Molecular Hashkeys: A Novel Method for Molecular Characterization and Its Application for Predicting Important Pharmaceutical Properties of Molecules

Anwar M. Ghuloum, Carleton R. Sage,* and Ajay N. Jain†

MetaXen, 280 East Grand Avenue, South San Francisco, California 94080

Received September 18, 1998

We define a novel numerical molecular representation, called the molecular hashkey, that captures sufficient information about a molecule to predict pharmaceutically interesting properties directly from three-dimensional molecular structure. The molecular hashkey represents molecular surface properties as a linear array of pairwise surface-based comparisons of the target molecule against a common 'basis-set' of molecules. Hashkey-measured molecular similarity correlates well with direct methods of measuring molecular surface similarity. Using a simple machine-learning technique with the molecular hashkeys, we show that it is possible to accurately predict the octanol−water partition coefficient, log *P*. Using more sophisticated learning techniques, we show that an accurate model of intestinal absorption for a set of drugs can be constructed using the same hashkeys used in the aforementioned experiments. Once a set of molecular hashkeys is calculated, its use in the training and testing of property-based models is very fast. Further, the required amount of data for model construction is very small. Neural network-based hashkey models trained on data sets as small as 30 molecules yield statistically significant prediction of molecular properties. The lack of a requirement for large data sets lends itself well to the prediction of pharmaceutically relevant molecular parameters for which data generation is expensive and slow. Molecular hashkeys coupled with machine-learning techniques can yield models that predict key pharmacological aspects of biologically important molecules and should therefore be important in the design of effective therapeutics.

Introduction

Computational techniques for structure-based drug design have shown utility in the design and identification of ligands of therapeutic targets.[1] Iterative structure-based drug design has yielded potent and specific ligands for therapeutically relevant targets.[2−6] However, techniques that have been applied to prospective design of compounds with improved pharmacological profiles have not shared the same degree of success. Optimization of pharmacological properties of lead molecules is the bottleneck in the development of a clinical candidate drug. Most potential drugs fail because of deficiencies in the pharmacological profile of the molecule. The development of computational methods to accurately predict pharmacological properties would speed clinical drug candidate development and more importantly increase the probability of success in the clinic. This approach allows for the integrated optimization of potency/selectivity in parallel with the pharmacological profile.

Underlying the processes of absorption, distribution, metabolism, and excretion (ADME) lie at least two different phenomena: interactions of molecules with specific proteins (e.g., in metabolism and active transport) and interactions of molecules with bulk solvent-like systems (e.g., partitioning into a cell membrane). In the first case, a molecule's behavior will tend to be dominated by its propensity to adopt a specific conformation and align itself with respect to complementary protein binding sites. In the second case, the molecule's behavior will be dependent on its ensemble behavior in a bulk fluid system. In both cases, however, the interactions occur at the junction of molecular surfaces, suggesting that as a sensible first approximation, a three-dimensional surface-based representation may exist that satisfies the requirements of computational models for predicting ADME properties.

In a purely theoretical sense, it should be possible to predict ADME properties on the basis of molecular structure alone, since molecules elicit a specific response from animals of a particular species. For an organism in a given genetic and environmental state, there is enough information in a molecular structure for the organism to "decide" the molecule's fate. However, two questions arise. First, how much data (molecule/value pairs) will be necessary in order to build predictive models of ADME properties? Second, how can one best represent the essential characteristics of molecules in a manner that leads to efficient generalization from machine-learning techniques?

The definition of a molecular representation that can capture important aspects of molecules is a critical aspect in the application of computational methods to drug design. Currently used methods include two-dimensional topological descriptors, energetic descriptors, quantum mechanical descriptors, and three-dimensional field descriptors (see review in ref 7). Two other methods have been described which describe molecules based on either the behavior of a molecule in

* Corresponding author. Fax: (650) 553-8101. E-mail: carleton_sage@metaxen.com.
† Current address: Iconix Pharmaceuticals, 850 Maude Ave, Mountain View, CA 94043.

10.1021/jm980527a CCC: $18.00
Published on Web 04/24/1999

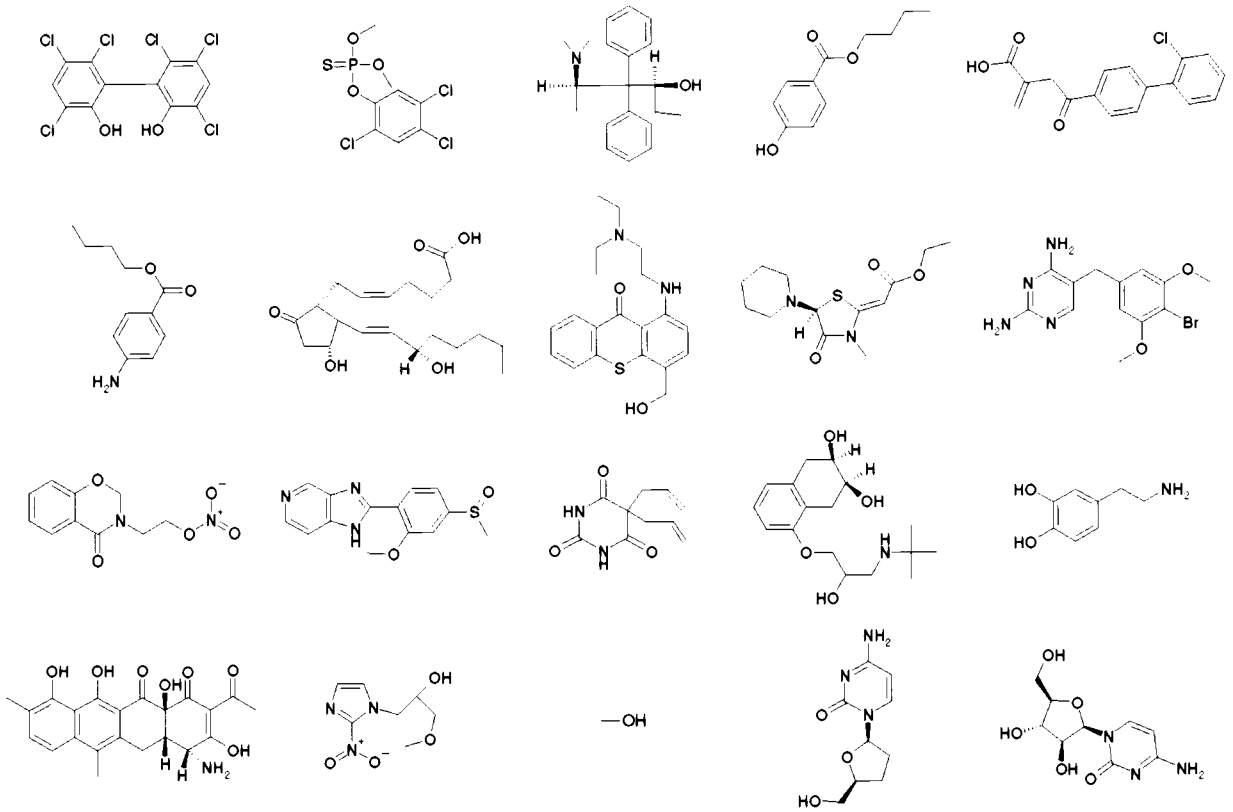

Figure 1. Twenty basis molecules used for molecular hashkeys arranged by high to low measured log *P*.

