Assessment of the Consistency of Medicinal Chemists in Reviewing Sets of Compounds

Michael S. Lajiness,* Gerald M. Maggiora,[§] and Veerabahu Shanmugasundaram[#]

Computer-Aided Drug Discovery, Pharmacia Corporation, Kalamazoo, Michigan 49008

Received April 5, 2004

Medicinal chemists are frequently asked to review lists of compounds to assess their drug- or leadlike nature and to evaluate the suitability of lead compounds based on their “attractiveness” and/or synthetic feasibility as a basis for launching a drug-discovery campaign. It is often felt that one medicinal chemist’s opinion is as good as any other, but is it? In an attempt to answer this question, an experiment was performed in conjunction with a recent compound acquisition program (CAP) conducted at Pharmacia. Historically, the CAP included a review of many thousands of compounds by medicinal chemists who eliminate anything deemed undesirable for any reason. In a review conducted in 2002, about 22 000 compounds requiring review by medicinal chemists were broken down into 11 lists of approximately 2000 compounds each. Unknown to the medicinal chemists, a subset of 250 compounds, previously rejected by a very experienced senior medicinal chemist, was added to each of the lists. Most of the 13 medicinal chemists who participated in this process reviewed two lists, although some only reviewed a single list and one reviewed three lists. Those compounds that were deemed unacceptable were recorded and tabulated in various ways to assess the consistency of the reviews. It was found that medicinal chemists were not very consistent in the compounds they rejected as being undesirable. The inconsistency arises from the subjective analysis that all humans utilize when considering “data sets” of any kind. This has important implications for pharmaceutical project teams where individual medicinal chemists review lists of primary screening hits to identify those compounds suitable for follow-up. Once a compound is removed from a list, it and other structurally similar compounds are effectively removed from further consideration. This can also have an impact on computational chemists who are developing models for assessing the desirability or attractiveness of different classes of compounds for lead discovery.

Introduction

As part of the 2002 Pharmacia compound acquisition program (CAP) approximately 62 000 high-quality, structurally diverse compounds were selected for potential acquisition using a heuristic approach developed inhouse for this purpose.[1] On the basis of an analysis of these compounds, approximately 22 000 were identified as requiring additional review by medicinal chemists.[1] It should be emphasized that these compounds had already passed a number of standard compound filters designed to eliminate chemically and/or pharmaceutically undesirable compounds.[1] Thirteen medicinal chemists volunteered to review the compounds and to reject those they felt were unsuitable for purchase but were given no specific guidelines on how to accomplish the task.

The ∼22 000 compounds were divided into 11 lists of about 2000 compounds each. It was decided to assess the degree of consistency of chemists’ rejections by adding to each list a set of 250 compounds rejected earlier by a senior medicinal chemist with over 30 years of experience (reviewer 1) who had often participated in the selection of compounds in past years. Of the 13 medicinal chemists who volunteered, 8 reviewed two lists, 1 reviewed three lists, and 4 reviewed one list.

Once the reviewers made selections, the compounds rejected were recorded. Various tabulations were used to compare and contrast the results. The choice was made to focus on rejections rather than acceptances because the former can have a greater influence on research than the latter. Once a compound is rejected, it is effectively eliminated from further consideration; this is somewhat akin to a false negative in high-throughput screening. In addition, compounds similar to a rejected compound may also be, directly or indirectly, removed from further consideration. An accepted compound, on the other hand, will be investigated further; its efficacy can be determined and the compound can either be carried forward or eliminated from further consideration.

The analysis described in this paper focuses on two data sets. The first is related to the consistency among medicinal chemists in rejecting compounds from the same 250-compound set, which is embedded in each of the eleven 2000-compound sets described above. Although the 250-compound set was chosen from the set of compounds rejected by an experienced medicinal chemist (vide supra), this does not introduce significant bias into the results because the analysis is based on pairwise comparisons among all of the medicinal chem-

* To whom correspondence should be addressed. Address: Lilly Research Laboratories, Indianapolis, IN 46285. Phone: 317-651-4079. Fax: 317-276-2441. E-mail: LajinessMS@Lilly.com.

[§] Department of Pharmacology and Toxicology, University of Arizona College of Pharmacy, Tucson, AZ 85721.

