Automated Selection of Compounds with Physicochemical Properties To Maximize Bioavailability and Druglikeness

Taiji Oashi, Ashley L. Ringer, E. Prabhu Raman, and Alexander D. MacKerell, Jr.*

Department of Pharmaceutical Sciences, School of Pharmacy, University of Maryland, Baltimore, 20 Penn Street, Baltimore, Maryland 21201, United States

Received September 9, 2010

Adequate bioavailability is one of the essential properties for an orally administered drug. Lipinski and others have formulated simplified rules in which compounds that satisfy selected physiochemical properties, for example, molecular weight (MW) $\leq$500 or the logarithm of the octanol−water partition coefficient, log P(o/w) $<$ 5, are anticipated to likely have pharmacokinetic properties appropriate for oral administration. However, these schemes do not simultaneously consider the combination of the physiochemical properties, complicating their application in a more automated fashion. To overcome this, we present a novel method to select compounds with a combination of physicochemical properties that maximize bioavailability and druglikeness based on compounds in the World Drug Index database. In the study four properties, MW, log P(o/w), number of hydrogen bond donors, and number of hydrogen acceptors, were combined into a 4-dimensional (4D) histogram, from which a scoring function was defined on the basis of a 4D dependent multivariate Gaussian model. The resulting equation allows for assigning compounds a bioavailability score, termed 4D-BA, such that chemicals with higher 4D-BA scores are more likely to have oral druglike characteristics. The descriptor is validated by applying the function to drugs previously categorized in the Biopharmaceutics Classification System, and examples of application of the descriptor are given in the context of previously published studies targeting heme oxygenase and SHP2 phosphatase. The approach is anticipated to be useful in early lead identification studies in combination with clustering methods to maximize chemical and structural diversity when selecting compounds for biological assays from large database screens. It may also be applied to prioritize synthetically feasible chemical modifications during lead compound optimization.

INTRODUCTION

Despite advances in pharmaceutical sciences, drug discovery and development is still an expensive, time-consuming process. This process is, in part, hampered by difficulties in identifying pharmacologically active compounds with the appropriate pharmacokinetic properties that are ultimately successful in the clinic. Thus, more efficient methods to search for pharmacologically active compounds that also have desirable absorption, distribution, metabolism, and excretion (ADME) properties are needed.[1] Of the ADME properties, absorption and distribution, or bioavailability, are central to the successful development of orally administered drugs.

Numerous studies have addressed the issue of relating bioavailability to the physiochemical properties of druglike molecules. From these efforts a number of physicochemical properties have been shown to correlate with bioavailability, including, but not limited to, molecular weight (MW), lipophilicity measured as the logarithm of the octanol−water partition coefficient (log P(o/w)), intrinsic aqueous solubility (log S_w), number of hydrogen bond (H-bond) donors (HDO) and acceptors (HAC), molar refractivity, number of rings (RNG), and number of rotatable bonds (RTB). Among those properties, MW is related to intestinal and blood−brain barrier permeability, where molecules with larger MW tend to have less permeability.[2,3] Lipophilicity is related to absorption,[4] while an excessive amount of H-bond donors lowers permeability across lipid bilayers.[5,6] Vieth and co-workers have shown that oral drugs tend to have lower MW, fewer H-bond donors and acceptors, and rotatable bonds compared with other classes of drugs.[7]

Lipinski and co-workers chose 2245 compounds from the World Drug Index (WDI) database that they thought are likely to have superior physicochemical properties.[8] The WDI is a compound database that contains pharmacologically active compounds with comprehensive medical data including usage, drug target, mechanism of action, activity keywords, and adverse side effects.[9] For those 2245 compounds, the MW, log P(o/w), and number of H-bond donors and acceptors were calculated. Analysis revealed that approximately 90% of the compounds had values of the listed physiochemical properties that fell in a range of 5 or a multiple of 5 of several physicochemical properties. These include a MW under 500 (89% satisfied), log P(o/w) not more than 5 (90% satisfied), the number of H-bond donors not more than 5 (92% satisfied), and the number of H-bond acceptors not more than 10 (88% satisfied).[8] Combination of any two properties was even more selective with 1% of the compounds outside of both the MW and log P(o/w) criteria, 4% outside of the MW and H-bond donors criteria, 7% outside of the MW and H-bond acceptor criteria, and

* Corresponding author phone: (410) 706-7442; fax: (410) 706-5017; e-mail: alex@outerbanks.umaryland.edu.

10.1021/ci100359a
Published on Web 12/06/2010

10% outside of the H-bond donor and acceptor criteria. Thus, compounds out of these ranges are likely to have poor absorption and/or permeation with the exception of compounds that are substrates for biological transporters.[10] These observations are now referred to as Lipinski's rule of 5. Similar concepts have been pursued by Oprea and co-workers, in which they define criteria for leadlike compounds as MW $\leq$ 460, $-4 \leq \log P(\text{o/w}) \leq 4.2$, $\log S_w \geq -5$, RTB $\leq$ 10, RNG $\leq$ 4, HDO $\leq$ 5, and HAC $\leq$ 9.[11] Recently, Ohno and co-workers considered two physicochemical properties simultaneously to evaluate druglikeness using multivariate non-normal distributions.[12]

In the present paper we extend Lipiniski's rule of 5 and related concepts to allow for automatic selection of compounds with an appropriate combination of physicochemical properties to maximize bioavailability. The work takes advantage of almost all of the approximately 50 000 compounds in the 1999 version of the WDI database, versus the 2245 compounds previously studied, allowing for development of a scoring function based on a 4-dimensional (4D) probability distribution of the physicochemical properties MW, log *P*(o/w), HDO, and HAC. It should be noted that other descriptors, such as the number of rotatable bonds or polar surface area, could be included in such an analysis though we focus on using MW, log *P*(o/w), HDO, and HAC to keep the dimensionality of the solution tractable. The presented scheme can be combined with known approaches that evaluate the diversity of compounds by including clustering based on the chemical structure during the selection process.

The approach is anticipated to be particularly useful for rapid evaluation of potential bioavailability in early lead identification when large numbers of compounds may be selected from, for example, in silico database screening. In the later phases of lead optimization, possible chemical modifications to increase the binding affinity to the target may be prioritized in the context of bioavailability. In both cases, the presented method automatically selects those compounds with the highest possibility of having the best bioavailability/druglikeness from a collection of compounds selected on the basis of criteria other than bioavailability.

MATERIALS AND METHODS

Compound Databases. The WDI database (1999 edition) contains approximately 50 000 compounds consisting of all marketed drugs and pharmacologically active compounds as well as comprehensive medical data including usage, drug target, mechanism of action, activity keywords, and adverse drug effects.[9] Additional analysis was performed on three publically available chemical databases from commercial vendors. Databases from Maybridge (Thermo Fisher Scientific Inc., Waltham, MA), ChemBridge (San Diego, CA), and ChemDiv (San Diego, CA) contain approximately 62 000, 455 000, and 761 000 compounds, respectively. The 2009 version of each of these databases was used.

Four physicochemical properties, MW, log *P*(o/w), HDO, and HAC, were calculated for all compounds using the program Molecular Operating Environment (MOE) (Chemical Computing Group Inc.). To develop the scoring function, a 4D histogram was then constructed on the basis of MW, log *P*(o/w), HDO, and HAC for all compounds in the WDI database, allowing for consideration of the combination of the four physicochemical properties. The following bin sizes were used in creation of the 4D histogram: 50 for MW and 1 for log *P*(o/w), HDO, and HAC. The number of compounds in each bin were normalized as probabilities, which were then fit to a 4D variance−covariance matrix, as described below. The resulting model was then used to calculate the score of a given compound as a predictor of bioavailability. These scores will be referred to as the 4D-BA (i.e., BA indicates bioavailability) score for the remainder of the paper.

