Prospective Identification of Biologically Active Structures by Topomer Shape Similarity Searching

Richard D. Cramer,*,† Michael A. Poss,‡ Mark A. Hermsmeier,‡ Thomas J. Caulfield,‡ Mark C. Kowala,‡ and Maria T. Valentine‡

Tripos, Inc., 1699 South Hanley Road, St. Louis, Missouri 63366, and Bristol-Myers Squibb, Princeton, New Jersey 08546

Received April 1, 1999

The principle of bioisosterism—similarly shaped molecules are more likely to share biological properties than are other molecules—has long helped to guide drug discovery. An algorithmic implementation of this principle, based on shape comparisons of a single rule-generated "topomer" conformation per molecule, had been found to be the descriptor most consistently predictive of similar biological properties, in retrospective studies, and also to be well-suited for searching large ($>10^{12}$) "virtual libraries" of potential reaction products. Therefore a prospective trial of this shape similarity searching method was carried out, with synthesis of 425 compounds and testing of them for inhibition of binding of angiotensin II (A-II). The 63 compounds that were identified by shape searching as most similar to any of four query structures included all of the seven compounds found to be highly active, with none of the other 362 structures being highly active ($p < 0.001$). Additional consistent relations ($p < 0.05$) were found, among all 425 compounds, between the degree of shape similarity to the nearest query structure and the frequency of various levels of observed activity. Known "SAR" (rules specifying structural features required for A-II antagonism) were also regenerated within the biological data for the 63 shape similar structures.

Introduction

Despite many advances, the notorious inefficiency of drug[1] discovery research persists, with tens of thousands of molecules tested for every agent that finds clinical use. Most advances in computer-aided molecular design methodologies are intended to improve the selection of molecules for synthesis and testing, by improving the analyses of all relevant experimental information. However, such experimental information, whether a receptor structure[2] or previous test results (QSAR analysis),[3] remains expensive and time-consuming to acquire. Furthermore, converting these analysis results into actual structural targets is often an uncertain task, crucially dependent on the imagination of an individual chemist.[4] Finally, although these computational advances have contributed to the discovery of several valuable drugs, comparative studies[5] that permit a critical assessment of their efficacy are mostly lacking.

In particular, established computational methodologies cannot help in one of the most critical chemical decisions for a drug discovery project,[6] the "follow-up" to the initial discovery of a "hit".[7] What compounds can be obtained that would best exploit this "hit"? The very continuance of a discovery project may depend on the answer, because any delay and disappointment, caused by unfortunate synthetic choices that fail to generate additional biological activity, may soon overcome the initial organizational commitment. But the "underlying information" that accompanies an initial hit will seldom be more than "no actives found" while screening mostly unrelated structures. Thus the chemist often has only the single piece of information, "this structure is good", on which to make these critical decisions. In this situation the usual strategy is to "make small changes to the good structure", hoping from the results of testing to distinguish the structural features that are required for activity from those whose modification may improve properties. Underlying this strategy of "small changes" is the assumption, tacit but justified by a century of discovery experience, of "neighborhood behavior": structures which appear "similar" (to a medicinal chemist) are much more likely to have similar biological properties than are structures which appear different.[8]

Yet such a "neighborhood behavior"—tendency for "similar" molecules to have similar properties—is not associated with every molecular descriptor.[9] Thus for several years scientists at one of our organizations have been explicitly investigating this phenomenon. A methodology was devised for "validation"—estimating the magnitude and consistency of any neighborhood behavior—of any class of molecular descriptors.[10,11] Application of this methodology to 20 published data sets identified 2 of 11 classes of molecular descriptors as distinctly superior: the "2D fingerprint"[12] (qualitatively resembling a synthetic view of structures) and the 3D shape of a specific "topomeric" conformation[13] (qualitatively resembling bioisosterism[14]).

It then also became apparent that these two superior descriptor classes could enable extraordinarily rapid searching for structures "in the neighborhood" of a query structure, particularly within "virtual libraries" comprising all products obtainable by applying a particular reaction scheme to all appropriate and available reagents.[15] Such a capability seemed especially timely and useful because of the ongoing "high throughput revolu-

† Tripos, Inc.
‡ Bristol-Myers Squibb.

10.1021/jm990159q CCC: $18.00
Published on Web 09/03/1999

tion" in the experimental methods of drug discovery,[16] with its implied need for 100-fold increases in the rates of selecting structures for acquisition and testing. Therefore a technology known as ChemSpace[17] is being developed for building, managing, and searching large virtual libraries (>10^{12} structures), in support of two major applications. The first application, the design of experimentally efficient screening libraries by requiring that every included structure be separated from every other by *at least* the "neighborhood distance" for all descriptors established as having a neighborhood behavior,[18] will not be considered further here. We instead are concerned with the second application of hit follow-up, involving selection for acquisition and testing only those structures that are separated from the hit by *less than* some "neighborhood distance". For this hit follow-up application, "topomeric shape" is strongly preferred as the similarity descriptor having the most consistent neighborhood behavior, the one easiest to work with computationally, and the one providing the most unexpected and interesting structural output.

Before relying on shape similarity searching to guide hit follow-up, however, the critical question of efficacy arises. *Are the structures identified only by their similarity in topomeric shape to a biologically interesting query molecule at least as likely to be active as structures that are "similar" in the eyes of a chemist?* While the molecular descriptor validation studies had provided encouragement, they also relied on an unfamiliar, difficult, and retrospective interpretation of perhaps unrealistically "polished" data sets. Prospective library design experiments, involving "real world" synthesis and testing of a substantial number of compounds, might provide much more convincing evidence of efficacy. We here describe such an experiment.

Experimental Section

Selection of the Structural Series. Because promptly publishable results were the major goal of this experiment, we selected as the "hit" to be followed up a structural area of recent but not current interest to BMS. These were the "sartans", whose high therapeutic value as angiotensin II (A-II) receptor antagonists has resulted in several commercial antihypertensive drugs, four of which are shown in Figure 1. Each of these four drug structures was considered to be a "hit" for the purposes of this study.

Preparation of the Virtual Library. As exemplified in Figure 1, most sartans comprise a biphenyl ring system linked by a methylene to a heterocycle. Therefore, a targeted virtual library was constructed for this project based on a typical sartan synthetic sequence that had been reduced to practice at BMS, shown in Figure 2. As summarized in Table 1, the three variation sites in this library, corresponding to the ***a***, ***b***, and ***c*** rings as labeled in Figure 2, were populated by substructure searches of the Available Chemicals Directory[19] (ACD), using as queries[20] the fragments in line 1 of Table 1, to yield 2.6 billion potential products. Variation counts found at the three individual sites are shown in line 2 of Table 1.

Shape Similarity Searching. In general, the difference in shape between two structures is defined to be the smallest root sum of the squared differences in the shapes of corresponding fragments. ("Smallest" implies, for example, that a structure A−B is to be compared to a structure a−b not only as A−a vs B−b but also as A−b vs B−a). To compute the difference in shape between any two monovalent fragments, first a single characteristic "topomeric" conformation is built for each fragment, by rule-based adjustments of the appropriate torsion angles and chiralities of its Concord-generated 3D structure,[21] and the two resulting conformations are positioned

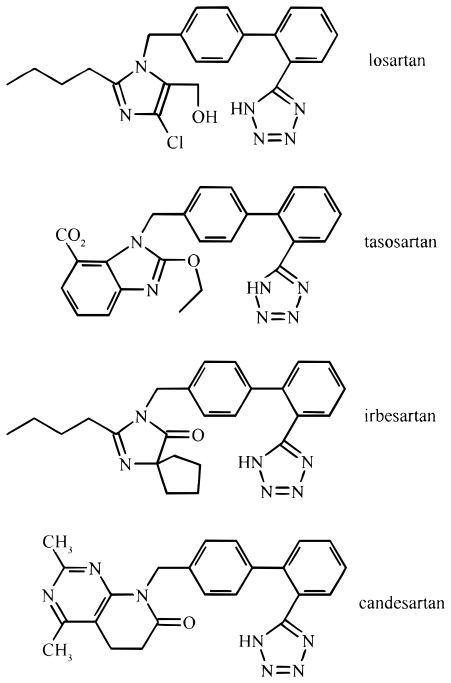


Figure 1. The four structures used as queries in the shape similarity searches. Each "sartan" is an antagonist of angiotensin II used clinically for treatment of hypertension.

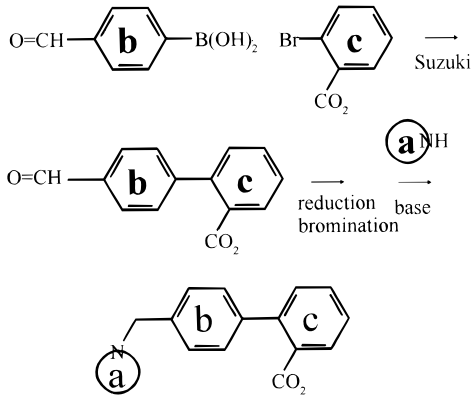


Figure 2. The synthetic scheme underlying construction of the "sartan" virtual library used for shape similarity searching. Its three independently variable elements are labeled with ***a***, ***b***, and ***c*** for later reference.

to superimpose the two attachment valences. (Details of this procedure were published previously.[13]) The difference in fragment shapes is then computed as the root sum of the squared differences in field values over the intersections of a 3D lattice, where the fields may be steric or "hydrogen-bonding".[22] (Thus, each field is calculated exactly as in CoMFA, except that the more the rotatable bonds separating any atom from the attachment valence, the "softer" the steric field of that atom.)[23]

There are two refinements of this procedure when comparing polyvalent fragments. First the squared fragment differences are summed over all possible attachment valences rather than one. Second, the nonsuperimposed valences are capped with methyl groups, and the squared differences in the six coordinates of the end atoms of each of the resulting bonds, weighted by 100.0, are added to the running sum of squared fragment differences.