a variety of biochemical assays[8] or the molecular docking of a target molecule to a variety of different protein sites[9] to generate "fingerprints" which were used to determine the relationships between molecules. Molecular representations have been used to determine quantitative structure−activity/property relationships[10] as well as to predict physicochemical properties of molecules such as partition coefficients and solubilities.[11−13] Recently, surface property-based descriptors have been applied to the prediction of intestinal absorption.[14]

In this work we define a novel numerical molecular representation, termed the molecular hashkey, which represents a general and compact method to represent the surface properties of molecules. A hashkey may always be computed for a molecule, unlike some other modeling systems that use fragment-based representations (e.g., ClogP). After hashkey calculation, the use of the hashkey in model generation and prediction is very fast. We show that molecular hashkeys can be used to predict a measure of molecular surface similarity that is well-correlated with specific binding to proteins.[15,16] To examine the power of the molecular hashkey in the prediction of pharmacological properties of molecules, we used data available in MDL's Comprehensive Clinical Medicinal Chemistry[17] database to generate models to predict the octanol−water partition coefficient (log *P*) as accurately as currently available methods. Finally, using a structurally diverse set of drugs selected because they are predominantly absorbed through passive processes and are not subject to early metabolism,[14] we used molecular hashkeys to predict human intestinal absorption.

Theory

In this section, we will present an abbreviated description of molecular hashkeys along with the applications to predicting molecular similarity and log *P*. The Experimental Section will give details of the computational methodology.

Molecular Hashkeys. A molecular hashkey[18] is a real-valued vector of fixed dimension that captures information about the surface properties of a molecule. A molecular hashkey has the property that molecules with similar hashkeys will appear similar based on observation of their surfaces. Molecules with identical surface properties will have identical hashkeys, independent of the underlying atomic scaffolding.

Given a molecule *M*, its *N*-dimensional hashkey $(H_1, H_2, ..., H_N)$ is computed by calculating its molecular surface similarity to a set of *N* basis molecules. The basis molecules are in low energy-fixed conformations. *M* is flexibly aligned to each B_i of the set of basis molecules $(B_1...B_N)$ to maximize molecular surface similarity, and the best match yields the surface similarity value that becomes H_i.

The molecular surface similarity computation used here is similar to published methods[16,19,20] and will be described in detail in the Experimental Section. The surface similarity values range from 0 to 1.0, with 0 denoting maximal dissimilarity and 1.0 denoting identity. In this work, we used no other representation besides molecular hashkeys to represent the molecules whose properties we wanted to predict.

Initial Hashkey Representations. Figure 1 shows the 20 molecules used as the basis set in initial tests of the molecular hashkey approach. These were selected randomly from the set of CMC[17] molecules that had

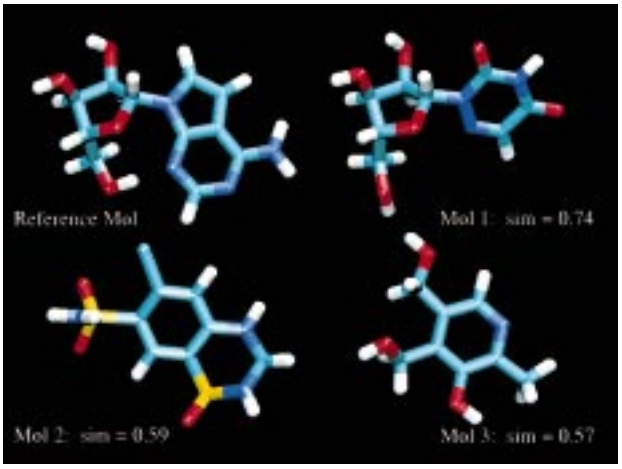

Figure 2. Four molecules from the CMC data set with similar hashkeys. Surface similarity measurements are shown for each molecule as compared to the reference molecule.

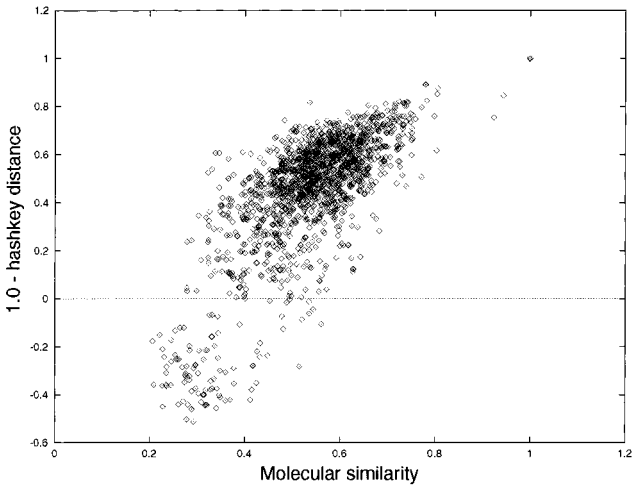

Figure 3. Plot of molecular surface similarity versus (1.0 − hashkey distance) for 1600 pairs of molecules chosen at random from the CMC data set.

measured log *P*'s (referred to as the full CMC set). It should be noted that they represent a wide variety of molecular shapes, sizes, polarities, hydrophobicities, and other physicochemical properties. Figure 2 shows a molecule (upper left) with three molecules (from the full CMC set) that have the most similar hashkeys. The molecules shown appear to share structural similarity both in terms of overall shape and in terms of substituents with similar physicochemical properties (e.g., hydrophilic moieties such as hydroxyls and sulfonamides).

Experimental Results

Four sets of experiments were performed. First, to demonstrate that molecular hashkeys correlate with specific binding to protein active sites, using our initial 20-member basis set, a comparison of molecular hashkey distances was made to directly computed molecular similarities. Second, to demonstrate that the representation is suitable for computing properties related to solvent interactions, again using our initial 20-member basis set, log *P* values were predicted using varying numbers of molecules in the training sets for model building. Third, experiments were performed in which the number and composition of the basis set members were varied in order to evaluate the importance of basis set size and composition in using molecular hashkeys as a surface-based descriptor in the prediction of log *P*. Last, to demonstrate the utility of the representation in predicting ADME properties of molecules, we used molecular hashkeys to predict human intestinal absorption for a structurally diverse set of drugs.

Molecular Similarities. Beyond interactions with solvent systems, for molecular hashkeys to be useful in predicting ADME properties, it must also be possible to compute values correlated with specific binding events. As detailed in the Experimental Section, the surface similarity function that underlies the molecular hashkeys is related to other surface-based molecular representations of similarity.[16,19,20] These computational methods use three-dimensional metrics to model the three-dimensional relationships or similarities between molecules. The results of the computational approaches have been confirmed in protein−ligand systems where the relationships between topologically different molecules have been determined experimentally in X-ray crystallographic studies,[21] suggesting that three-dimensional metrics can effectively model the three-dimensional properties that play a role in molecular recognition. To the extent that molecular hashkey distance correlates with a direct measure of molecular surface similarity of the sort used here, we gain confidence that the hashkey representation captures information relevant to specific protein binding events.

