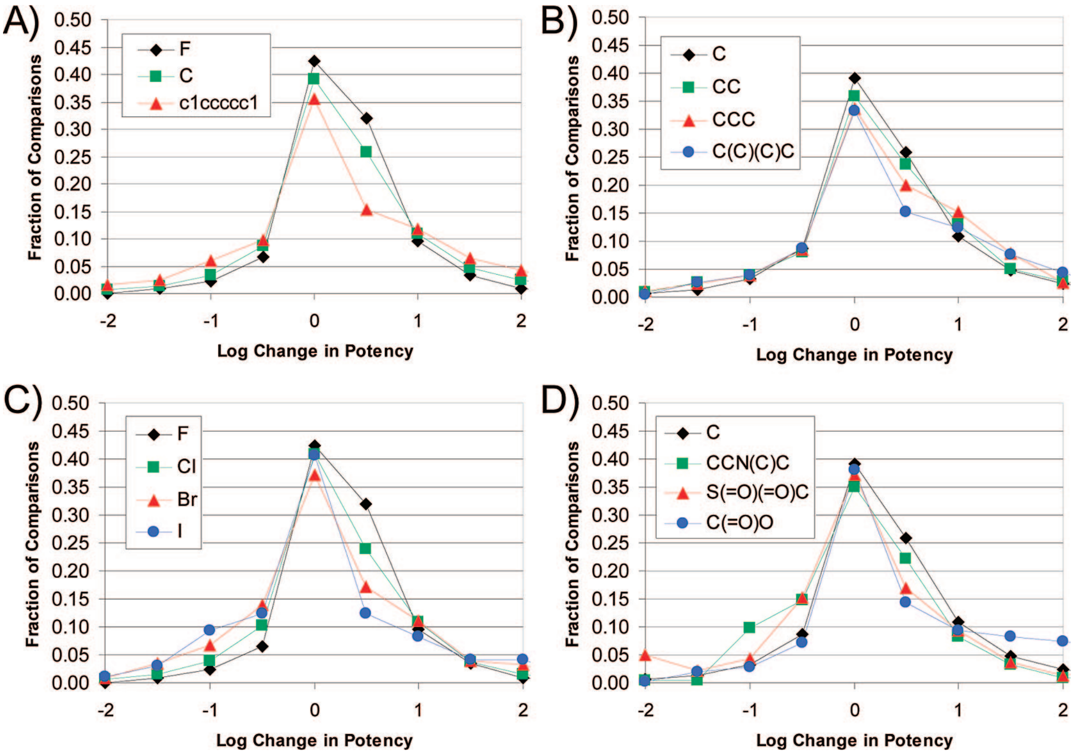

Figure 2. Potency distributions for a subset of substituents from Tables 1−4. Shown on the *x*-axis is the base-10 logarithm of the potency of the daughter compound (with the substituent) divided by the parent compound (without the substituent). Thus, an increase in potency (lower IC_{50} or K_I value) results in a negative value. Distributions are shown for (A) fluoro (entry **1**, black diamonds), methyl (entry **5**, green squares), and phenyl (entry **53**, red triangles) groups; (B) methyl (entry **5**, black diamonds), ethyl (entry **6**, green squares), propyl (entry **7**, red triangles), and *tert*-butyl (entry **12**, blue circles) groups; (C) fluoro (entry **1**, black diamonds), chloro (entry **2**, green squares), bromo (entry **3**, red triangles), and iodo (entry **4**, blue circles) groups; (D) methyl (entry **5**, black diamonds), ethyl dimethylamino (entry **32**, green squares), methyl sulfone (entry **50**, red triangles), and carboxylate (entry **41**, blue circles) groups.

distribution parameters on the change in ClogP imparted by the modification (see below) gives some evidence that perhaps this is not a significant concern with the present data set. To further evaluate this possibility, the data set was subdivided into three categories based on the ClogP of the parent compound (ClogP < 3, ClogP between 3 and 5, and ClogP > 5), and the distribution parameters were recalculated for each subset. As shown in Figure 5, for a subset of well-represented modifications, most substituents exhibit only modest changes in the F(−1.0) parameter when subdivided into separate bins, with a mean absolute change of 1.7 percentile points. However, there does appear to be a small but meaningful trend, in that 6 of the 24 substituents show net increases in the F(−1.0) value of more than 2 percentiles in the context of a parent molecule with ClogP < 3, while 10 of the 24 substituents show net decreases in the F(−1.0) value of more than 2 percentiles in the context of a parent molecule with ClogP > 5. This may indicate a small but real bias resulting from very lipophilic compounds.

Group Additions. Tables 1 and 2 contain data on 84 examples of single group additions to an otherwise identical parent molecule. While the average effects for most of these additions are near zero, significant variations in the widths of the distributions were observed, as illustrated graphically in Figure 2. For example, in Figure 2A, the standard deviation (SD) for the distributions becomes increasingly larger as one moves from fluoro (entry **1**, SD = 0.52) to methyl (entry **5**, SD = 0.71) to phenyl (entry **53**, SD = 1.0), even though the net effect on potency for these substituents is very small. For comparison, it is observed that adding a methyl group (entry **5**) has a 5.3% and 9.2% frequency of increasing or decreasing potency, respectively. In contrast, a phenyl group (entry **53**) has a statistically significant increased chance of both increasing and decreasing potency (10.9 and 15.8%, respectively) relative to a methyl group, whereas a fluoro group (entry **1**) exhibits decreased frequencies (3.2 and 4.7%). This makes intuitive sense as the bulkier substituents have the greatest potential for large potency gains, but also incur the largest potency losses when inappropriately placed on a core. A similar, though less dramatic, trend is observed for the series methyl (entry **5**, SD = 0.71), ethyl (entry **6**, SD = 0.79), *n*-propyl (entry **7**, SD = 0.86), and *t*-butyl (entry **12**, SD = 1.19; Figure 2B). For the halogens (Figure 2C), there is a small but distinct leftward (i.e., increasing potency) shift from fluoro (entry **1**) to chloro (entry **2**) to bromo (entry **3**) to iodo (entry **4**). While the shift is small, it is statistically significant. In fact, 11.6% of all compounds exhibited a potency gain of at least 10-fold when a bromo group was added to the molecule, as compared to only 3.2% upon the addition of a fluoro group, which is a statistically significant change ($p < 0.001$, Table 1).

A certain number of substituents exhibited striking and unexpected deviations from average effects. For example, as shown in Figure 2D and Table 1, a dimethyl amino group (entry **29**) exhibited a roughly average frequency of increasing potency by 1 log unit (F(−1.0) = 0.052), unless it was spaced from the core compound by an ethylene (entry **32**, F(−1.0) = 0.130) or propylene (entry **33**, F(−1.0) = 0.121) linker, where statistically significant increases were observed. Compared to a methyl group (entry **5**, F(−1.0) = 0.053), these cationic groups were twice as likely to yield potency gains of at least 10-fold when optimally spaced from the reference core. The sulfur-containing thiomethyl (entry **23**, F(−1.0) = 0.156), sulfone (entry **50**, F(−1.0) = 0.130), and sulfonamide (entry **51**, F(−1.0) = 0.157)