Structural Clustering. Structural clustering of compounds was performed using the BIT-packed version of MACCS structural keys (BIT MACCS), which encodes 2D structural features, and the Tanimoto coefficient.[13] Fingerprints, linear bit strings, based on the 166 MACCS keys that represent a small substructure in a given molecule, were assigned to each compound, where the combination of the assigned fingerprints characterizes a given compound. Fingerprint clustering was performed to group compounds into clusters of structurally similar compounds; similarity and overlap criteria were adjusted to obtain approximately 350−400 clusters of varying size as done in our previous studies.[14,15] These calculations were performed using MOE.

RESULTS AND DISCUSSION

In the present work we extend the efforts of Lipinski, Opera, and others to estimate potential bioavailability in the context of a more automated approach. Instead of focusing on 2245 of the approximately 50 000 compounds in the WDI database, we performed the analysis on the full database on the basis of the assumption that the pharmacologically active compounds will generally have the appropriate physicochemical properties required for bioavailability. Use of all compounds in the WDI allows for the range of physicochemical properties to be treated in terms of distributions, extending to a 4D distribution as defined by the MW, log *P*(o/w), HDO, and HAC of each compound as predictors of bioavailability. The 4D distribution may then be quantified into 4D "voxels", allowing for the frequency of the existence of a bioavailable compound with a specific range of physiochemical parameters to be determined. While these frequencies alone could be used to define a scoring function on the basis of a table lookup procedure, we considered it advantageous to determine a function describing the 4D distribution, thereby allowing for rapid score assignment and interpolation of the scores beyond the bin sizes characterizing the initial distribution. Importantly, this approach allows for ranking of ligands that fall both within and beyond the range specified by Lipiniski's rule of 5 as well as ready automation of the assignment process, thereby facilitating the drug discovery process. Having the method assign 4D-BA scores to compounds beyond the range specified by Lipinski's rule of 5 is considered important as a number of druglike molecules do indeed fall outside that range and we did not want our method to exclude such compounds.

MW, log *P*(o/w), HDO, and HAC Distributions in Four Compound Databases. To compare general physicochemical properties among four different compound databases (WDI, Maybridge, ChemBridge, and ChemDiv), MW, log *P*(o/w), HDO, and HAC were calculated for all four databases and are summarized in Table 1. The probability

Table 1. Mean MW, log *P*(o/w), HDO, HAC, and Standard Deviations in the WDI, Maybridge, ChemBridge, and ChemDiv Compound Databases[a]

database	MW	log *P*(o/w)	HDO	HAC	no. of compds
WDI	363.4 ± 146.8	2.7 ± 2.6	1.8 ± 1.9	3.5 ± 2.7	52 638
Maybridge	323.7 ± 89.4	3.3 ± 1.8	0.9 ± 0.9	2.5 ± 1.4	61 619
ChemBridge	361.8 ± 72.7	3.4 ± 1.5	1.0 ± 0.8	3.0 ± 1.3	454 960
ChemDiv	406.7 ± 76.3	3.7 ± 1.6	1.0 ± 0.8	3.5 ± 1.4	760 648

[a] Each value is shown as the average ± standard deviation.

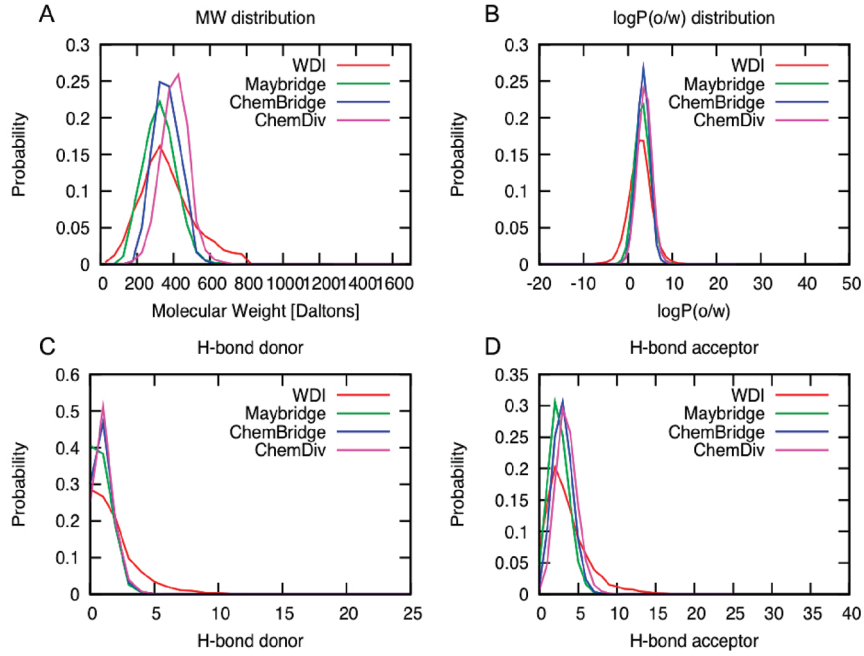


Figure 1. Probability distributions of physicochemical properties in all four compound databases. (A) MW, (B) log *P*(o/w), (C) HDO, (D) HAC. Data are shown for the WDI (red), Maybridge (green), ChemBridge (blue), and ChemDiv (magenta) databases.

distributions of each physical property in all four databases are shown in Figure 1. Notably, the distributions are all approximately Gaussian in nature, a property that facilitated the selection of functions from which to derive an analytical scoring function. Comparison of the distributions shows them to have similarities, though the WDI database covers wider ranges than any of the other databases. In particular, the distributions in the WDI go to higher values for MW, HDO, and HAC and lower values for log *P*(o/w). Similar distributions were observed by Ohno and co-workers.[12] This is due to the inclusion of peptides and other natural products in the WDI, compounds that have larger sizes and more H-bonding moieties as compared to the commercial databases, while also being more soluble in aqueous environments. While the removal of these particular compounds from the WDI during further analysis was considered due to known stability and bioavailability issues, we instead eliminated compounds from model fitting on the basis of physiochemical criteria, as described below.

To better characterize the combinations of MW, log *P*(o/w), HDO, and HAC, 2-dimensional (2D) histograms of all combinations of MW, log *P*(o/w), HDO, and HAC were created (Figure 2). Analysis of the plots reveals varying levels of correlation between the different properties. For example, the correlation between H-bond acceptors and donors is high ($r^2 = 0.51$), while the correlation between log *P*(o/w) and H-bond acceptors is relatively low ($r^2 = 0.06$). Similar correlations were also observed in the study by Ohno and co-workers.[12] These results point toward the need to include this correlation during function development, an observation that was subsequently validated as described below.

Scalar Representation of Bioavailability Based on a 4D Combination of MW, log *P*(o/w), HDO, and HAC in Marketed Drugs and Pharmacologically Active Compounds. To develop a scalar quantity that is representative of bioavailability, it is necessary to simultaneously take into account the four physicochemical properties of interest (MW, log *P*(o/w), HDO, and HAC). Step 1 of this process was to calculate a 4D histogram of those properties on the basis of the marketed drugs and pharmacologically active compounds as found in the WDI to determine the impact of bin size on the probabilities obtained in the histogram. The ranges of the distributions are shown in Figures 1 and 2, while the impact of the bin sizes on the maximum number of compounds in the most populated bin are listed in Table 2. From Table 2 the balance between the resolution of the data, as determined by the bin size, and the number of counts is significant. A final selection of using a 4D histogram with a bin size for MW of 50, for log *P*(o/w) of 1, for HDO of 1, and for HAC of 1 was associated with the need to keep the binning of the number of donors and acceptors at 1 and adjust the resolution of MW and log *P*(o/w) values that yielded a significant number of counts in the highest occupied bin. The identities of the top five most populated 4D bins using the selected bin sizes are listed in Table 3. Comparison of those results with those in Figure 1 and Table 1 show that