We measured the correlation between the Euclidean distance of the hashkeys of molecule pairs to molecular surface similarity. Figure 3 shows a plot relating hashkey distance to direct molecular surface similarity for 1600 randomly chosen pairs of molecules (chosen from the 863 used in the log *P* experiment). The *x*-axis is molecular surface similarity, and the *y*-axis is 1.0 less the hashkey distance (so that 1.0 indicates a perfect match in both cases). The correlation, while being somewhat spread out toward the lower surface similarity values, is highly statistically significant ($p \ll 0.01$, based on PRCC[16]).

This correlation supports the notion that information in molecular hashkeys can be related to specific binding to protein active sites. At the very least, in terms of a practical application of molecular hashkeys, one can speed up three-dimensional similarity searches of large databases significantly. By precomputing hashkeys for molecules in a large database and then computing the hashkey for a probe molecule, one can very rapidly identify a small subset of molecules in the database that may have high molecular surface similarity to the probe molecule. In this example, one can eliminate 80% of the computationally expensive molecular surface similarity comparisons performed and still retain 80% of the molecules that score better than 0.70 in terms of direct molecular surface similarity. More significant reductions in computational time are possible where higher surface similarity values are desired. Further optimization of the hashkey basis set (both size and composition) will enhance this benefit.

Log *P* Prediction. Computational approaches to

Table 1. Performance of Log *P* Prediction Using Molecular Hashkeys with a Weighted KNN Classifier ($K = 5$)

hashkey	*N* training	*N* testing	mean error (log units)	PRCC (1.0 log)
KNN cross-validation	862	1 (863 times)	0.94	0.85
KNN 400 × 2	400	463 (2 times)	1.03	0.83
KNN 200 × 4	200	663 (4 times)	1.15	0.80
KNN 100 × 8	100	763 (8 times)	1.23	0.77

predicting octanol−water partition coefficients typically work by one of two general methods: by analyzing the fragments of a molecule in a two-dimensional topological manner[11] or by calculating parameters which represent specific molecular properties[12,13] and using mathematical models to fit the data. Molecular hashkeys are sensitive only to three-dimensional surface properties. Of the molecules in the CMC,[17] 863 compounds with measured log *P* values had molecular structures that were amenable to the computations involved (see the Experimental Section for details). For each molecule, the hashkey computation was performed once, and this hashkey was used in all subsequent experiments.

We employed a very simple machine-learning tool called a weighted *K* nearest neighbor (KNN) classifier[22] to construct computational models of log *P* using molecular hashkeys as input. The KNN method was chosen initially for ease of application, and if a simple learning method can develop accurate models for a particular parameter, then it is likely that more sophisticated learning methods will perform even better. A KNN classifier holds a collection of training data consisting of vectors that represent objects along with associated values. The classifier assigns a value to an input object based on the "votes" of the *K* nearest "neighbors" to the input object from the training data. So, if $K = 3$, and the two closest examples to an input object are of class A and one is from class B, the input object is assigned to class A. A *weighted* KNN classifier can be used for generating real-valued output. With this modification, the classifier returns the weighted output values of the *K* nearest neighbors to the input object, where the weight is simply the inverse of the distance to each neighbor (plus a small constant to avoid division by zero). The output is normalized by the sum of the weights.

Table 1 summarizes the results of log *P* prediction using KNN classifiers constructed with various amounts of data. The first row shows the result for a leave-one-out cross-validation, where 863 classifiers were constructed, each by withholding one data point that was then predicted. The optimum value for *K* was 5 and was determined empirically. The mean predicted error of log *P* was 0.94, with a pair rank correlation coefficient (PRCC) of 0.85, which is highly statistically significant ($p \ll 0.01$, see the Experimental Section for details on the PRCC and statistical significance). The next rows show the degradation of performance as less data is used. It should be noted that performance degrades very little; just 200 molecules yield a model of log *P* with a mean error of 1.15.

As an independent test of log *P* prediction using a fragment-based method, we chose to use the widely used ClogP program.[11,23] Table 2 summarizes the performance of the ClogP[11,23] program on the 799 molecules (of the original 863) for which the computation terminated successfully. The overall performance had a 0.51 log unit of mean error and a PRCC of 0.96, which may appear to be significantly better than the hashkey technique. However, 635 of the 799 molecules in the set were part of the ClogP[11,23] tuning set. Performance on those 635 had a mean error of 0.39 log unit. Performance on the 134 that were novel had a 0.97 log unit mean error, with a PRCC of 0.89. This level of performance is indistinguishable from that of the hashkey technique for molecules not used in model construction. Figures 4 and 5 show plots of measured versus predicted log *P* for the 863 hashkey-based predictions from the cross-validation experiment and the 134 ClogP[11,23] predictions on nontraining molecules.

This experiment illustrates several points. The hashkey technique is able to perform as well as ClogP on

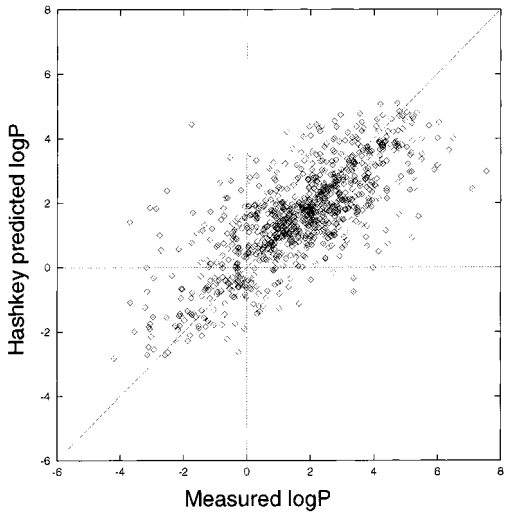

Figure 4. Plot of measured log *P* versus predicted log *P* for 863 molecules using the molecular hashkey method with a basis set size of 20 molecules.

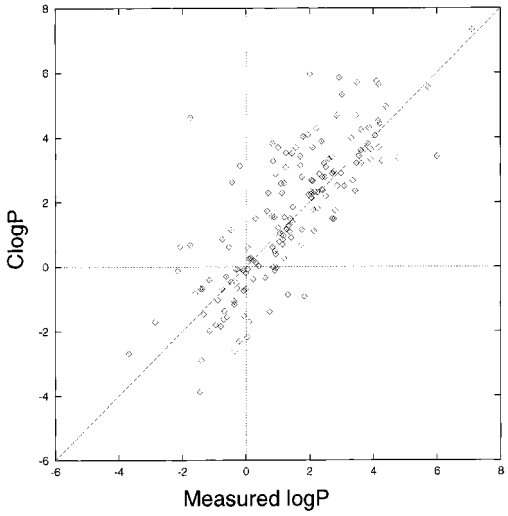

Figure 5. Plot of measured log *P* versus ClogP for 134 molecules, which were not part of the ClogP training set.

