Mapping Human Metabolic Pathways in the Small Molecule Chemical Space

Antonio Macchiarulo,*,† Janet M. Thornton,†,‡ and Irene Nobeli*,§

Dip. Chimica e Tecnologia del Farmaco, Faculty of Pharmacy, University of Perugia, Via del Liceo 1, 06123 Perugia, Italy, EMBL Outstation - Hinxton, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, U.K., and Institute of Structural and Molecular Biology, School of Crystallography, Institute of Structural and Molecular Biology, Department of Biological Sciences, Birkbeck, University of London, Malet Street, London WC1E 7HX, U.K.

Received June 1, 2009

The work presented here is a study of human metabolic pathways, as projected in the chemical space of the small molecules they comprise, and it is composed of three parts: a) a study of the extent of clustering and overlap of these pathways in chemical space, b) the development and assessment of a statistical model for estimating the proximity to a given pathway of any small molecule, and c) the use of the above model in estimating the proximity of marketed drugs to human metabolic pathways. The distribution, overlap, and relationships of human metabolic pathways in this space are revealed using both visual and quantitative approaches. A set of selected physicochemical and topological descriptors is used to build a classifier, whose aim is to predict metabolic class and pathway membership of any small molecule. The classifier performs well for tightly clustered, isolated pathways but is, naturally, much less accurate for strongly overlapping pathways. Finally, the extent of overlap of a set of known drugs with the human metabolome is examined, and the classifier is used to predict likely cross-interactions between drugs and the major metabolic pathways in humans.

INTRODUCTION

The traditional view of metabolism was small-molecule-based: metabolic pathways drawn in textbooks as reactions of substrates and products, occasionally annotated by the type of enzyme carrying out each reaction. The advent of genomics, on the other hand, brought with it a macromolecule-centered view of the cell, where gene products dominated every aspect of biological function, including metabolism. The important entities in metabolic pathways became the enzymes that carried out the work, not the molecules they were acting on. More recently, the need to consider all molecular entities in the cell for a holistic view of the cell's machinery has led to a renewed interest in the structures, properties, and classification of biologically important small molecules,[1−4] especially those we refer to as endogenous metabolites, defined here as molecules we encounter in the metabolic pathways of organisms. The availability of metabolic pathway databases that include information on small molecules[5,6] or indeed databases specializing in small molecules and their biological roles[7,8] has paved the way for a more comprehensive exploration of metabolism. At the same time the notion of chemical space has been employed in drug discovery to envisage and explore the vast set of all possible chemical structures for biologically relevant small molecules.[9−11]

In this work, we employ chemical informatics to link the human metabolic network to the structural and physicochemical properties of the small molecules that are part of this network and to form tentative links between synthetic molecules and human metabolism. This study focuses on human metabolism, even though the genomes of small model organisms like *E. coli* and *S. cerevisiae* have been more extensively functionally characterized, and hence their metabolic pathways are inevitably more complete. This is because one of our ultimate goals is to explore and uncover links between human metabolism and the effects of marketed drugs.

The work presented here is composed of three parts. In the first part we examine the distribution and relationships of human metabolic pathways in the chemical space, as defined by the small molecules involved in these pathways. The aim of this part is to shed light on the extent of possible cross-reactivity between pathways. We start from the simple observation that within a typical pathway small molecules tend to look similar, as they are related to each other through stepwise chemical transformations. Hence, in the chemical space defined by the properties and structures of molecules involved in metabolism one would expect pathways to appear as clusters of the small molecules they comprise. In addition, due to the fact that pathways often share the same small molecules, or share molecules of significant chemical similarity, one would expect to observe an overlap between certain pathways. Furthermore, this overlap would not likely be distributed uniformly among all pairs of pathways, thus leading to clusters of pathways in chemical space. We test the hypothesis that metabolic pathways cluster in chemical space by using the structural and physicochemical properties of their constituent small molecules. By structure we mean the atom and bond connectivity information for each

* Corresponding author e-mail: antonio@chimfarm.unipg.it (AM), i.nobeli@bbk.ac.uk (IN).
† University of Perugia.
‡ EMBL Outstation - Hinxton, European Bioinformatics Institute.
§ University of London.

10.1021/ci900196u CCC: $40.75
Published on Web 10/01/2009

molecule. As this 2D structure might be a poor descriptor for similarity, we use in addition physicochemical and topological descriptors, usually employed in Quantitative Structure Activity Relationship (QSAR) studies.

In the second part of the work, we address an issue that has been fueled by progress in the experimental science of metabolomics.[12] As more metabolites are being detected and identified using spectrometric analyses, at least some will be "orphan" molecules, i.e. molecules not belonging to any known metabolic pathways. Placing these molecules in the context of known metabolic pathways would help us understand the processes they might be involved in and could hint at the presence of as yet unidentified gene products that may be catalyzing relevant reactions. Thus, the aim of this part of the work is to develop and assess a statistical model to predict for any given molecule the pathways that it would lie closest to. This means that we need to be able to represent pathways using the small molecules they comprise, and, in addition, we need a quantitative way of either predicting the distance of a molecule from a pathway or classifying this molecule as "belonging" to one of the existing pathways. For the classification task we found that the *random forest* classifier[13] performed well in our case. Random forests are robust, can be used with a large number of variables, and do not overfit. In addition, they have a built-in prediction of their accuracy, and they provide measures of the importance of descriptors and can also provide a measure of similarity for the objects that are being described. All these advantages have recently made random forests a popular method in many chemoinformatics studies: In an early chemoinformatics application, Svetnik et al. tested them in regression and classification tasks for small molecules,[14] and other studies have followed since, including the classification of prohibited substances used for doping in sport,[15] the identification of molecular targets for active ingredients in Chinese herbs,[16] and the classification of small molecules into metabolites and nonmetabolites.[17] The random forest classifier in this study is built using as response variables either the pathway classes or individual KEGG pathways, and as independent variables the 32 QSAR descriptors introduced by Labute[18] (relating to the atomic contributions to van der Waals surface area, octanol/water partition coefficient, molar refractivity and partial charge), as they summarize well the breadth of important properties in a biological context.

In the final part of this work, we address the issue of relationships between human metabolites in the context of pathways and drugs. This is motivated by the observation that many of the drugs in the marketplace are analogues of, or inspired by, natural products.[19] The comparison of the physicochemical properties of natural products and their typical structural features with those of bioactive molecules and synthetic molecules have so far been pursued with the aim of inspiring the design of new drugs.[4,17,20−25] Such studies have benefited significantly from modern chemoinformatics methods for comparing small molecules and statistical techniques for exploring and dividing the multivariate descriptor chemical space. One of the conclusions from these studies is that the overlap between natural and synthetic molecules appears to be dependent on the definition of the chemical space in which they are compared, e.g. Gupta and Aires-de-Sousa[17] showed that overlap was least between the two groups when global descriptors, such as molecular weight and the number of OH groups, were used, instead of 3D or 2D-structure-based descriptors. However, some overlap (occasionally strong[22]) is always observed between drugs and natural metabolites, not surprisingly, as many drugs were developed to mimic natural products with known biological activity.