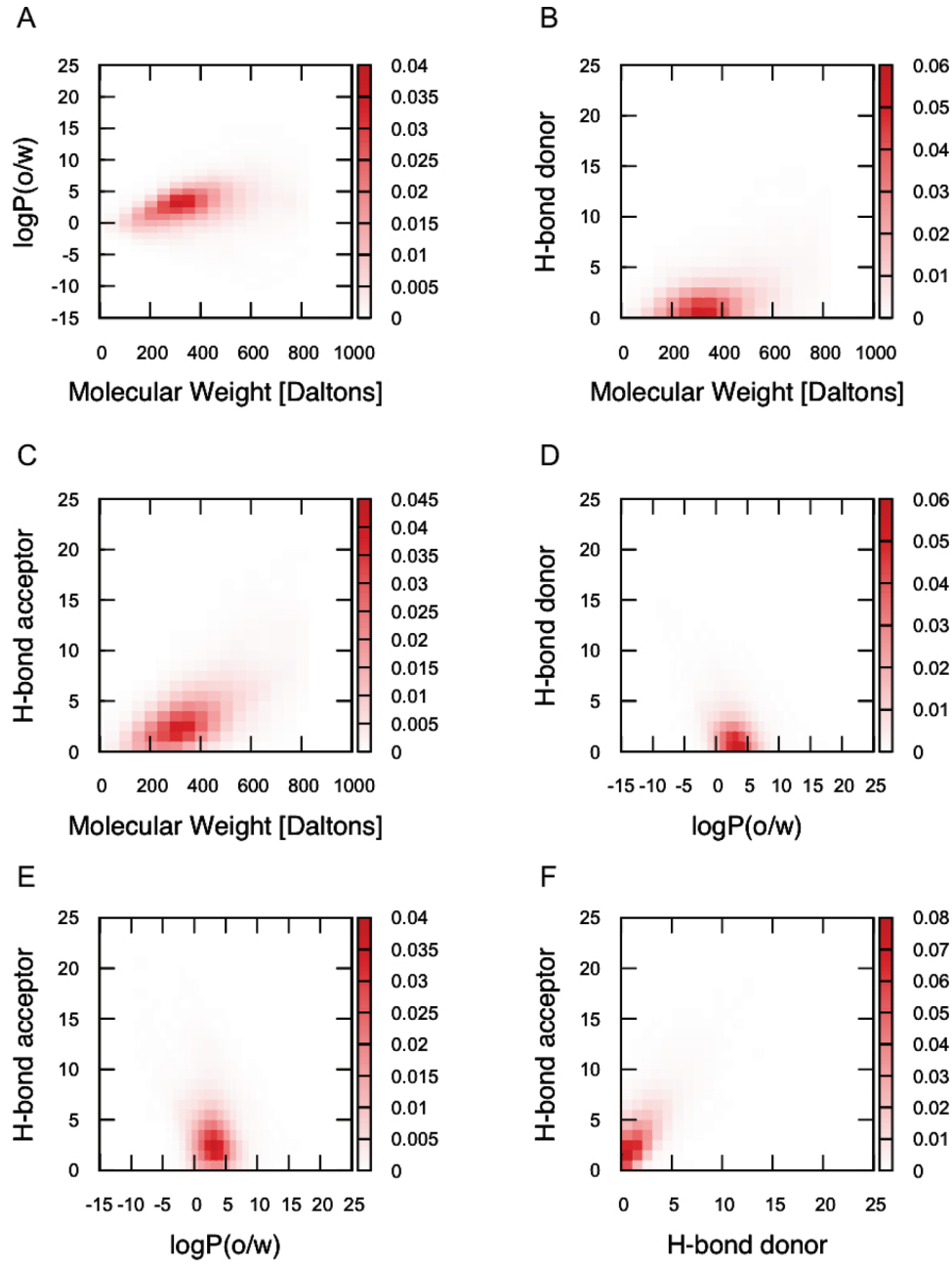


Figure 2. 2D probability distribution of two physicochemical properties among MW, log *P*(o/w), HDO, and HAC in the WDI database: (A) MW vs log *P*(o/w), (B) MW vs HDO, (C) MW vs HAC, (D) log *P*(o/w) vs HDO, (E) log *P*(o/w) vs HAC, (F) HDO vs HAC. The correlations of determination, r^2, for the respective plots were 0.12, 0.19, 0.40, 0.10, 0.06, and 0.51.

Table 2. Impact of Bin Sizes for MW, log *P*(o/w), HDO, and HAC on the Maximum Number of Compounds in a 4D Bin

MW	log *P*(o/w)	HDO	HAC	max no. of compds in bin
25	0.5	1	1	104
25	0.5	2	2	177
50	1	1	1	348 (used for fitting)
50	1	2	2	620

Table 3. Top Five Most Populated Bins from the 4D Histogram of the WDI Database with Bin Sizes of 50 for MW and 1 for log *P*(o/w), HDO, and HAC

MW	log *P*(o/w)	HDO	HAC	no. of compds
325	3.5	1	2	348
275	3.5	0	1	278
275	3.5	0	2	243
325	3.5	0	2	239
325	4.5	0	1	236

the log *P*(o/w), MW, and HDO values approximately correspond with highly populated regions of the respective 1D distribution and the average values. However, with the HAC, the highly populated bins are at values lower than the maximum of the 1D distribution and the average values. While this difference in not large, it suggests that the HAC term is making a smaller contribution to bioavailability than the remaining physicochemical properties.

Initial attempts to develop a function to describe the 4D probability distribution involved a 4D Gaussian distribution model based on independent Gaussians for each property. This was motivated by the distributions shown in Figure 1,

where each property has an approximately normal (Gaussian) distribution in 1D space, as described by the following equation, where $f(x)$ is the actual count in each bin, coefficient a is the height of the Gaussian, μ is the mean, and σ is the standard deviation of each property:

$$f(x) = ae^{-[(x - \mu)^2/2\sigma^2]} \quad (1)$$

Instead of using coefficient a, eq 1 can be written using the probability density function (PDF), where $f(x)$ is the probability density of each bin, as in the following equation:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-[(x - \mu)^2/2\sigma^2]} \quad (2)$$

Extending eq 2 to treat any two of the four properties, as shown in the 2D distributions in Figure 2, one obtains a PDF as described by the following equation, where $f(x,y)$ is the probability density of the occupancy of each 2D, μ_x and μ_y are the means of properties x and y, and σ_x and σ_y are the standard deviations of properties x and y:

$$f(x, y) = \frac{1}{\sqrt{2\pi\sigma_x^2}}\frac{1}{\sqrt{2\pi\sigma_y^2}}e^{-[(x - \mu_x)^2/2\sigma_x^2]-[(y - \mu_y)^2/2\sigma_y^2]} \quad (3)$$

Equation 3 assumes that the two variables x and y are independent of each other. Thus, our first attempt to fit data to develop a scoring function was a product of four independent Gaussian functions as described by the following equation, where four physicochemical properties, log P(o/w), MW, HAC, and HDO, are represented as w, x, y, and z, respectively, $f(w,x,y,z)$ is the probability density of each 4D bin, μ_w, μ_x, μ_y, and μ_z are the means of properties w, x, y, and z, and σ_w, σ_x, σ_y, and σ_z are the standard deviations of properties w, x, y, and z:

$$f(w, x, y, z) = \frac{1}{4\pi^2\sigma_w\sigma_x\sigma_y\sigma_z} \times e^{-[(w - \mu_w)^2/2\sigma_w^2]-[(x - \mu_x)^2/2\sigma_x^2]-[(y - \mu_y)^2/2\sigma_y^2]-[(z - \mu_z)^2/2\sigma_z^2]} \quad (4)$$