Structure Fragmentation. Within the virtual library to be searched, the underlying synthetic reactions define a unique fragmentation pattern for every product structure. In contrast, usually the query structure must be fragmented in every way that might result in shape matches to any fragmentation

Table 1. Properties of the Sartan Virtual Library, of the "Neighbors" Found by Shape Similarity Searching, and of the Compounds Actually Synthesized and Tested[a]

line	property	total	***a*** variations	***b*** variations	***c*** variations
1	substructure query (SLN)[b]		NH~[R]Any	O=C[is=CC](H)\ XBr[is=BrC:=Any]	BrC:=C
2	virtual library	2 574 728 802	10 363	54	4 601
3	shape similar "neighbors"[c]	33 511	631	37	477
	to losartan	11 744	110	37	477
	to candesartan	5 439	146	9	142
	to irbesartan	317	9	4	103
	to tasosartan	16 751	448	30	320
4	synthesized and tested	425	44	2	8[f]
	shape similar "neighbors"[c] (if only the one variation)[d]	63	44[e]	2	4[f]
	shape similar "neighbors"[c] (within actual structures)[d]	63	36	1	3

[a] The sartan virtual library consisted of all commercially offered variations of each of three sites ***a***, ***b***, and ***c***, corresponding to those labels in Figure 2. [b] These SLNs (Sybyl line notations) define these substructure queries, respectively, as follows: ***a*** = any nitrogen contained within a ring and attached to a hydrogen; ***b*** = any bromoaldehyde, where the aldehyde is attached to a carbon and the bromine is aromatic or vinyl; ***c*** = any aromatic or vinyl bromide. (Structures containing more than one instance of a substructure query are excluded.) [c] Structures differing by 120 or fewer shape units from at least one of the four query structures. [d] Most of the synthesized structures actually varied at more than one site. The more sites varied, of course, the greater the overall shape difference. Thus, for example, the two variations of ***b*** exemplified among the 425 synthesized compounds are by themselves less than 120 shape units different from the corresponding fragment in the four query structures. Indeed, one of the two variations is identical to the query fragment. But the other variation appeared in synthesized structures only in combination with changes at one or more of the other two sites, and therefore it was not part of any of the 63 synthesized structures that were shape similar to any of the query structures. [e] One of the 44 synthesized variations of ***a*** (structure 44 in Figure 4) used a building block that was not listed in ACD and thus could not have been included among the 631 ***a*** variations sent by Tripos to BMS. [f] Four of the eight synthesized variations of ***c*** (structures c, f, g, h, l, j, m in Figure 5) used building blocks that were not listed in ACD and thus could not have been included among the 477 variations sent by Tripos to BMS. Furthermore, these four variations by *themselves* differed by more than 120 units from the corresponding ***c*** fragment in the query structures so that *no structure containing such a variation was a shape similar "neighbor".*

pattern within the virtual library, and an independent shape similarity search is performed for each of the resulting fragment combinations. However, for this particular study, the only virtual library reaction to be searched was the sartan library just described. Therefore the four query sartans were fragmented only by breaking the bonds formed by the synthetic sequence of Figure 2 to yield the fragments labeled there as ***a***, ***b***, and ***c***. (As can be seen in Figure 1, only the ***a*** fragment differed among the four query structures, since the four ***b*** and ***c*** fragments are structurally identical.)

In a "hit follow-up" search like this, the searcher must also choose a "neighborhood distance", the largest acceptable difference, here in "topomer shape", between the query structure and any acceptable structural answer. The validation studies had suggested that the majority of structures whose topomer shape difference was less than 90 shape difference units[24] from a query structure would share the biological properties of the query. (For mental calibration, a shape difference of 60 units results from replacing −H by $-CH_3$, a change that indeed disappoints whenever it suppresses biological activity.) However, for these shape similarity searches, in order to obtain a larger number of synthetic targets, the search radius used was 120 units rather than 90 units.[25]

The four query structures were submitted separately, and the union of the four sets of search results was then formed, yielding the counts of "shape similar" total product structures and individual fragments shown in Table 1 by line 3 and its subheadings. Thus these four searches, which together took a few hours on an SGI R4K, eliminated roughly 99.999% of the 2.6 billion possible product structures, yielding 33511 synthetic targets that were shape similar by 120 units or less to at least one of the four query structures. These design results were transmitted from Tripos to the BMS group, as the starting point for synthesis and testing. No further discussions took place until the experimental work had been completed.

Selection of Synthetic Targets. There was agreement between Tripos and BMS scientists to place the highest priority on the goal of ensuring rapid and successful completion of the experiment. Therefore the targets selected for actual synthesis were limited to those structural combinations for which laboratory protocols already existed.

To implement this policy, the BMS chemists decided to base the synthesis on an existing inventory of 15 subassemblies,

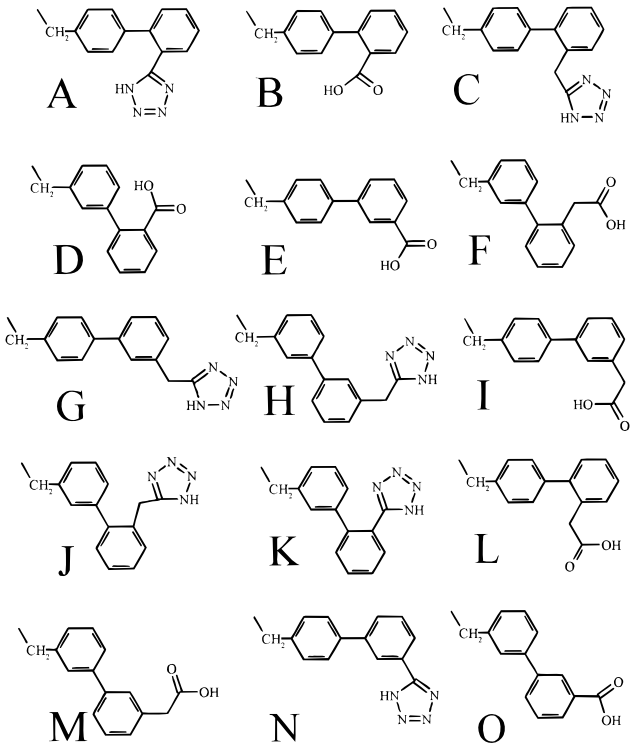


Figure 3. The structures of the 15 biphenyls comprising the ***b*** and ***c*** moieties of the synthesized "sartans". Most of these structures were by themselves dissimilar by more than 120 shape units from structure A (the ***b***−***c*** moiety within the four query structures), and therefore their "sartan" products were expected to have a very low occurrence of biological activity.

each containing the ***b*** and ***c*** fragments already joined. As can be seen from Figure 3, their structures differed in the relative positioning of fragments much more than in the fragments themselves.[26] Two of these subassemblies (structures A and B in Figure 3) are well-known to confer activity, with structure A of course comprising the ***b***−***c*** moiety within the commercial

drug query sartans of Figure 1. However, the shape differences of the other 13 subassemblies from that query ***b*−*c*** moiety are quite large, so that the use of these available starting materials produced a rather large proportion of negative control experiments among the overall results. Another effect of using the existing inventory of ***b*−*c*** subassemblies was to decrease the number of shape similar candidate structures from the original 33 511 (including variation at ***a***, ***b***, and ***c***) to 631 (including the ***a*** ring only). On the other hand, these decisions helped to ensure that at least some compounds would be successfully synthesized and tested, thus minimizing the risk of a inconclusive experiment.

The majority of the 631 building blocks, originally chosen from ACD only on the basis of a substructural query and filtered only on the shape similarity criterion, were either very expensive or not readily available. Furthermore, many of the obtainable ***a*** variations did not undergo the necessary alkylation in high enough yield or purity. The structures of the 44 ***a*** synthons occurring in any of the 425 products that were finally synthesized and tested are shown in Figure 4.

Thus the final set of synthetic targets can be described simply as the array defined by combining all ***b*-*c*** subassemblies at hand with all commercially available and usable ***a*** variations that were no more than 120 topomeric shape units different from at least one of the four query ***a*** moieties.

General Synthetic Procedures. All of the compounds were prepared by solid-phase synthesis using the IRORI microkan technology.[27] Actual synthesis of the compounds occurred as two separate runs. The first run was a 1 × 46 array using template O and a set of 46 heterocycles. The second run was a 14 × 48 array using the remaining 14 templates (A..N) and a set of 48 heterocycles. In total, the synthesis of 718 compounds was attempted. Next, 425 of the 718 attempted products were obtained in sufficient amount and a purity of ≥70%, showed a correct mass spectrum, and were therefore tested in the biological assay. Details of the synthetic sequence, as summarized in Figure 5, are now described for the 672 (14 × 48) compound array.

Loading of the IRORI Microkans. The biphenyl carboxylic acids were attached to polystyrene beads through a *tert*-butyl ester linkage. The biphenyl tetrazoles were attached to polystyrene beads through an *ortho*-methoxy benzyl linkage. The loading of the resins ranged from 0.85 mmol/g to 0.91 mmol/g. Slurries of the 14 different resins were prepared from approximately 1.5 g of each resin in dichloromethane and heptane (20 mL; 19:1). Each slurry was added to 48 microkans utilizing an eight-channel pipet. Approximately 31 mg of resin was transferred to each of the 672 microkans. At the end of the transfers, less than 100 mg of each resin remained in the respective source containers. Radio frequency tags were then added to each microkan, and the reactors were capped. Next, products were assigned to each of the microkans using an R_f scanner and the synthesis software.

Reduction of Benzaldehyde to Benzyl Alcohol. The 672 microkans were placed in a three-neck 2 L round-bottom flask. THF (700 mL) was added, and stoppers were placed over the two side neck openings. Next, the flask was repeatedly evacuated and flushed with nitrogen while shaking vigorously to "degas" the microkans. The THF was drained from the reaction vessel, and additional THF (400 mL) and ethanol (350 mL) was added. Next, sodium borohydride (5.67 g) was added. Again, the flask was repeatedly evacuated and flushed with nitrogen while shaking vigorously. The flask was then secured to the deck of a ThermoLyne BiggerBill shaker and agitated at 150 rpm at room temperature. After 16 h, the reaction solvent was drained, and the microkans were washed with THF (2 × 700 mL), THF/ethanol (1:1, 3 × 700 mL), acetic acid/water/THF (1:1:8, 5 × 700 mL), and CH_2Cl_2 (4 × 700 mL). After the final wash, the microkans were dried under vacuum at room temperature overnight.