Here, we take the approach that the comparison of drugs and human metabolites should ultimately prove useful in uncovering and understanding possible links between metabolism and the biological actions of drugs, especially in cases where little is known about the drug's mechanism of action. A drug lying in the vicinity (in chemical space) of metabolites with well-defined biological activities may have a pharmacological profile that relates to these activities. Moreover, a significant number of drugs fail at the clinical trials stage because of unwanted side effects and thus need to be abandoned at a time when too much time and money has already been invested in their development. Clearly, many of these side effects arise from off-target interference with the protein regulatory networks or other cellular mechanisms, rather than directly with metabolism. However, some drugs have been shown to interfere with enzymes in metabolic pathways: Zanamivir (an anti-influenza drug) inhibits not only the viral sialidases but also was recently shown to have significant activity against human sialidases, enzymes involved in sialic acid metabolism.[26] Occasionally, side effects are due not to the drug hitting a different target but are the results of blocking an enzyme that is responsible for multiple metabolic routes. For example, statin drugs reduce cholesterol levels by blocking HMG-CoA reductase, but they are also responsible for cases of myopathy among patients, as inhibition of this enzyme results in depletion of farnesyl pyrophosphate, an intermediate in the production of coenzyme Q10.[27] Apart from the fact that drugs can and do affect metabolism, our understanding of metabolic networks is far superior to that of regulation, and so we believe that at this stage it is useful to concentrate on metabolic networks first. We hope that increasing our understanding of how drugs may interfere with metabolic pathways could lead to more efficient computational prescreening of lead compounds and a way to obtain early warnings about potential damaging interference with important metabolic pathways.

There are, of course, caveats to our approach, and we briefly summarize the most important ones here. A major issue is that pathways in any database are artificial constructs created by humans to organize the complex biochemical network of metabolism into easily managed modules. Although some of these modules may be justified on the basis of the organization of the related proteins into large complexes, or the coexpression of the relevant genes, or even evidence of their proximity in the genome and their lateral transfer across organisms as a whole, the point remains that where a pathway starts and where it ends is largely an arbitrary decision. This is reflected in the fact that metabolic pathway databases do not always agree on the length and constitution of even the best-studied metabolic pathways. Nevertheless describing pathways in terms of such modules is widely accepted and facilitates our understanding of what is in reality a very complex network.

Furthermore, the model of chemical space we have built here ignores spatial and temporal constraints that are

inevitably imposed in complex eukaryotic organisms by compartmentalization and differential gene expression. At present we lack the necessary information to build models that will account for these effects. Hence, our notion of "overlap" of pathways must be interpreted as the potential to overlap, given the right conditions and circumstances. Indeed, it is likely that where significant overlap could be present in theory, the organism would probably find a way of avoiding any cross-reactivity problems by actively regulating the construction and operation of potentially overlapping pathways. It is known that more complex organisms with larger genomes require more regulatory mechanisms,[28,29] and it is possible that simpler unicellular organisms may exhibit comparatively fewer potentially overlapping pathways. However, we do not think that the data are available yet to test this hypothesis using our methodology. Further caveats in this work are discussed in more detail following the presentation of our results.

METHODS

The Data Set of Metabolites. The data set of human metabolites was built using information from the KEGG database (release 32 with updates from release 36).[30] Each compound entry (i.e., anything with a KEGG id starting with "C") was then processed in the following way: the software CACTVS[43] was used to create SMILES strings of the initial KEGG MDL mol files, and the CACTVS editor was used to examine each entry for problems arising from missing (undefined) stereochemistry, charges, or incorrect atom and bond types. Stereochemistry was manually assigned, where missing, following the reaction information in the pathway in which the compound was involved. If it was not obvious what the stereochemistry should be, it was assigned randomly (this was the case for approximately 13% of the molecules we examined). The online CORINA server (which currently exists only as a demo version on the Molecular Networks Web site at www.molecular-networks.com/online_demos/corina_demo.html) and the LigPrep program (LigPrep version 2.1, Schrödinger, LLC, New York, NY 2005) were used to produce energy-minimized conformers for all molecules. We should note here that the descriptors reported in this work rely only on the connectivities of the molecules, and hence 3D conformers and stereochemistry information were actually not used in the current study. However, the use of software to calculate 3D minimized conformations provides a further automatic checkpoint to help detect and correct problems with the 2D diagrams.

Of the initial 156 pathways and 1909 compounds listed in the *hsa*.cpd* files, we produced a data set of 7 metabolic classes, 61 pathways, and 881 compounds after the following filtering: The pathways belonging to the classes: glycan biosynthesis and metabolism, biosynthesis of polyketides and nonribosomal peptides, biosynthesis of secondary metabolites, and xenobiotics degradation and metabolism were removed in their entirety. These pathways contain molecules that cannot be automatically processed (they are polymers of unspecified length), or they are pathways that are either involved in the process of xenobiotics (and so the substrates are not endogenous to the human metabolome), or they appear to comprise human enzymes but the pathways are unlikely to be present in humans. In addition, all compound entries with unspecified R groups (where an assumption of the length of the chain and the type of the group could not be made) and all DNA/RNA-containing molecules were removed. Finally, a pathway was removed if it contained fewer than 3 enzymes (usually an indication that this pathway is not active in humans, or at least, there is insufficient evidence for its presence).

The filtering resulted in 61 pathways comprising 1205 small molecule - pathway relationships and 881 small molecules with unique KEGG ids. We refer to this data set as *human_all*. We then built an additional data set for classification using only compounds that had unique pathway annotations (i.e., were found in a single pathway in the first data set). We call this new data set *human_unique*, and it comprises 681 metabolites distributed among 52 pathways. The two data sets are provided as lists in the Supporting Information (Data set 1 and Data set 2).

Molecules from the *human_all* data set were imported in the Molecular Operating Environment (MOE 2006.08 release, http://www.chemcomp.com) software for the calculation of 2D descriptors. The molecules were "washed" and ionized at physiological pH, and partial charges were calculated using the default force field (MMFF94x). All 2D descriptors available in MOE were calculated for this data set, and they were exported into an ASCII file for further analysis.