[#] Computer-Assisted Drug Discovery, Pfizer Global Research and Development, Ann Arbor, MI 48105.

10.1021/jm049740z CCC: $27.50
Published on Web 08/21/2004

Table 1. Number of Compounds Rejected by Chemists Reviewing the 250-Compound Subset

Reviewer	R1-L1	R1-L9	R2-L1	R2-L7	R2-L10	R3-L11	R3-L8	R4-L11	R4-L2	R5-L3	R5-L7	R6-L4	R6-L5	R7-L2	R8-L10	R9-L3	R9-L6	R10-L4	R10-L6	R11-L8	R12-L5	R12-L10	R13-L1
R1-L1	250																						
R1-L9	192	192																					
R2-L1	64	57	64																				
R2-L7	56	50	38	56																			
R2-L10	90	80	47	50	90																		
R3-L11	76	70	28	23	40	76																	
R3-L8	36	30	15	10	19	25	36																
R4-L11	62	54	24	22	34	37	15	62															
R4-L2	43	35	18	19	29	21	8	38	43														
R5-L3	43	35	15	15	26	23	14	17	10	19													
R5-L7	45	40	19	20	25	22	12	25	18	13	45												
R6-L4	50	45	30	27	39	25	13	22	17	20	15	50											
R6-L5	79	71	40	34	50	39	25	32	21	11	28	43	79										
R7-L2	40	40	18	12	21	24	11	24	17	22	22	11	24	40									
R8-L10	88	71	32	30	50	39	24	36	25	13	24	25	46	23	88								
R9-L3	33	30	18	19	27	19	9	20	19	16	15	15	19	14	24	33							
R9-L6	50	44	24	23	34	25	12	26	23	21	19	23	28	16	31	31	50						
R10-L4	111	100	35	30	56	53	20	38	23	29	27	32	48	22	54	18	28	111					
R10-L6	129	112	37	33	63	63	25	45	27	28	32	35	51	27	57	21	33	101	129				
R11-L8	132	113	44	40	66	54	23	47	32	13	34	37	56	32	63	29	35	70	82	132			
R12-L5	55	51	28	22	33	31	18	26	20	12	23	21	37	22	38	19	23	37	43	40	55		
R12-L10	49	45	26	21	35	25	14	21	17	38	18	22	34	21	35	17	21	35	36	38	36	49	
R13-L1	190	157	58	51	82	73	35	61	43	19	39	47	73	36	76	32	49	98	113	112	51	43	190

ists participating in the study. While it is true that different sets of compounds could lead to different results, it is assumed that they will not differ significantly from those reported here and thus should be reasonably representative of the results one would obtain in general. However, this is clearly an assumption, which could be tested in future studies.

The second data set assesses the consistency among medicinal chemists reviewing the same 2000-compound set. Interestingly, regardless of which data set is examined or how consistency is evaluated, the results obtained are basically the same, namely, that *chemists are inconsistent in the compounds they reject*. While this is not a surprising result, since humans in general have difficulty with tasks that involve visual pattern recognition of large and complex data sets, it nevertheless is the first analysis that attempts to address the issue on a quantitative basis.

Results

250-Compound Set Comparisons. The consistency of medicinal chemists with respect to their reviews of the 250-compound set was evaluated from the results summarized in Table 1. The lower triangular part of the table lists the number of compounds rejected by any two medicinal chemists. The row and column headings are designated by a reviewer-list code. For example, R1-L9 indicates that reviewer 1 (the most senior medicinal chemist) reviewed list 9. Individual cell entries in the table can be used to determine how consistent any two medicinal chemists are with respect to the compounds they rejected in the 250-compound set. For example, the value 34 in the cell in row R12-L10 and column R6-L5 indicates that reviewers 12 and 6 both agree to reject the same 34 compounds out of the 250 compounds that were embedded in lists 10 and 5, respectively. The diagonal entries indicate how many compounds out of the 250-compound set a given medicinal chemist rejected. These values are identical to the values in column R1-L1, which is due to the fact that reviewer 1 rejected all 250 compounds from list 1. The values located in the highlighted boxes lying along the main diagonal of the table correspond to the internal consistency of a given reviewer with his or her own rejections. Consider, for example, reviewer 3, who reviewed lists 8 and 11. When reviewing list 8, reviewer 3 rejected 36 compounds whereas 76 compounds were rejected when reviewing list 11. Only 25 compounds were common to both rejection sets!