Table 2. Performance of ClogP Prediction

ClogP	*N* training, *N* testing	mean error (log units)	PRCC (1.0 log)
all compounds	>9000, 799	0.51	0.96
training compounds	>9000, 665	0.39	0.98
testing compounds	>9000, 134	0.97	0.89

Table 3. Performance of Log *P* Prediction Using Molecular Hashkeys with a Neural Network with 3 Hidden Units and Training Sets Selected for Structural Diversity

hashkey	*N* training, *N* testing	mean error (log units)	PRCC (1.0 log)
NN 30	30, 833	1.21	0.80
NN 50	50, 813	1.29	0.77
NN 70	70, 793	1.21	0.80
NN 100	100, 763	1.18	0.81

this population of biologically relevant molecules and will give a prediction for every molecule in the set, in contrast to the ClogP method which fails if a particular fragment is missing. In addition, using the hashkey technique, it is not necessary to use a tremendous amount of data to make accurate predictions. A small number of hundreds of molecules are sufficient to train an accurate model, whereas ClogP uses over 9000 molecules in its training set. This observation is important because in most medicinal chemistry drug development projects it is unlikely that with the relatively low throughput of most ADME assays, there would be enough time or resources to generate the large amounts of data (e.g., 9000 values) to create a model that could be implemented in the project. Furthermore, it is likely that the surface-sensitive approach embodied by the hashkey technique is capturing the relevant information about molecules with respect to interaction with solvent systems, since no complex parametrization of the molecules with respect to conjugation, intramolecular hydrogen bonding, shielding, or other effects was necessary to achieve this level of performance.

The ability to build high-quality predictive models using relatively small and carefully selected data sets can be enhanced using more sophisticated machine-learning techniques. Models of log *P* were constructed using a neural network[24] and relatively small log *P* data sets selected from the CMC[17] data set, using only structural diversity as a selection criteria. This roughly approximates an environment in which early data on a relatively diverse, small set of molecules from a pharmacology group would be used to build initial models in a drug design project. Table 3 shows the results of these experiments over a range of small training sets. Using data sets as small as 30 molecules from the CMC,[17] neural network models were constructed that were nearly as accurate as KNN-based models, which used an order of magnitude more training data. The neural networks used were 20−3−1 back-propagation networks[24] with sigmoidal activation functions. The learning rate for training was 0.2, with a momentum term of 0.1. Models were trained to within an error threshold of 0.8 log unit for all training examples. All parameters were empirically determined.

Effects of Basis Set Size and Composition. The size and composition of the basis set may influence the representational power of the hashkeys. If a small basis set is chosen, the basis molecules may not contain enough structural information to model relevant physicochemical phenomena. Further, if the basis set members are chosen in a contrived manner to bias against the discrimination of certain critical molecular properties, their utility in building predictive models will be reduced. For example, if one were to select basis set members completely lacking in polar features, models using hashkeys would be seriously hampered in their ability to model phenomena in which hydrophilicity or numbers of acceptors or donors play important roles. Likewise, a very high degree of redundancy in the basis set would be tantamount to selecting a very small basis set and would thus limit the predictive power of models using hashkeys by missing important areas of chemical space.

Table 4. Performance of Log *P* Prediction Using Rigid Molecular Hashkeys with a Weighted KNN Classifier ($K = 5$) and Randomly Chosen Basis Sets

basis set		train, test set size: 400 (×3), 463	train, test set size: 100 (×3), 763
1	PRCC (1.0 log)	0.79	0.75
	mean error (log units)	1.13	1.24
2	PRCC (1.0 log)	0.80	0.73
	mean error (log units)	1.13	1.3
3	PRCC (1.0 log)	0.81	0.76
	mean error (log units)	1.12	1.22
4	PRCC (1.0 log)	0.80	0.74
	mean error (log units)	1.10	1.27
5	PRCC (1.0 log)	0.80	0.74
	mean error (log units)	1.12	1.27

An ideal selection of a basis set would maximize the orthogonality of the structural properties of the molecules. However, the strength of the results obtained with the initial basis set shows that the use of hashkeys tolerates well the degree of redundancy or bias that may arise from picking molecules at random from the CMC. To examine the effects of altering the composition of the basis sets, several other basis sets were selected randomly from the CMC, varying in size up to 30 molecules. As an approximation of the standard hashkey computation, which utilizes fully flexible molecular alignments to determine surface similarity, hashkeys were computed for these secondary basis sets using rigid alignments of minimized conformations. Despite that, the results of using these basis sets demonstrate the reproducibility of the ability of hashkeys to generate statistically significant models for log *P*. Table 4 shows the results of five KNN-trained models of log *P* using five different and nonoverlapping randomly chosen basis sets of size 20. The results show that the independent models deviate only slightly from each other for each basis set and that a randomly chosen set of basis set molecules allows sufficient encoding of molecular surface properties to predict log *P*. The results also demonstrate that despite using a less accurate alignment technique, the performance of the models on the alternative basis sets is still comparable.

Since a randomly chosen basis set is sufficient to encode the surface properties of a particular molecule in the molecular hashkey, experiments were performed to examine the consequences of altering the size of the hashkey basis set. Figure 6A shows log *P* model performance using 400 molecules in the training set and hashkeys created with a rigid hashkey computation where basis set size was varied from 3 to 30 members. The experiments demonstrate that, for the purposes of computing log *P*, hashkey basis set size was important, but model performance as measured by changes in mean error and PRCC reached a plateau well short of 20 molecules for the rigid hashkey computation. Figure 6B shows log *P* model performance using 400 molecules in the training set and hashkeys created using a flexible

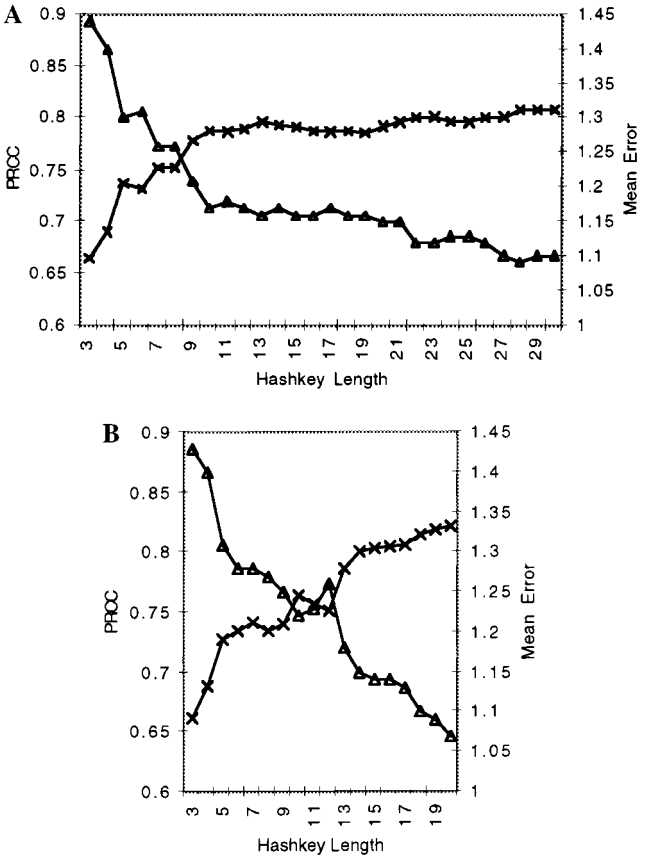

Figure 6. A. Plot of PRCC (×) and mean error (△) in predicted log *P* (relative to measured log *P*) versus basis set size for rigid hashkeys computed using basis sets ranging from 3 to 30 members. B. Plot of PRCC (×) and mean error (△) in predicted log *P* (relative to measured log *P*) versus basis set size for flexible hashkeys computed using basis sets ranging from 3 to 20 members.