Table 1. Potency Distribution Descriptors for Acyclic Substituents

No.	modification[a]	*N*[b]	*M*[c]	avg[d]	std dev[e]	F(−1.0)[f]	*p*[g]	F(1.0)[h]	*p*[i]	ΔMW[j]	ΔClogP[k]	ΔPSA[l]	D[m]	B[n]
1	F	2587	30	0.054	0.518	0.032	***	0.047	***	18.0	0.1	0.0	1	
2	Cl	3885	30	0.004	0.657	0.064	*	0.061	***	34.4	0.7	0.0	1	
3	Br	1048	29	−0.081	0.812	0.116	***	0.082		78.9	0.9	0.0	1	1
4	I	95	21	−0.006	0.929	0.137	***	0.116		125.9	1.1	0.0		
5	C	9867	30	0.093	0.708	0.053	ref	0.092	ref	14.0	0.5	0.0		
6	CC	1425	29	0.090	0.785	0.079	***	0.100		28.1	1.0	0.0	1	
7	CCC	503	28	0.170	0.863	0.076	*	0.139	***	42.1	1.6	0.0		
8	CCCC	233	26	0.122	0.979	0.094	**	0.155	**	56.1	2.1	0.0		
9	CCCCC	73	18	0.145	0.838	0.082		0.178	*	70.2	2.6	0.0		
10	C(C)C	528	29	0.162	0.950	0.104	***	0.133	**	42.1	1.4	0.0		
11	CC(C)C	172	25	0.146	1.054	0.134	***	0.180	***	56.1	2.0	0.0		1
12	C(C)(C)C	251	27	0.194	1.186	0.076		0.163	***	56.1	1.8	0.0		
13	C(F)(F)F	1141	29	0.067	0.848	0.115	***	0.123	***	68.0	0.9	0.0	1	1
14	C═C	161	20	0.019	0.926	0.124	***	0.124		26.0	0.7	0.0		1
15	CC═C	84	19	0.171	1.056	0.107	*	0.214	***	40.1	1.1	0.0		1
16	OC	2941	30	0.102	0.623	0.041	*	0.083		30.0	−0.1	9.2	1	
17	OCC	195	27	0.023	0.718	0.062		0.077		44.1	0.4	9.2		
18	OC(C)C	57	16	−0.021	1.033	0.140	**	0.105		58.1	0.8	9.2		1
19	OCCOC	50	17	0.181	0.703	0.060		0.120		74.1	−0.2	18.5		
20	OC(F)(F)F	245	27	−0.010	0.851	0.122	***	0.127		84.0	1.0	9.2		1
21	COC	221	26	0.136	0.642	0.045		0.077		44.1	−0.2	9.2		
22	CCOC	130	21	0.179	0.638	0.038		0.115		58.1	−0.1	9.2		
23	SC	128	23	−0.180	0.804	0.156	***	0.055		46.1	0.6	0.0		1
24	O	1447	30	0.193	0.792	0.054		0.135	***	16.0	−0.7	20.2		
25	CO	490	27	0.210	0.747	0.029	*	0.108		30.0	−1.0	20.2		
26	CCO	211	24	0.082	0.609	0.043		0.071		44.1	−0.8	20.2		
27	CCCO	65	16	0.093	0.564	0.046		0.062		58.1	−0.4	20.2		
28	N	652	27	0.089	0.673	0.057		0.083		15.0	−1.2	26.0		
29	N(C)C	324	29	0.153	0.695	0.052		0.086		43.1	0.2	3.2	1	
30	CN	77	22	0.189	0.824	0.052		0.104		29.1	−1.0	26.0		
31	CN(C)C	243	25	0.197	0.761	0.070		0.148	**	57.1	−0.2	3.2	1	
32	CCN(C)C	215	20	−0.150	0.801	0.130	***	0.060		71.1	0.0	3.2		1
33	CCCN(C)C	66	15	−0.202	0.672	0.121	*	0.045		85.2	0.4	3.2		
34	CCN(CC)CC	58	15	0.209	0.840	0.034		0.103		99.2	1.0	3.2		
35	C(═O)N	305	25	0.136	0.874	0.069		0.161	***	43.0	−1.5	43.1		
36	C(═O)NC	53	15	0.241	0.817	0.019		0.245	***	57.1	−1.3	29.1		
37	C(═O)N(C)C	94	19	0.080	0.862	0.106	*	0.117		71.1	−1.5	20.3		
38	C(═O)NCC	88	17	0.053	0.850	0.080		0.125		71.1	−0.8	29.1		
39	CC(═O)N	75	17	0.077	0.652	0.027		0.093		57.1	−1.7	43.1		
40	NC(═O)C	172	25	0.184	0.785	0.058		0.174	***	57.1	−1.0	29.1		
41	C(═O)O	498	26	0.389	1.023	0.056		0.247	***	44.0	−0.3	37.3		
42	CC(═O)O	133	21	0.372	0.828	0.023		0.165	**	58.0	−0.7	37.3		
43	C(═O)OC	333	27	0.182	0.756	0.051		0.138	**	58.0	0.0	26.3		
44	C(═O)OCC	193	27	0.351	0.849	0.026		0.166	***	72.1	0.5	26.3		
45	CC(═O)OCC	55	15	0.249	0.901	0.073		0.182	*	86.1	0.2	26.3	1	
46	C═O	58	20	0.141	0.689	0.069		0.103		28.0	−0.6	17.1	1	
47	C(═O)C	467	29	0.046	0.748	0.071		0.105		42.0	−0.6	17.1		
48	C≡N	679	30	0.033	0.767	0.078	**	0.097		25.0	−0.6	23.8		
49	CC≡N	75	20	0.275	0.797	0.040		0.187	**	39.0	−0.6	23.8		
50	S(═O)(═O)C	277	26	−0.138	0.951	0.130	***	0.083		78.1	−1.6	34.1	1	1
51	S(═O)(═O)N	51	16	−0.261	0.648	0.157	**	0.039		79.1	−1.8	60.2		1
52	S(═O)(═O)N(C)C	65	15	0.498	0.773	0.000		0.308	***	107.1	−0.8	37.4	1	

[a] A comprehensive listing of each modification along with representative structures is given in the Supporting Information. [b] Number of pairwise comparisons used in the analysis. [c] Number of targets represented in the comparisons. [d] Average value for the distribution. [e] Standard deviation of the distribution. [f] Cumulative frequency of achieving at least a 10-fold gain upon modification. [g] *p*-Value of the F(−1.0) descriptor from a 2 × 2 contingency test relative to addition of a methyl group, where *p*-values less than 0.05, 0.01, and 0.001 are denoted by one, two, and three asterisks, respectively. [h] Cumulative frequency of achieving at least a 10-fold loss upon modification. [i] *p*-Value of the F(1.0) descriptor from a 2 × 2 contingency test relative to addition of a methyl group. [j] Change in molecular weight upon modification. [k] Change in ClogP upon modification, as described in the Methods section. [l] Change in polar surface area upon modification, as described in the Methods section. [m] Denotes whether the substituent was chosen as part of the 24-compound diversity set, as described in the text. [n] Denotes whether the substituent was chosen as part of the 24-compound biased set, as described in the text.

groups also exhibited significant average increases in potency, while the dimethylated sulfonamide (entry **52**, F(−1.0) = 0.0) did not. Addition of a carboxylate group (e.g., entries **41**, F(1.0) = 0.247, and **42**, F(1.0) = 0.165) or an amide group (e.g., entries **35**, F(1.0) = 0.161, and **36**, F(1.0) = 0.245) significantly increased the chances of *losing* potency by more than a log unit compared to a methyl group (entry **5**, F(1.0) = 0.092).

Regiospecific Phenyl Substitutions. While the substituents shown in Tables 1 and 2 are for additions to a molecule without regard to regioselectivity, we also investigated the effects of multiple site-specific substitutions on a phenyl group (see Table 3). While only three types of substituents (methyl, methoxy, and chloro groups) were significantly represented in the database, the trends observed for these substituents given in Table 1 are maintained. For example, the frequencies of achieving 10-fold gains in potency for dimethyl substitutions (entries **85–89**) are generally higher than that observed for dimethoxy substitutions (entries **90–95**) but lower than dichloro substitutions (entries **96–101**). Interestingly, there is a general trend that 2,4- and 3,4-substitution patterns yield higher values for F(−1.0) than the others, regardless of the type of substituent.

Table 2. Potency Distribution Descriptors for Cyclic Substituents[a]

No.	modification	*N*	*M*	avg	std dev	F(−1.0)	*p*	F(1.0)	*p*	ΔMW	ΔClogP	ΔPSA	D	B
53	c1ccccc1	1395	30	0.129	1.008	0.109	***	0.158	***	76.1	1.9	0.0		1
54	c1ccccn1	107	26	−0.002	0.715	0.084		0.065		77.1	0.6	12.9		
55	c1cccnc1	186	23	−0.033	0.955	0.145	***	0.129		77.1	0.4	12.9		1
56	c1ccncc1	152	27	−0.101	0.814	0.118	***	0.079		77.1	0.4	12.9	1	1
57	c1ccco1	53	16	−0.416	1.184	0.226	***	0.057		66.1	1.3	13.1	1	1
58	c1cccs1	102	23	−0.001	0.957	0.176	***	0.127		82.1	1.7	0.0	1	1
59	c1ccsc1	62	18	−0.148	1.326	0.290	***	0.194	**	82.1	1.5	0.0	1	
60	Cc1ccccc1	593	30	0.014	1.011	0.152	***	0.138	***	90.1	2.1	0.0		1
61	CCc1ccccc1	90	25	−0.011	0.934	0.144	***	0.100		104.2	2.4	0.0		1
62	Oc1ccccc1	219	26	0.198	0.838	0.023		0.128		92.1	2.1	9.2		
63	OCc1ccccc1	180	26	0.373	0.890	0.056		0.244	***	106.1	1.7	9.2		
64	C(=O)c1ccccc1	110	23	0.001	0.919	0.118	**	0.118		104.1	1.0	17.1		
65	C(=O)Nc1ccccc1	72	23	−0.140	1.136	0.222	***	0.139		119.1	0.5	29.1	1	1
66	C(=O)OCc1ccccc1	93	20	0.654	0.919	0.043		0.419	***	134.1	1.8	26.3		
67	S(=O)(=O)c1ccccc1	65	18	−0.214	1.138	0.215	***	0.092		140.2	0.3	34.1		1
68	Cc1ccccn1	65	20	0.013	0.742	0.092		0.062		91.1	0.6	12.9	1	
69	Cc1cccnc1	74	18	0.129	0.888	0.081		0.149		91.1	0.6	12.9		
70	Cc1ccncc1	65	20	−0.018	0.836	0.123	*	0.092		91.1	0.6	12.9		
71	C1CC1	121	23	0.008	0.767	0.107	**	0.058		40.1	0.9	0.0		1
72	C1CCCC1	91	23	0.040	1.081	0.154	***	0.154	*	68.1	2.1	0.0		1
73	C1CCCCC1	139	24	−0.042	0.876	0.137	***	0.101		82.2	2.6	0.0		
74	CC1CC1	83	16	0.270	0.943	0.084		0.133		54.1	1.5	0.0		
75	CC1CCCCC1	56	18	−0.401	1.629	0.339	***	0.179	*	96.2	3.1	0.0		
76	N1CCCC1	87	21	0.112	0.863	0.080		0.115		69.1	0.3	3.2	1	
77	N1CCCCC1	64	20	0.150	1.194	0.078		0.109		83.2	0.8	3.2		
78	N1CCOCC1	182	23	0.136	0.882	0.071		0.137	*	85.1	−0.5	12.5	1	
79	N1CCN(C)CC1	61	19	0.156	0.901	0.066		0.131		98.2	0.0	6.5	1	
80	CN1CCCC1	70	18	0.051	0.782	0.100		0.100		83.2	0.5	3.2	1	
81	CN1CCCCC1	66	16	0.046	0.607	0.015		0.091		97.2	1.0	3.2		
82	CN1CCOCC1	183	26	0.205	0.719	0.049		0.137	*	99.2	−0.3	12.5	1	
83	CCN1CCOCC1	152	20	0.066	0.696	0.053		0.112		113.2	0.0	12.5		
84	C(=O)N1CCOCC1	65	16	0.135	0.645	0.031		0.092		113.1	−1.4	29.5		

[a] Column headings are as described for Table 1.