To develop a scoring function in the form of eq 4, the 1D distributions obtained from the WDI database were used to obtain values for variables σ and μ in eq 4. Prior to determination of the values for the variables σ and μ, outliers in the WDI database were eliminated by using the following ranges of the four properties, resulting in 51 957 compounds in total, from which the mean and standard deviation for each property were calculated: $-6 < w \leq 10$ for log P(o/w), $0 < x \leq 800$ for MW, $0 \leq y \leq 15$ for HAC, and $0 \leq z \leq 10$ for HDO. Since we use a bin size of 50 for MW, eq 4 becomes the following, which was used for fitting:

$$f(w, x, y, z) = \frac{50}{4\pi^2\sigma_w\sigma_x\sigma_y\sigma_z} \times e^{-[(w - \mu_w)^2/2\sigma_w^2]-[(x - \mu_x)^2/2\sigma_x^2]-[(y - \mu_y)^2/2\sigma_y^2]-[(z - \mu_z)^2/2\sigma_z^2]} \quad (5)$$

Using eq 5 with the means and standard deviations directly computed from the truncated WDI database without outliers (51 957 compounds), $\mu_w = 2.7$, $\sigma_w = 2.4$, $\mu_x = 360.4$, $\sigma_x = 144.7$, $\mu_y = 3.5$, $\sigma_y = 2.6$, $\mu_z = 0$ (0 is the maximum of the H-bond donor distribution as shown in Figure 1C), $\sigma_z = 1.8$, the probability was computed for the compounds in the truncated WDI database and then compared with the actual probability from the 4D histogram. This model gave a poor predictive model with an r^2 of 0.27. Accordingly, it is evident that the 4D independent Gaussian model does not capture well the observed probability from the 4D histogram.

Closer examination of the plots in Figure 2 suggests one possibility for the limitation in the independent model. If Gaussian distributions based directly on the mean and standard deviation from the WDI database (Table 1) are compared to the distributions in Figure 1, the shape of the distribution is not well reproduced and in particular the highest scoring values are consistently missed (Figure S1, Supporting Information). To overcome this, in the second approach the calculated mean and standard deviation values were used as initial guesses and then optimized to maximize reproduction of the 4D histogram obtained directly form the WDI. Fitting was performed using the FindFit utility in Mathematica, which is based on the Levenberg−Marquardt algorithm (LMA). The resulting optimized means and standard deviations are as follows: $\mu_w = 3.2$, $\sigma_w = 1.8$, $\mu_x = 302.6$, $\sigma_x = 97.9$, $\mu_y = 2.2$, $\sigma_y = 1.5$, $\mu_z = 0.6$, $\sigma_z = 1.3$. While this yielded an improved model with an r^2 value of 0.57 (observed vs predicted), this model was considered to not be of sufficient predictive power to be of utility for predicting potential oral bioavailability (Figure S2, Supporting Information).

A product of independent Gaussian distributions assumes that each of the four properties is not correlated with each other, which is not the case as suggested by Figure 2 and then verified on the basis of the model development in the preceding section. Accordingly, the appropriate model needed to take into account the correlation between the properties. The approach selected to do this was a multivariate normal distribution using a matrix of variances and covariances for the four properties. Using such an approach, eq 2 can be rewritten as follows:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-[(1/2)(x - \mu)^T(\sigma^2)^{-1}(x-\mu)]} \quad (6)$$

The exponent is similar to the Mahalanobis square distance formula as follows:

$$(x - \mu)^T\mathbf{S}^{-1}(x - \mu)$$

where T represents a transposed matrix, $\mathbf{S}$ is a variance−covariance matrix, μ is the mean, σ^2 is the variance, and x is the independent variable. Equation 6 can thus be extended into 4D as described in the following equation, which is known as a multivariate normal distribution:

$$f(x) = \frac{1}{(2\pi)^{N/2}\sqrt{\det(\mathbf{S})}}e^{-[(1/2)(x - \mu)^T\mathbf{S}^{-1}(x-\mu)]} \quad (7)$$

Since we use a bin size of 50 for MW, eq 7 becomes the following, which was used for fitting:

$$f(x) = \frac{50}{(2\pi)^{N/2}\sqrt{\det(\mathbf{S})}}e^{-[(1/2)(x - \mu)^T\mathbf{S}^{-1}(x-\mu)]} \quad (8)$$

For the N-dimensional cases, the exponent for a multivariate normal distribution is described by the matrix shown in the

$$\mu_{initial} = \begin{pmatrix} 2.7 \\ 360.4 \\ 3.5 \\ 0 \end{pmatrix} \quad S_{initial} = \begin{pmatrix} 5.9 & 129.4 & -1.2 & -1.2 \\ 129.4 & 20938.7 & 237.9 & 110.1 \\ -1.2 & 237.9 & 6.8 & 3.2 \\ -1.2 & 110.1 & 3.2 & 3.2 \end{pmatrix}$$

$$\mu_{optimized} = \begin{pmatrix} 3.2 \\ 292.0 \\ 1.8 \\ 0.5 \end{pmatrix} \quad S_{optimized} = \begin{pmatrix} 4.4 & 128.2 & -0.9 & -0.9 \\ 128.2 & 12650.7 & 103.2 & 35.9 \\ -0.9 & 103.2 & 3.5 & 1.7 \\ -0.9 & 35.9 & 1.7 & 2.0 \end{pmatrix}$$

Figure 3. Initial guess and optimized mean matrices, $\boldsymbol{\mu}$, and the variance−covariance matrices, **S**. Note $\mu_4 = 0$ in the initial mean matrix as 0 is the maximum of the H-bond donor distribution as shown in Figure 1D.

following equation, where **x** and $\boldsymbol{\mu}$ are both vectors of length N, 4 in the present analysis, and **S** is an $N \times N$ matrix of the variance and covariance of the data set:

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{pmatrix} \quad \mathbf{S} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} \end{pmatrix} \tag{9}$$

In our case, **x** consists of four elements, x_1 for log P(o/w), x_2 for MW, x_3 for HAC, and x_4 for HDO. $\boldsymbol{\mu}$ is composed of four elements, each of which represents the mean of four properties, μ_1 for log P(o/w), μ_2 for MW, μ_3 for HAC, and μ_4 for HDO. **S** is the variance−covariance matrix, where σ_{ii} is the variance of property i, σ_{ij} is the covariance of properties i and j, and $\sigma_{ij} = \sigma_{ji}$ (i and j are any two of the four properties). The elements of the variance−covariance matrix were calculated directly on the basis of the truncated WDI database based on the data in Table 3 and the correlation coefficients corresponding to the 2D plots in Figure 2. The resulting matrix equation model when applied to all the compounds in the truncated WDI yielded a predictive power of $r^2 = 0.55$; detailed analysis of the predicted vs observed values showed the probabilities of the high-score compounds to be consistently underestimated (not shown). Thus, the model yields no significant improvement over the fitted 4D independent Gaussian product model. This lack of improvement is based on the use of the means in matrix $\boldsymbol{\mu}$ and the variance and covariance values in matrix **S** directly from the WDI-based probability distribution.

Consequently, analogous to the approach used to improve the independent Gaussian model, a multivariate Gaussian model was developed in which the means in matrix $\boldsymbol{\mu}$ and the elements of the variance−covariance matrix **S** were optimized in the least-squares fitting procedure. The total number of optimized parameters is 14, consisting of the 4 means of the four properties and 10 from the **S** matrix, σ_{11}, σ_{12}, σ_{13}, σ_{14}, σ_{22}, σ_{23}, σ_{24}, σ_{33}, σ_{34}, σ_{44}, considering the symmetry of the matrix. In Figure 3 are given $\boldsymbol{\mu}_{\text{initial}}$, the mean matrix, $\mathbf{S}_{\text{initial}}$, the variance−covariance matrix used as an initial guess for fitting, $\boldsymbol{\mu}_{\text{optimized}}$, the optimized mean matrix, and $\mathbf{S}_{\text{optimized}}$, the optimized variance−covariance matrix. Upon fitting, the predictive power of the model was significantly improved over that of the nonoptimized multivariate model, yielding an r^2 value of 0.89. The optimized mean values for each property are [log P(o/w)] 3.2, (MW) 292.0, (HAC) 1.8, and (HDO) 0.5.