The Data Set of Drugs. The data set of all small molecule drugs was downloaded from the DrugBank[44] online database (02/01/2008 update). SMILES strings were extracted from the flat file and were imported in the MOE. The molecules were "washed" and ionized at physiological pH, and partial charges were calculated using the default force field (MMFF94x). This data set is referred to as the *all_drugs* data set and comprises 4464 molecules. All 2D descriptors available in MOE were calculated for this data set, and they were exported into an ASCII file for further analysis using the R statistical software.[45] In addition the data set of all drugs tagged "experimental" (3103 molecules) was downloaded from DrugBank (02/01/2008 update), and following the same procedure as with the *all_drugs* data set their descriptors were calculated in MOE. Finally, to obtain all drugs flagged as withdrawn (64 molecules), the "Data Extractor" facility on the DrugBank Web site was used setting the "Drug Type" to "withdrawn". The results were downloaded, and the DrugBank ids for these molecules were extracted.

Calculation of Small Molecule Similarity. Small molecule similarity was calculated using three different types of fingerprints: a) Hashed fingerprints produced with the Fingerprinter class of the open source Chemistry Development Kit[46] (with fingerprint size=1088 bits and depth search = 7 bonds), b) the MACCS structural keys calculated using the MOE software, and c) the chemical fingerprints calculated using the JChem library's (JChem version 3.2, 2006, ChemAxon, www.chemaxon.com) GenerateMD program with parameters set to length = 1024, bondCount = 7, and bitCount = 3. In all cases, a pairwise similarity matrix was calculated using the Tanimoto coefficient.[47] The pairwise dissimilarity is simply calculated as (1-similarity).

Between-Pathway-Similarity and within-Pathway Dissimilarity. We calculate the between-pathway-similarity of two pathways P_A and P_B as the mean of all pairwise

comparisons $t_{i,j}$ (Tanimoto scores) for all molecules i and j belonging to the two pathways, i.e.

$$\bar{S}_{P_A,P_B} = \frac{\sum_{i\in P_1, j\in P_2} t_{i,j}}{N_A \times N_B} \quad (1)$$

where N_k is the number of molecules in pathway P_k. We assume here that the average similarity between two pathways is independent of pathway size.

The calculation of within-pathway-dissimilarity is less straightforward. Here, there is a clear dependence between pathway size and the expected dissimilarity, so the mean within-pathway-dissimilarity cannot be simply calculated as the average of all possible pairwise comparisons for all molecules within the same pathway. As an example, assume for simplicity that pathways are linear and that each reaction step alters a molecule by the same amount, λ (so that each product differs from each substrate of the reaction by λ). Then the simple case of a pathway with three steps and four metabolites M_k, would look like

$$M_1 \xrightarrow{\lambda} M_2 \xrightarrow{\lambda} M_3 \xrightarrow{\lambda} M_4 \quad (2)$$

In this simple case, there would be six pairwise small molecule comparisons, three equal to λ, two equal to 2λ, and one equal to 3λ. The mean dissimilarity would then be

$$\frac{3\times 1\lambda + 2\times 2\lambda + 1\times 3\lambda}{6} = \frac{5}{3}\lambda \quad (3)$$

More generally, for N metabolites and N-1 steps, we would have

$$\sum_{k=1}^{N-1} k = \frac{N(N-1)}{2} \quad (4)$$

comparisons, and an expected mean within-dissimilarity of

$$\bar{D}^{w}_{\mathrm{A}} = \frac{\sum_{k=1}^{N-1}(N-k)k}{\sum_{k=1}^{N-1} k}\lambda = \frac{N+1}{3}\lambda \quad (5)$$

Clearly λ is reaction-dependent and not all pathways are linear, but to a first approximation this is a reasonable weighting scheme.

Hence, we see that the within-pathway-dissimilarity of a pathway of N small molecules is proportional to N, assuming linearity and a constant λ. In other words, longer pathways are expected to have higher values of within-pathway dissimilarity than shorter pathways. In addition, if λ is constant across pathways, then we can calculate an expected value for the within-pathway dissimilarity of pathway P_{A} relative to a standard pathway P_{std} as

$$\bar{D}^{w}_{\mathrm{A,\,expected}} = \frac{N_{\mathrm{A}}+1}{N_{\mathrm{std}}+1}\bar{D}^{w}_{\mathrm{std}} \quad (6)$$

The difference between the actual and expected values for pathway P_{A} gives us an estimate of how tight or diffuse this pathway is. Defining the standard pathway is, of course, not straightforward, but any linear set of reactions that do not involve major structural rearrangements (such as the addition of a cofactor molecule) is a reasonable choice. Ultimately, the choice of the standard is less critical because the values for the pathways will be relative to the same standard. Here, we choose a linear set of five reaction steps from the fundamental pathway of glycolysis (KEGG map: hsa00010) that transform D-glyceraldehyde 3-phosphate to pyruvate. The six molecules involved and corresponding KEGG codes are as follows: D-glyceraldehyde 3-phosphate (C00118), 3-phospho-D-glyceroyl-phosphate (C00236), 3-phospho-D-glycerate (C00197), 2-phospho-D-glycerate (C00631), phosphoenolpyruvate (C00074), and pyruvate (C00022).

Using the MACCS fingerprints, the average dissimilarity for these six molecules/five steps is 0.325. This results in a lambda value of 0.139. In other words, two molecules separated by seven reaction steps or more are expected to be sharing no detectable similarity using this similarity coefficient (the Tanimoto score would drop to zero). Substituting these values in eq 6 we can calculate the expected mean within-pathway dissimilarity value for any given pathway. One problem with this approach is that actual dissimilarity values are constrained between zero and one, whereas values calculated from eq 6 are unconstrained. In practice, using the standard pathway proposed here, a pathway with 21 or more molecules would already have a mean dissimilarity exceeding one, so we set the within-pathway dissimilarity of any pathway with more than 20 molecules automatically to an expected value of 1.

We can also briefly examine the effect of a pathway not being linear on the calculations above. In the case of a cyclic pathway, it is easy to see that after the first half set of reactions molecules presumably converge again in terms of their structure toward the starting point of this pathway. Hence, our measure from eq 5 will be an overestimate of the within-pathway dissimilarity, as any long-distance comparison (longer than half the total number of reactions in the pathway) will be replaced by a shorter-distance one. The second type of nonlinear pathways, branched pathways, can vary enormously in their arrangement of substrates and products, but it is obvious again that the qualitative effect will be one of lowering the mean dissimilarity, as calculated for a linear pathway comprising the same number of molecules. The higher node connectivity of such networks results in a lower average path length between all pairs of molecules in the network and hence a lower mean dissimilarity of the molecular structures.