hashkey computation where the basis set size was varied from 3 to 20 members. For the flexible hashkey computation, it is observed that the mean error and PRCC begin to plateau at 20 molecules at a slightly higher level of accuracy. Since less information is encoded in each hashkey basis set member in the rigid hashkey models, the information contribution of each additional basis set member is quickly diminished. Thus, the rigid hashkey models plateau much earlier than the flexible hashkey model. The overall difference in maximum accuracy for log *P* in the range of basis set sizes examined is fairly small. However, for less noisy data and more complex properties, one would expect the difference in accuracy to increase substantially.

Predicting Intestinal Absorption of Drugs. The foregoing experiments showed that the hashkeys contain sufficient information to predict molecular similarities (which correlate well with specific binding) as well as log *P* (a bulk property of molecules). Hashkeys implicitly contain information about properties of molecules which may otherwise have to be discovered before a significant correlation to their physicochemical or ADME properties can be established. For example, Palm et al.[14] have determined that an excellent nonlinear correlation existed between the dynamic polar surface area descriptor and the fraction absorbed (FA) in humans for a structurally diverse set of drugs. FA for this set ranged from 0.3 to 100%. The development of this model (and previous models) required the discovery of an appropriate molecular descriptor or ad-hoc combinations of theoretical descriptors. Using the same data set used by Palm et al.,[14] it can be demonstrated that hashkeys implicitly contain information relevant to predicting intestinal absorption for this set of compounds.

The drugs selected by Palm et al.,[14] shown in Figure 7, are predominantly absorbed through passive processes and are not subject to early metabolism. Furthermore, the molecules are all relatively soluble. We performed a leave-one-out cross-validation experiment using hashkeys (as computed above) as the molecular representation and trained a 20−1−1 back-propagation neural network using the reported FA data. We used the PRCC as the measure of the performance of the model and assumed that relative variances of 10% were acceptable (reported deviations for the measured FA ranged from 0.3 to 20%, with a mean of 8.75%). The PRCC for this experiment, whose predicted versus measured values are shown in Table 5 and plotted in Figure 8, was 0.9 (with $p \ll 0.001$, a highly statistically significant result). While the accuracy of the predictions is not as good as that for a fitted function (as one would expect; measuring accuracy using molecules whose data was used to fit the function is analogous to training and testing a model using the same molecules), the relative ranking of the drugs based on predicted FA, shown in Table 6, is very accurate. This is not surprising since Palm et al. selected these drugs with a bias toward transepithelial transport phenomena in which bulk surface properties were the critical factor. We believe that by breaking down transport phenomena into discrete factors influencing transport (i.e., Will the molecule passively diffuse through the lipid bilayer of the cell membrane? Is it a substrate for an efflux pump?), we can build more accurate predictive models. These simpler models can then be composed to predict biological phenomena for more complex systems involving a number of these simpler mechanisms. This experiment is a promising indicator of success.

Discussion

In the Introduction it was proposed that the information contained in a small molecule is sufficient to yield a deterministic result in a particular species with respect to a physiological response. The critical issue in many machine-learning problems is data representation. It is theoretically possible to construct predictive models independent of the representation of the input objects. However, if the remapping process is sufficiently complex as to dwarf the signal related to the learning problem, it will be difficult to generate a good model or it will require an inordinate amount of data.

The smaller the data set required to generate a predictive model, the more likely it is that the input representation is capturing the information relevant to the learning task in a compact form. For a typical small molecule of therapeutic relevance, approximately 30 bytes is required in order to fully represent two-dimensional molecular structure (using a form of SMILES strings with a reduced alphabet). It would take approximately 500 bytes to represent the three-dimen-