Conversion of Benzyl Alcohol to Benzyl Bromide. The 672 microkans were placed in a three-neck 2 L round-bottom flask. Dry CH_2Cl_2 (700 mL) was added, and the flask was repeatedly evacuated and flushed with nitrogen. The CH_2Cl_2 was drained from the flask, and more dry CH_2Cl_2 (700 mL) was added. *N*,*N*′-Diisopropylethylamine (36.6 mL) was added, and the flask was repeatedly evacuated and flushed with nitrogen again. Next, dibromotriphenylphosphorane (44.32 g) was added in three equal weight batches. The flask was repeatedly evacuated and flushed and then secured to the deck of a ThermoLyne BiggerBill shaker and agitated at 180 rpm at room temperature. After 18 h, the reaction solvent was drained, and the microkans were washed with CH_2Cl_2 (3 × 700 mL), THF (4 × 700 mL), and CH_2Cl_2 (3 × 700 mL). After the final wash, the microkans were dried under vacuum at

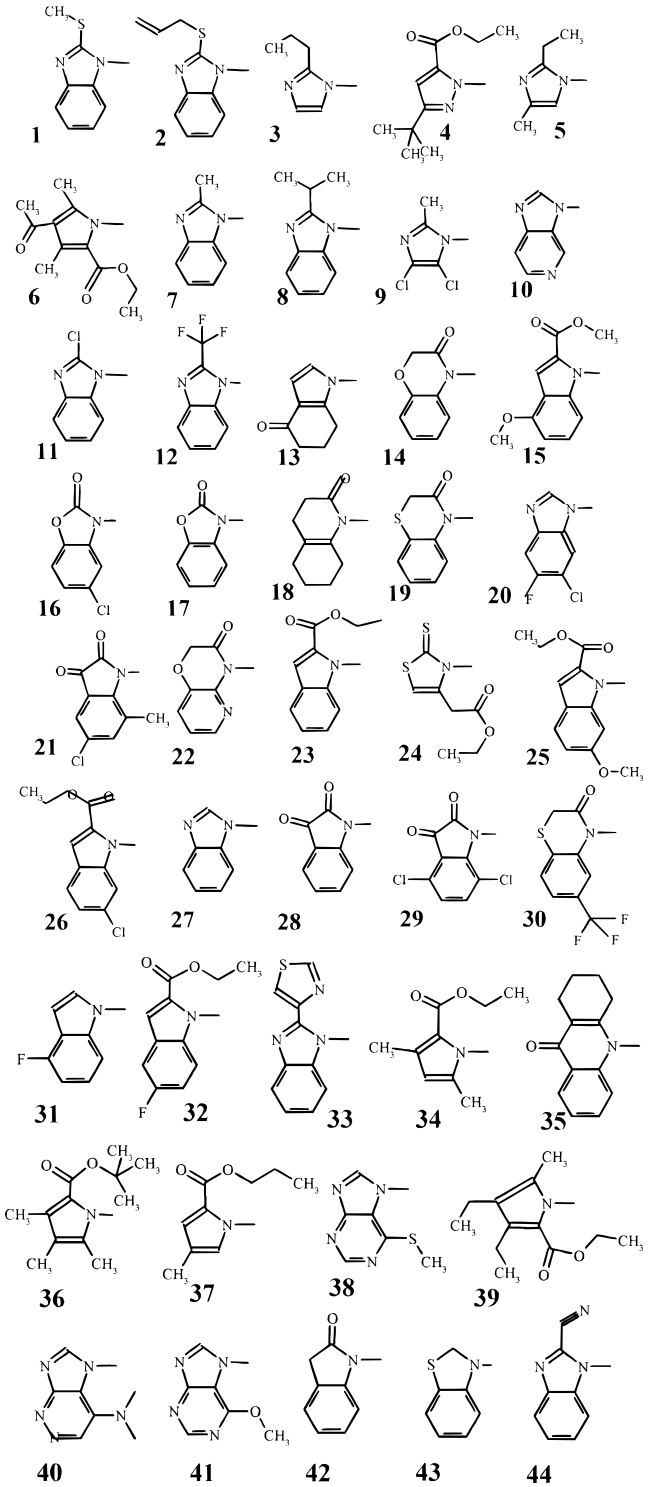


Figure 4. The structures of 44 nitrogen heterocycles comprising the ***a*** moieties of the synthesized "sartans". All of these structures are similar by 120 shape units or less to an ***a*** ring within at least one of the four query sartan structures.

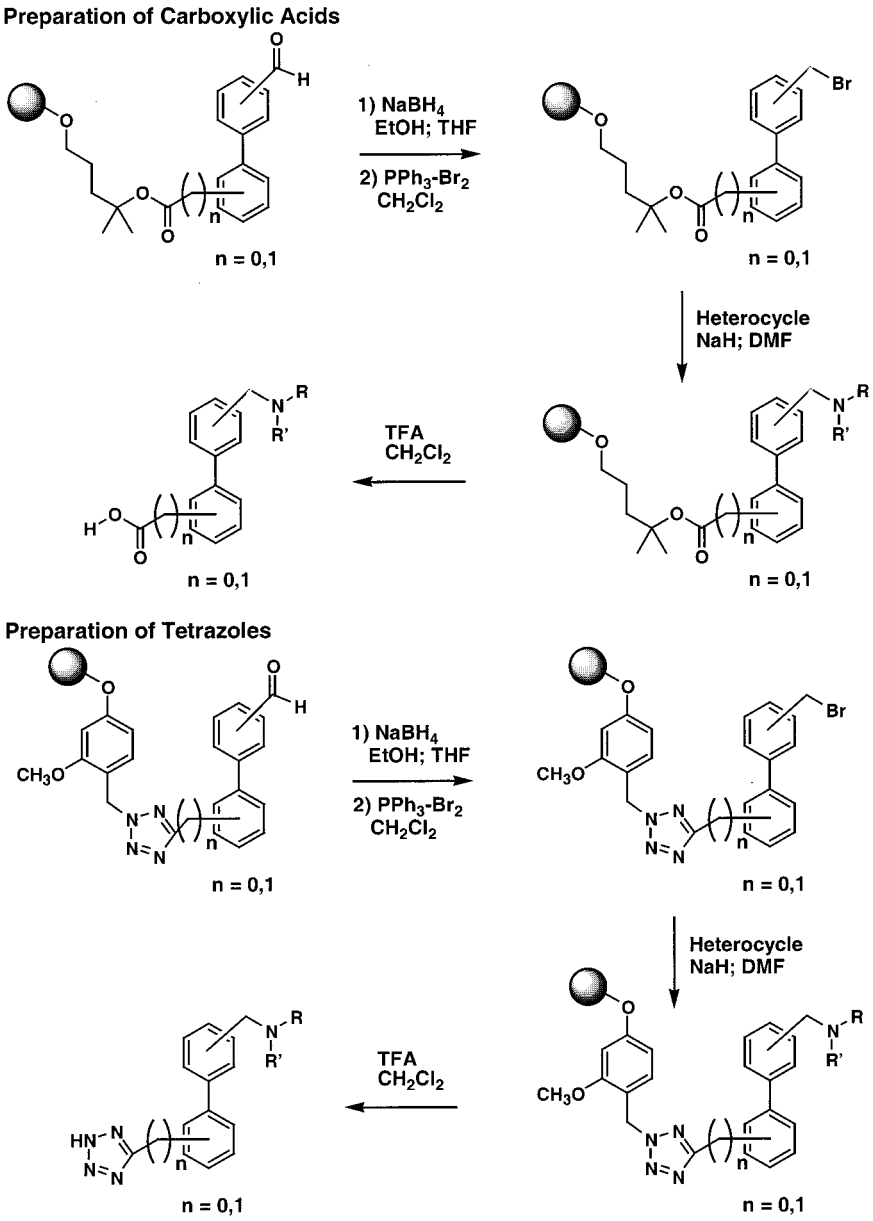


Figure 5. The synthetic sequences used to prepare the sartans.

room temperature for 3 h, and then they were sorted into 48 different groups of 14 microkans each for reaction with the 48 respective heterocycles.

Heterocycle Displacement of Benzyl Bromide. Each of the 48 heterocycles (4.5 mmol) was weighed into a 25 × 95 mm^2 glass vial. Dry DMF (15 mL) was added to each vial followed by sodium hydride (90 mg, 2.25 mmol). The vials were agitated and then capped after the hydrogen gas evolution ceased to be rapid. After 1 h, the appropriate 14 microkans were added to each vial, and the vials were shaken vigorously to remove air bubbles. Next, tetra-*n*-butylammonium iodide (~2 mg) was added to each reaction vial. The vials were placed in an aluminum heating block and heated at 40 °C. The aluminum block was placed on a Lab Line shaker and agitated at 265 rpm. After 20 h, the vials were shaken by hand and placed back in the aluminum block. The temperature of the heating block was raised to 45 °C, and the vials were agitated at 265 rpm for an additional 44 h. Next, the vials were allowed to cool to room temperature, and the reaction solvent was drained from each vial. Additional DMF (15 mL) was then added to each vial. The vials were agitated briefly, and the solvent was drained. Each set of 14 microkans was transferred to a 50 mL round-bottom flask and washed with DMF (2 × 15 mL) and THF (15 mL). After the THF wash, all of the microkans were collected in a 2 L three-neck flask and washed with DMF (3 × 750 mL), DMF/water (1:1, 4 × 700 mL), DMF (2 × 700 mL), THF (4 × 700 mL), and CH_2Cl_2 (3 × 700 mL). After the final wash, the microkans were dried under vacuum for 16 h and sorted into 14 groups based on the starting resin.

Cleavage of the Products from the Resin. The microkans were sorted into individual polypropylene tubes fitted with a frit and stopper at the bottom. Trifluoroacetic acid (0.80 mL) was added to each tube containing a microkan, and the tubes were agitated at 600 rpm for 90 min. The reaction solution was drained into a tared tube. Dichloromethane (0.80 mL) was then added to each tube, and the tubes were agitated for 15 min. The solvent was drained from each tube and combined with the previous TFA solution of product. Racks of the collection tubes were placed in a SpeedVac and evaporated. After 5 h, the tubes were removed from the SpeedVac, and CH_2Cl_2 (0.75 mL) was added to each tube. After brief agitation, the racks of tubes were placed back in the SpeedVac and concentrated again. Each product sample was weighed and analyzed by HPLC and mass spectroscopy. As described earlier, products having sufficient weight, a purity of ≥70%, and a correct mass spectrum were tested in the biological assay.

Biological Assay. [^{125}I]Sar, Ile-angiotensin II binding assays to rat AT_1 receptors were performed in a total volume

of 250 μL in a 96-well format. The incubation mixture contained 50 μL of radioligand (at a concentration of 0.2 nM for competition assays), 50 μL of competing drug or 50 μL of angiotensin II (10 μM final concentration) to define nonspecific binding, and 150 μL of buffer (Dulbecco's Modified Eagles Medium pH 7.4, containing 5 mM $MgCl_2$, 1 mM EDTA, and 0.1% BSA). The 96-well plates (Costar, Corning, NY) were seeded with rat aortic smooth muscle cells (RASMCs), and 24 h later, the binding reaction was initiated by the addition of [^{125}I]Sar, Ile-Angiotensin II (NEN), drug, or angiotensin II (Sigma). The plates were incubated at 25 °C for 2 h with mild shaking. The RASMCs were washed twice with PBS to remove unbound isotope and solubilized with 1% Triton X plus 0.1% BSA for 15 min at room temperature on a shaker. Lysed RASMCs were collected in individual tubes, and radioactivity was measured on a γ counter (Packard Cobra II, Downers Grove, IL). Competition binding data were analyzed by XL fit (IDBS, U.K.).