Table 4 summarizes the predictive power of the four developed models. The independent model based on means and variances yields the poorest predictive power; allowing the values of the mean and variance to be optimized, targeting the WDI data, leads to a significant improvement, though the predictive power is still considered insufficient. Accordingly, a matrix model that allowed for inclusion of correlation between the parameters, in part motivated by the correlation plots shown in Figure 2, was implemented. On the basis of the mean, variance and covariance values directly from the WDI still led to a model with low predictive power; however, by fitting those terms, a significant increase in the predictive power is gained. This is evident in Figure 4 comparing probability values directly from the WDI with those based on the final matrix model. Given the high predictive power of the model, this was selected for use in the prediction of the 4D-BA scores of compounds to facilitate drug design.

Table 4. r^2 Values of the 4D Gaussian Distribution Models

model	database param	fitted param
independent	0.27	0.57
dependent	0.55	0.89

To evaluate the possibility of overfitting, the WDI data set was divided randomly into equally sized training and test sets. The 4D multivariate correlated Gaussian model was built using the training data set following the same procedure described above for the full data set. Using the resulting fitted values of the means and the variance−covariance matrix (which did not differ significantly from the ones fitted from the complete data set), the predictability of the model was evaluated. An r^2 value of 0.87 (test set observed vs test set predicted) indicates that the 4D multivariate Gaussian model is robust and does not overfit to the data (see Figure S3 in the Supporting Information). The mean values and variance−covariance matrix presented in this paper are the ones fitted from the entire truncated WDI data set to provide the best possible estimates of the parameters intended for new compound selection.

An additional test for quality control was performed by comparing the three probability distributions for the four physicochemical properties considered in this study. The distribution of observed probabilities directly obtained from the WDI database was compared with the distribu-

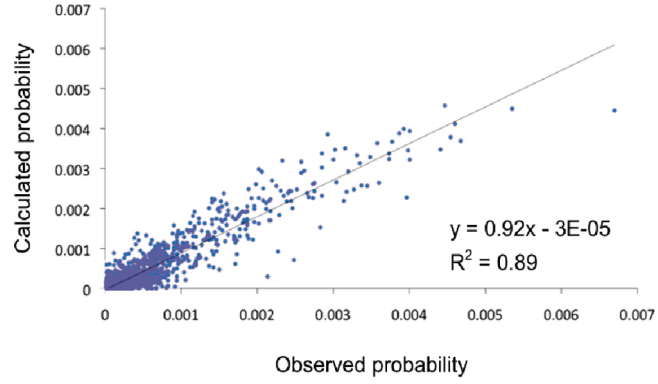


Figure 4. Comparison between the observed probability from the WDI database and the calculated probability based on the improved 4D multivariate dependent Gaussian model. Linear regression analysis was performed to obtain the linear function $y = 0.92x - (3.0 \times 10^{-5})$ with a square of the correlation coefficient, r^2, of 0.89.

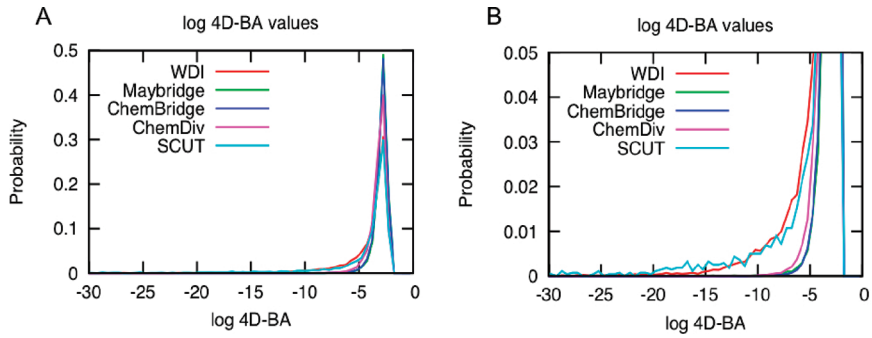


Figure 5. Probability distributions of the log 4D-BA values in all five compound databases, WDI, Maybridge, ChemBridge, ChemDiv, and SCUT: (A) entire range and (B) low-scoring range. 4D-BA values were calculated on the basis of four properties, MW, log *P*(o/w), HDO, and HAC, using the 4D Gaussian dependent model ($r^2 = 0.89$ model) shown in Figure 4.

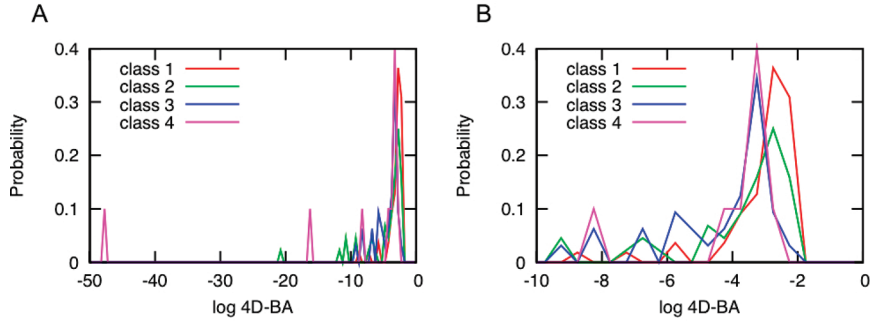


Figure 6. Probability distributions of the log 4D-BA values in all four BCS classes of compounds in the WHO essential medicines list: (A) whole range and (B) high-scoring range. 4D-BA values were calculated on the basis of four properties, MW, log *P*(o/w), HDO, and HAC, using the 4D Gaussian dependent model ($r^2 = 0.89$ model) shown in Figure 4. BCS classes 1, 2, 3, and 4 are shown in red, green, blue, and magenta, respectively.

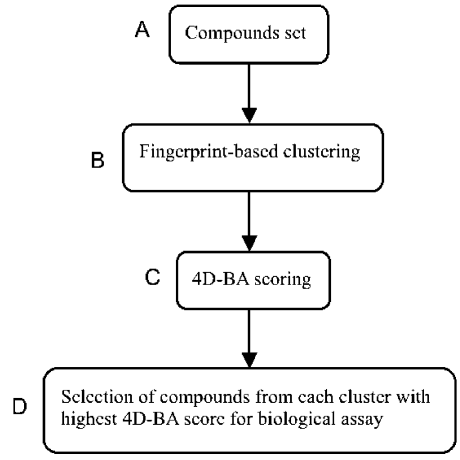
Table 5. BCS Compounds with log 4D-BA Scores of Less Than −10

	log 4D-BA	MW	log *P*(o/w)	HDO	HAC	comment
			Class 2			
azithromycin	−10.7	749.0	3.43	5	13	macrolide antibiotic
digoxin	−10.7	780.9	3.32	6	13	cardiac glycoside
cyclosporin A	−20.6	1202.6	5.29	5	12	cyclic peptide
rapamycin	−11.9	914.2	6.87	3	12	macrolide immunosuppresant
			Class 4			
amphotericin B	−16.3	924.1	2.80	10	14	polyene antifungal
colistin	−47.7	1035.3	−7.38	12	12	cyclic peptide

tions of the calculated probabilities using the 4D dependent multivariate Gaussian model (Figure S1, Supporting Information). The improved 4D dependent multivariate Gaussian model captures the overall qualitative features found in the WDI database, suggesting that this could be a reasonable model to predict the relative bioavailability of new compounds.

Application of the 4D-BA Descriptor to FDA-Approved Drugs. The SCUT database is an in-house database maintained by Ekins, Swaan, and co-workers that contains 2815 FDA-approved drugs in clinical use in the United States derived from the *Clinician's Pocket Drug Reference*.[16] Thus, it is of interest to apply the 4D-BA descriptor to the SCUT database to see whether the 4D-BA distribution is similar to that from the WDI. 4D-BA values for the SCUT database were therefore calculated using the 4D Gaussian dependent model. For comparison, 4D-BA values were also calculated for four other databases, WDI, Maybridge, ChemBridge, and ChemDiv, with the probability distributions shown in Figure 5 on the logarithmic scale. Notably, a number of compounds in WDI and SCUT have log 4D-BA scores of less than −5. This is again due to the presence of peptidic drugs and other natural products that have larger MW and more H-bond moieties as compared to compounds in the commercial databases.