A similarity measure can be used to quantitatively assess the degree of consistency among the sets of rejections. In this case, each of the 250 bits of the "compound-review" fingerprint encodes the rejection (bit = 1) or acceptance (bit = 0) of a particular compound.

Reviewer A	..	0	1	1	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	..
Reviewer B	..	1	1	0	0	0	1	0	1	0	0	0	1	0	1	1	1	1	1	..

The Tanimoto similarity coefficient, S_{Tan}, is then computed using the formula

$$S_{\mathrm{Tan}}(\mathrm{A,B}) = \frac{\mathrm{a\&b}}{a + b - \mathrm{a\&b}}$$

where a is the number of rejections from reviewer A, b is the number of rejections from reviewer B, and a&b is the number of rejections in common between reviewers A *and* B. This is identical in form to the Tanimoto similarity coefficient commonly used to assess the structural similarity between pairs of molecules[2,3] represented by molecular fingerprints. Stated in words,

$$S_{\mathrm{Tan}}(\mathrm{A,B}) = \frac{\text{no. of compounds rejected by both reviewers A and B}}{\text{no. of compounds rejected by either reviewer A or B}}$$

For example, the similarity of the rejections made by reviewer 1 on list 9 (R1-L9) compared to those of

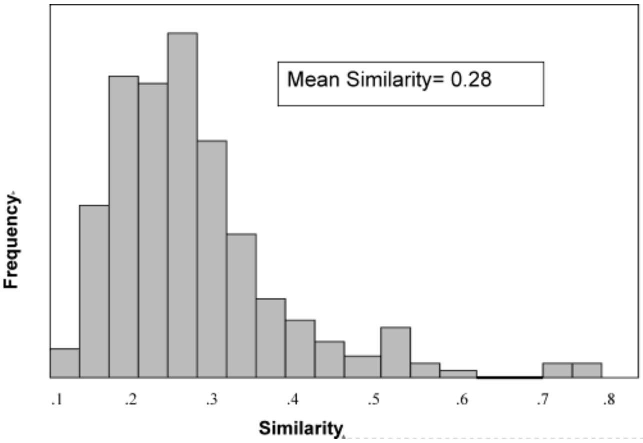

Figure 1. Distribution of similarities among chemists with respect to the compounds 13 chemists agreed should be rejected out of the 250-compound subset.

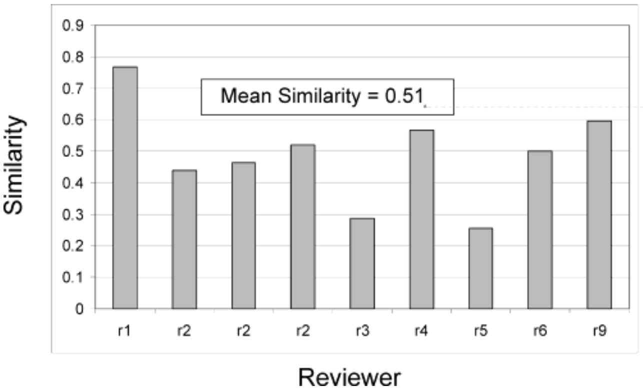

Figure 2. Similarity of chemists to themselves when reviewing the same set of 250 compounds in different lists.

reviewer 2 on list 1 (R2-L1) is given by

$$S_{\mathrm{Tan}}(\text{R1-L9,R2-L1}) = \frac{57}{192 + 64 - 57} = 0.26$$

Calculation of this similarity value for the lower triangular data in Table 1 and plotting the resultant frequency distribution produces the histogram shown in Figure 1. As can be seen in the figure, the pairwise similarities range from a maximum of 0.77 to a minimum of 0.11. The observed mean similarity in rejecting compounds was calculated to be 0.28. This can also be expressed as a percentage. Thus, one would expect that chemists would agree to reject the same compounds about 28% of the time. In other words, if 100 compounds were rejected by either of two chemists they would agree on only 28 of those rejections. Moreover, the frequency distribution of similarity values is skewed toward lower values, indicating that very few consistent results were obtained. It is also interesting to note that reviewers 1 and 2 each have over 25 years of experience in medicinal chemistry and still agree only about 28% of the time.