Calculation of Shape Differences between Synthesized Structures and Query Structures. Most of the structures actually synthesized contained ***b***−***c*** subassemblies that could not have been selected by the original shape similarity searches (the underlying building blocks not being commercially offered in ACD). Therefore shape differences were explicitly recalculated between each of the 425 synthesized compounds and each of the four sartan query structures. For this purpose all structures were divided into three fragments, as suggested in Figure 2, by breaking the bond between the ***b*** and ***c*** rings and also by breaking the bond attaching the ***a*** ring to the rest of the structure. The shape difference between any pair of compounds was then found as the root sum of the squared shape differences over the corresponding three fragment sets, as detailed above and in the references. For subsequent purposes, the final "shape difference" for any of the synthesized compounds was then taken to be the smallest of its four differences from the query sartans.

Results

The experimental results are presented in Table 2 as the actual combinatorial array, with the 44 synthesized variations in the ***a*** ring as its rows and the 15 variations in the ***b***−***c*** subassemblies as its columns. Each cell therefore corresponds to a particular structure, with the experimental percentage displacement value appearing in the cell for each compound that was actually synthesized and tested. The structures of each of the ***b***−***c*** and ***a*** variations appeared in Figures 3 and 4, respectively. The gray scale shading of a data-containing cell gives an impression of the shape difference of that structure from the most shape similar query structure.

Table 3 is a complement of Table 2, in which the numerical values record the shape differences between the synthesized structure and its most shape similar query sartan, and the gray scale gives an impression of the experimental percentage displacement value. The first column of Table 3 identifies by a two letter abbreviation the most shape similar of the four query sartans to each of the ***a*** building blocks used in synthesis.

As stated in the Introduction, this experiment primarily addresses a single question: *Are the structures identified by shape similarity searching as being close "neighbors" of a biologically interesting query molecule indeed at least as likely to be active as molecules which are "structurally similar"(and therefore worthy of synthesis and testing) in the eyes of a chemist?* To answer this question, the values in Table 3 were arbitrarily partitioned into five shape difference bins, the middle three being 30 shape units wide. Within each of these bins, the counts of structures and the counts and percentages of "active structures" could be tabulated from the values in the corresponding cells of Table 2.

An attendant difficulty is that any definition of "activity" is rather arbitrary. Usually "active" compounds are those which demand the commitment of resources to further investigations. For the A-II assay in current use at BMS, it is "displacement of >75% of radioactive A-II by a 1 μM concentration" that defines "activity" and triggers IC_{50} determination.[28] However, for purposes of this experiment, a less restrictive definition of "activity" might be more appropriate, such as "effecting any detectable displacement". The lowest "detectable" level of displacement in any assay depends of course on the assay variability, which from the magnitudes of the (meaningless) negative values recorded in Table 2 seems for this assay to be at least ±20%.[29] Therefore the relation between shape similarity and the intensity of displacement was examined for several different definitions of "activity": displacement > 75%; displacement > 40%; displacement > 30%; displacement > 20%; and the average of "percentage displaced".

The relation that emerged between shape similarity and the frequency of observed activity is indeed very strong, as listed in block A of Table 4. The top five lines of data in that block summarize the results for those compounds whose shape differences from their nearest query structure fell in a particular range of values, shown in the first column. (Note that the greater the shape similarity, the smaller this shape difference value.) The bottom two lines provide parameters appropriate for statistical significance tests of a nonrandom data distribution within the corresponding column. The second column is the number of synthesized compounds within that range. The right-hand column shows the average of the percentage displacement values for those compounds. The remaining columns show, for the various definitions of activity, the number of those compounds whose binding value equaled or exceeded the activity thresholds shown, expressed both as a count and as a percentage.

Note first that only 63 of the 425 compounds finally synthesized and tested were acceptable by the shape similarity criterion used in the original search. These are the 63 compounds referenced by the first two lines of data in block A. (Recall that the shape difference threshold used in the four original searches was "<120 shape units".)

The most dramatic result occurred when the activity was the ">75% binding" criterion used in the laboratory. Only seven such active compounds were reported among the 425 tested. All of those seven were included among the 63 shape similarity selections, while none of the other 362 structures synthesized showed any activity. Such a distribution of activity within the 425 structures is absurdly unlikely to have arisen by chance; the associated χ^2 value of 59.67 is almost three times the highest tabulated threshold, for $p < 0.001$.[30] Thus, it is a virtual certainty that shape similarity selection was a better predictor of activity than was the synthetic rationale for any of the other 362 molecules.

Yet, in many ways the more impressive result, because of its extraordinarily consistency among all of the synthesized compounds, is found throughout the

Table 2. Observed Percentage Displacements of Angiotensin II for All Synthesized Compounds[a]

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	99	35	38	34	28		3	11	15	1	-2			-23	
2	86	34	29	29	28	8	9	16	1	26	19	23	16	24	14
3	85	37	54	34	38	9	-3	20		14	15	4		-4	17
4	83	39	36	21	31	-13	-5	23		7	10	-3	12	4	15
5	83		51	2	33	11		-16		20	-17	26		7	5
6	77	19	40	27	27	35	20	8		29	14	-12	17	14	7
7	76	29	45	31	36	35	11	-10	8	29	16	2	7		
8	69		32	17	47	8	-12	6	28	-2	-12	-9	-5	2	17
9	46		38	10	30	23	-11	4		6	9	-7		-3	8
10	39	6	26	9	36	-8	11	34	11	17	16	-21	16	12	14
11	34	7	31	15		-1	-20	7		26	-4	22		17	14
12	32	23	32	-3	33	11	9	16		-62	26	3		1	
13	27	19	21	18	28	14		-3		-20	25	-8		22	22
14	26	27	34	53	42	37	21	-1	15	25	20	-28	22	2	23
15	18	26	23	11		28		8		14	-7	5		7	
16	17	29	42	24			23	-13		6	2	23	19		1
17	16	36	13	9		-15	19	37	35	31	11	-16	19	12	9
18	15	3		11	49		16	33		7				17	
19	11	11	5	13	34	30	-1	12	15	-5	29			-9	14
20	10	9	27	10	38	-16	12		12	24	-6	16	3	7	19
21	9	19		31	32		24	18		18				21	
22	8	4	34	34		8	7	21	28	32	8	19	22	25	
23	8	13		21		30			26		18	-12		20	8
24	6	14	16	3	35	37	-5	20	23	32	10	9		20	2
25	5	23		32		37		3	35	4	22	-7	20	-25	
26	1	31	30	19				18			-5	22	11	9	14
27	-1	24	59	9	36	-15	0	24	14	19	10	-15	28	15	20
28	-3	30	18	28		-8	1	-20		12	20	-10		17	6
29	-8	18	28	16	31	-5	19	23	2	-48	5	8		5	
30		38	24	6	34	11	15	23	28	23		28	7		14
31		31													
32	3	23	26	-1		19		7	16	0	31	-20		28	9
33		20	35	0	36	12	-6	26		-1	3	23		-4	
34		16	25	15							13				
35				15		36		34		8	12				
36			15				25			5	1	25	-5		
37				19		-13					9	18			
38				0											
39				-23											
40				47		28		12			-2	-13			
41									19						
42										26					
43										6					
44													27		

Bold < 90 shape similarity to nearest query structure
90 - 120 shape similarity to nearest query structure
120 - 150 shape similarity to nearest query structure
Bold 150 - 180 shape similarity to nearest query structure
>180 shape similarity to nearest query structure

[a] The structure for any cell can be obtained by joining the fragments from Figures 3 and 4 that match its row and column labels. Empty cells were not synthesized. The shading of a cell indicates its shape similarity to the most similar of the four query structures.

other results in block A. In every case considered, the farther in shape difference a structure is from its nearest query structure (the further down any list in block A), the less likely that compound was to show "activity", however defined. For example, if the activity criterion is ">30% displacement", then among the compounds having a shape similarity better than 90 units, 66.7% were found active, while a similarity between 90 and 120 units yielded only 28.4% of actives. The activity percentage continues to decline as the dissimilarities of the remaining three groups increase. To provide a statistical significance test for this apparent trend within each column, the tabulated percentage values were regressed on the midpoint shape similarity values for the bins (end bins being assigned their shape similarity values by simple linear extrapolation of the mid-bin values). The resulting $F(1,3)$ ratios are shown as the last line in block A of Table 4.[31]

Figure 6 summarizes graphically and perhaps more understandably the results just presented. Each of the

Table 3. Shape Similarity Values to the Nearest of the Four Query Sartans for All Synthesized Compounds[a]