The Random Forest Classifier. We classified the uniquely annotated 681 metabolites using as response variable either a) the 52 pathways in our data set or b) the 7 classes to which these pathways correspond (as defined in KEGG). We built two different classifiers using as variables: i) the 184 2D descriptors available from the MOE package and ii) the 32 descriptors suggested by Labute[18] as a small set of relatively uncorrelated descriptors that have been shown to perform reasonably well in describing the properties of molecules. The Random Forest classifier was applied using the *R* package "randomForest". After some initial testing, the number of trees was set to 5000, and the *mtry* parameter was optimized using the *tuneRF* function provided in this package. Random forests include an out-of-bag estimate of

error that is calculated by constructing trees with a different bootstrap sample of the data, so there is no need for cross-validating the model.

The plot in Figure S7: Classification errors from using all 2D MOE descriptors and the 32 Labute descriptors shows the comparison of classification errors from the two random forests (using as response the 52 pathways) and shows that for the same number of trees (5000) and optimized *mtry* parameters, the set of Labute descriptors performs on average slightly better than the set of all descriptors. Hence, all reported results in this study refer to the random forest built with the 32 Labute descriptors.

To assess the quality of the classifier we have calculated for each class the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), assuming that the classification was binary in each case. We have then calculated the Matthews correlation coefficient (MCC) for each class using

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (7)$$

The MCC is recognized as a more representative measure of the quality of a classification, as it takes into account the fact that classes may vary significantly in size.

RESULTS

A data set of small molecules involved in metabolic reactions catalyzed by enzymes encoded by the human genome was built using information from the KEGG database.[30] As in our experience all small molecule databases contain several entries where molecules are incompletely (and occasionally incorrectly) described for computational processing, we have manually curated every molecule in our data set, to deal with common problems such as the lack of hydrogens, inconsistent inclusion of charges, lack of stereochemistry definitions for chiral centers, and the presence of unspecified *R*-groups or undefined repeated motifs within a molecule. The final data set, which we will refer to as *human_all*, contained 881 small molecules from 61 pathways. 190 of these 881 molecules appear in more than one pathway. Removing molecules with multiple pathway annotation (for the purpose of training the classifier) and, in addition, whole pathways comprising fewer than 3 small molecules, results in a reduced data set of 681 small molecules distributed among 52 pathways (which we refer to as *human_unique*). Following the KEGG classification of metabolic pathways (see www.genome.ad.jp/kegg/pathway.html), each pathway in our data set belongs to one of seven classes: carbohydrate metabolism (abbreviated here to "CM", comprising 14 pathways), energy metabolism ("EM", 2), lipid metabolism ("LM", 8), nucleotide metabolism ("NM", 2), amino-acid metabolism ("AM", 12), other amino-acid metabolism ("OAM", 4), and cofactor and vitamin metabolism ("CVM", 10). A summary of our data sets is given in Table 1. The KEGG small molecule codes for both data sets in this study are provided as Data set 1 and Data set 2 in the Supporting Information.

To obtain an overview of the properties of the human metabolome, we have calculated a series of topological and physicochemical descriptors for each molecule. Figure 1 depicts the distribution of six common descriptors across the seven pathway classes in our *human_all* data set. The lipids (LM) and cofactors and vitamins metabolism (CVM) stand out as the classes with the largest, most hydrophobic molecules. The LM and CVM classes also exhibit higher chemical complexity, as evidenced by larger Wiener path numbers and larger atomic connectivity indices. These classes are populated with several molecules likely to have emerged later in evolution (synthesized by combination of simpler scaffolds) with the appearance of multicellular organisms requiring more lipid-based membranes for cell compartmentalization.

Table 1. Summary of the Data Sets Used in This Study

pathway class	no. of pathways in class	no. of molecule-pathway associations in class (no. of unique molecules in class)
	Human_all	
carbohydrate metabolism (CM)	14	274 (199)
lipid metabolism (LM)	10	266 (225)
cofactors and vitamins metabolism (CVM)	11	124 (121)
amino acid metabolism (AM)	14	340 (280)
other amino acid metabolism (OAM)	6	67 (64)
nucleotide metabolism (NM)	2	101 (99)
energy metabolism (EM)	4	33 (32)
total	**61**	**1205 (881)**
	Human_unique	
carbohydrate metabolism (CM)	14	112
lipid metabolism (LM)	8	167
cofactors and vitamins metabolism (CVM)	10	93
amino acid metabolism (AM)	12	193
other amino acid metabolism (OAM)	4	30
nucleotide metabolism (NM)	2	78
energy metabolism (EM)	2	8
total	**52**	**681**

Part I: The Distribution of Metabolic Pathways in Chemical Space. Intuitively, one would think that each pathway forms a cluster in chemical space, as the small molecules it comprises are likely to lie close to each other in that space. Here we examine to what extent this intuitive view is true for each class of pathway in our data set. To define the chemical space in a way that will allow us to map pathways onto it, we describe small molecules in two alternative ways: using their 2D connectivity information (which we refer to as 2D structure) and using a series of physicochemical/topological descriptors.

Inter- and Intrapathway Similarities Using 2D Structure and Descriptors. We calculated the pairwise similarity matrix for the *human_all* data set, i.e. 881 small molecules from 61 human metabolic pathways (see Methods for details). Small molecule similarity was assessed using the Tanimoto score on three different types of binary fingerprints, as described in the Methods: a) the CDK hashed fingerprints, b) the MOE MACCS structural keys, and c) the JChem chemical fingerprints. These three fingerprint descriptors allow us to compare metabolites based on the 2D connectivity information. Having an all by all similarity matrix, we can then obtain a qualitative view of the overall distribution of pathways in chemical space with the help of multidimen-