Table 3. Potency Distribution Descriptors for Regiospecific Phenyl Substitutions[a]

No.	modification	*N*	*M*	avg	std dev	F(−1.0)	*p*	F(1.0)	*p*	ΔMW	ΔClogP	ΔPSA	D	B
85	2,3-dimethyl	56	16	−0.138	0.468	0.054		0.000	*	28.1	0.9	0.0	1	
86	2,4-dimethyl	84	16	−0.097	0.618	0.107	*	0.024	*	28.1	1.0	0.0		
87	2,5-dimethyl	83	17	0.087	0.506	0.024		0.036		28.1	1.0	0.0		
88	3,4-dimethyl	87	18	−0.082	0.632	0.103	*	0.023	*	28.1	0.9	0.0		
89	3,5-dimethyl	74	18	0.005	0.458	0.027		0.027		28.1	1.0	0.0		
90	2,3-dimethoxy	57	16	0.128	0.464	0.000		0.035		60.1	−0.3	18.5		
91	2,4-dimethoxy	98	22	0.090	0.580	0.051		0.061		60.1	0.0	18.5		
92	2,5-dimethoxy	83	21	0.229	0.581	0.036		0.108		60.1	0.0	18.5		
93	3,4-dimethoxy	125	23	0.158	0.564	0.024		0.096		60.1	−0.3	18.5		
94	3,5-dimethoxy	77	19	0.173	0.490	0.026		0.039		60.1	0.0	18.5		
95	3,4,5-trimethoxy	72	20	0.365	0.689	0.028		0.153		90.1	−0.7	27.7		
96	2,3-dichloro	111	21	−0.085	0.809	0.126	***	0.072		68.9	1.3	0.0		1
97	2,4-dichloro	129	24	−0.057	0.744	0.124	***	0.039	*	68.9	1.4	0.0		1
98	2,5-dichloro	73	17	−0.052	0.672	0.082		0.014	*	68.9	1.4	0.0		
99	2,6-dichloro	57	17	0.015	0.661	0.088		0.105		68.9	1.4	0.0		
100	3,4-dichloro	158	24	−0.025	0.678	0.101	**	0.070		68.9	1.3	0.0		
101	3,5-dichloro	87	20	0.124	0.775	0.080		0.161	*	68.9	1.4	0.0		

[a] Column headings are as described for Table 1.

Group Transformations. Table 4 contains data on 26 of the most common chemical transformations that were represented in the database. Many of these transformations exhibited very narrow distributions (as evidenced by small values for the standard deviation, F(−1.0), and F(1.0)), consistent with the conservative nature of the changes. For example, the distributions for changing an ether to a thioether (entry **107**), a sulfur to an ethylene (entry **109**), a bromo to a trifluoromethyl (entry **117**), or a benzene to thiophene (entry **124**) all exhibited standard deviations less than 0.5 log units (as compared to 0.71 log units for addition of a methyl group). In contrast, replacing a nitrogen with a carbon in either an aromatic (entry **103**) or nonaromatic (entry **102**) context exhibited a significantly broader distribution (standard deviations greater than 0.75 log units) and a roughly equal chance of either increasing or decreasing potency, while replacing a nitrogen with an oxygen (entry **104**) was twice as likely to decrease potency by 10-fold rather than gain potency (realizing that nearly 85% of these nitrogen to oxygen changes maintained potency values within 1 log unit of the parent). Interestingly, while the transformation from benzene to thiophene (entry **124**) was highly conservative, more than 40% of the conversions from pyridine to thiazole (entry **126**) resulted in at least a 10-fold decrease in potency.

Physical Basis for Substituent Effects. As illustrated in Figure 6, there is no strong correlation between the distribution descriptors (e.g., average F(−1.0) and F(1.0) values) and simple physicochemical properties such as molecular weight and ClogP. While weak trends can be observed within specific series, as described above (e.g., bulkier substituents such as the acyclic alkyl groups have slightly broader distributions), this is not a

Table 4. Potency Distribution Descriptors for Group Transformations[a]

No.	modification	*N*	*M*	avg	std dev	F(−1.0)	*p*	F(1.0)	*p*	ΔMW	ΔClogP	ΔPSA
102	c_to_n_aliphatic	800	30	−0.016	0.761	0.093	***	0.080		1.0	−2.2	12.0
103	c_to_n_aromatic	2592	30	−0.027	0.751	0.088	***	0.075	**	1.0	−1.5	12.9
104	nh_to_o	353	25	0.035	0.716	0.062		0.091		1.0	0.4	−2.8
105	nh_to_s	59	16	−0.075	0.408	0.034		0.017		17.0	1.4	−12.0
106	ether_to_methylene	839	30	−0.092	0.586	0.079	**	0.046	***	−2.0	1.8	−9.2
107	ether_to_thioether	261	29	0.030	0.491	0.027		0.038	**	16.1	1.0	−9.2
108	hydroxy_to_carbonyl	147	26	0.100	0.511	0.034		0.041	*	−2.0	−0.2	−3.2
109	sulfide_to_ethylene	70	19	0.146	0.485	0.014		0.057		−18.0	0.8	0.0
110	sulfide_to_sulfone	169	25	0.173	0.655	0.036		0.130		32.0	−2.6	34.1
111	amide_to_ester	93	20	−0.147	0.691	0.118	**	0.065		1.0	1.3	−2.8
112	amide_to_sulfonamide	296	28	0.150	0.569	0.017	**	0.091		36.1	0.1	17.1
113	amide_to_urea	269	26	−0.009	0.711	0.078		0.078		1.0	0.4	12.0
114	amide_to_retroamide	390	23	0.048	0.505	0.021	**	0.044	**	0.0	0.3	0.0
115	olefin_saturation	514	29	0.081	0.564	0.033	*	0.068		2.0	0.5	0.0
116	olefin_to_amide	77	17	0.293	0.654	0.026		0.130		17.0	−3.1	29.1
117	bromo_to_trifluoromethyl	357	27	0.073	0.480	0.011	***	0.048	**	−10.9	0.0	0.0
118	methyl_to_trifluoromethyl	602	29	0.115	0.572	0.033	*	0.085		54.0	0.4	0.0
119	nitro_to_trifluoromethyl	191	26	0.077	0.515	0.031		0.047	*	23.0	1.1	−43.1
120	carbamate_to_urea	76	17	−0.081	0.827	0.118	*	0.092		−1.0	−0.5	2.8
121	carbonyl_to_sulfone	75	20	−0.169	0.591	0.080		0.027		36.1	−1.2	17.1
122	carboxylate_to_amide	191	23	0.017	1.202	0.037		0.031	**	−1.0	−1.2	5.8
123	benzene_to_cyclohexane	410	29	0.152	0.648	0.037		0.117		6.1	1.2	0.0
124	benzene_to_thiophene	785	30	−0.027	0.439	0.018	***	0.028	***	6.0	−0.4	0.0
125	benzene_to_thiazole	218	27	0.223	0.689	0.032		0.124		7.0	−1.7	12.9
126	pyridine_to_thiazole	56	16	0.835	1.188	0.036		0.411	***	6.0	−0.2	0.0
127	thiophene_to_thiazole	84	23	0.207	0.581	0.036		0.095		1.0	−1.3	12.9

[a] Column headings are as described for Table 1.

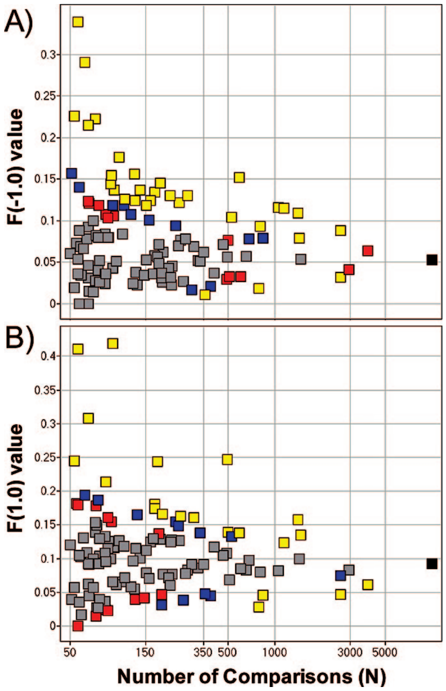

Figure 3. Dependence of the (A) F(−1.0) and (B) F(1.0) values for each descriptor on the number of comparisons identified in the data set. Substituents that achieve *p*-values of less than 0.05, 0.01, and 0.001 (as listed in Tables 1−4) are colored red, blue, and yellow, respectively. The data point for a methyl group is colored black in both panels.