Nb[a]		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	*ca*	**87.75**	98.24	120.7	129.3	**141.4**		151.9	173.6	141.4	147.1	121.5			144.1	
2	*lo*	**91.55**	101.6	**123.5**	**128.8**	143.7	140.7	154.1	175.5	143.7	**149.4**	124.2	**112.8**	169.5	**146.4**	143.7
3	*lo*	**98.42**	107.9	128.7	136.7	148.2	145.2	158.3	**179.2**		153.7	129.4	118.5		150.8	148.2
4	*ir*	**89.03**	99.38	121.7	**130.1**	142.2	139	152.7	**174.2**		147.9	122.4	110.8	168.2	144.9	142.2
5	*lo*	**108.8**		136.8	144.4	155.3	152.5		185.1		**160.6**	137.5	**127.3**		157.8	155.3
6	*lo*	**93.59**	103.5	125.1	**133.3**	**145**	142	**155.4**	176.6		**150.6**	125.8	114.5	170.6	147.7	145.1
7	*ca*	**113.9**	**122.2**	140.9	148.3	158.9	156.1	168.4	188.2	158.9	**164.1**	141.5	131.6	182.6		
8	*lo*	116.3		142.8	150.1	160.6	157.9	170	189.6	**160.6**	165.7	143.4	133.7	184	163	160.6
9	*lo*	110		137.8	145.3	158.2	**153.3**	165.8	185.8		161.4	138.4	128.3		158.6	156.2
10	*ca*	114.9	123.1	**141.7**	149	159.6	156.9	169.1	188.8	159.6	164.7	142.3	132.5	183.2	162.1	159.6
11	*lo*	107.4	116.1	135.7	143.3		151.5	**164.1**	184.3		**159.6**	136.4	**126**		156.8	154.3
12	*ta*	112.4	**120.8**	139.7	147.1	157.9	155	167.4	187.3		163	**140.3**	130.3		160.3	
13	*ta*	**96.19**	105.8	**127**	135.1	**146.7**	143.7		178		152.3	**127.7**	116.6		**149.4**	**146.7**
14	*ca*	**106.3**	**115.1**	134.9	142.5	153.6	150.7	**163.4**	183.7	153.6	**158.9**	**135.5**	125.1	**177.9**	156.1	**153.6**
15	*ta*	112.7	**121.1**	**140**	147.4		**155.3**		187.5		163.2	140.6	130.6		160.5	
16	*ta*	102.9	**112**	132.2	**140**			**161.1**	181.7		156.6	132.8	122.2	175.9		151.2
17	*ta*	111.8	120.2	139.2	146.6		154.6	167	186.9	157.4	162.6	139.8	129.8	181.2	159.9	157.4
18	*ta*	84.18	95.06		126.9	139.2		149.9	171.8		145				141.9	
19	*ta*	104.5	113.4	133.4	141.1	152.3	149.4	162.2	182.6	152.3	157.6	**134.1**			154.8	152.3
20	*ta*	101.9	111.1	**131.4**	139.2	150.5	147.6	160.5		150.5	**155.9**	132.1	121.4	175.3	153.1	150.5
21	*ta*	105.1	114		141.6	152.8		**162.6**	183		158.1				**155.3**	
22	*ta*	105.4	114.3	134.1	141.9		150.1	162.8	**183.2**	**153**	158.3	134.8	124.4	**177.4**	**155.5**	
23	*lo*	97.58	107.1		**136.1**		144.7			**147.7**		128.8	117.8		**150.3**	147.7
24	*lo*	112.8	121.7	140	147.4	158.2	155.4	167.7	**187.5**	**158.2**	163.3	140.7	130.7		**160.6**	158.2
25	*ca*	102.8	**111.9**		139.9		148.2		181.7	151.2	156.5	**132.8**	122.2	**175.9**	153.7	
26	*lo*	112.6	121	139.9	147.3				187.4			140.5	**130.5**	181.8	160.5	158
27	*ta*	111.6	**120**	139	146.5	157.3	154.4	166.8	**186.8**	157.3	162.4	139.7	129.6	**181.1**	159.7	**157.3**
28	*ta*	104.1	113.1	133.1	**140.8**		149.1	161.9	182.4		157.4	**133.7**	123.2		154.6	152
29	*ta*	101.5	110.7	**131.1**	139	150.3	147.4	160.3	**180.9**	150.3	155.7	131.8	121.1		152.9	
30	*ta*		100.4	**122.5**	130.9	142.8	139.7	153.3	**174.8**	**142.8**	**148.5**		111.7	168.8		142.8
31	*ta*		114.4				150.1									
32	*lo*	97.74	**107.3**	**128.2**	136.2		144.8		178.8	147.8	153.3	128.9	117.9		150.4	147.8
33	*ta*		**108.1**	128.9	136.9	150.5	145.4	158.5	**179.4**		153.9	129.6	**118.7**		151	
34	*lo*		98.11	**120.6**	129.2							121.4				
35	*ca*				150.8		158.6		190.2		166.4	144.2				
36	*lo*			131.4				**160.5**			156	132.1	**121.4**	175.3		
37	*lo*				132.5		141.2					124.9	113.5			
38	*lo*				148.7											
39	*ca*				138.2											
40	*lo*				139		**146.6**		180.9			131.8	121.1			
41	*ca*									156.9						
42	*ta*										**155.1**					
43	*ta*										161.4					
44	*ca*													**178.3**		

Bold Displacement > 75%
Displacement 40% - 75%
Displacement 30% - 39%
Bold Displacement 20% - 29%
Displacement <20%

[a]This column identifies the most shape-similar of the four query structures. (*ca*=candesartan, *ir*=irbesartan, *lo*=losaratan, *ta*=tasosartan.)

[a] The structure for any cell can be obtained by joining the fragments from Figures 3 and 4 that match its row and column labels. Empty cells were not synthesized. The shading of a cell indicates its observed displacement of angiotensin II. [b] This column identifies the most shape similar of the four query structures (*ca* = candesartan, *ir* = irbesartan, *lo* = losaratan, *ta* = tasosartan).

five lines of data in Figure 6, excerpted from the first five lines of block A of Table 4, is linked to the appropriate shell in the "similarity target" by an arrow.

For comparison with the topomer shape difference descriptor, blocks B−E in Table 4 report results from repeating this analysis with four other molecular descriptors used in diversity designs. Higher values of Tanimoto coefficients from "2D fingerprints" are widely accepted as predictors of similar biological effects.[10,12] For these structures, although 2D fingerprint Tanimoto coefficients were statistically significant predictors of "displacement > 75%", there was no significant association among all 425 compounds between 2D fingerprint Tanimoto and the probability of activity. Furthermore the maximal 2D fingerprint Tanimoto value of 0.76, surprisingly low for these structurally similar compounds, is lower than the usually accepted threshold of 0.85 or higher.[32] The atom pair descriptor, also assessed

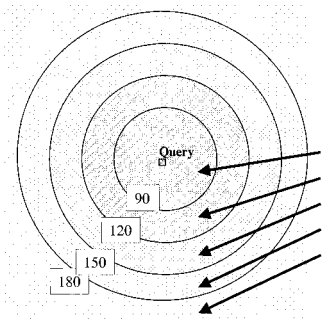
Table 4. Dependence of Probability of Biological Activity (A-II Displacement) on Similarity to a Known Active Molecule, for Various Thresholds of "Biological Activity" and Various Descriptors of "Molecular Similarity"

		A-II displacement								
		>75%		>40%		>30%		>20%		
	no. cpds	no. cpds	% tot	no. cpds	% tot	no. cpds	% tot	no. cpds	tot	avg total
A. Topomeric Shape Similarity										
<90	3	2	66.7	2	66.7	2	66.7	2	66.7	65.7
90−120	60	5	8.3	7	11.7	17	28.3	27	45	22.1
120−150	178	0	0	9	5.1	35	19.7	77	43.3	16.2
150−180	156	0	0	2	1.3	24	15.4	57	36.5	13.5
>180	28	0	0	0	0	3	10.7	8	28.6	11.2
χ^2		59.74[a]		22.63[a]		6.6		2.9		
F-test, nonzero slope			4.4		5.6		10.1[c]		23.3[b]	6.1[c]
B. Unity "2D Fingerprints" (Tanimoto Coefficient)										
>0.70	7	2	28.6	3	42.9	3	42.9	3	42.9	32.6
0.60−0.70	52	3	5.8	6	11.5	12	23.1	15	28.9	16.6
0.50−0.60	141	2	1.4	5	3.5	21	14.9	57	40.4	18.6
0.40−0.50	161	0	0	2	1.2	28	17.4	65	40.4	14.8
<0.40	64	0	0	3	3.1	17	26.6	31	48.4	18.3
χ^2		22.2[a]		23.9[a]		5		3.9		
F-test, nonzero slope			5.9[c]		5.1		1.3		1	2.5
C. Atom Pairs (Tanimoto Coefficient)										
>0.70	15	3	20	5	33.3	7	46.6	8	53.3	34.6
0.60−0.70	81	3	3.7	6	7.4	13	16	32	39.5	15.3
0.50−0.60	121	1	0.8	4	3.3	19	15.7	45	37.2	14.7
0.40−0.50	163	0	0	5	3.1	31	19	68	41.7	15.2
<0.40	45	0	0	0	0	11	24.4	18	40	17.5
χ^2		8.6		24.4[a]		7.25		1.05		
F-test, nonzero slope			6.2[c]		6.9[c]		1.2		1.9	2.2
D. Electrotopological ("EState") Descriptor Difference										
<1.0	108	1	1	5	4.6	19	17.6	50	46.3	17
1.0−2.0	85	1	1.2	2	2.4	16	18.8	34	40	16.8
2.0−4.0	83	1	1.2	3	3.6	10	12	24	28.9	11.9
4.0−7.0	85	1	1.2	3	3.5	19	22.4	34	40	14
>7.0	64	3	4.7	7	10.9	17	26.6	29	45.3	21.3
χ^2		2.1		4.98		4.6		5.9		
F-test, nonzero slope			9.3[c]		16.3[b]		4.4		0.3	2.2
E. Molecular Weight Difference										
<5.0	123	3	2.4	3	2.4	22	17.9	51	41.5	16
5.0−15.0	78	1	1.3	2	2.6	16	20.5	32	41	14.1
15.0−25.0	46	0	0	2	4.3	5	10.9	16	34.8	12.8
25.0−40.0	66	0	0	3	4.5	11	16.7	27	40.9	15.8
>40.0	112	3	2.7	10	8.9	27	24.1	48	42.9	19
χ^2		0.8		5.01		3.24		0.49		
F-test, nonzero slope			0		19.1[b]		0.4		0.1	0.8

[a] $P < 0.001$. [b] $P < 0.05$. [c] $P < 0.10$.

by means of a Tanimoto coefficient, behaved similarly to the 2D fingerprint Tanimoto, as summarized in block C. These results are consistent with the use of 2D fingerprints and atom pairs in similarity searching, though the results are not as compelling as those for topomeric shape. On the other hand, neither of the final two descriptors, total electrotopological index and molecular weight, exhibited any neighborhood behavior for these data, as shown in blocks D and E, where the modestly significant associations are actually backward[33] and hence surely artifacts.

A concern about these results is that the four query structures are not typical of newly discovered leads, because each is the highly potent outcome of a lead optimization program. To address this possibility of a "privileged starting point" bias, we have repeated the tabulations that generated Table 4 for many of the newly synthesized structures, to ask repeatedly, in essence, "How good a predictor of activity is topomeric shape similarity to this particular structure?" Such tabulations were performed for all 81 compounds displacing more than 30% of bound A-II and for 14 randomly chosen compounds displacing between 20% and 30% of A-II. Representative results of these studies appear in Table 5 (complete results may be found in the Supporting Information). Each row in Table 5 presents χ^2 tests of significance for a nonuniform association between frequency of activity and topomeric shape similarity (to the compound labeling that row). This row of values is exactly comparable to the fifth row of values within each of the four blocks of Table 4. The last two values in each row are counts and degree-of-significance-weighted counts of the statistically significant χ^2 values in that row. It is evident that significantly nonrandom associations exist between frequency of activity and topomeric shape similarity to many of the structures analyzed (overall 40 of 95, or 42%). Furthermore, the more strongly bound a compound, the more likely it is to show such significant neighborhood behaviors. (Using the same bins, the results are as follows: of those binding >75%, 6 of 7 (86%); from 40% to 75%, 7 of 13 (54%); from 30% to 40%, 24 of 61 (39%); from 20% to 30%, 3 of 14 (21%); $p < 0.10$). There is no *further* significant tendency for neighborhood behavior to be more likely for structures containing any particular ***b*−*c*** fragments.