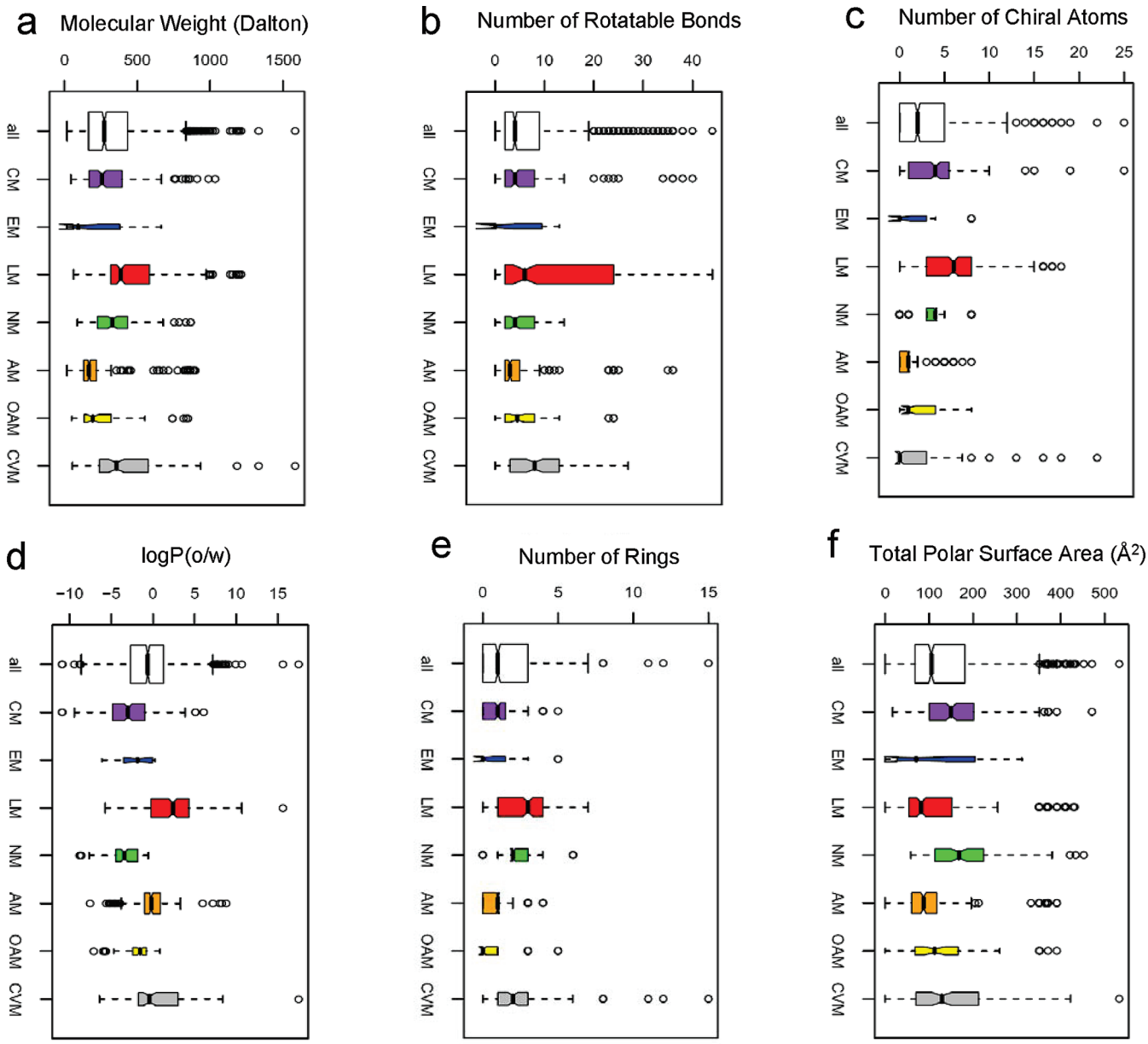

Figure 1. Common descriptor distributions for the pathway classes of the *human_all* data set. Box-and-whisker plots of the distribution of values for six common physicochemical descriptors calculated for the pathway classes of the *human_all* data set (molecular weight (a), number of rotatable bonds (b), number of chiral atoms (c), octanol−water partition coefficient, logP (d), number of rings (e), and polar surface area (f)). Abbreviations for the pathway classes defined in KEGG are as follows: CM = carbohydrate metabolism, EM = energy metabolism, LM = lipid metabolism, NM = nucleotide metabolism, AM = amino acid metabolism, OAM = other amino acid metabolism, CVM = metabolism of cofactors and vitamins. In these boxplots, the box contains 50% of the distribution (from the first to the third quartile), and the whiskers extend to the most extreme values of the distribution that are still within the inner fences (that is, one-and-a-half times the width of the box). Anything outside these whiskers is considered an outlier and it is represented by circles in these plots. The vertical line inside the box represents the median of the distribution. The width of the box is proportional to the square-root of the number of observations in the group. Notches are drawn in each side of the box to give roughly a 95% confidence interval for the difference in the medians of these plots (see the *R* boxplots manual for details). Plots were drawn using *R*.

sional scaling. Figure 2a shows that using, for example, the MACCS fingerprints, there is a tendency for classes of pathways (as defined in KEGG) to occupy different areas in structural space, although clearly there is some overlap not only between pathways but also between the broader pathway classes. There are also at least some outliers occupying a space outside that of the main cluster of every class. Figure 2b-d additionally shows that the amino acid and other amino acid classes overlap strongly and that the cofactors/vitamins and energy classes are more widely spread in chemical space. Results are similar for the CDK and JChem fingerprints (plots not shown), and although the details vary between the different methods, the overall picture remains the same. One should also expect the overlap between pathways and classes of pathways to be overestimated in this figure, as the effect of scaling the multidimensional space down to three (but effectively two) dimensions would be to make the data points appear more "squashed" than they really are.

A different way to visualize these data is to use a graph-based approach such as the one employed by the Cytoscape[31] software (Cytoscape version 2.4.1, http://www.cytoscape.org). In such plots metabolites are represented as nodes, with an edge drawn, if the pairwise similarity is greater than a given cutoff. The plot in Figure S1 (Network of *human_*