Scheme 1. Flowchart of the Compound Selection Scheme To Maximize Bioavailability and Druglikeness as Well as Chemical and Structural Diversity



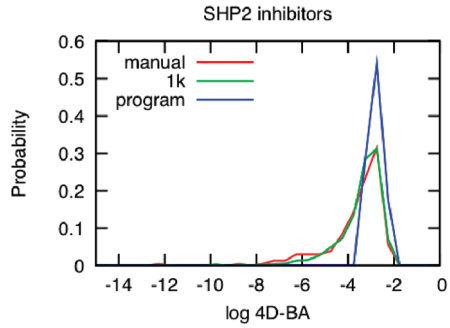


Figure 7. Probability distributions of candidate SHP2 PTP domain inhibitors as a function of the log 4D-BA score. The green line represents the top 1000 compounds selected from database screening on the basis of energy scoring. The red line represents 235 compounds selected manually from the top 1000 compounds. The blue line represents the top 235 compounds selected from the same 1000 compounds using our automated scheme shown in Scheme 1. A bin size of 0.5 for the log 4D-BA value was used.

Importantly, the low-scoring compounds are well predicted by the 4D Gaussian dependent model (Figure 4).

Validation of the 4D-BA Descriptor. The Biopharmaceutics Classification System (BCS), developed by Amidon and co-workers,[17] is the scheme used by the FDA to categorize the orally administered drugs into four groups on the basis of aqueous solubility and gastrointestinal permeability, properties that strongly influence the rate and extent of drug absorption. Class 1 includes compounds with high solubility and high permeability that are over 90% absorbed, class 2 are compounds with low solubility and high permeability, class 3 compounds have high solubility and low permeability, and class 4 compounds have low solubility and low permeability.[18,19] As this classification scheme is based on in vivo experimentally measured properties of known drugs, it represents an ideal set of data by which to validate the 4D-BA descriptor. This was performed using the BCS compounds listed by Wu and Bennett,[19] which include 55 class 1 compounds, 44 class 2 compounds, 32 class 3 compounds, and 10 class 4 compounds, with 11 compounds in more than one category as reported in different studies.[18,19] 4D-BA was calculated for all compounds in the four classes, with the results presented as probability distributions of log 4D-BA (Figure 6). The majority of class 1 drugs (92.7%) have log 4D-BA values bigger than −5, with the distributions for the highly permeable class 1 and 2 drugs shifted to values higher than those of classes 3 and 4. This simple analysis supports the ability of the 4D-BA descriptor to capture the bioavailability characteristics of orally administered drugs.

Of note are class 2 and 4 compounds that have scores lower than −10. Table 5 lists those compounds along with their physiochemical properties. All the compounds are large with MWs of 700 or more, with large numbers of hydrogen bond acceptors, consistent with their being large natural products that include carbohydrate or peptidic moieties. While low-scoring class 4 compounds are expected, the presence of class 2 compounds emphasizes the limitation of any scoring scheme, whereby compounds predicted to have poor bioavailability are actually orally available. These types of compounds contribute to the low 4D-BA scores seen in the WDI and SCUT databases presented above.

Practical Application of the 4D-BA. To place the presented 4D-BA scoring approach in a more practical context, we present a flow diagram in Scheme 1 which may be applied, for example, to a collection of compounds (Scheme 1A) selected from a database on the basis of docking against a target receptor followed by energy scoring.[14,15] Those compounds are then subjected to fingerprint-based clustering to identify structurally similar compounds (Scheme 1B). In studies in our laboratory, typically 350−400 clusters are obtained from 1000 compounds. These range from a few large clusters containing 30 or more

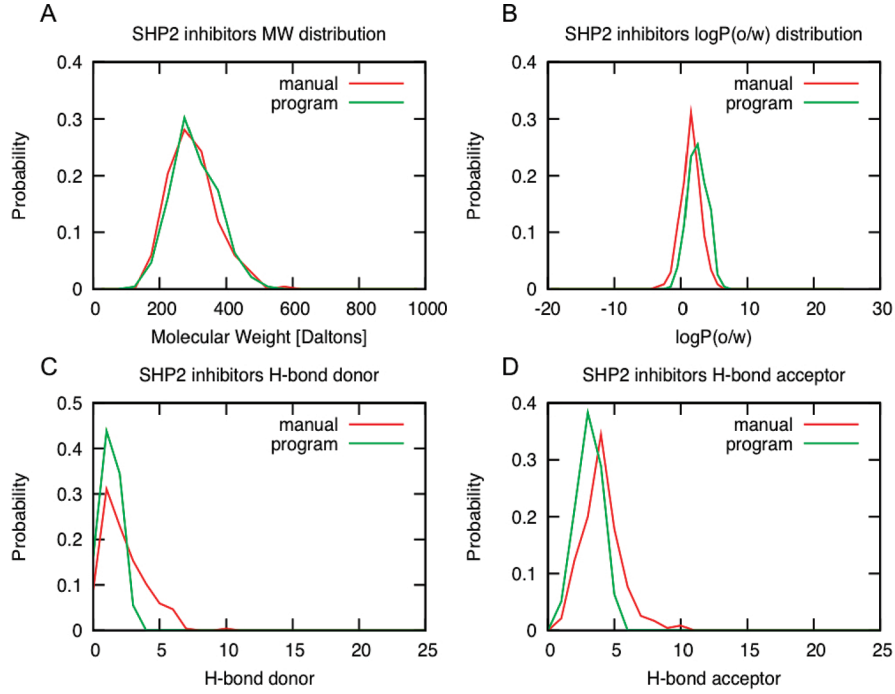


Figure 8. Probability distributions of four physicochemical parameters in 235 compounds of candidate SHP2 PTP domain inhibitors selected manually or by the presented automated procedure: (A) MW, (B) log *P*(o/w), (C) HDO, (D) HAC. Red lines represent the manual selection, while green lines represent the selection by the automated procedure.

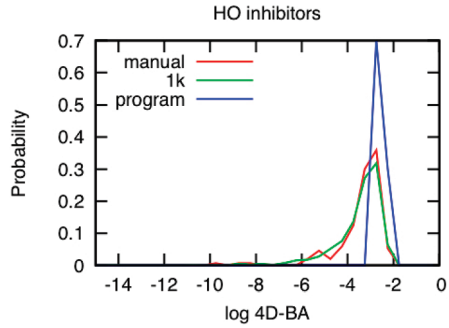


Figure 9. Probability distributions of candidate *nm*-HO inhibitors as a function of the log 4D-BA score. The green line represents the top 1000 compounds selected from database screening on the basis of energy scoring. The red line represents 153 compounds selected manually from the top 1000 compounds. The blue line represents the top 153 compounds selected from the same 1000 compounds using the automated scheme. A bin size of 0.5 for the log 4D-BA score was used.

compounds to a large number of "clusters" with only a single molecule. The 4D-BA score is then determined for all the compounds (Scheme 1C), with the compound with the highest score in each cluster selected for biological assay (Scheme 1D). Typically, compounds from the single-molecule clusters are subjected to initial inspection from which a subset are selected for assay on the basis of structural features, physiochemical properties, and scores from the database screen.