Also in this study, nine chemists reviewed more than one 2000-compound set. This provided the opportunity to calculate the similarity of a chemist's rejections when reviewing the same 250-compound set embedded in different 2000-compound sets. Thus, S_{Tan} was computed for compounds rejected by the same chemist when viewing them in different lists, and the results are plotted in Figure 2. The similarities ranged from a maximum of 0.77 for reviewer 1 to a minimum of 0.25 for reviewer 5. Reviewer 1 is the most internally consistent of the reviewers. The mean similarity of rejection when chemists review the same list is 0.51 Thus, when a chemist looks at the same set of compounds repeatedly they will tend to reject the same compounds only about 50% of the time.

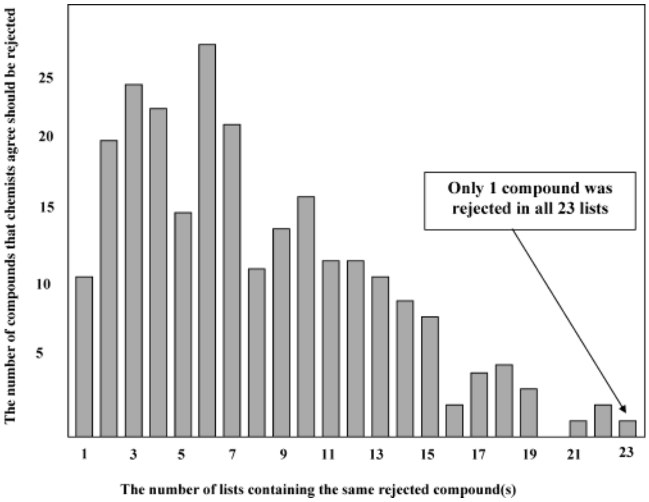

Figure 3. Frequency distribution indicating how often a compound was rejected out of 23 lists.

Another way to look at the results of this 250-compound study is to consider how many times a specific compound was rejected in the 23 reviews that were carried out. Figure 3 depicts a histogram illustrating this. In this figure, one can see for instance that 15 compounds were rejected 5 times or 4 compounds were rejected 17 times. Of more interest is the observation that only 1 compound is rejected in all 23 reviews. Compounds that were consistently rejected at a higher frequency (defined as those rejected in 15 or more reviews) are shown in Figure 4.

2000-Compound Set Comparisons. Ten of the eleven 2000-compound sets were reviewed by two or more medicinal chemists. Sets 1 and 10 were reviewed by three medicinal chemists, set 9 was reviewed by one, and all of the remaining eight sets were reviewed by two medicinal chemists (see Table 2). These compound sets afford the possibility for assessing the number of agreements among rejections by medicinal chemists with respect to much larger sets of compounds. The results are summarized in Table 2. The same reviewer-list code is used as was used in Table 1. For example, the value of 142 in the cell defined by row R13-L1 and column R2-L1 indicates that reviewers 1 and 13 reviewed list 1 and agreed to reject the same 142 compounds. The diagonal values in the table give the number of compounds rejected by each medicinal chemist. There are sometimes vast differences in the number of compounds rejected by different chemists when reviewing the same list. For example, in the review of list 1, reviewer 2 rejected 179 compounds while reviewer 13 rejected 960.

Calculation of S_{Tan} for the rejection data located in the lower triangular region along the diagonal in Table 2 and plotting the resultant frequency distribution yield the graph shown in Figure 5. The pairwise similarities range from a maximum of 0.42 to a minimum of 0.14, with an observed mean similarity of 0.23. Thus, on the