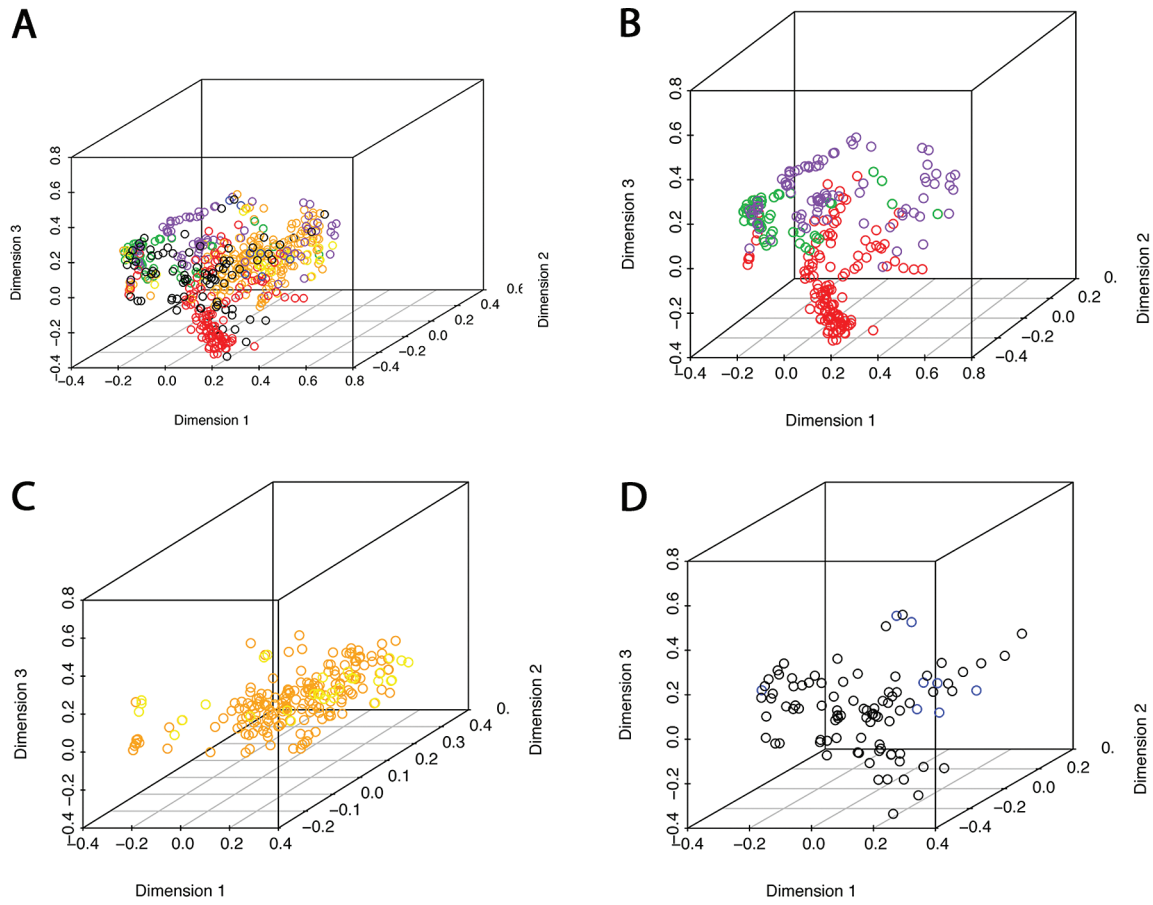

Figure 2. The uniquely annotated metabolites distributed in space according to their pairwise similarity. The plots are produced using multidimensional scaling (in three dimensions) on the dissimilarity matrix (defined as 1-Tanimoto_score) resulting from a pairwise comparison of all MACCS fingerprints. The three axes represent the three dimensions resulting from the scaling. The molecules are colored according to their pathway class, rather than the pathway they belong to, as there are too many individual pathways to color and distinguish easily. Colors are as follows: purple = CM, blue = EM, red = LM, green = NM, orange = AM, yellow = OAM, black = CVM. These plots were drawn with the *scatterplot3d* package in *R*. a) All 681 uniquely annotated metabolites, b) metabolites from the lipid (red), nucleotide (green) and carbohydrate (purple) metabolism classes, c) metabolites from the amino acid (orange) and other amino acid classes (yellow), and d) metabolites from the cofactors and vitamins (black) and energy (blue) classes.

unique metabolites) depicts the network of human metabolites using Cytoscape and supports the observations from Figure 2 that the pathway classes are clustered in chemical space.

As an alternative (and possibly more relevant to cross-reactivity) description of metabolic pathways in chemical space we calculated a series of physicochemical and topological descriptors (all 1D/2D descriptors available in the MOE software) for the *human_all* data set. We then applied Principal Component Analysis (PCA) on the scaled values of these descriptors in order to visualize the human metabolites in the space of the first three principal components (see Figure 3a). This plot offers a qualitative view of the distribution of pathways in physicochemical (rather than structural/connectivity) space that largely agrees with that offered by Figure 2, e.g. the EM and CVM classes appear more spread and less clustered. A further observation is that several metabolites that are found on the edge of these plots are interesting in different ways. We annotate some of these in Figure 3b, and their chemical structure is reported in Figure 4. For example, beta-carotene (**1** - C02094), chitin (**2** - C00461), and iron (**3** - C00023) appear at extremes of the plot, and neither of them is actually synthesized by humans, although humans have enzymes to catabolize these molecules such as in the retinol and amino-sugars metabolism pathways for **1** and **2**, respectively (KEGG pathway codes: hsa00830 and hsa00530). Others, like the inositol derivatives (**4** - C01204 and **5** - C01284), simply do not cluster with other compounds because their structures and properties are not shared among many other metabolites. This effect is accentuated in this case by the fact that in our original data set these two compounds were "at the end" of a reaction path, meaning that some of the compounds surrounding them in the reference pathway of KEGG were not recognized as belonging to the human metabolome (the genes for the catalytic transformations were not yet known or incorporated in these maps). This situation has partially changed (see an example in Figure S2: The inositol phosphate metabolism pathway from two different KEGG releases). As more genes become annotated and the pathways become more complete several of these outlier molecules in the plot might end up in bigger clusters. Finally, several of these outliers although present in the KEGG release that this study was based on, are now removed from the KEGG human metabolic pathways. This is the case for the two vitamin B12 molecules (**6** - C00194 and **7** - C00853) and 2-octaprenyl 6-hydroxyphenol (**8** - C05811). The fact that many of the outliers can be rationally explained reinforces the significance and potential usefulness of these plots.