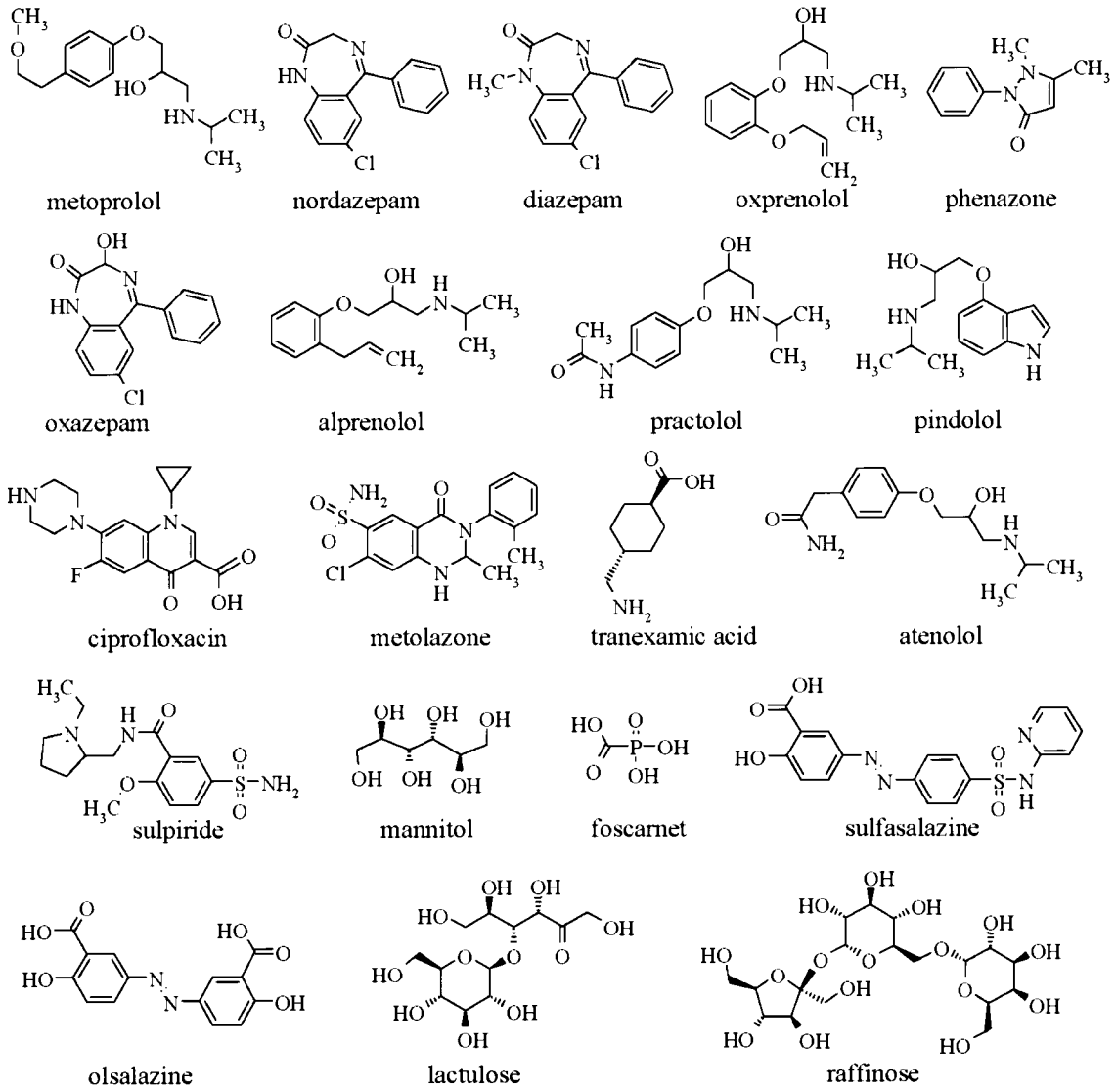

Figure 7. Drugs used in intestinal absorption model construction and verification.

Table 5. Measured versus Predicted Fraction Absorbed

	measured (%)	predicted (%)
metoprolol	102	59.4
nordiazepam	99	83.9
diazepam	97	99.0
oxprenolol	97	98.1
phenazone	97	97
oxazepam	97	78.5
alprenolol	96	100
practolol	95	70.6
pindolol	92	86.9
ciprofloxacin	69	62.5
metolazone	64	45.6
tranexamicacid	55	74.4
atenolol	54	74.7
sulpiride	36	54.5
mannitol	26	37.7
foscarnet	17	26.3
sulfasalazine	12	37.9
olsalazine	2.3	62.2
lactulose	0.6	20.7
raffinose	0.3	3.3

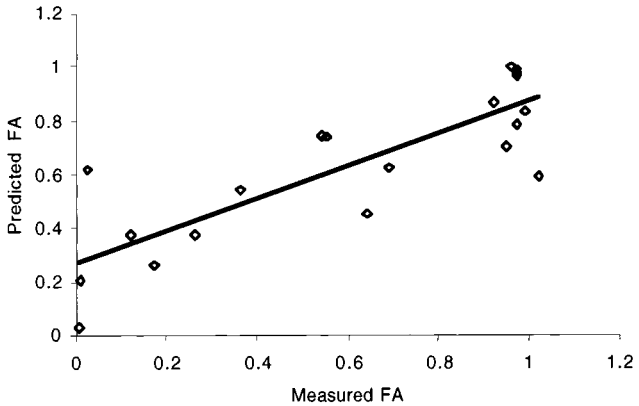

Figure 8. Plot of measured fraction absorbed of drugs after administration to humans versus fraction absorbed predicted via hashkey-based models constructed in leave-one-out cross-validation.

sional structure of the same molecule using a coordinate-based system. The molecular hashkeys use approximately 50 bytes per molecule, which is on the same order as the amount required to capture two-dimensional structure. Molecular hashkeys appear to efficiently remap molecular structure information to a representation that is dependent on observable surface properties. They apparently yield a propitious starting point from which to do machine learning.

Two other groups have suggested somewhat similar approaches to the molecular hashkey technique reported

Table 6. Measured versus Predicted Rank Order of Drugs Based on Fraction Absorbed, from Highest to Lowest

measured	predicted
metoprolol	alprenolol
nordiazepam	diazepam
diazepam	oxprenolol
oxprenolol	phenazone
phenazone	pindolol
oxazepam	nordiazepam
alprenolol	oxazepam
practolol	atenolol
pindolol	tranexamicacid
ciprofloxacin	practolol
metolazone	ciprofloxacin
tranexamicacid	olsalazine
atenolol	metoprolol
sulpiride	sulpiride
mannitol	metolazone
foscarnet	sulfasalazine
sulfasalazine	mannitol
olsalazine	foscarnet
lactulose	lactulose
raffinose	raffinose

here. Kauvar et al.[8] described molecules as a "fingerprint" of activities derived from a panel of biochemical assays. Briem and Kuntz[9] based their representations on molecular docking of small molecules to a variety of different protein sites to form "fingerprints" and used them to compute molecular similarities; their results parallel those reported here to a degree. However, their system underperformed a two-dimensional-based similarity technique on the benchmark they used; this could be related to the scoring function employed to evaluate the quality of molecular dockings.

An important feature of molecular hashkey-based models, one that differentiates this work from that of Palm et al.[14] and others,[11−13,25,26] is that it required no explicit descriptor of molecular structure other than a hashkey. Hashkeys are reinforced by these experiments as a viable representation for analyses across a range of pharmaceutically relevant properties. Rather than requiring significant time investment in exploring ad-hoc descriptors and descriptor combinations, building usable, prototype- and production-quality structure/surface-based models of ADME or physicochemical properties using data sets of tractable size is simply a matter of using hashkeys in a machine-learning system.

In the drug discovery process, database screening can be significantly accelerated by using precomputed hashkeys to subselect relevant compounds from large databases. Furthermore, direct models of physicochemical processes (certainly log *P* but perhaps solubility, passive diffusion, paracellular transport, etc.) can be applied simultaneously to speed the iterative stages of drug lead optimization.

Conclusions

The molecular hashkey technique captures sufficient information about small molecules to predict properties that depend on two of the principle underlying processes in the ADME profile of a drug: specific binding to protein active sites and interaction with solvent systems. Further, it does so in a representation that is amenable to unsophisticated machine-learning techniques. The hashkey technique should yield significant benefits in direct application to specific drug discovery projects. Additional refinement of the technique coupled with exploration of larger data sets of ADME parameters using more sophisticated machine-learning systems may yield insight into the underlying phenomena that currently constitute a largely impenetrable "black box" of bioavailability in drug discovery.

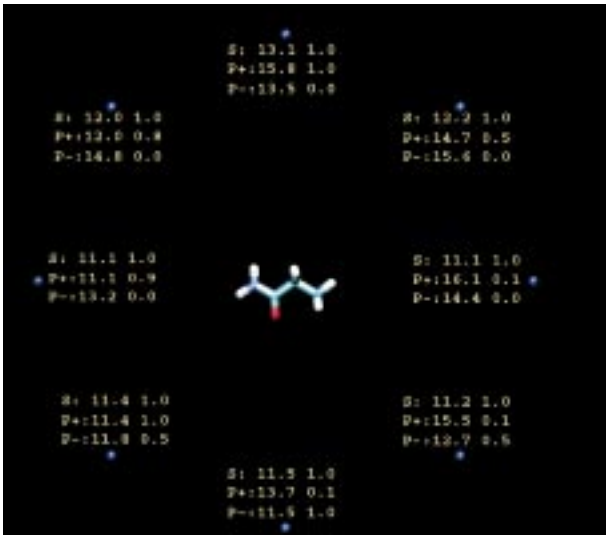

Figure 9. Molecular features for a simple molecule (2D slice). At each point around the molecule, values are reported in pairs for steric (S), polar positive (P+), and polar negative (P−) features of the molecule. The first number in each pair represents the inverse field strength of the feature, and the second number in each pair represents the directionality component of the feature.