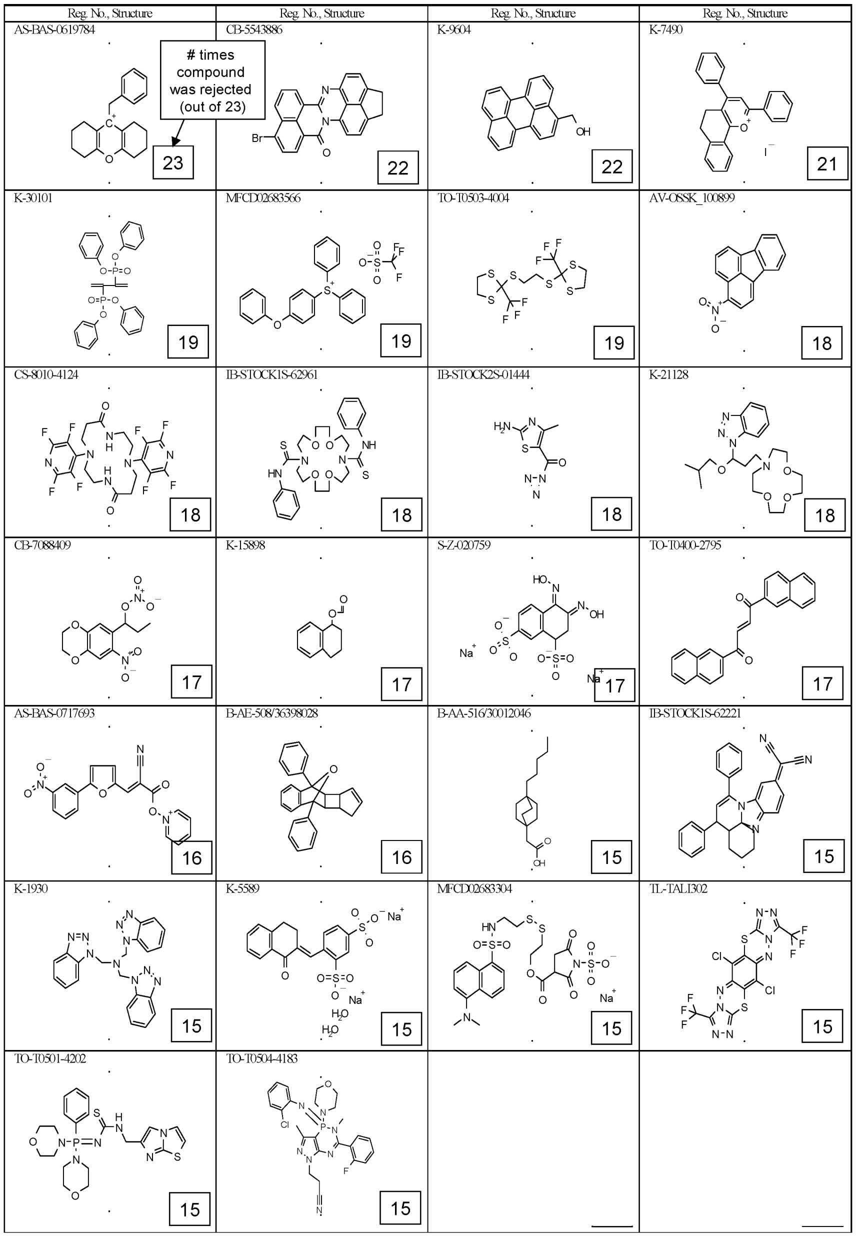

Figure 4. Selected structures of commercially available compounds rejected in 15 or more lists.

basis of a more extensive comparison in terms of numbers of compounds, two medicinal chemists on average only agree to reject the same compounds approximately 23% of the time. A histogram illustrating the similarity of the rejections among medicinal chemists reviewing the same 2000-compound lists is given in Figure 6.

Discussion

As noted earlier, it was decided to focus on the consistency of the rejections among medicinal chemists rather than the consistency of their acceptances or their total agreement (treating matching acceptance or rejection decisions as an agreement). There are a number of reasons for this, but the most compelling reason is the fact that medicinal chemists are often called upon to review lists of active compounds obtained from HTS campaigns, eliminating those deemed unsuitable for further evaluation. In many cases, because of scientific resource limitations, only one or two medicinal chemists are involved. From the results presented here it is clear that the level of consistency observed for rejected compounds is low. This can have serious implications in hit follow-up or lead optimization studies because the choice of which compounds to take forward depends on who is doing the review. As noted earlier, potentially good lead compounds can be lost because of a given medicinal chemist's predilection for or against particular classes of compounds. In addition, the number of compounds accepted is considerably greater than the number of compounds rejected. Thus, including the number of acceptances into a measure of consistency would unduly bias the results, thereby obscuring the differences of opinion that matter most.