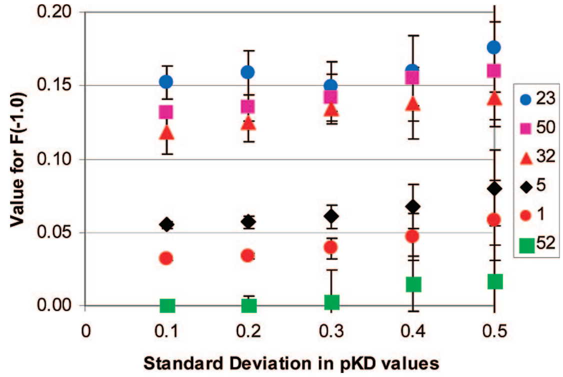

Figure 4. Dependence of the F(−1.0) values for six representative modifications on the level of noise in the assay results. For each modification, assay noise was simulated by changing the pK_D values using a Gaussian error distribution with a defined standard deviation. A total of 10 sets of simulated data were generated for each value of the standard deviation (ranging from 0.1 to 0.5 pK_D units), and the average (symbols) and standard error (error bars) in F(−1.0) was derived from these 10 sets. Shown are data for SC (**23**, blue circles), S(=O)(=O)C (**50**, magenta squares), CCN(C)C (**32**, red triangles), C (**5**, black diamonds), F (**1**, red circles), and S(=O)(=O)N(C)C (**52**, green squares).

general effect. Polarity and conformational flexibility likely play a role in the differences observed for the alkyl and alkoxy series. For example, compared to the ethyl (entry **6**), *n*-propyl (entry **7**), and *n*-butyl (entry **8**) substituents, the corresponding alkoxy groups (e.g., entries **16**, **17**, and **22**, respectively) all have significantly narrower distributions (see Table 1). It is also interesting to note the trends for the halogen series {F, Cl, Br, I}. Similar to the alkyl series (e.g., **5–9** in Table 1), the standard deviations and F(−1.0) values increase with increasing size and lipophilicity. However, unlike the alkyl series, the distributions for the halogen series move distinctively leftward (toward increasing potency) as one progresses from F to Cl to Br (see Table 1 and Figure 2C). This suggests that the larger halogens have a greater intrinsic propensity to interact with protein targets, and fits nicely with the energetic trend in potential halogen binding for this series.[18,19]

The data for the cationic groups are intriguing and appear to be somewhat generalized effects that are independent of both target type and the exact nature of the tertiary amino group. The physical basis for the progressive shift to increased potency gains as the cationic group is extended from the core becomes even more striking when it is realized that, in many cases, these groups are solvent exposed and do not make direct contact with the target protein. In fact, while many of these groups were

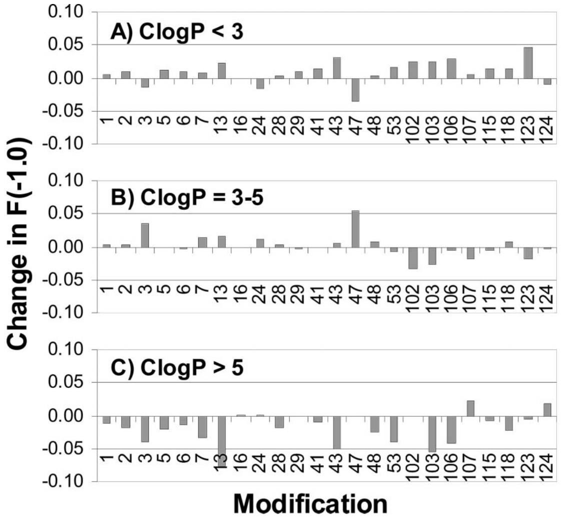

Figure 5. Dependence of the F(−1.0) values for a subset of descriptors on the calculated ClogP of the parent compound. Shown is the change in the F(−1.0) value for the subset as compared to the data set as a whole. Data were subdivided into compounds with ClogP < 3 (19327 pairwise comparisons), Clog P between 3 and 5 (20474 pairwise comparisons), and ClogP > 5 (10326 pairwise comparisons). Modifications are defined by number, as listed in Tables 1−4, and data are only shown for modifications that were represented at least 50 times across at least 15 different targets in each subset.

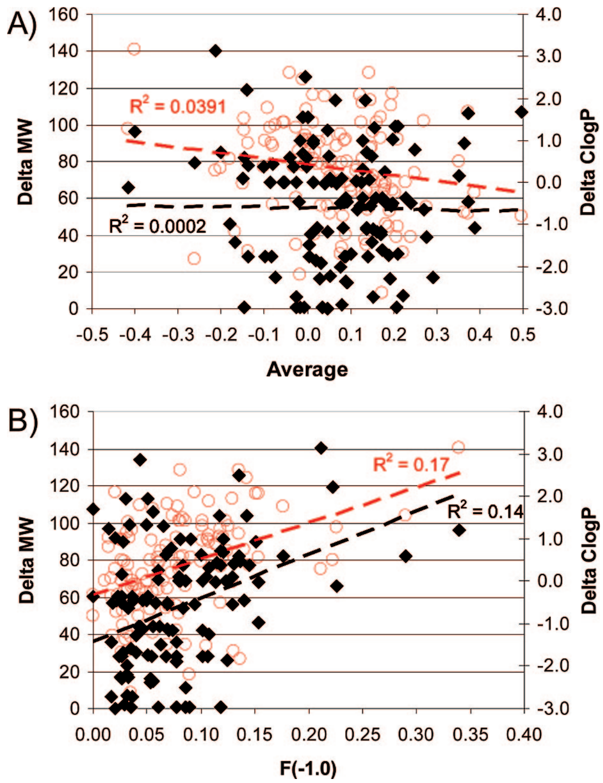

Figure 6. Dependence of the (A) average and (B) F(−1.0) distribution descriptors on the change in molecular weight (filled black diamonds) and ClogP (open red circles) for the modifications shown in Tables 1−4. A linear regression is shown for each case. MW and ClogP calculations were performed as described in Methods.

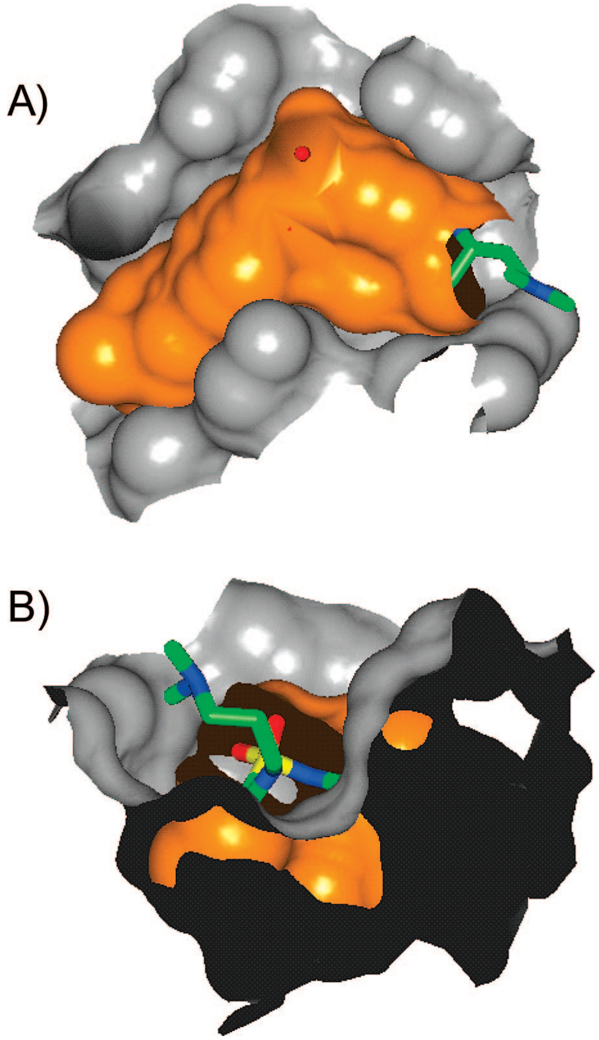

Figure 7. Surface representation of inhibitors (with surfaces shown in orange) that contain tertiary amino groups complexed to (A) Bcl-xL[20,21] and (B) MetAP-2.[22] These inhibitors exhibited significant potency gains (>10-fold) over the parent upon the addition of the cationic substituent (surface removed and colored by atom type). In each case, the dimethylamino substituent is highly solvent exposed.

incorporated with the express purpose of increasing compound solubility, both increased solubility and potency were realized. This is shown in Figure 7 for inhibitors complexed to Bcl-xL[20,21] and MetAP-2,[22] where the tertiary amino group was rationally incorporated to reduce albumin binding but also imparted improved potency. These results can potentially be rationalized by sophisticated electrostatic analyses, in which large and positive electrostatic potentials are formed by the protein in the solvent region. The long-range nature of the electrostatic attraction between the protein and the tertiary amine thus provides a substantial increase in binding energy. However, more generalized and complex physical effects may also play a role. These tertiary amino groups may significantly and beneficially alter the desolvation and intramolecular energies of the ligand and thus play a major role in the final binding energy without actually making contact with the protein.[23] Such a phenomenon is supported by analyzing the context dependence observed for the cationic substituents. For example, for the ethyl dimethylamino group **32**, 13% of the 215 pairs exhibited at least a 10-fold gain in potency when the cationic group was added to the molecule. However, as shown in Figure 8, there was a striking dependence of the potency change on the net charge present on the reference compound. In fact, 27% of all comparisons where the parent molecule carried a formal negative charge exhibited a 10-fold potency gain upon addition of the amino group, whereas only 7.7% of neutral or positively charged molecules exhibited this change.

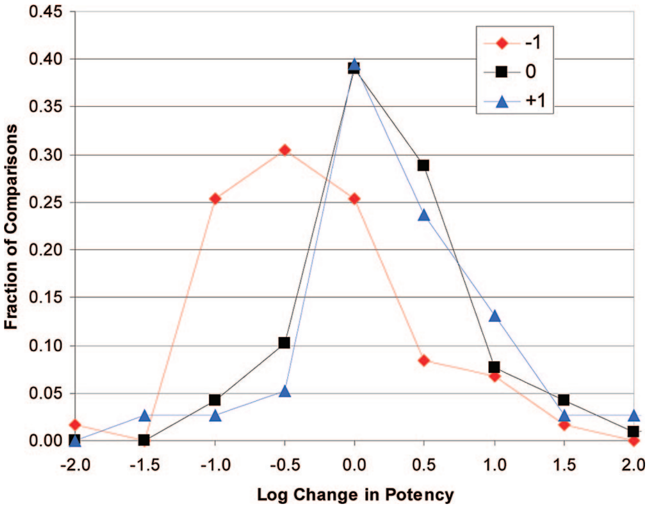

Figure 8. Dependence of the potency distribution profile for the ethyl dimethyl amino substituent (entry **32**) on the net charge of the reference compound lacking the substituent. Parent molecules having a formal net charge of –1, 0, and +1 are represented by red diamonds, black squares, and blue triangles, respectively.