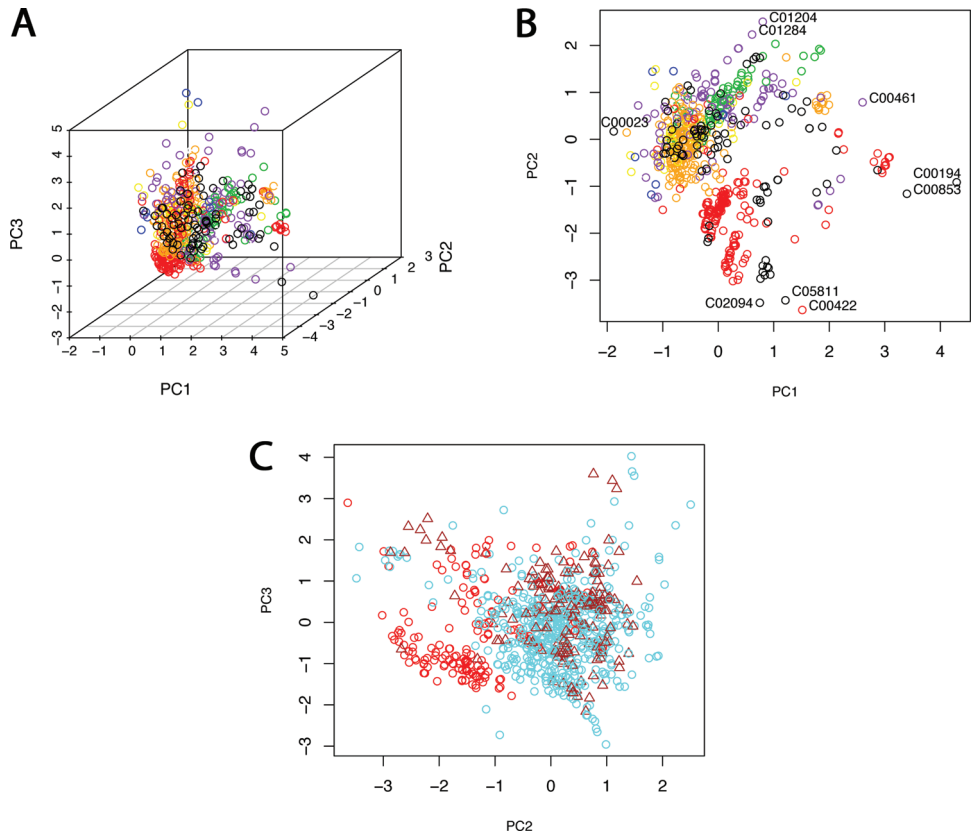

Figure 3. Human metabolites distributed in physicochemical space. Plots of all metabolites from the *human_unique* data set in the space of the first three principal components obtained from PCA on all scaled 1D/2D descriptors calculated with MOE. Each colored dot is a small molecule. a) Metabolites in the space of the first three principal components, accounting for 47%, 60%, and 67% of the cumulative proportion of variance in the data set (plot drawn with the *scatterplot3d* package). PC1 relates mainly to size and PC2 to hydrophobicity, and PC3 has large contributions from descriptors relating to partial charges and molar refractivity. b) PC2 vs PC1 with annotation of metabolites at "extreme" positions (using KEGG codes). c) PC3 vs PC2, including both metabolites with unique annotations (cyan and red) and metabolites with multiple annotations (brown triangles). Metabolites from the lipid class are highlighted in red. The coloring scheme follows that of other figures: purple = CM, blue = EM, green = NM, red = LM, orange = AM, yellow = OAM, black = CVM. Molecules represented as red circles in plot (c) belong mostly to steroid-metabolizing pathways. The KEGG codes in plot (b) correspond to the following compounds: adenosyl cobalamin (C00194), cob(I)alamin (C00853), myo-inositol-hexakis-phosphate (C01204), inositol 1,3,4,5,6-pentakisphosphate (C01284), iron (C00023), chitin (C00461), beta-carotene (C02094), 2-octaprenyl 6-hydroxyphenol (C05811), and triacylglycerol (C00422). Plots were drawn using *R*.

In Figure 3c, we have included metabolites with multiple annotations (annotated as brown triangles). At first glance, it appears that these metabolites largely overlap with those uniquely annotated in our data set, suggesting that these two categories of metabolites might not be easily distinguishable from their physicochemical properties. A notable exception is a cluster of small molecules from the lipid metabolism class (red circles in Figure 3c), which do not overlap significantly with others and are uniquely annotated. These belong mostly to steroid pathways, so they comprise lipid-soluble molecules, able to traverse membranes and used in signaling. Indeed 94 of the 96 molecules in the three main steroid pathways (hsa00100, hsa00140, and hsa00150) are uniquely annotated, a proportion much larger than the average 77% across all pathways. This observation agrees with the recent study of Gutteridge et al.[32] who showed that many essential metabolites (such as pyruvate) have a central role in bacterial but not in human regulation, where molecules that act in regulatory roles tend to be used primarily in signaling, and would not be expected to be part of multiple metabolic reactions and pathways.

Pathway Tightness. Although the above are reasonable ways to visualize a multidimensional space, it would be more useful to have a quantitative way of assessing both pathway tightness and pathway overlap in chemical space. Starting with the assumption that each pathway is a predefined cluster in chemical space, pathway tightness may be assessed by comparison of the calculated within-pathway mean dissimilarity to an *expected* mean dissimilarity score that grows linearly with pathway length. The expected score takes into account the length of the pathway, but not the type of reactions involved, or the extent of pathway branching (see Methods for details). The rationale behind this score is that one would intuitively expect longer pathways to be associated with larger mean dissimilarities, simply because at every reaction step a change in the metabolite structure occurs, resulting in more diverse structures, the further away from the starting point of the pathway one is. To estimate the relative tightness of a pathway, we compare it to a standard pathway, chosen here as a linear set of five reaction steps from the fundamental pathway of glycolysis (see Methods). The red line in Figure 5a depicts the expected mean dissimilarity value for a "standard" pathway of given length. Following our method, any pathway with more than 20 molecules is expected to have a mean dissimilarity value of 1, so the line representing the standard is set to $y = 1.0$ at a pathway length of 20 molecules.