Shape Difference	# Cpds	# Active Cpds >75% Displ	# Active Cpds >30% Displ
< 90	3	2(67%)	2(67%)
90-120	60	5(8%)	17(28%)
120-150	178	0	35(20%)
150-180	156	0	24(15%)
>180	28	0	3(11%)

Figure 6. Summary of the results. The 425 tested compounds were divided into five groups, according to the shape difference between each structure and the most similar of the four query structures. For each of the five groups, the numbers of compounds tested and the number of compounds found to be "active", where activity is taken as displacing either 75% or 30% of the radioactive angiotensin II, are tabulated. The arrows pointing to the corresponding shells in the target help in visualizing the overall result: as predicted, the greater the shape similarity of a molecule to a biologically active structure, the more likely is that molecule to exhibit the same activity.

To help in visualizing the generation of the topomeric shape similarity descriptor, Figure 7 shows the topomeric conformations for three of the seven most active ***a*** side chains, overlayed on the topomeric conformations of the ***a*** rings from their most shape similar query structures.

Discussion

Relation between Shape Similarity and the Frequency of Activity. As emphasized in the Introduction, this experiment had only one objective: to evaluate whether a shape similarity technology proposed for selecting synthetic targets from large virtual libraries could indeed identify active structures. The outcome, that all of the highly active molecules were suggested by shape similarity and that no highly active molecules at all were found among the 5 times more numerous structures synthesized for other reasons, seems to be rather substantial evidence for the efficacy of the shape similarity searching methodology. The most optimistic extrapolation of these results would suggest that assiduous application of the technology in lead follow-up might identify the same number of actives with 6-fold less synthesis and testing.

It can be objected that our data set is better described as 15 groups of structures differentiated by their ***b***−***c*** building block (Figure 3) than as 425 individual structures, since all of the "highly active" compounds included the ***b***−***c*** building block that the four query structures also included. Finding all seven actives among these 30 structures (first column in Table 2 or 3) is an even more significant enrichment than is finding the seven among the 63 that are shape similar to the query structure. Indeed we agree that in hindsight it would have been preferable for the synthesized molecules to have included more shape similar variations in ***b***−***c***. Fortunately, more could be done with the existing data. We removed these 30 biassing compounds from the 425 and repeated the tabulation that generated Table 4, with the first shape similarity bin deleted. Of course, since no ">75% displacement" structures remain, the χ^2 test of the distribution of highly active structures is no longer applicable. Nevertheless the *F*-test for a nonzero slope when correlating "frequency of >30% displacement" with "shape similarity" can still be performed, and the resulting $F(1,2)$-ratio of 47.2 ($p < 0.05$) is actually larger than any shown in Table 4. We conclude that the hypothesis fundamental to the use of shape similarity in lead follow-up remains statistically validated by the results: the more similar in shape a structure was to that of a known active query molecule, the more likely that structure was to be active. The persistence of this relationship even after removal of the most influential group of structures is one reason we place as much weight on the consistency of the overall results as on the dramatic segregation of high activity within the shape similarity selections.

How important was the topomer shape similarity descriptor used? Would some other diversity descriptor have performed as well or better? The results for the four other descriptors, in blocks B−E of Table 4, show retrospectively that our prospective focus on the topomer descriptor was completely appropriate. Although the 2D fingerprint Tanimoto showed some neighborhood behavior (block B), the Tanimotos of all the synthesized structures with respect to any of the queries were very low (maximum of 0.76), well below the accepted "neighborhood radius" of 0.85 for this descriptor. So, according to this descriptor, all of the synthesized structures are too dissimilar to any of the queries to be active, and the experiment should not have been performed. The electrotopological index and molecular weight were even worse than useless, with the relation between descriptor similarity to a query and the probability of activity being *negative* (blocks D and E). The descriptor remaining, atom pairs, does exhibit a significant neighborhood behavior (block C), although the tendency for structures atom-pair-similar to any of the four query structures to have an enhanced likelihood of activity is not as strong as for topomeric shape.

There may be a tendency, especially among those already familiar with structure−activity relationships for A-II antagonism, to interpret all these positive results as trivial consequences of structural similarities that are obvious (to a chemist, retrospectively). However, if these structural similarities are not correctly recognized by three of four other popular diversity descriptors examined, can the prospective effectiveness of the topomer shape descriptor really be considered a trivial matter?

How important was the starting point? The four query structures are all much more potent, by at least 2 orders of magnitude, than are any of the synthesized structures. To be useful in lead follow-up, shape similarity to weakly active query structures must also be predictive of activity. However, the structures at hand were not selected with weakly active structures as the queries. Unlike the four query structures which generated these structures, the "query structures" referenced in the rows of Table 5 are not at the center of one of four shape similarity defined hyperspheres, but rather more or less closer to one of the hypersphere surfaces. Nevertheless, as indicated in Table 5 and further demonstrated in the Supporting Information, an enhanced frequency of activity was very often found for shape similarity to weakly active compounds, as well as for the shape similarity to the most potent compounds.[34]

Table 5. χ^2 Tests of Significance for the Dependence of Frequency of Biological Activity (A-II Displacement) on Topomeric Shape Similarity to Various Active Compounds, for Various Activity Ranges[a]

compd			χ^2 values[a]					
BC	A	activity (% displ)	>75%	40−75%	30−40%	20−30%	n^c signif	wgt n^d signif
A	8	69	14.62	19.5	6.19	5.51	2	8
B	3	37	35.08	55.19	27.46	18.35	4	19
C	1	38	6.78	17.95	19.8	12.8	3	11
C	11	31	1.38	2.28	7.18	8.23	1	1
C	33	35	2.09	0.4	6.75	7.21	0	0
D	40	47	3.51	5.49	3.32	4.11	0	0
E	12	33	0.57	0.22	13.67	5.16	1	3
E	29	31	1.22	2.94	13.21	9.06	2	3
F	23	30	2.5	3.12	3.31	1.34	0	0
H	22	21	0.97	0.84	1.96	1.52	0	0

[a] Representative results. The Supporting Information includes results for 95 compounds, including all displacing >30% of A-II. [b] Each χ^2 value tests the probability that the distribution of the stated range of activity among all 425 compounds is related to their topomeric shape similarity to the compound labeling that row. In other words, each line of this table corresponds to any sixth line of data in each block of Table 4. [c] The number of χ^2 values in this row which are statistically significant ($p < 0.10$). [d] The number of χ^2 values in this row which are statistically significant, weighted by the degree of significance (weights are as follows: $p < 0.10$, 1; $p < 0.5$, 2; $p < 0.025$, 3; $p < 0.01$, 4; $p < 0.005$, 5; $p < 0.001$, 6).

To summarize these supporting analyses, the efficacy of similarity searching in identifying active structures depended very much on how similarity is described, but much less on the degree of activity of the query structure.

Structural Novelty. The structural novelty among the shape similar selections that were actually synthesized is admittedly limited. (As shown in Table 3, the structures having a shape difference less than 120 combine 35 of the 44 variations of the ***a*** ring with 3 of the 14 variations in the ***b***−***c*** moiety). This lack of novelty results from the conservative bias in the selection of synthetic targets (in order to minimize the risk of not producing enough data to analyze[35]), as well as vagaries in the availability and reactivity of building blocks. Yet the initial similarity search results too were input constrained, first by searching only within the single virtual library synthesized by the route in Figure 2 and second by considering only products that can be formed from building blocks offered commercially or otherwise at hand. The first of these constraints no longer exists. Studies have recently been performed in which "typical lead structures" became queries for the successful searching of multiple virtual libraries.[36] Furthermore, the resulting hits almost always promise "lead hopping" or "lead crossing", which we might define as "finding other 2D structures ('chemotypes') that produce the same 3D shape, hence biological effect, as exhibited by a lead structure of interest". The second constraint might better be characterized as an inherent limitation in the variety of structures accessible by combinatorial chemistry (i.e., short sequences of reactions applicable to a wide variety of readily available building blocks).[37]

Moreover, structural novelty is not always of highest priority, particularly in the lead exploration stages where similarity searching has the relatively greatest value. True, it is often stated that when evaluating the output of any computer-aided molecular design effort, only structures that would not otherwise have been considered by the medicinal chemist are of great value. But, as drug discovery research evolves from a craft to a process, and as the quantity of SAR data explodes, this belief requires reassessment. Specifically, the value of techniques such as shape similarity searching can also reside in the following:

- superior ability to objectively indicate structures most likely to be active (as reported herein);
- much greater speed, both in processing time itself and in lessened demands on experienced medicinal chemists, when selecting the next synthetic targets;
- higher organizational efficiency, in taking fullest advantage of cumulative synthetic experiences.

Structure−Activity Relationships. Along with more active structures, it was mentioned that a successful "hit follow-up" study will also distinguish the structural features that are required for activity from those whose modification may improve properties. With respect to the sartans, the two groups[38,39] who have published the most SAR data both concluded that the critical features are the acidic group at the ortho position of the ***c*** ring, and in the ***a*** ring an H-bond accepting group (typically aromatic nitrogen) adjacent to a short and unbranched lipophilic side chain. Our experimental results yield identical conclusions, most strikingly for the ***a*** ring where shape similarity alone selected the structures. Of the 44 ***a*** ring variants investigated, 10 (1, 2, 3, 5, 7, 8, 9, 11, 33, 44) have the aromatic nitrogen adjacent to a side chain. These include 5/7 (0.714) of the ">75%" compounds and 13/20 (0.650) of the ">40%" compounds, which are obviously much higher frequencies than the 2/418 (0.005) and 7/405 (0.017) frequencies of activity among the corresponding sets of compounds lacking either of these critical features. Indeed this discrimination among ***a*** ring structures exceeds the discrimination in favor of *ortho* acidic groups (COOH or tetrazole) in the ***c*** ring, where the corresponding "activity enrichments" would be 7/30 (0.233) and 11/61 (0.180) compared to 0/395 (0.000) and 9/364 (0.025). (In a next round of synthesis and testing, such SAR observations could be used to refine shape similarity searching, either implicitly simply by using the seven hits as new queries or explicitly by using other searching techniques to filter the shape similarity search results.) Thus both of the structural features believed to be important for displacement of angiotensin II were clearly revealed within the results of an essentially automatic design process using a completely different selection criterion.