The sulfone-containing substituents are equally intriguing. The sulfone (entry **50**), sulfonamide (entry **51**), and sulfonylphenyl (entry **67**) all exhibit net increases in potency (negative average change in pK_D) and large values for F(−1.0). In many of these cases, the available structural data would suggest that these groups are interacting with at least the periphery if not the core of the binding pocket. These effects are clearly due to the sulfonyl group, as the corresponding ketone (entry **47**), carboxamide (entry **35**), and phenylketone (entry **64**) substituents did not have any remarkable effects on potency relative to either a methyl or phenyl group. The unique properties of sulfur can also be illustrated by comparing the ethyl (entry **6**, F(−1.0) = 0.079), methoxy (entry **16**, F(−1.0) = 0.041), and thiomethoxy (entry **23**, F(−1.0) = 0.156) substituents, where the sulfur-containing thiomethoxy group exhibits a 2-fold increase in the frequency of increasing potency by more than 1 log unit relative to the ethyl group. Most surprising is the clearly opposing effects of adding a sulfonamide (entry **51**, which tends to increase potency) versus a dimethylsulfonamide (entry **52**, which on average decreases potency).

Target Dependence. The distribution descriptors shown in Tables 1−4 are averaged values from all 30 targets represented in the data set. However, it can be expected that certain substituents will be more or less appropriate based on a particular protein family. To investigate this dependence, we again subdivided the data into three target families (9 GPCR targets, 7 kinase targets, and 14 other targets) and recalculated the distribution descriptors. The results for a subset of well-represented modifications are shown in Table 5. Many trends are maintained regardless of the target type. For example, the increasing trend in the F(−1.0) value for {F, Cl, Br} is maintained, even though the absolute magnitude of the value differs between targets. The same trend in F(−1.0) values is also observed for the alkyl series (modifications **5**−**12**), in which larger values for F(−1.0) are generally observed for larger substituents. However, it is clear that the results for certain modifications are highly dependent on the particular protein family that is being targeted. For example, the sulfone substituent (**50**) maintains a high F(−1.0) value for GPCRs (0.20) and other targets (0.15), but is not significantly different from a methyl group (**5**) for kinases. The same is true for the trifluoromethyl (**13**), thiomethyl (**23**), and cyclopropyl (**71**) groups. In contrast, nitrogen (**28**) and morpholino (**78**) substituents appear to have higher F(−1.0) values (especially relative to a methyl group in the same target category) for kinases than the other target types. Thus, while many of the general trends for the modifications hold across many targets, class-dependent differences do exist that may influence the application of these results to any particular protein.

Substituent Correlation. The current analysis simply compares pairs of compounds for effects on potency, without regard to chronology of synthesis or the available SAR that could have influenced the choice of modifications. In reality, lead optimization progresses through rounds of synthesis, analysis of the activity data, choice of new modifications, and so on, such that the results of one modification may strongly influence the choice of the next. This in fact is the basis for the Topliss[24,25] and other analyses that attempt to guide the design of subsequent compounds based on the SAR observed for smaller sets of modifications. To assess the influence of substituent correlation on these results, compound triplets were identified for a set of six highly represented modifications (methyl, fluoro, chloro, hydroxyl, ethyl, and methoxy). A compound triplet is defined here as a parent molecule plus two compounds that have been modified *at the same place* but with different substituents. For example, if a compound has been modified at the same position with both a methyl and a chloro group, the triplet would comprise the parent molecule (where R = H) plus the chloro- and methyl-modified analogs. Table 6 lists the results of this analysis for a subset of modifications where at least 20 triplet comparisons could be identified. To assess the degree of correlation, a simple Pearson correlation coefficient (R) between the two modifications was calculated. Overall, only modest correlation was observed, with an average Pearson correlation coefficient of ~0.45 (corresponding to a coefficient of determination, r^2, of ~0.20). This may simply reflect the intrinsically steric nature of protein binding sites, such that no modifications will be allowed at certain regions (e.g., causing a steric clash), while other positions can tolerate a wide range of substituents. However, certain groups are clearly highly correlated. For example, the effects on potency with a methyl substitution are highly correlated with those observed for thiomethoxy (**23**, R = 0.80) and ethyl (**6**, R = 0.70). As shown in Figure 9, a chloro substitution is highly correlated with addition of a bromo (**3**, R = 0.78) and iodo (**4**, R = 0.78) group, but less correlated with a fluoro group (**1**, R = 0.54). Surprisingly, there are very few anticorrelations, where success with one substituent would predict failure with another (as might be expected for polar vs nonpolar groups). Admittedly, this simple pairwise analysis does not capture the fact that medicinal chemists consider a wide range of SAR data when designing new molecules, such that multiple prior modifications guide the selection of new substituents. Nonetheless, the low overall correlations observed for these triplet combinations suggest that the pairwise comparisons used in this work are not *overly* biased by prior knowledge of structure–activity relationships.

Implications for Lead Optimization. The most obvious finding from our study is that no substituent is perfectly biased to always result in potency gains or losses. Instead, our results illustrate that most substituents exhibit nearly symmetrical, normal distributions centered at or near zero potency change. When substituents do cause a change in potency, the change is usually small, with an exponentially decreasing likelihood of causing larger gains or losses in potency. In fact, considering the data as a whole, the probability of achieving 10-fold gains

Table 5. Dependence of Various Distribution Descriptors on Target Type[a]

		all targets[b]				GPCRs[c]				kinases[d]				other[e]			
No.	modification	*N*	*M*	F(−1.0)	F(1.0)	*N*	*M*	F(−1.0)	F(1.0)	*N*	*M*	F(−1.0)	F(1.0)	*N*	*M*	F(−1.0)	F(1.0)
1	F	2587	30	0.032	0.047	942	9	0.038	0.068	642	7	0.014	0.037	1003	14	0.037	0.033
2	Cl	3885	30	0.064	0.061	1572	9	0.076	0.082	790	7	0.030	0.051	1523	14	0.070	0.044
3	Br	1048	29	0.116	0.082	435	8	0.106	0.117	260	7	0.050	0.062	353	14	0.178	0.054
4	I	95	21	0.137	0.116	32	5	0.250	0.156	31	7	0.000	0.129	32	9	0.156	0.062
5	C	9867	30	0.053	0.092	3591	9	0.079	0.141	2260	7	0.019	0.069	4016	14	0.050	0.061
6	CC	1425	29	0.079	0.100	442	9	0.118	0.124	356	7	0.039	0.096	627	13	0.073	0.086
7	CCC	503	28	0.076	0.139	168	8	0.119	0.214	104	7	0.019	0.135	231	13	0.069	0.087
8	CCCC	233	26	0.094	0.155	70	8	0.129	0.243	33	5	0.030	0.303	130	13	0.092	0.069
10	C(C)C	528	29	0.104	0.133	135	9	0.185	0.148	117	7	0.043	0.128	276	13	0.091	0.127
11	CC(C)C	172	25	0.134	0.180	48	7	0.271	0.208	27	6	0.000	0.148	97	12	0.103	0.175
12	C(C)(C)C	251	27	0.076	0.163	67	8	0.209	0.149	84	7	0.024	0.226	100	11	0.030	0.120
13	C(F)(F)F	1141	29	0.115	0.123	367	9	0.131	0.191	295	7	0.031	0.132	479	13	0.154	0.065
16	OC	2941	30	0.041	0.083	962	9	0.054	0.111	810	7	0.028	0.063	1169	14	0.038	0.073
17	OCC	195	27	0.062	0.077	66	7	0.045	0.076	45	7	0.067	0.044	84	12	0.071	0.095
20	OC(F)(F)F	245	27	0.122	0.127	71	8	0.070	0.211	44	7	0.068	0.091	130	12	0.169	0.092
21	COC	221	26	0.045	0.077	44	8	0.114	0.091	106	6	0.009	0.047	71	12	0.056	0.113
23	SC	128	23	0.156	0.055	38	9	0.263	0.079	36	5	0.028	0.083	54	9	0.167	0.019
24	O	1447	30	0.054	0.135	349	9	0.057	0.249	411	7	0.046	0.039	687	14	0.057	0.135
25	CO	490	27	0.029	0.108	93	8	0.032	0.269	215	7	0.028	0.047	182	12	0.027	0.099
26	CCO	211	24	0.043	0.071	44	6	0.000	0.091	108	7	0.046	0.056	59	11	0.068	0.085
28	N	652	27	0.057	0.083	146	8	0.048	0.192	263	7	0.087	0.049	243	12	0.029	0.053
29	N(C)C	324	29	0.052	0.086	85	8	0.047	0.153	125	7	0.056	0.064	114	14	0.053	0.061
31	CN(C)C	243	25	0.070	0.148	33	8	0.121	0.121	127	6	0.055	0.181	83	11	0.072	0.108
32	CCN(C)C	215	20	0.130	0.060	24	6	0.125	0.167	74	6	0.068	0.095	117	8	0.171	0.017
35	C(=O)N	305	25	0.069	0.161	41	7	0.000	0.366	131	7	0.046	0.122	133	11	0.113	0.135
40	NC(=O)C	172	25	0.058	0.174	42	6	0.048	0.286	59	6	0.051	0.068	71	12	0.070	0.197
41	C(=O)O	498	26	0.056	0.247	38	7	0.053	0.368	131	7	0.069	0.183	329	12	0.052	0.258
43	C(=O)OC	333	27	0.051	0.138	53	8	0.075	0.226	87	7	0.023	0.103	193	12	0.057	0.130
44	C(=O)OCC	193	27	0.026	0.166	41	8	0.024	0.195	79	7	0.000	0.127	73	12	0.055	0.192
47	C(=O)C	467	29	0.071	0.105	105	9	0.057	0.124	172	7	0.035	0.081	190	13	0.111	0.116
48	C≡N	679	30	0.078	0.097	251	9	0.076	0.147	181	7	0.039	0.066	247	14	0.109	0.069
50	S(=O)(=O)C	277	26	0.130	0.083	85	7	0.200	0.082	75	7	0.027	0.080	117	12	0.145	0.085
53	c1ccccc1	1395	30	0.109	0.158	484	9	0.132	0.246	362	7	0.041	0.110	549	14	0.133	0.113
55	c1cccnc1	186	23	0.145	0.129	61	7	0.279	0.148	78	7	0.051	0.154	47	9	0.128	0.064
56	c1ccncc1	152	27	0.118	0.079	35	8	0.143	0.029	72	7	0.069	0.097	45	12	0.178	0.089
60	Cc1ccccc1	593	30	0.152	0.138	168	9	0.202	0.190	131	7	0.107	0.115	294	14	0.143	0.119
62	Oc1ccccc1	219	26	0.023	0.128	55	7	0.018	0.145	70	7	0.000	0.129	94	12	0.043	0.117
63	OCc1ccccc1	180	26	0.056	0.244	44	7	0.091	0.227	73	7	0.027	0.315	63	12	0.063	0.175
71	C1CC1	121	23	0.107	0.058	30	7	0.167	0.100	34	5	0.000	0.059	57	11	0.140	0.035
78	N1CCOCC1	182	23	0.071	0.137	38	5	0.026	0.316	67	6	0.075	0.075	77	12	0.091	0.104