Experimental Section

This section will give details of the molecular surface similarity computation as well as the particulars of computing the molecular hashkeys used in the study.

Similarity Definition. The surface similarity function is a normalized Gaussian-like function of the differences between the molecular features of two molecules. Features are measured on the surface of a sphere surrounding a molecule. The radius of the sphere is 2 Å over the maximum possible distance between two atoms in the molecule. Figure 9 illustrates the molecular features for a small molecule in two dimensions. At each point, six values are reported, two each for the steric, polar positive, and polar negative features. The first number reported is the molecule's inverse field strength (defined below), which can be thought of as the distance to an electron density isosurface of the molecule in Angstroms. The second number is the degree of directional match of the molecule with respect to the feature reference point. Note the relationship between directionality of the donor atoms and the acceptor atoms with respect to the feature point locations.

For a particular conformation and alignment of a molecule (a pose), the features are formally defined as follows. For the pure shape component, a molecular field is defined, which falls off exponentially with distance from the molecular surface (approximated as the van der Waals surface). The field values at a set of points placed uniformly on a sphere form the shape component of the feature representation. For convenience, inverse field strength values are used, as they can be thought of as having units of Angstroms. Let a_i denote the position of atom i and r_i denote its van der Waals radius. Let p_i denote the position of feature point i. Also, let d denote the Euclidean distance between two point positions. Then:

$$F_i(m) = \frac{-\log(F_i(m))}{\alpha}$$

$$F_i(m) = \sum_j e^{-\alpha D_{ij}}$$

$$D_{ij} = d(p_i,a_j) - r_j$$

where D is simply the distance from a point to the van der Waals surface of a molecule. For the experiments reported here, $\alpha = 10.0$. The function f_i is the sum of the exponentially decreasing fields of each of j atoms, and F_i is the inverse field function which yields an Angstrom-like unit. Note that F_i is continuous and differentiable with respect to the pose of m. This is a critical feature since many optimization techniques rely on well-defined gradients.

Defining the polar features is slightly more complicated. While it is possible to define a polar field in the same manner as the steric field based on the subset of atoms that have the proper polarity, this ignores the important contribution of directionality in many polar interactions. If we imagine our feature reference points to be observers looking for a particular polarity emanating from a particular direction, it is possible to define two values: one corresponding to field strength and one corresponding to directional compatibility.

The polar field value at a feature point is defined similarly to the shape field, except that, instead of distances relative to each atomic surface dominating the measured field, directional match to each polar atom is used. At each feature point, both positive and negative polar features are computed. The following describes the computation for positive polar features. The computation for negative polar features is analogous. Each atom j of m that can participate as a hydrogen-bond donor or a positively charged salt-bridging moiety is identified. For each of these, a direction is defined corresponding to a unit length vector pointing from the centroid of atoms connected to j to j itself (denoted u_j below). For each feature point i, v_i denotes a unit length vector pointing at the center of the sphere on which the feature point resides. The final unit length vector w_{ij} points from p_i to a_j. The polar field strength S at point i is defined as follows:

$$S_i^d(m) = \frac{-\log(S_i^d(m))}{\alpha}$$

$$S_i^d(m) = \sum_j e^{-\alpha P_{ij}}$$

$$P_{ij} = -\frac{1}{1 + e^{-\alpha((v_i w_{ij})(-u_j \cdot w_{ij})-\beta)}}$$

For the experiments reported here, $\beta = 0.6$. At each feature point, the polar field value is dominated by the most favorable interaction on the molecule from a directional perspective. This value ignores proximity. As discussed above, proximity must be measured relative to the dominating interaction. This value is the distance to each polar atom weighted by its contribution to the field strength at p_i. The distance is normalized by the total strength at p_i. The result is as follows:

$$F_i^d(m) = \frac{\sum_j e^{-\alpha P_{ij}} D_{ij}}{\sum_j e^{-\alpha P_{ij}}}$$

This yields a weighted distance, with the most important polar interaction dominating the computation. The analogous computation is done for acceptors. For convenience in defining the surface similarity function, we define S for the steric component of the feature representation to be 1 for all feature points.

To summarize, the feature representation for a molecule m in a particular pose is a set of six values per feature point: three pairs of feature values and strengths with one pair for each of pure shape, positive polar, and negative polar components. Each of the feature values can be thought of as having units of Angstroms, and each of the strengths is between 0 and 1.

The surface similarity function is very simple. For convenience, let F_i denote the full set of feature values (including steric, positive polar, and negative polar) and S_i denote the full set of strength values. Then, the surface similarity s of molecules m_1 and m_2 in a particular relative alignment and in particular conformations is:

$$s(m_1,m_2) = \frac{\sum_i \max(S_i(m_1),S_i(m_2)) \cdot g(F_i(m_1) - F_i(m_2),\sigma_1) \cdot g(S_i(m_1) - S_i(m_2),\sigma_2)}{\sum_i \max(S_i(m_1),S_i(m_2))}$$

$$g(x,\sigma) = e^{-x^2/\sigma}$$

So, if the poses of two molecules have identical feature values and strengths, the function s will return a value of 1.0. Any deviations from noncoincidence will decrease the value of s, with a minimum of 0. Places where no polar strength is observed received little weight with regard to concordance of polar feature values. Note that the metric is symmetrical with respect to the molecules. For the experiments reported here, $\sigma_1 = 2.0$ and $\sigma_2 = 1.0$.

Molecular Hashkey Computation. A hashkey is computed by first sampling the conformations of a molecule. For this work, stochastic sampling with a maximum of 20 conformations was used, beginning from protonated, MM3[27]-minimized Concord[28]-generated three-dimensional structures. The protonation states for acidic and basic groups were assigned using heuristics to reflect the state of the molecule at neutral pH. Rings were not searched, and bump relaxation was utilized in lieu of full minimization of each of the resulting conformations. Each H_i in the hashkey of molecule M is simply the surface similarity of the best-matching conformation of M in its maximally similar alignment to B_i (basis molecule i (see Figure 1)).

All computations were performed on desktop Silicon Graphics workstations, with processor configurations including R4400, R5000, and R10000, each with 128 Mb of RAM. Each surface similarity computation takes 5−8 s per conformer. Each hashkey computation thus requires 30−45 processor minutes. Note, however, that the hashkey computation itself is highly parallel in nature, for both individual molecules and collections of molecules. The speedup realized by adding more processors scales linearly. For significantly larger sets of molecules, farms of inexpensive compute engines (e.g., midrange personal computers) can be easily deployed to provide a computational throughput proportional to the number of processors.