"Lead Hopping". Often a newly discovered hit will prove to have undesirable properties, such as low potency or selectivity, unacceptable pharmacodynamic

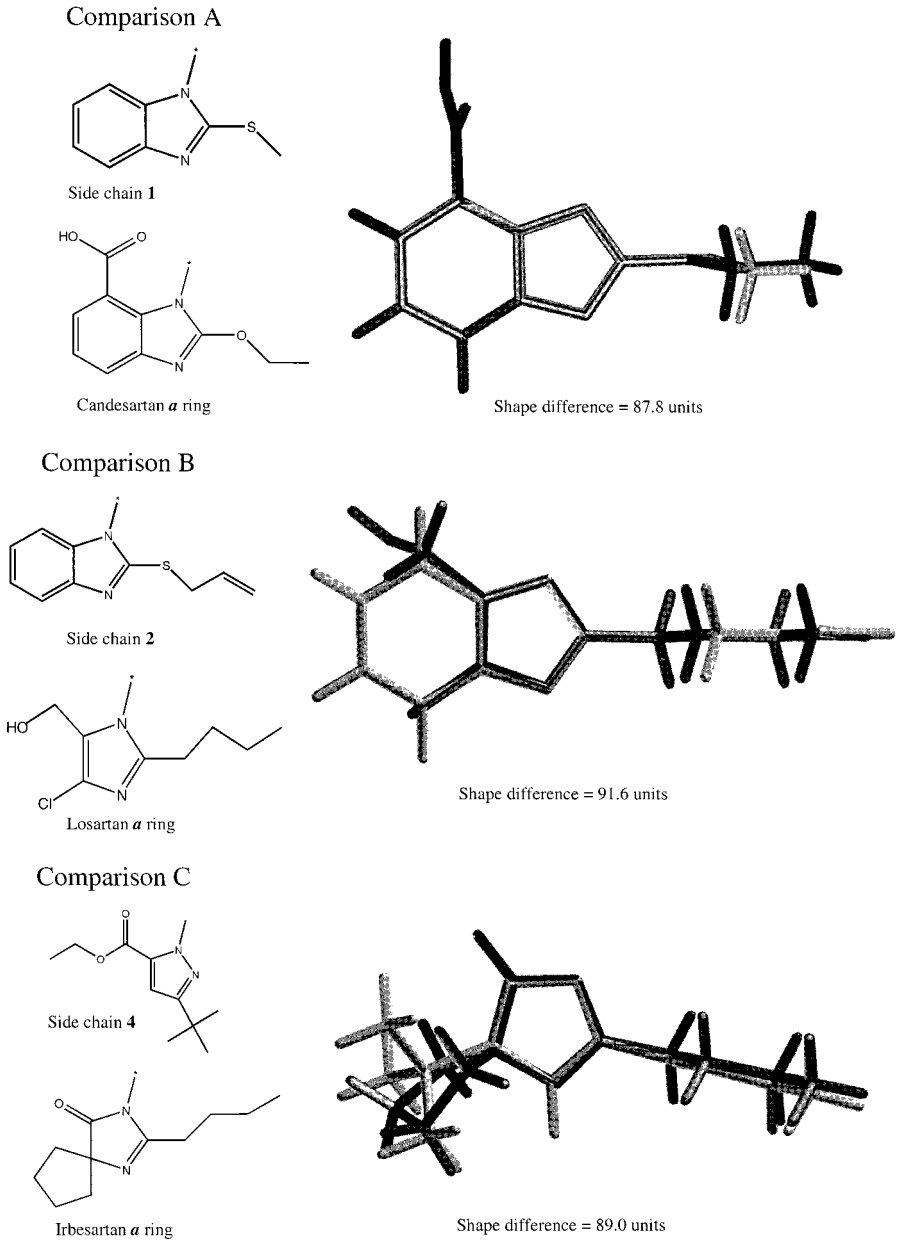


Figure 7. Overlays of three topomeric conformations. Each overlay compares the ***a*** ring from one of the query "sartan" structures with an ***a*** ring from one of the seven "highly active" structures reported in this work. For reference, structural drawings of the two ***a*** rings appear to the left of the overlaid conformations.

properties, or competitive patent coverage. Particularly in the case of an adverse patent position, it then becomes necessary to "lead hop", i.e., to seek structures which are, on one hand, sufficiently similar in their 3D features to be active at the same receptor and, on the other hand, sufficiently different in their 2D features (the basic vocabulary from which patent claims are constructed) to become an independent intellectual property. By its nature, shape similarity searching promises to be a very powerful technology for identifying possible "lead hopping" structures.

However, consider an apparent counter-example, the original four query structures of Figure 1, as outcomes of successful "lead hopping" (since each was efficacious enough to become a commercial product and each is independently patented). The shape similarities formed by each pairing of the four structures are as follows: losartan vs tasosartan, 147; losartan vs irbesartan, 121;[40] losartan vs candesartan, 147; tasosartan vs irbesartan, 149; candesartan vs tasosartan, 141; irbesartan vs candesartan, 166. Evidently none could have been identified by a single round of shape similarity searching as likely to share the biological activity of another. How then might shape similarity searching have helped to achieve these particular examples of actual lead hopping? In principle this could involve some mix of the following three strategies:

1. Iterative Searching, by Repetitive Rounds of Successful Neighborhood Searching and Synthesis. (To visualize how this process can work, recall that not all of the neighbors of my neighbor's home are also neighbors of my home.) For example, there is a published account[41] of the discovery of ranitidine (Zantac) as a lead hopping alternative to cimetidine (Tagamet). Although ranitidine and cimetidine are not themselves neighbors, it was found[42] that every structure in the

sequence of structural changes that generated ranitidine from burimamide was a shape neighbor of its predecessor.

2. Sampling More Distant Neighborhoods. Our results show that the incidence of activity continued to fall off as shape similarity decreased, even beyond the original search radius of 120. Therefore, when successful "lead hopping" is mandatory to a discovery program, the screening of structures which are only roughly shape similar to the lead but otherwise acceptable should be more productive than completely random screening (rather as, geographically, islands are most commonly found relatively near to other islands). Typically the number of hits increases sharply with the shape radius,[43] so that the neighborhood-based techniques used in general screening library design may also be useful here in trimming an excessively large set of synthesis candidates to a manageable number. Conventional 3D pharmacophore searching among such a large set of roughly shape similar structures is another obvious stratagem, one which would have been especially productive for "lead hopping" among A-II antagonists, judging from the extensive published literature.[44]

3. Following a Trend. Shape similarity searching seems to imply that, on average, all shape changes are both equivalent and bad, but of course the objective of a medicinal chemist is to identify particular kinds of shape change that are not bad but good. As mentioned above, the current shape searching methodology could be extended to actively detect and exploit trends. First, the regions of space could be recognized where shape changes are compatible with or, better, favorable to biological activity.[45] Second, this information could be exploited during shape similarity searching by favoring rather than penalizing shape change at the appropriate grid points.

Despite the conservative synthetic strategy, these experimental results do include a few "lead hopping" possibilities. In particular, consider comparison C of the shape overlays shown in Figure 7 between a commercially available pyrazole having a particular substitution pattern and the corresponding ***a*** ring of irbesartan. The 2D structures of these two fragments are very different. Yet, as can be visualized from the shape overlay in Figure 7 and as reported in Table 3, the shape similarity of these two substituted ***a*** rings to each other is the third greatest among commercially available ***a*** rings to the ***a*** ring in any of the four query structures. Also noteworthy about this pyrazole is its lack of the H-bond accepting feature usually believed to be a critical feature for potent A-II receptor binding. Considering also the apparent absence of any pyrazole feature from the extensive published literature on candidate A-II antagonists (though not from the patent literature[46]), it seems reasonable to propose this result as a realized example of "lead hopping". If this structure instead is "intuitively obvious", it is remarkable that, despite its evident synthetic accessibility and high potency in a widely performed assay, a technique such as topomeric shape similarity searching seems to have been required for its discovery.

Relation to Other Methods. As discussed in the Introduction, similarity searching has its greatest value in following up hits, especially when no additional information such as a receptor structure or other SAR data is available. Topomer shape similarity searching has the particular virtue, unique as far as is known, of being so fast as to make searching even among "all readily accessible drug-like structures" a realistic long-range goal.[47] Another value of shape as a similarity metric is compatibility with other methods. The effects on shape of any structural modification are well-understood, and the methodological pathway from shape-based similarity to receptor-based or SAR-based design, which will be traveled as more information emerges in progressing from hit follow-up to lead optimization, is clear and continuous.

Conclusions

The automated selection process of topomer shape similarity searching within large virtual libraries was found to be highly effective for identifying other active compounds, in an unambiguously prospective experiment. Also noteworthy about this selection process are its great speed and scope, thoroughness, and compatibility with the fundamental "high throughput" changes underway in the drug discovery process. While more examples of successful application are highly desirable, all available information supports the proposition that topomer shape similarity searching might usefully enhance the effectiveness of the overall drug discovery process.

Acknowledgment. We thank David Floyd for leadership and David Patterson, Robert Clark, Fred Soltanshahi, Michael Lawless, Stefan Guessregen, Kathe Andrews-Cramer, and the remarkably able referees for technical support and valuable discussions.

Supporting Information Available: χ^2 tests of significance for the dependence of frequency of biological activity (A-II displacement) on topomeric shape similarity to various active compounds, for various activity ranges. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) "Drug" is intended herein as a convenient abbreviation referring to any substance having a useful effect on any biological system.
(2) Babine, R. E.; Bender, S. L. *Chem. Rev.* **1997**, *47*, 1359−1472.
(3) (a) Hansch, C. *Acc. Chem. Res.* **1969**, *2*, 232. (b) Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967. Many excellent discussions of current issues are in (c) *3D-QSAR in Drug Design: Recent Advances* and (d) *3D-QSAR in Drug Design: Ligand-Protein Interactions and Molecular Similarity*; Kubinyi, H., Folkers, G., Martin, Y. C., Eds.; KLUWER/ESCOM, Kluwer Academic Publishers: Dordrecht, The Netherlands, 1997.
(4) Two exceptions to this generalization are "3D database searching", applicable to collections of specific compounds containing no more than 10^6 structures (a good leading reference is Wang, T.; Zhou, J. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 71−77), and "de novo design", generation of (often synthetically challenging) structures under constraints, usually to fit well into an experimentally determined structure of a receptor cavity (Rotstein, S. H.; Murcko, M. A. GroupBuild: a fragment-based method for de novo drug design. *J. Med. Chem.* **1993**, *36*, 1700; DeWitte, R. S.; Shaknovich, E. I. SMoG: de novo design method based on simple, fast, and accurate free energy estimates. 1. Methodology and supporting evidence. *J. Am. Chem. Soc.* **1996**, *118*, 11733−11737.).
(5) Apart from the current study, to the best of our knowledge the only other *prospective* medicinal chemistry experiment reported, where evaluation of a general structure selection methodology

was the primary objective, is Kick, E. K.; Roe, D. C.; Skillman, A. G.; Liu, G.; Ewing, T. J. A.; Sun, Y.; Kuntz, I. D.; Ellman, J. A. Structure-Based Design and Combinatorial Chemistry Yield Nanomolar Inhibitors of Cathepsin-D. *Chem. Biol.* **1997**, *4*, 297−307.

(6) However, for exploring aromatic substituent variations in this situation, the Topliss tree continues to be a valuable tool, which though used manually is based on "classical QSAR" principles. Topliss, J. G. *J. Med. Chem.* **1972**, *15*, 1006.