[a] Column headings are as described for Table 1. Descriptions are only shown for those modifications that were represented at least 20 times across at least 5 different targets within each category. [b] Descriptors for all 30 targets are as listed in Tables 1 and 2 (derived from 50127 pairwise comparisons). [c] Descriptors for a subset of nine GPCR targets (derived from 16284 pairwise comparisons). [d] Descriptors for a subset of seven kinase targets (derived from 12884 pairwise comparisons). [e] Descriptors for the remaining 14 targets (derived from 20959 pairwise comparisons).

in potency is 8.5%, and the probability of exceeding 100-fold gains with a single substituent is less than 1%. Interestingly, consistent with earlier reports,[26] this suggests an inherent limit on the maximal impact that a single group can have on the binding energy. For example, the average molecular weight of the substituents used in this analysis is 66 Da. If the probability of achieving a 10-fold gain is taken as a reasonable estimate of success as to whether a given substituent can be incorporated into the parent molecule at the appropriate place and in the appropriate way, then it can be expected that a compound with an increase in potency of 10-fold can be obtained with average increase in mass of 66 Da. This is very close to the estimate of 64 Da per log unit derived from a retrospective analysis of 18 highly optimized inhibitors[26] and also in line with optimal ligand efficiency guidelines of 0.3 kcal/mol per heavy atom.[27]

On the basis of these results, it is possible to construct a set of substituents for parallel synthesis campaigns based not on chemical diversity but on propensity to yield compounds with enhanced activity against protein targets. As an example of this approach, we constructed two sets of 24 substituents each, one based on maximizing chemical diversity and the other based on maximizing the potential for 10-fold or greater potency gains (biased toward the highest F(−1.0) values). The substituents used in each set are indicated in Tables 1−3, and the potency profiles for each set as compared to the average results for all substituents are shown in Figure 10. As illustrated in this figure, there is little difference between the average distribution and the diversity set, where the cumulative chance of achieving 10-fold gains in potency is 8.5 and 9.8%, respectively. For the biased set, there is a distinct (and statistically significant) shift toward increased potency, with a cumulative chance of achieving 10-fold gains in potency of 14.2%. These gains over the average of diversity set can be more clearly visualized in Figure 10B, where it is observed that the biased set exhibits a 40% improvement in achieving at least 10-fold gains in potency, while the chances of increasing potency by 100-fold more than doubles. In the context of high-throughput organic synthesis and lead optimization campaigns involving the synthesis of hundreds of analogs, this improvement can result in a substantial decrease in the number of compounds that need to be synthesized to explore the potential for potency gains at a specific site.

In addition to influencing monomer selection in high-throughput organic synthesis campaigns, it is anticipated that the chemical modifications listed in Tables 1−4 can serve as guides to medicinal chemists during hit-to-lead and lead optimization exercises. During early exploration of potency

Table 6. Correlations between Different Modifications

		fluoro (**1**)		chloro (**2**)		methyl (**5**)		ethyl (**6**)		methoxy (**16**)		hydroxy (**24**)	
No.	modification	#[a]	*R*[b]	#	*R*	#	*R*	#	*R*	#	*R*	#	*R*
1	F	2587	1.00	792	0.54	617	0.45	70	0.57	517	0.37	231	0.52
2	Cl	792	0.54	3885	1.00	838	0.63	115	0.68	663	0.61	185	0.36
3	Br	315	0.54	502	0.78	429	0.66	75	0.70	352	0.65	119	0.23
4	I	26	0.80	42	0.78	31	0.54	-	-	31	0.54	-	-
5	C	617	0.45	838	0.63	9867	1.00	584	0.70	697	0.58	324	0.37
6	CC	70	0.57	115	0.68	584	0.70	1425	1.00	119	0.69	64	0.29
7	CCC	-[c]	-	-	-	211	0.54	169	0.74	44	0.47	27	0.16
8	CCCC	-	-	22	0.54	111	0.65	84	0.54	27	0.64	-	-
9	CCCCC	-	-	-	-	28	0.57	34	0.37	-	-	-	-
10	C(C)C	36	0.22	54	0.55	177	0.64	121	0.80	78	0.52	38	−0.12
11	CC(C)C	-	-	-	-	47	0.33	42	0.51	-	-	-	-
12	C(C)(C)C	35	0.02	53	0.36	89	0.50	42	0.58	53	0.60	22	−0.04
13	C(F)(F)F	368	0.42	487	0.70	384	0.64	79	0.58	333	0.59	114	0.13
14	C=C	21	0.53	49	0.52	96	0.69	62	0.80	22	0.67	-	-
16	OC	517	0.37	663	0.61	697	0.58	119	0.69	2941	1.00	352	0.49
17	OCC	58	0.32	64	0.48	74	0.21	35	0.68	128	0.64	45	0.08
18	OC(C)C	-	-	28	0.28	27	0.24	-	-	39	0.49	22	0.74
19	OCCOC	-	-	-	-	27	0.48	-	-	31	0.83	24	0.38
20	OC(F)(F)F	133	0.54	163	0.71	135	0.59	28	0.77	152	0.47	48	−0.31
21	COC	-	-	-	-	80	0.44	47	0.61	-	-	-	-
23	SC	49	0.39	59	0.84	64	0.80	34	0.81	78	0.72	36	0.38
24	O	231	0.52	185	0.36	324	0.37	64	0.29	352	0.49	1447	1.00
25	CO	45	0.65	42	0.73	219	0.54	61	0.56	61	0.84	100	0.79
26	CCO	-	-	-	-	37	0.40	21	0.78	-	-	-	-
28	N	113	0.27	158	0.35	202	0.43	21	−0.27	131	0.30	145	0.65
29	N(C)C	84	0.18	114	0.41	107	0.47	30	0.68	146	0.55	94	0.55
31	CN(C)C	-	-	-	-	60	0.65	26	0.62	-	-	-	-
35	C(=O)N	43	0.26	49	0.45	84	0.13	-	-	57	0.35	66	0.40
37	C(=O)N(C)C	-	-	-	-	25	0.05	-	-	22	0.07	-	-
40	NC(=O)C	55	−0.01	57	−0.02	59	0.33	-	-	85	0.43	39	0.27
41	C(=O)O	53	0.12	78	0.46	116	0.34	26	0.34	70	0.38	100	0.61
43	C(=O)OC	50	0.64	71	0.51	83	0.66	-	-	71	0.50	55	0.54
44	C(=O)OCC	-	-	-	-	37	0.30	27	0.29	29	0.55	-	-
47	C(=O)C	51	0.57	65	0.68	53	0.48	-	-	61	0.32	25	−0.08
48	C≡N	183	0.58	277	0.46	225	0.31	48	0.48	242	0.50	118	0.68
50	S(=O)(=O)C	60	0.40	70	0.41	55	0.34	-	-	68	0.16	39	0.61
53	c1ccccc1	105	0.28	207	0.38	436	0.51	141	0.58	137	0.51	120	0.53
55	c1cccnc1	-	-	31	0.51	61	0.29	-	-	21	0.48	-	-
56	c1ccncc1	-	-	23	0.69	47	−0.02	-	-	21	0.37	-	-
58	c1cccs1	25	0.49	37	0.44	67	0.47	23	0.66	-	-	-	-
59	c1ccsc1	-	-	26	0.62	26	0.57	-	-	-	-	-	-
60	Cc1ccccc1	-	-	-	-	106	0.64	47	0.55	23	0.67	32	0.33
62	Oc1ccccc1	96	0.38	105	0.46	130	0.40	28	0.55	108	0.34	52	0.13
63	OCc1ccccc1	54	0.07	59	0.70	87	0.51	-	-	79	0.55	72	0.27
71	C1CC1	-	-	-	-	47	0.54	48	0.92	-	-	-	-
73	C1CCCCC1	105	0.28	207	0.38	436	0.51	141	0.58	137	0.51	120	0.53
75	CC1CCCCC1	-	-	-	-	106	0.64	47	0.55	23	0.67	32	0.33
76	N1CCCC1	-	-	-	-	29	0.36	-	-	28	0.53	-	-
78	N1CCOCC1	24	0.32	44	0.40	63	0.28	30	0.38	56	0.59	57	0.70
82	CN1CCOCC1	-	-	-	-	55	0.61	-	-	21	0.58	-	-

[a] Total number of triplet combinations used in the analysis. [b] Pearson correlation coefficient. [c] Less than 20 triplet combinations identified.

space and the development of SAR, those substituents that appear to be biased toward protein targets should be preferentially incorporated into lead molecules in the hopes of rapidly generating meaningful potency gains. However, it is not the purpose of this work to specify exactly which chemical modifications should be made to a given ligand for a particular target. For a given protein, the specific molecular context of the binding site will ultimately dictate the nature of the chemical matter that will be tolerated. Nonetheless, since all of the targets used in this analysis are proteins, it is not unreasonable to expect that there might be some general (albeit weak) preferences for certain chemical groups over others. These biases certainly become weaker as you move from specific target to protein family to "general protein", but they can still be usefully exploited by preferentially incorporating these groups in library design.