SHP2. Src homology 2 (SH2) domain-containing phosphatase 2 (SHP2) is a protein tyrosine phosphatase (PTP) that is involved in a variety of cell signaling events.[20] Hyperactivation of the catalytic activity has been found in the developmental disorder Noonan syndrome,[21] several childhood leukemias,[22,23] and sporadic solid tumors.[24] Thus, selective inhibitors for SHP2 activity are of great interest as novel therapeutic candidates for these diseases. Our laboratory conducted a database screen in which compounds were docked into the catalytic site of SHP2.[14] From that screen 1000 compounds were selected on the basis of energy criteria from approximately 1.3 million druglike small molecules. Chemical clustering was then performed on the top 1000 compounds, yielding 376 clusters, from which 235 compounds were manually selected for biological experiments on the basis of the Lipinski and Oprea empirical rules[8,11] as well as qualitative chemical diversity considerations beyond those obtained by the clustering procedure. The present automated procedure was applied to select 235 compounds from the same 376 clusters on the basis of the 4D-BA scores; clusters associated with low 4D-BA values below those in the top 235 compounds were not considered. Distributions of the log 4D-BA scores for the top 1000 compounds, the manually selected 235 compounds, and the automatically selected 235 compounds are shown in Figure 7. A total of 142 compounds were overlapped between the manually and automatically selected sets of compounds. The result shows that the set of automatically selected compounds consistently has more compounds in the higher scoring region, though significant overlap is present. The presence of lower scoring compounds in the manually selected set was due to compounds being selected to maximize structural diversity beyond that from the BIT-MACCS similarity-based clustering. While such an additional selection criterion may be considered desirable, the availability of the present automated 4D-BA score would have facilitated the selection process. Additional considerations of chemical diversity are beyond the scope of the present study.

Probability distributions of the four physicochemical properties are shown for the manually and automatically selected 235 compounds in Figure 8. The results verify that compounds selected by the automated procedure typically do not violate the Lipinski rule of 5, although exceptions

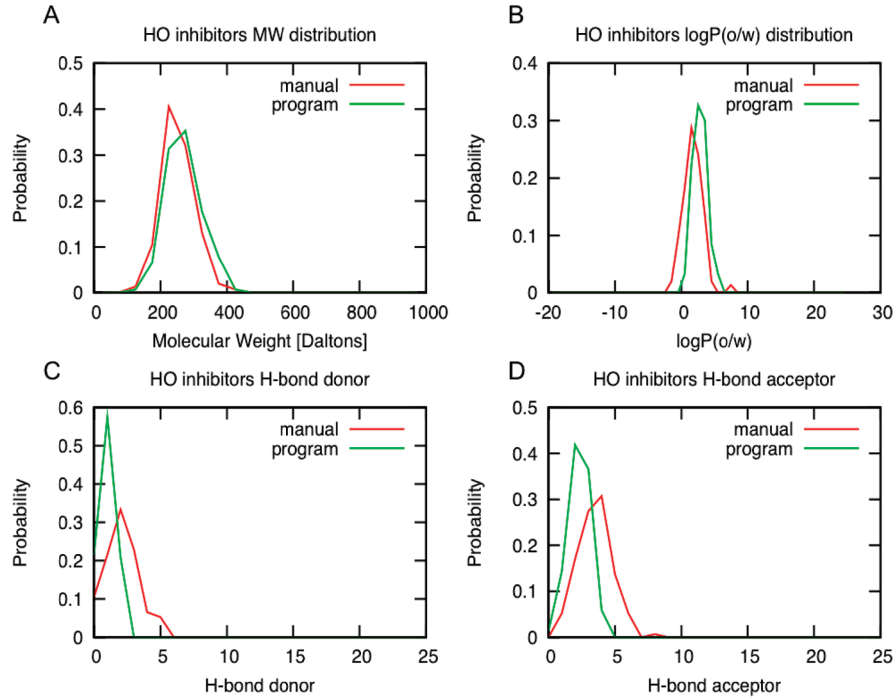


Figure 10. Probability distributions of four physicochemical parameters in 153 compounds of candidate *nm*-HO inhibitors selected manually or by the automated procedure: (A) MW, (B) log *P*(o/w), (C) HDO, (D) HAC. Red and green lines represent the compound set from manual selection and automated selection using our procedure.

are present. In addition, almost all of these compounds also satisfy stricter rules defined by Oprea and others: MW ≤ 460, −4 ≤ log *P*(o/w) ≤ 4.2, HDO ≤ 5, HAC ≤ 9.[11] The exceptions with respect to the criteria of Lipinski or Oprea are primarily due to the inclusion of structural and chemical diversity in the final compounds. However, as diversity may be considered an attribute in drug discovery, and known drugs do indeed fall out of the range of the Lipinski or Oprea rules, the ability of our procedure to include exceptions with respect to ideal bioavailability may be considered desirable.

Heme Oxygenase. Iron acquisition is essential for bacterial pathogens, such as *Neisseria meningitidis*,[25] *Haemophilus influenzae*,[26] *Vibrio cholerae*,[27,28] and *Shigella dysenteriae*,[29,30] to survive and reveal infectivity. Heme oxygenase (HO) plays a critical role in oxidative cleavage of the porphyrin macrocycle to biliverdin and carbon monoxide to release iron for the last step in heme utilization in a number of bacterial pathogens.[31,32] Therefore, selective inhibitors for bacterial HO are possible therapeutic candidates for diseases caused by those pathogens. Our laboratory applied a database screen targeting the *N. meningitidis* HO (*nm*-HO). From that screen 1000 compounds were identified on the basis of docking into the heme binding site and energy criteria from approximately 0.8 million druglike small molecules.[15] Chemical-fingerprint-based clustering was then performed on the top 1000 compounds, yielding 435 clusters, from which 153 compounds were manually selected for biological assays,[15] again considering bioavailability and diversity criteria. The automated procedure was used to select the top 153 compounds from the same 435 clusters, again selecting the compounds on the basis of the highest 4D-BA score from each cluster to include maximum chemical and structural diversity. The distributions of the top 1000 manually and automatically selected 153 compounds are compared in Figure 9. A total of 63 compounds overlapped between the manually and automatically selected sets. As expected, compounds selected using the automated scheme typically have higher scores with less compounds in the lower scoring region. The automatically selected compounds again obey the empirical criteria of Lipinksi or Oprea (Figure 10), while a larger number of exceptions are observed with the manually selected compounds. The impact of including additional diversity during the manual selection is again evident.

CONCLUSIONS

Presented is a novel descriptor of bioavailability of druglike compounds, termed 4D-BA. The descriptor builds on previous studies by Lipinski, Oprea, and others by simultaneously taking into account the MW, log *P*(o/w), number of hydrogen bond donors, and number of hydrogen bond acceptors using a 4D multivariate dependent Gaussian model. The resulting descriptor is convenient in that it offers a scalar quantity to predict bioavailability that goes beyond the four individual descriptors listed above. Validation of the descriptor is performed by showing that the 4D-BA descriptor ranks drugs in a manner consistent with their published BCS ratings. The utility of the 4D-BA descriptor is then illustrated in the context of in silico database screening, where the 4D-BA values of compounds selected from the screen and clustered on the basis of structural similarity are used to facilitate the selection of final compounds for biological assay. It is anticipated that the 4D-BA descriptor will also be of utility for the facilitation of lead optimization studies where it can help prioritize compounds for chemical synthesis and biological testing.

ACKNOWLEDGMENT

We acknowledge helpful discussions with Drs. James Polli and Peter Swaan and financial support from the NIH (Grant CA120215), the Samuel Waxman Cancer Foundation, and the University of Maryland Computer-Aided Drug Design Center.