The results presented here are very striking and cannot be easily dismissed. Chemists involved in the study had a minimum of 3 years of experience and a maximum of over 25 years. Interestingly, experience had little to do with consistency of opinion. Both reviewers 1 and 2 each had over 25 years of experience,

Table 2. Number of Compounds Rejected by Chemists Reviewing the ~2000-Compound Lists

Reviewer	R1-L1	R2-L1	R13-L1	R4-L2	R7-L2	R5-L3	R9-L3	R6-L4	R10-L4	R6-L5	R12-L5	R9-L6	R10-L6	R2-L7	R5-L7	R3-L8	R11-L8	R1-L9	R2-L10	R8-L10	R12-L10	R3-L11	R4-L11
Total # in list	2143	2143	2143	2189	2189	2174	2174	2174	2174	2176	2176	2175	2175	2173	2173	2199	2199	2173	2173	2173	2173	2169	2169
Total # rejected	320	179	960	220	231	199	265	323	380	406	317	399	253	354	506	477	846	713	599	408	371	267	284
R1-L1	320																						
R2-L1	78	179																					
R13-L1	239	142	960																				
R4-L2				220																			
R7-L2				62	231																		
R5-L3						199																	
R9-L3						68	265																
R6-L4								323															
R10-L4								111	380														
R6-L5										406													
R12-L5										118	317												
R9-L6												399											
R10-L6												81	253										
R2-L7														354									
R5-L7														125	506								
R3-L8																477							
R11-L8																279	846						
R1-L9																		713					
R2-L10																			599				
R8-L10																			240	408			
R12-L10																			175	216	371		
R3-L11																						257	
R4-L11																						107	284

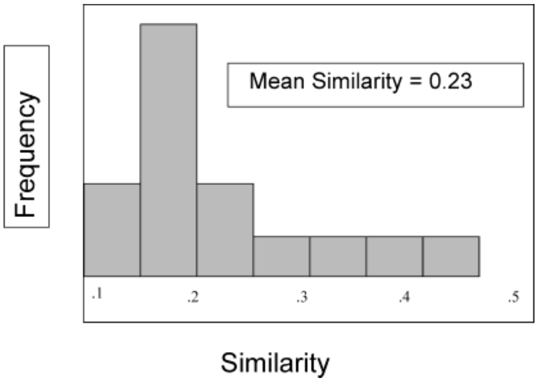

Figure 5. Distribution of similarities among chemists with respect to the compounds they agree should be rejected out of the identical ~2000-compound lists.

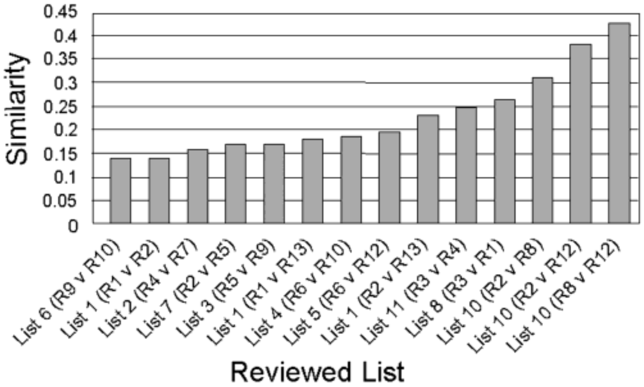

Figure 6. Similarity among chemists' rejections when reviewing the same ~2000-compound lists.

but there was very little consistency in their selections. For example, in the 250-compound study, reviewer 2 only rejected 64 of the same compounds that reviewer 1 rejected, yielding a Tanimoto similarity value of $S_{\mathrm{Tan}} = 0.20$. The results are even less consistent considering how these two medicinal chemists compared in their examination of list 1. Reviewer 1 rejected a total of 320 compounds, while reviewer 2 rejected 179, agreeing on only 78 compounds out of ~2000, yielding $S_{\mathrm{Tan}} = 0.16$.