As described in the Introduction, several approaches already exist that attempt to catalog chemical transformations that have been successfully applied to protein inhibitors.[12,13] The biased modifications listed in this work can be viewed as additional examples that can be implemented to increase compound potency. Of course, potency is a necessary but insufficient requirement for a drug, and other properties must be considered, especially as optimization progresses. Thus, the integration of potency information with other calculated or experimental properties will be a powerful aid to drug design. This can be easily performed with calculated molecular properties (as listed for molecular weight, ClogP, and polar surface area in Tables 1−4). However, the statistical analyses recently described by the groups from AstraZeneca[10,11] can be used in combination with the current work to further aid the medicinal chemist to simultaneously optimize multiple parameters that impact successful drug discovery.

Conclusions

In summary, we have presented a statistical analysis of the effects that different chemical substituents have on compound

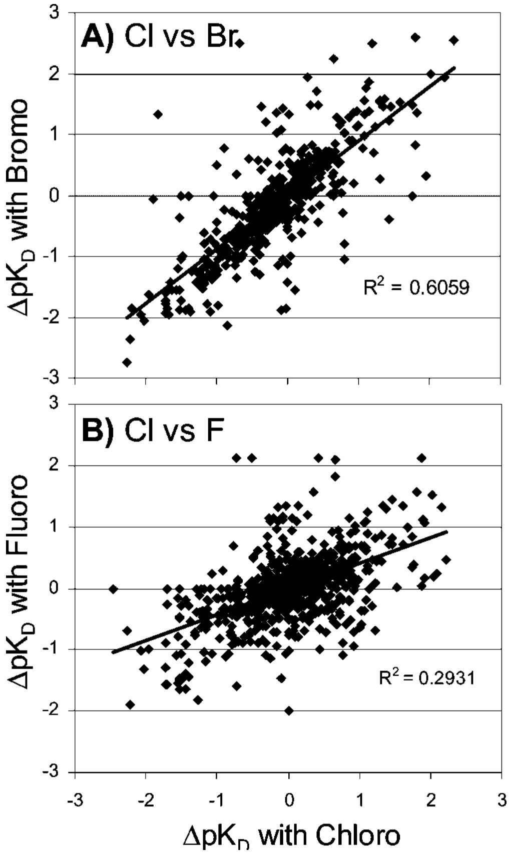

Figure 9. Plot of potency changes observed upon substitution with a chloro group (**2**) vs (A) a bromo group (**5**) and (B) a fluoro group (**1**) at the same position on the parent molecule.

potency. While a nearly normal distribution of potency changes is observed for all substituents, the widths of the distributions vary significantly between different substituents. While 10-fold gains in potency were frequently observed for a single substituent (8.5% of all comparisons), larger gains were very rare. Intriguingly, the average potency change for certain substituents did appear to bias for or against gains in compound potency, suggesting the existence of preferred and nonpreferred chemical groups for lead optimization. This information can be used to construct biased substituent sets that maximize the probability of achieving potency gains with a minimum number of chemical syntheses.

Methods

Data Collection and Curation. Data were derived from our corporate electronic database for the following 30 protein targets: 5-HT1A, ACC-2, Akt-1, Bcl-xL, CB2, Chk-1, Cot kinase, D4, DPP-4, FBPase, farnesyl transferase, Ghrelin, glucocorticoid receptor, HCV polymerase, HCV protease, HDAC, 11b-HSD, Jnk-1, KDR, Lck kinase, MCH, MetAP2, MK2, NNR, P2X7, PARP-1, PTP1B, survivin, V1b, and VR1. All data were obtained from in vitro enzymatic or ligand-competition assays, and IC$_{50}$ and K_I values were used consistently within a given target data set, but no distinction was made between data sets. An average of 2818 compounds were obtained per target, resulting in a total of 84526 compounds with potencies ranges from 0.1 nM to 100 μM. A complete listing of each modification with corresponding reference structures is given in the Supporting Information, Table S1, while a file containing all 50127 transformations used to derive the data in Tables 1−4 is given in Supporting Information, Table S2.

Substituent Identification. Within each target data set, exhaustive pairwise comparisons were performed with the *findsub* routine available from Daylight, version 4.83,[14] (www.daylight.com) in which two SMILES strings were compared and a listing of R-groups generated. When only a single substituent differentiated the two input SMILES strings, the SMILES pair was saved, along with the potency information and substituent data. This analysis was performed individually for each target data set, and then all data sets were concatenated for further analysis. In total, 93824 pairs of compounds were identified that differed by the addition of a single substituent. For regiospecific substitutions on aromatic rings and group transformations, SMIRKS strings were created that defined a list of 828 potential chemical transformations, many of which were derived from our internal Drug GURU application.[13] Each input SMILES string was then exhaustively modified using all 828 transformations and product molecules were searched for in the data set. When a match was found for the product, the SMILES pair was saved, along with the potency information and SMIRKS conversion. This analysis was performed individually for each target data set and then all data sets were concatenated for further analysis. In total, 96269 pairs of compounds were identified that were related by a defined chemical transformation. All Daylight calculations were performed on a Silicon Graphics Octane workstation.

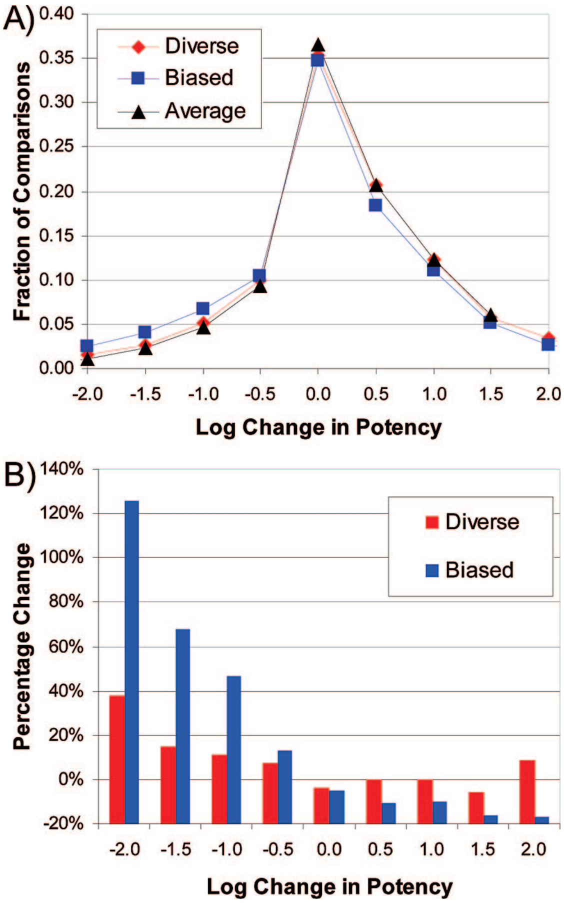

Figure 10. (A) Potency distributions for the average of all substituents (average, black diamonds), a subset of 24 substituents chosen to maximize chemical diversity (diverse, red diamonds), and a subset of 24 substituents chosen to maximize the probability of achieving greater than 10-fold gains in potency (biased, blue squares). (B) Percent changes in the frequency of achieving specified potency changes using a subset of 24 substituents chosen to maximize chemical diversity (red) and a subset of 24 substituents chosen to maximize the probability of achieving greater than 10-fold gains in potency (blue squares) as compared to the average over all substituents.

Data Analysis. For each substituent or transformation identified in the analysis, potency distributions were generated. The potency change was represented as the base-10 logarithm of the ratio of the IC$_{50}$ or K_I value of the test compound (containing the substituent) over the reference compound (lacking the substituent). For each modification, potency changes that were larger than four standard deviations from the average were discarded to minimize the influence of outliers (using this criterion, approximately 1.5% of the data was omitted). To ensure that individual targets did not bias the distributions, potency profiles for each substituent were iteratively calculated after removing a single target from the data set and re-examining the potency distributions. If removal of data from a single target produced a statistical change ($p < 0.05$) in the F(−1.0) value for the overall distribution, then the data for that particular target was removed from the analysis. This resulted in the removal of data for a single target for five modifications: piperidine (entry **77**), *t*-butyl (entry **12**), phenoxy (entry **62**), carboxylate to amide (entry **122**), and n-to-c aromatic (entry **103**). The number of tests (*N*) and targets (*M*) listed for these entries in Tables 1−4 reflects the remaining number of comparisons after discarding the data from the biasing target. Substituents were not included in the analysis if there were not at least 50 comparisons derived from a minimum of 15 different targets. The resulting distributions were analyzed to derive the average, standard deviation, and the cumulative probability of increasing (F(−1.0)) or decreasing (F(1.0)) the potency by 1 log unit relative to the parent compound. Statistically significant changes in the F(−1.0) or F(1.0) values relative to that of a methyl group were assessed using 2 × 2 contingency tests. The resulting *p*-values from these analyses were then corrected for random discovery using the Benjamini−Hochberg method.[15] Only the corrected *p*-values are listed in Tables 1−4.