Pair Rank Correlation Coefficient and Statistical Significance. The PRCC is a nonparametric measure of rank order correlation, related to Kendal's tau, but with the addition of a real-valued notion of "ties" in rank.[16] Given two lists, the first the target values of a prediction and the second the predicted values, the PRCC is simply the number of correctly ranked pairs divided by the number of pairwise comparisons. A parameter Δ gives the difference between two target values that is considered sufficient to warrant a comparison of rank. So, in the example of PRCC of similarities correlated with hashkey distances, Δ was 0.05, which is the level at which two surface similarity values should be considered different. For the oral bioavailability data, Δ was 0.0, and for the log P data, Δ was 1.0. All computations of p values were done numerically using the precise distribution of the target values by generating large sets of random correlations to estimate likelihood of observing high PRCC values by chance. This yields a very accurate measurement of statistical significance.

Acknowledgment. The authors thank Drs. Michael J. Ross and Lutz B. Giebel for supporting this work as well as for many valuable discussions.

References

(1) Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual Screening − an overview. *Drug Discovery Today* **1998**, *3*, 160−178.
(2) Bodian, D. L.; Yamasaki, R. B.; Buswell, R. L.; Stearns, J. F.; White, J. M.; Kuntz, I. D. Inhibition of the fusion-inducing conformational change of influenza hemagglutinin by benzoquinones and hydroquinones. *Biochemistry* **1993**, *32*, 2967−2978.
(3) Ring, C. S.; Sun, E.; McKerrow, J. H.; Lee, G. K.; Rosenthal, P. J.; Kuntz, I. D.; Cohen, F. E. Structure-based inhibitor design by using protein models for the development of antiparasitic agents. *PNAS* **1993**, *90*, 3583−3587.
(4) Rutenber, E.; Fauman, E. B.; Keenan, R. J.; Fong, S.; Furth, P. S.; Demontellano, P. R. O.; Meng, E.; Kuntz, I. D.; Decamp, D. L.; Salto, R.; Rose, J. R.; Craik, C. S.; Stroud, R. M. Structure of a non-peptide inhibitor complexed with HIV-1 Protease: Developing a cycle of structure-based drug design. *J. Biol. Chem.* **1993**, *268*, 15343−15346.
(5) Shoichet, B. K.; Stroud, R. M.; Santi, D. V.; Kuntz, I. D.; Perry, K. M. Structure-based discovery of inhibitors of thymidylate synthase. *Science* **1993**, *259*, 1445−1450.
(6) Jones, T. R.; Varney, M. D.; Webber, S. E.; Lewis, K. K.; Marzoni, G. P.; Palmer, C. L.; Kathardekar, V.; Welsh, K. M.; Webber, S.; Matthews, D. A.; Appelt, K.; Smith, W. W.; Janson, C. A.; Villafranca, J. E.; Bacquet, R. J.; Howaland, E. F.; Booth, C. L. J.; Herrmann, S. M.; Ward, R. W.; White, J.; Moomaw, E. W.; Bartlett, C. A.; Morse, C. A. Structure-based design of lipophilic quinazoline inhibitors of thymidylate synthase. *J. Med. Chem.* **1996**, *39*, 904−917.
(7) 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.
(8) Kauvar, L. M.; Higgins, D. L.; Villar, H. O.; Sportsman, J. R.; Engqvist-Goldstein, A.; Bukar, R.; Bauer, K. E.; Dilley, H.; Rocke, D. M. Predicting ligand binding to proteins by affinity fingerprinting. *Chem. Biol.* **1995**, *2*, 107−118.
(9) Briem, H.; Kuntz, I. D. Molecular similarity based on dock-generated fingerprints. *J. Med. Chem.* **1996**, *39*, 3401−3408.
(10) Brown, R. D.; Martin, Y. C. The information content of 2D and 3D structural descriptors relevant to ligand−receptor binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1−9.
(11) Leo, A. J. Calculating log Poct from Structures. *Chem. Rev.* **1993**, *93*, 1281−1306.
(12) Bodor, N.; Gabanyi, Z.; Wong, C. K. A new method for the estimation of partition coefficient. *J. Am. Chem. Soc.* **1989**, *111*, 3783−3786.
(13) Moriguchi, I.; Hirono, S.; Liu, Q.; Nakagome, Y.; Matsushita, Y. Simple method of calculating octanol/water partition coefficient. *Chem. Pharm. Bull.* **1992**, *40*, 127−130.
(14) Palm, K.; Stenber, P.; Luthman, K.; Artursson, P. Polar molecular surface properties predict the intestinal absorption of drugs in humans. *Pharm. Res.* **1997**, *14*, 568−571.
(15) Jain, A. N.; Koile, K.; Chapman, D. Compass: Predicting Biological Activities from Molecular Surface Properties. Performance Comparisons on a Steroid Benchmark. *J. Med. Chem.* **1994**, *37*, 2315−2327.
(16) Jain, A. N.; Harris, N. L.; Park, J. Y. Quantitative binding site model generation: Compass applied to multiple chemotypes targeting the 5HTla receptor. *J. Med. Chem.* **1995**, *38*, 1295−1307.
(17) *Comprehensive Medicinal Chemistry Database*. CMC, 1997.
(18) The term "hashkey" is taken from Computer Science. A hashkey is a compact numerical representation of an object that is used to solve indexing problems by storing objects using their hashkeys as memory addresses.
(19) Blaney, F.; Finn, P.; Phippen, R.; Wyatt, R. Molecular surface comparison: Application to drug design. *J. Mol. Graph.* **1993**, *11*, 98−105.
(20) Blandon, P. A rapid method for comparing and matching the sperical parameter surfaces of molecules and other irregular objects. *J. Mol. Graph.* **1989**, *7*, 130−137.
(21) Dunn, W. J.; Hopfinger, A. J.; Catana, C.; Duraiswami, C. Solution of the conformation and alignment tensors for the binding of trimethoprim and its analogues to dihydrofolate reductase: 3D-quantitative structure−activity relationship study using molecular shape analysis, 3-way partial least squares regression, and 3-way factor analysis. *J. Med. Chem.* **1996**, *39*, 4825−4832.
(22) Duda, R. O.; Hart, P. E. In *Pattern Classification and Scene Analysis*; Duda, R. O., Hart, P. E., Eds.; John Wiley & Sons: New York, 1973.
(23) MacLogP 2.0; Biobyte, 1997.
(24) Rumelhart, D. E.; McClelland, J. L. *Parallel Distributed Processing*; The MIT Press: Cambridge, MA, 1986; Vol. 1.
(25) Sutter, J. M.; Jurs, P. C. Prediction of aqueous solubility for a diverse set of heteroatom-containing organic compounds using a quantitative structure−property relationship. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 100−107.
(26) Dressman, J. B.; Amidon, G. L.; Fleisher, D. Absorption potential: estimating the fraction absorbed for orally administered compounds. *J. Pharm. Sci.* **1985**, *74*, 588−589.
(27) Allinger, N. L.; Yuh, Y. H.; Lii, J.-J. Molecular mechanics. The MM3 force field for hydrocarbons I. *J. Am. Chem. Soc.* **1989**, *111*, 8551−9556.
(28) Rusinko, A. I.; Skell, J. M.; Balducci, R.; McGarity, C. M.; Pearlman, R. S. *CONCORD: a program for the rapid generation of high quality 3D molecular structures*: Tripos Associates: St. Louis, MO.

JM980527A