(7) While the nomenclature associated with drug discovery is imprecise at best, a "hit" is here taken to be the reproducible observation of an exceptional biological effect (observed in fewer than one in 10^3 tests of randomly chosen molecules) caused by a known concentration of a defined substance.

(8) In the absence of *any* such tendency for "similar" compounds to have similar properties, no rational strategy for compound selection *can* exist.

(9) For example, molecular weight, though very useful in separation, has no "neighborhood behavior" for biological properties. Accordingly, few chemists would "follow up" a "hit" structure having a molecular weight of 375.2 by eliminating from consideration structures having a molecular weight, for example, less than 325.2 or greater than 425.2.

(10) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood behavior: a useful concept for validation of molecular diversity descriptors. *J. Med. Chem.* **1996**, *39*, 3049−3059.

(11) Matter, H. Selecting optimally diverse compounds from structure databases: a validation study of two-dimensional and three-dimensional molecular descriptors. *J. Med. Chem.* **1997**, *40*, 1219−1229.

(12) Brown, R. D.; Martin, Y. C. Use of structure−activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.

(13) Cramer, R. D.; Clark, R. D.; Patterson, D. E.; Ferguson, A. M. Bioisosterism as a molecular diversity descriptor: steric fields of single topomeric conformers. *J. Med. Chem.* **1996**, *39*, 3060−3069.

(14) Patani, G. A.; LaVoie, E. J. Bioisosterism: a rational approach in drug design. *Chem. Rev.* **1996**, *96*, 3147−3176.

(15) Cramer, R. D.; Patterson, D. E.; Clark, R. D.; Soltanshahi, F.; Lawless, M. S. Virtual libraries: a new approach to decision making in molecular discovery research. *J. Chem. Inf. Comput. Sci.* **1998**, *6*, 1010−1023.

(16) Carr, G. The pharmaceutical industry: the alchemists. *The Economist*, Feb. 21, 1998, special section, pp 1−5.

(17) ChemSpace is a licensed trademark of Tripos, Inc.

(18) Ferguson, A. M.; Patterson, D. E.; Garr, C. D.; Underiner, T. L. Designing Chemical Libraries for Lead Discovery. *J. Biomol. Screening* **1996**, *1*, 65−73.

(19) This compendium of all commercially offered compounds may be obtained from MDL Information Systems, Inc., 140 Catalina Street, San Leandro, CA 94577.

(20) The actual query fragment for ***c*** contained a −COOH instead of the tetrazole found in the query sartans. As would be expected for such a widely recognized bioisosteric pair (shape difference of 38 units), however, this difference would have had very little effect on the search results. However, the shape differences critical to this work, those used to evaluate the results shown, are based on the corrected structure, the one containing the tetrazole.

(21) Rusinko, A., III; Skell, J. M.; Balducci, R.; McGarity, C. M.; Pearlman, R. S. Program available from Tripos, Inc., St. Louis, MO.

(22) Because this sum over grid points is a root square sum, not a simple sum, these shape differences increase less rapidly than does a volume difference, such as that produced by the SYBYL MVOLUME command. For example, the shape difference between pentyl and methyl is twice the shape difference between ethyl and methyl, not 4 times.

(23) The lattice used in all topomer calculations has Angstrom extents of −4 to 14, −12 to 6, and −8 to 10 along the *X*, *Y*, and *Z* axes respectively, the fragments being rooted at the origin and conformation generation biased to emphasize negative *Y* and especially positive *X* space.

(24) The nominal units of "steric shape unit" are the same as those of CoMFA steric fields, kcal/mol per lattice point. In all topomeric shape difference calculations, we use the "standard CoMFA" defaults for calculating the steric fields. Major implications of these defaults on the magnitudes of shape similarity values include a maximum steric field difference of 30.0 units/lattice point and a lattice point density of 1 point per 8 Å^3. However, in standard CoMFA, field values at multiple lattice points are combined essentially by a straight summation, rather than by the root-mean-square calculation used here with topomeric fields.

(25) The inherent uncertainty in any steric shape difference calculated by these methods is no less than 20−30 units. This much uncertainty arises from the coarseness of the lattice point grid (2.0 Å spacing) over which the field value differences are summed.

(26) It must be emphasized that the medicinal chemists did *not* believe that useful A-II antagonism was very likely to result from 13 of these side chains, based on their own previous experience in synthesizing prospective A-II antagonists by conventional means, as well as the published literature. However, our intent was to simulate a hit follow-up experiment, in which such additional SAR information would be lacking. In this situation, such a repositioning of a critical functional group (here the acidic carboxyl or tetrazole) has been a very common tactic. Such a tactic also facilitates the reuse of experience (the results of validation or "rehearsal" studies) gained with the first synthesis and therefore becomes even more attractive in combinatorial syntheses.

(27) Technology available from IRORI, Torrey Pines Science Park, 11149 N. Torrey Pines Rd., La Jolla, CA 92037.

(28) The IC_{50} values reported for the first seven compounds in column A of Table 3 are, in order: 141, 171, 158, 206, 131, 540, and 324 nM. The four query structures are roughly 100 times more potent.

(29) It was not possible to obtain a more direct assessment of the systematic error for this assay. An isolated erroneous value, such as the single −48% in Table 3, was almost certainly a localized failure in correctly performing the assay, caused, for example, by some malfunction of the robot liquid handler.

(30) The URL http://www.math.gatech.edu/~spruill/m4216/projsf97/ChiSquare.html reports a probability threshold of 9.2×10^{-7} for a χ^2 value of 60 with 4 degrees of freedom.

(31) An apparent anomaly in block A, that the lowest *F*-ratios are associated with the highest χ^2 values, is because a single straight line does not well describe the flat part of the curve, the values at or near 0%, which dominate the leftmost data sub-blocks.

(32) Perhaps the reason for observing "neighborhood behavior" among such low Tanimoto coefficients is an unusually small number of common bits set as the sartan 2D fingerprints are generated, because of the almost "undecorated" biphenyl moieties common to all structures.

(33) The apparently "significant" observations in blocks D and E of Table 4 actually have the wrong sign (by suggesting that *increasing* electrotopological index and molecular weight differences are associated with *decreasing* similarities in biological properties).

(34) Furthermore, the shape similar neighbors of weakly active structures were frequently more active. Of course, this result is almost a logical necessity, since the similarity of A to B is the same as the similarity of B to A, so B can lead to A or A to B, regardless of which is the more potent or which is first identified.

(35) Even with a tested protocol at hand, synthesis could be attempted for only 44 of 631 ChemSpace-designated ***a*** rings and was carried through successfully for about $^2/_3$ of the full combinatorial array. Therefore the original concern, that a less conservative synthetic strategy might not generate enough useful data, seems to have been justified.

(36) Andrews-Cramer, K.; Cramer, R. Manuscript in preparation. Note, however, that these "successes" are almost entirely computational, since the only available experimental information on the biological properties of the structures found by shape similarity searching is any other SAR information in the article from which the query was taken.

(37) Indeed, it is confidently predicted that the majority of the structures discussed in this issue of this Journal could not have been prepared by such combinatorial methods!

(38) Duncia, J. V.; Carini, D. J.; Chiu, A. T.; Johnson, A. L.; Price, W. A.; Wong, P. C.; Timmermans, P. B. M. W. The discovery of DuP 753, a potent, orally active nonpeptide angiotensin II receptor antagonist. *Med. Res. Rev.* **1992**, *12*, 149−191.

(39) Prendergast, K.; Adams, K.; Greenlee, W. J.; Nachbar, R. B.; Patchett, A. A.; Underwood. D. J. Derivation of a 3D pharmacophore model for the angiotensin-II site one receptor. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 491−511.

(40) The current version of the topomeric shape searching software actually generates a shape difference of 186 units between the ***a*** rings of irbesartan and losartan. This mistake occurs because the topomeric conformation rules rotate the ***a*** rings 180° away from their most similar relative superposition, with their butyl side chains ending up on opposite sides of the ***a*** rings. However, such a risk of a "false negative" result in topomeric shape similarity searching is currently being alleviated by generating additional "topomeric" conformations for the occasional query structures which have great asymmetry in shape but low asymmetry in numbers of heavy atoms.

(41) Ganellin, C. R. In *Medicinal Chemistry − The Role of Organic Chemistry in Drug Research*; Ganellin, C. R., Roberts, S. M., Eds.; Academic Press: London, 1993; pp 227−255.

(42) Internal study by S. Guessregen of Tripos. Many additional examples of topomer shape similarity among ligands binding to the same or similar receptors have very recently been generated (K. Andrews-Cramer, to be presented at the Fall American Chemical Society National Meeting, 1999, New Orleans).
(43) To a first approximation, the number of products increases by almost a factor of 10 for each increase of 10 units in the shape similarity searching radius.
(44) Bradbury, R. H.; Allott, C. P.; Dennis, M.; Fisher, E.; Major, J. S.; Masek, B. B.; Oldham, A. A.; Pearce, R. J.; Rankins, N.; Revill, J. R.; Roberts, D. A.; Russell, S. T. New nonpeptide angiotensin II receptor antagonists. 2. Synthesis, biological properties, and structure−activity relationships of 2-alkyl-4-(biphenylylmethoxy)quinoline derivatives. *J. Med. Chem.* **1992**, *35*, 4027−4038. Masek, B. B.; Merchant, A.; Matthew, J. B. Molecular shape comparison of angiotensin II receptor antagonists. *J. Med. Chem.* **1993**, *36*, 1230−1238. Lin, H.-S.; Rampersand, A. A.; Zimmerman, K.; Steinberg, M. I.; Boyd, D. B. Nonpeptide angiotensin II receptor antagonists: synthetic and computational chemistry of N-[[4-[2(2H-tetrazol-5-yl)-a-cycloalken-1-yl]phenyl]methyl]imidazole derivatives and their in vitro activity. *J. Med. Chem.* **1992**, *35*, 2658−2667. Perkins, T. D. J.; Dean, P. M. An exploration of a novel strategy for superimposing several flexible molecules. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 155−172. Also see ref 35.
(45) For more discussion of this problem, see: Martin, Y. C. in ref 3c, pp 3−23.
(46) Carini, D. J.; Wells, G. J.; Duncia, J. V. Eur. Pat. Appl. 323,-841, 1989. Ross, B. C.; Middlemiss, D.; Eldred, C. D.; Montana, J. G.; Shah, P. Eur. Pat. Appl. 446,062, 1991.
(47) Complete topomeric shape searches of 5×10^{12} "drug-like" structures routinely take a few hours on standard workstations. (For comparison, this number of structures is approximately 20 000 times greater than the number of structures currently registered by Chemical Abstracts Services.)

JM990159Q