In both the 250- and 2000-compound sets, the results obtained indicate there is very little consistency among medicinal chemists in deciding which compounds to reject. This study suggests that the expected agreement between chemists is on the order of only about 24%; that is, approximately 24% of the compounds rejected by one chemist will be rejected by another. Thus, the results presented clearly indicate that chemists do not agree among themselves as to which compounds should be rejected.

From this study it is also clear that medicinal chemists are not even internally consistent when reviewing compounds. It was shown that when medicinal chemists review the same compounds a second time but embedded within a different 2000-compound set, they reject the same compounds only about 50% of the time. Obviously, individual medicinal chemists make their selections based on their own personal set of guidelines. Thus, it seems fair to assume that even if a set of guidelines were established, consistency between different medicinal chemists would likely be less than 50% as seen by the average agreement between a pair of chemists of about 28%. It is unclear whether any set of guidelines could be established that would result in a high level of consistency. After the conclusion of the study, conversations with several reviewers indicated that some compounds may have been removed because they were similar to other compounds on a list. The potential impact of this on the study results presented here is unclear. A further study is needed to definitively clarify this issue.

Examination of the list of compounds most frequently rejected shows that there are clearly no obvious "good compounds" in the list that were mistakenly rejected by medicinal chemists. However, there are a number of undesirable compounds included in this study that really should be rejected. The reasons that these compounds were not rejected every time are unclear and are likely tied to the reasons behind the apparent inconsistency among medicinal chemists in reviewing compounds that this study highlights. Some of these reasons include personal bias, inattention, the inability of humans to deal with many pieces of complex and disparate data, and a lack of clear guidelines governing rejection criteria. The result that "good compounds" were not rejected seems to indicate that medicinal chemists are very good at identifying good compounds but are, understandably, unsure and hence inconsistent when it comes to compounds that fall within a "gray area". This may be one of the most important reasons for these results. If instead of a choice to accept or reject a compound medicinal chemists were also allowed to include compounds in an "undecided" or "uncertain" category, it is quite likely that this category would contain a significant number of compounds. Moreover, its presence would undoubtedly improve the consistency of chemists' picks, but this cannot be assessed in the current study because the data are not available. In addition, there are no objective or generally accepted criteria for assessing the actual quality of "good" compounds. Thus, we are left with using consensus chemists' opinion of what makes a compound attractive or conversely what makes a compound unattractive.

Conclusions

From the results presented here one must conclude that medicinal chemists are not consistent with themselves or compared to other medicinal chemists with respect to the compounds they find unacceptable. This inconsistency may have broader implications beyond the purchase of "chemically attractive or desirable" compounds. It should be emphasized that in this study, we are not making any judgments of who is right or wrong but only showing that their opinions often differ. Pharmaceutical and other companies often rely on subject matter experts, medicinal chemists in the present case, to decide which research compounds should be followed up and which ones should be passed by. Given the inconsistency demonstrated in the study described here, how then are appropriate, unbiased, and consistent decisions to be made?

Companies may want to consider alternative procedures to the more traditional methods of review of compounds by one or two medicinal chemists. One alternative is to utilize a team approach. In addition, more sophisticated computational rules could be developed to eliminate the need for manual review to a large extent if not entirely. It is suggested that the remedy is not as important as recognizing the fact that a problem exists.

Acknowledgment. The authors acknowledge the help and support provided by the many scientists at Pharmacia Corporation, Kalamazoo, who participated in this study. The authors also thank Mark Boys and John Mickelson at Pfizer, Ann Arbor; and Mike Genin, MariJean Eggan, and J. Craig Ruble at Eli Lilly and Company for their helpful comments.

References

(1) Maggiora, G. M.; Shanmugasundaram, V.; Lajiness, M. S.; Doman, T. N.; Schulz, M. W. A Practical Strategy for Directed Compound Acquisition. In *Cheminformatics Aspects in Drug Discovery*; Oprea, T., Ed.; Wiley-VCH: New York, in press.

(2) Willett, P.; Barnard, J. P.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

(3) Maggiora, G. M.; Shanmugasundaram, V. Molecular Similarity Measures. In *Methods in Molecular Biology; Chemoinformatics: Concepts, Methods and Tools for Drug Discovery*; Bajorath, J., Ed.; Humana Press: Totawa, NJ, 2004, Vol. 275, pp 1−50.

JM049740Z