Calculation of Structural Descriptors. ClogP values[28] were calculated using Biobyte software (www.biobyte.com) in the following manner: Each substituent was incorporated onto a simple biphenyl group and the ClogP of the resulting compound was calculated. Next, the ClogP of biphenyl (calcd ClogP = 4.03) was subtracted to yield a corrected ClogP value for the substituent itself. The analysis was performed in this manner to simulate the actual change in ClogP imparted by the various substituents when added to a core molecule, and a biphenyl was used to ensure regioselectivity for the phenyl substitutions. In a similar manner, hierarchical clustering was performed on the entire set of 101 unique substituents (Tables 1−3) attached to a simple biphenyl group. A final Tanimoto similarity of 0.66 was used to generate 24 clusters from which the parent of each cluster was chosen for inclusion in the diversity set.

Acknowledgment. The authors would like to thank Drs. Kent Stewart, Yvonne Martin, and Jonathan Greer for helpful discussions and critical reading of the manuscript.

Supporting Information Available: A table containing example structures for each modification listed in the text and a table containing all 50127 data points used to generate the distribution parameters. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Hinzen, B. Polymer-supported reagents: Preparation and use in parallel organic synthesis. *Methods Princ. Med. Chem.* **2000**, *9*, 209–237.

(2) Sauer, D. R.; Kalvin, D.; Phelan, K. M. Microwave-assisted synthesis utilizing supported reagents: A rapid and efficient acylation procedure. *Org. Lett.* **2003**, *5* (24), 4721–4724.

(3) Alanine, A.; Nettekoven, M.; Roberts, E.; Thomas, A. W. Lead generation—Enhancing the success of drug discovery by investing in the hit to lead process. *Comb. Chem. High Throughput Screening* **2003**, *6* (1), 51–66.

(4) Bleicher, K. H.; Bohm, H. J.; Muller, K.; Alanine, A. I. Hit and lead generation: Beyond high-throughput screening. *Nat. Rev. Drug Discovery* **2003**, *2* (5), 369–378.

(5) Kitchen, D. B.; Stahura, F. L.; Bajorath, J. Computational techniques for diversity analysis and compound classification. *Mini-Rev. Med. Chem.* **2004**, *4* (10), 1029–1039.

(6) Bohacek, R. S.; McMartin, C.; Guida, W. C. The art and practice of structure-based drug design: a molecular modeling perspective. *Med. Res. Rev.* **1996**, *16* (1), 3–50.

(7) Chen, G.; Zheng, S.; Luo, X.; Shen, J.; Zhu, W.; Liu, H.; Gui, C.; Zhang, J.; Zheng, M.; Puah, C. M.; Chen, K.; Jiang, H. Focused combinatorial library design based on structural diversity, druglikeness and binding affinity score. *J. Comb. Chem.* **2005**, *7* (3), 398–406.

(8) Abagyan, R.; Totrov, M. High-throughput docking for lead generation. *Curr. Opin. Chem. Biol.* **2001**, *5* (4), 375–382.

(9) Bursulaya, B. D.; Totrov, M.; Abagyan, R.; Brooks, C. L., 3rd. Comparative study of several algorithms for flexible ligand docking. *J. Comput.-Aided Mol. Des.* **2003**, *17* (11), 755–763.

(10) Haubertin, D. Y.; Bruneau, P. A database of historically-observed chemical replacements. *J. Chem. Inf. Model.* 2007.

(11) Leach, A. G.; Jones, H. D.; Cosgrove, D. A.; Kenny, P. W.; Ruston, L.; MacFaul, P.; Wood, J. M.; Colclough, N.; Law, B. Matched molecular pairs as a guide in the optimization of pharmaceutical properties; a study of aqueous solubility, plasma protein binding and oral exposure. *J. Med. Chem.* **2006**, *49* (23), 6672–6682.

(12) Sheridan, R. P. The most common chemical replacements in drug-like compounds. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (1), 103–108.

(13) Stewart, K. D.; Shiroda, M.; James, C. A. Drug Guru: A computer software program for drug design using medicinal chemistry rules. *Bioorg. Med. Chem.* **2006**, *14* (20), 7011–7022.

(14) Leo, A.; Weininger, A. *Daylight Chemical Information Systems, 3*; Daylight Chemical Information Systems, Inc.: Aliso Viejo, CA, 1995.

(15) Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc., Ser. B* **1995**, *57*, 289–300.

(16) Eastwood, B. J.; Farmen, M. W.; Iversen, P. W.; Craft, T. J.; Smallwood, J. K.; Garbison, K. E.; Delapp, N. W.; Smith, G. F. The minimum significant ratio: a statistical parameter to characterize the reproducibility of potency estimates from concentration-response assays and estimation by replicate-experiment studies. *J. Biomol. Screening* **2006**, *11* (3), 253–261.

(17) Iversen, P. W.; Eastwood, B. J.; Sittampalam, G. S.; Cox, K. L. A comparison of assay performance measures in screening assays: signal window, Z′ factor, and assay variability ratio. *J. Biomol. Screening* **2006**, *11* (3), 247–252.

(18) Politzer, P.; Murray, J. S.; Concha, M. C. Halogen bonding and the design of new materials: organic bromides, chlorides and perhaps even fluorides as donors. *J. Mol. Model.* **2007**, *13* (6–7), 643–650.

(19) Riley, K. E.; Merz, K. M., Jr. Insights into the strength and origin of halogen bonding: the halobenzene-formaldehyde dimer. *J. Phys. Chem. A* **2007**, *111* (9), 1688–1694.

(20) Oltersdorf, T.; Elmore, S. W.; Shoemaker, A. R.; Armstrong, R. C.; Augeri, D. J.; Belli, B. A.; Bruncko, M.; Deckwerth, T. L.; Dinges, J.; Hajduk, P. J.; Joseph, M. K.; Kitada, S.; Korsmeyer, S. J.; Kunzer, A. R.; Letai, A.; Li, C.; Mitten, M. J.; Nettesheim, D. G.; Ng, S.; Nimmer, P. M.; O'Connor, J. M.; Oleksijew, A.; Petros, A. M.; Reed, J. C.; Shen, W.; Tahir, S. K.; Thompson, C. B.; Tomaselli, K. J.; Wang, B.; Wendt, M. D.; Zhang, H.; Fesik, S. W.; Rosenberg, S. H. An inhibitor of Bcl-2 family proteins induces regression of solid tumours. *Nature* **2005**, *435* (7042), 677–681.

(21) Petros, A. M.; Dinges, J.; Augeri, D. J.; Baumeister, S. A.; Betebenner, D. A.; Bures, M. G.; Elmore, S. W.; Hajduk, P. J.; Joseph, M. K.; Landis, S. K.; Nettesheim, D. G.; Rosenberg, S. H.; Shen, W.; Thomas, S.; Wang, X.; Zanze, I.; Zhang, H.; Fesik, S. W. Discovery of a potent inhibitor of the antiapoptotic protein Bcl-xL from NMR and parallel synthesis. *J. Med. Chem.* **2006**, *49* (2), 656–663.

(22) Sheppard, G. S.; Wang, J.; Kawai, M.; Fidanze, S. D.; BaMaung, N. Y.; Erickson, S. A.; Barnes, D. M.; Tedrow, J. S.; Kolaczkowski, L.; Vasudevan, A.; Park, D. C.; Wang, G. T.; Sanders, W. J.; Mantei, R. A.; Palazzo, F.; Tucker-Garcia, L.; Lou, P.; Zhang, Q.; Park, C. H.; Kim, K. H.; Petros, A.; Olejniczak, E.; Nettesheim, D.; Hajduk, P.; Henkin, J.; Lesniewski, R.; Davidsen, S. K.; Bell, R. L. Discovery and optimization of anthranilic acid sulfonamides as inhibitors of methionine aminopeptidase-2: a structural basis for the reduction of albumin binding. *J. Med. Chem.* **2006**, *49* (13), 3832–3849.

(23) Sims, P. A.; Wong, C. F.; Vuga, D.; McCammon, J. A.; Sefton, B. M. Relative contributions of desolvation, inter- and intramolecular interactions to binding affinity in protein kinase systems. *J. Comput. Chem.* **2005**, *26* (7), 668–681.

(24) Topliss, J. G. Utilization of operational schemes for analog synthesis in drug design. *J. Med. Chem.* **1972**, *15* (10), 1006–1011.

(25) Topliss, J. G. A manual method for applying the Hansch approach to drug design. *J. Med. Chem.* **1977**, *20* (4), 463–469.

(26) Hajduk, P. J. Fragment-based drug design: How big is too big. *J. Med. Chem.* **2006**, *49*, 6972–6976.

(27) Hopkins, A. L.; Groom, C. R.; Alex, A. Ligand efficiency: A useful metric for lead selection. *Drug Discov Today* **2004**, *9*, (10), 430–431.

(28) Leo, A. J. Calculating logP(oct) with no missing fragments: The problem of estimating new interaction parameters. *Perspect. Drug Discovery Des.* **2000**, *18*, 19–38.

JM070838Y