Supporting Information Available: 1D probability distributions of physicochemical properties reproduced by the 4D dependent multivariate model with optimized mean values and the variance−covariance matrix, comparison between the observed probability from the WDI database and the calculated probability based on the improved 4D independent model, and comparison between the test set observed probability and the training set predicted probability. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Ekins, S.; Rose, J. In silico ADME/Tox: The state of the art. *J. Mol. Graphics Modell.* **2002**, *20*, 305–9.
(2) Navia, M. A.; Chaturvedi, P. R. Design principles for orally bioavailable drugs. *Drug Discovery Today* **1996**, *1*, 179–189.
(3) Pardridge, W. M. Transport of small molecules through the blood−brain barrier: Biology and methodology. *Adv. Drug Delivery Rev.* **1995**, *15*, 5–36.
(4) Testa, B.; Carrupt, P. A.; Gaillard, P.; Billois, F.; Weber, P. Lipophilicity in molecular modeling. *Pharm. Res.* **1996**, *13*, 335–43.
(5) Abraham, M. H.; Chadha, H. S.; Whiting, G. S.; Mitchell, R. C. Hydrogen bonding. 32. An analysis of water−octanol and water−alkane partitioning and the delta log P parameter of seiler. *J. Pharm. Sci.* **1994**, *83*, 1085–100.
(6) Paterson, D. A.; Conradi, R. A.; Hilgers, A. R.; Vidmar, T. J.; Burton, P. S. A non-aqueous partitioning system for predicting the oral absorption potential of peptides. *Quant. Struct.-Act. Relat.* **1994**, *13*, 4–10.
(7) Vieth, M.; Siegel, M. G.; Higgs, R. E.; Watson, I. A.; Robertson, D. H.; Savin, K. A.; Durst, G. L.; Hipskind, P. A. Characteristic physical properties and structural fragments of marketed oral drugs. *J. Med. Chem.* **2004**, *47*, 224–32.
(8) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3–25.
(9) *WDI: World Drug Index*, version 4/99; Derwent Information: London, 1999.
(10) Ekins, S.; Stresser, D. M.; Williams, J. A. In vitro and pharmacophore insights into CYP3A enzymes. *Trends Pharmacol. Sci.* **2003**, *24*, 161–6.
(11) Hann, M. M.; Oprea, T. I. Pursuing the leadlikeness concept in pharmaceutical research. *Curr. Opin. Chem. Biol.* **2004**, *8*, 255–63.
(12) Ohno, K.; Nagahara, Y.; Tsunoyama, K.; Orita, M. Are there differences between launched drugs, clinical candidates, and commercially available compounds? *J. Chem. Inf. Model.* **2010**, *50*, 815–21.
(13) Butina, D. Unsupervised data base clustering based on Daylight's fingerprint and Tanimoto similarity: A fast and automated way to cluster small and large data sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 747–750.
(14) Yu, W. M.; Guvench, O.; Mackerell, A. D., Jr.; Qu, C. K. Identification of small molecular weight inhibitors of Src homology 2 domain-containing tyrosine phosphatase 2 (SHP-2) via in silico database screening combined with experimental assay. *J. Med. Chem.* **2008**, *51*, 7396–404.
(15) Furci, L. M.; Lopes, P.; Eakanunkul, S.; Zhong, S.; MacKerell, A. D., Jr.; Wilks, A. Inhibition of the bacterial heme oxygenases from *Pseudomonas aeruginosa* and *Neisseria meningitidis*: Novel antimicrobial targets. *J. Med. Chem.* **2007**, *50*, 3804–13.

(16) Gomella, L.; Haist, S.; Adams, A. *Clinician's Pocket Drug Reference 2009*; McGraw-Hill: New York, 2009.
(17) Amidon, G. L.; Lennernas, H.; Shah, V. P.; Crison, J. R. A theoretical basis for a biopharmaceutic drug classification: The correlation of in vitro drug product dissolution and in vivo bioavailability. *Pharm. Res.* **1995**, *12*, 413–20.
(18) Lindenberg, M.; Kopp, S.; Dressman, J. B. Classification of orally administered drugs on the World Health Organization Model List of Essential Medicines according to the biopharmaceutics classification system. *Eur. J. Pharm. Biopharm.* **2004**, *58*, 265–78.
(19) Wu, C. Y.; Benet, L. Z. Predicting drug disposition via application of BCS: Transport/absorption/elimination interplay and development of a biopharmaceutics drug disposition classification system. *Pharm. Res.* **2005**, *22*, 11–23.
(20) Tonks, N. K. Protein tyrosine phosphatases: from genes, to function, to disease. *Nat. Rev. Mol. Cell Biol.* **2006**, *7*, 833–846.
(21) Tartaglia, M.; Mehler, E. L.; Goldberg, R.; Zampino, G.; Brunner, H. G.; Kremer, H.; van der Burgt, I.; Crosby, A. H.; Ion, A.; Jeffery, S.; Kalidas, K.; Patton, M. A.; Kucherlapati, R. S.; Gelb, B. D. Mutations in PTPN11, encoding the protein tyrosine phosphatase SHP-2, cause Noonan syndrome. *Nat. Genet.* **2001**, *29*, 465–8.
(22) Tartaglia, M.; Niemeyer, C. M.; Fragale, A.; Song, X.; Buechner, J.; Jung, A.; Hahlen, K.; Hasle, H.; Licht, J. D.; Gelb, B. D. Somatic mutations in PTPN11 in juvenile myelomonocytic leukemia, myelodysplastic syndromes and acute myeloid leukemia. *Nat. Genet.* **2003**, *34*, 148–50.
(23) Loh, M. L.; Vattikuti, S.; Schubbert, S.; Reynolds, M. G.; Carlson, E.; Lieuw, K. H.; Cheng, J. W.; Lee, C. M.; Stokoe, D.; Bonifas, J. M.; Curtiss, N. P.; Gotlib, J.; Meshinchi, S.; Le Beau, M. M.; Emanuel, P. D.; Shannon, K. M. Mutations in PTPN11 implicate the SHP-2 phosphatase in leukemogenesis. *Blood* **2004**, *103*, 2325–31.
(24) Bentires-Alj, M.; Paez, J. G.; David, F. S.; Keilhack, H.; Halmos, B.; Naoki, K.; Maris, J. M.; Richardson, A.; Bardelli, A.; Sugarbaker, D. J.; Richards, W. G.; Du, J.; Girard, L.; Minna, J. D.; Loh, M. L.; Fisher, D. E.; Velculescu, V. E.; Vogelstein, B.; Meyerson, M.; Sellers, W. R.; Neel, B. G. Activating mutations of the Noonan syndrome-associated SHP2/PTPN11 gene in human solid tumors and adult acute myelogenous leukemia. *Cancer Res.* **2004**, *64*, 8816–20.
(25) Zhu, W.; Hunt, D. J.; Richardson, A. R.; Stojiljkovic, I. Use of heme compounds as iron sources by pathogenic neisseriae requires the product of the hemO gene. *J. Bacteriol.* **2000**, *182*, 439–47.
(26) Sanders, J. D.; Cope, L. D.; Hansen, E. J. Identification of a locus involved in the utilization of iron by *Haemophilus influenzae*. *Infect. Immun.* **1994**, *62*, 4515–25.
(27) Henderson, D. P.; Payne, S. M. Characterization of the *Vibrio cholerae* outer membrane heme transport protein HutA: Sequence of the gene, regulation of expression, and homology to the family of TonB-dependent proteins. *J. Bacteriol.* **1994**, *176*, 3269–77.
(28) Henderson, D. P.; Payne, S. M. *Vibrio cholerae* iron transport systems: Roles of heme and siderophore iron transport in virulence and identification of a gene associated with multiple iron transport systems. *Infect. Immun.* **1994**, *62*, 5120–5.
(29) Mills, M.; Payne, S. M. Genetics and regulation of heme iron transport in *Shigella dysenteriae* and detection of an analogous system in *Escherichia coli* O157:H7. *J. Bacteriol.* **1995**, *177*, 3004–9.
(30) Mills, M.; Payne, S. M. Identification of shuA, the gene encoding the heme receptor of *Shigella dysenteriae*, and analysis of invasion and intracellular multiplication of a shuA mutant. *Infect. Immun.* **1997**, *65*, 5358–63.
(31) Ortiz De Montellano, P. R.; Wilks, A. Hemo oxygenase structure and mechanism. *Adv. Inorg. Chem.* **2000**, *51*, 359–402.
(32) Wilks, A. Hemo oxygenase: Evolution, structure, and mechanism. *Antioxid. Redox Signaling* **2002**, *4*, 603–614.

CI100359A