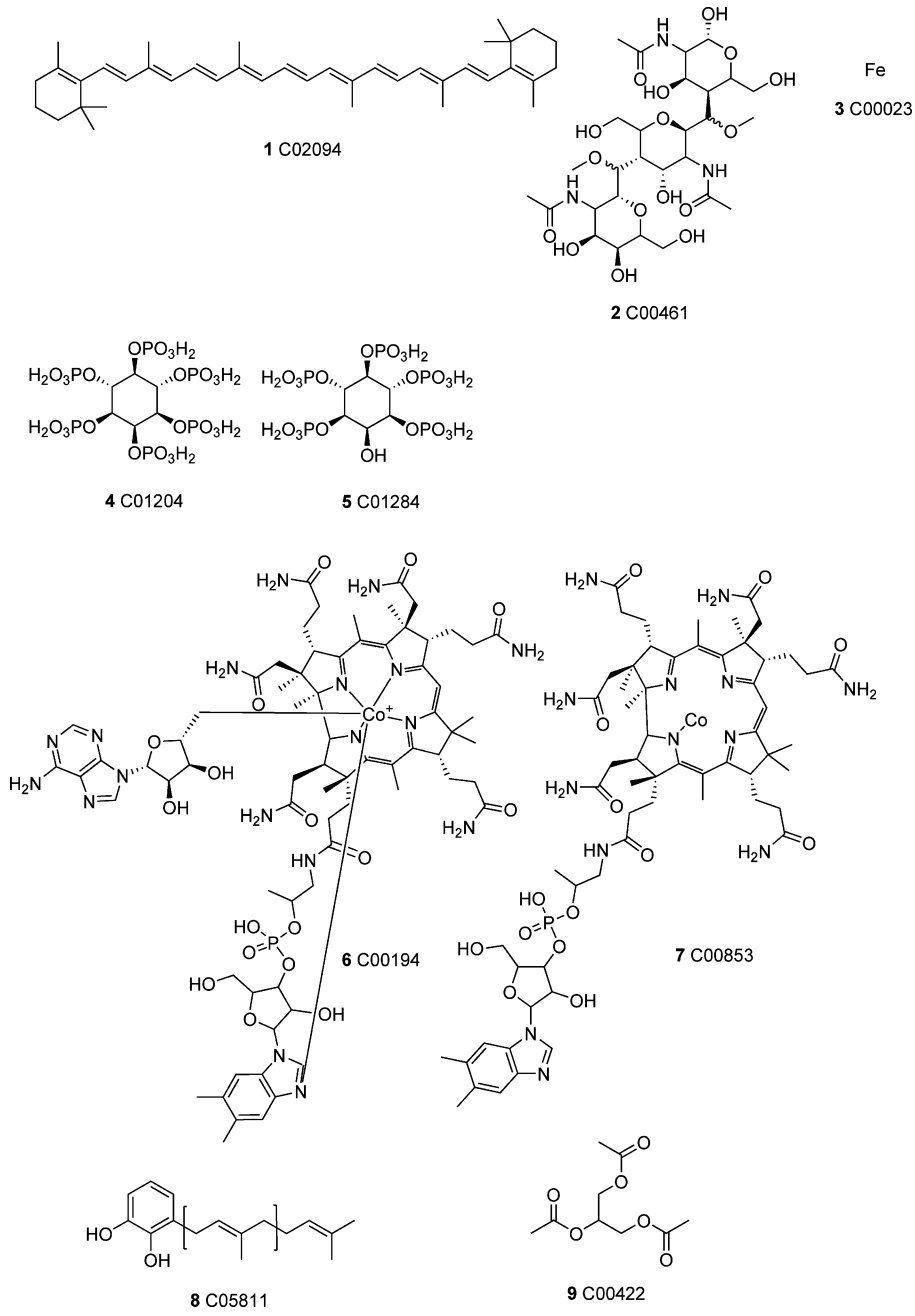

Figure 4. Chemical structures of metabolites at "extreme" positions of the chemical space.

The distribution of within-pathway dissimilarities is depicted in Figure 5a. Approximately one-third of all pathways (22 of 61) comprise more than 20 molecules, and hence are expected to have very high within-pathway mean dissimilarity values, as these values will reflect mostly comparisons between pairs of molecules separated by too many reactions to retain any significant similarity. In reality, these longer pathways are a lot tighter than expected in chemical space. Reasons for this include the fact that there are limited chemical scaffolds present in pathways so some molecular similarities are fortuitous, and many pathways are branched rather than linear, and hence a large number of molecules are closer in terms of reaction steps separating them than anticipated by our simple model. Interestingly, shorter pathways do not deviate far from the expected mean dissimilarity values (the red line in Figure 5a), and, in fact, approximately one-sixth of all pathways are less tight than expected (although the differences are small and may not be significant). The energy class pathways (in blue) are more scattered in chemical space, not surprisingly as these involve structures of metabolites that are unrelated among the protein complexes involved or that involve mainly attaching and releasing larger groups/cofactors to basic smaller molecules or making basic cofactors like PAP from sulfate (sulfur metabolism, hsa00920). Whereas most lipid pathways (in red) appear to be tighter than expected, the pathway for synthesis and degradation of ketone bodies (hsa00072) is less tight than expected due to the fact that some of the molecules in this pathway contain the rather large coenzyme A group, whereas others do not.

Pathway Overlap. We first calculate pathway overlap as the mean interpathway similarity (see Methods) for each pair of pathways. The matrix in Figure 5b depicts the results of the overlap calculation. Clearly the vast majority of pathways

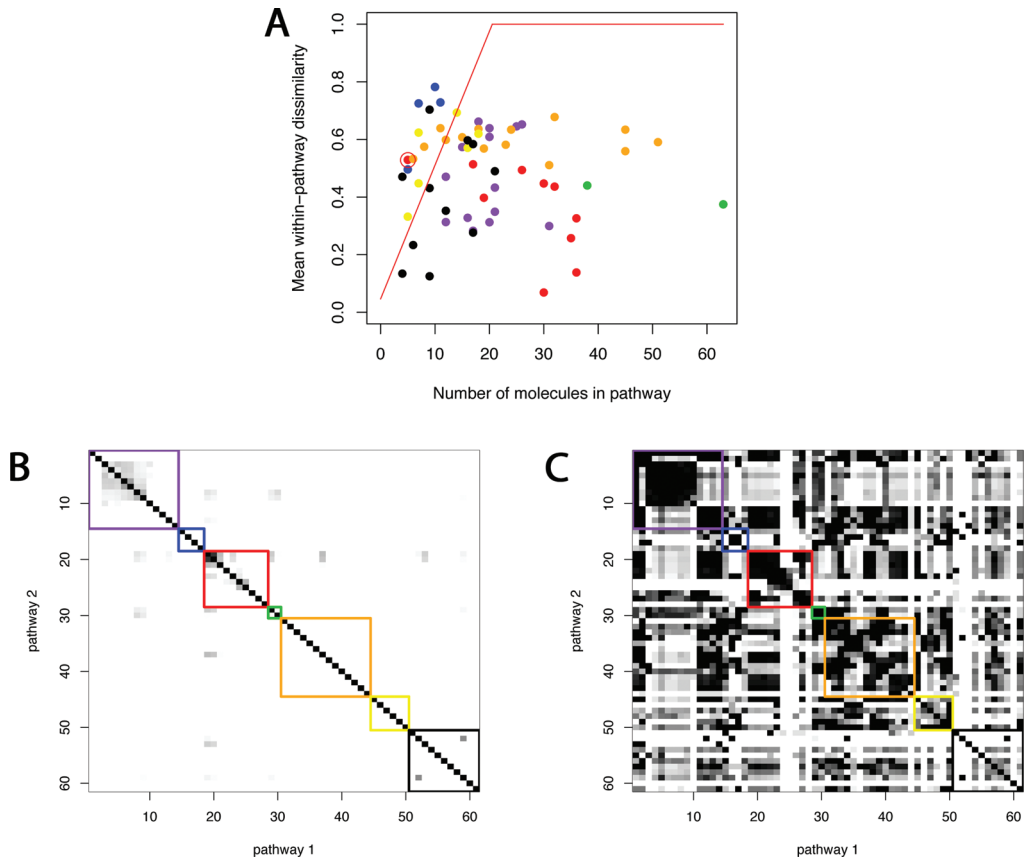

Figure 5. Human metabolic pathway tightness and overlap using the MACCS fingerprints. **a)** Within-pathway mean dissimilarities calculated for each pathway using the MACCS-based scores vs the number of molecules in the pathway. The red line represents the expected within-pathway mean dissimilarity for a pathway, given the "standard" pathway (see Methods), and the number of molecules in the pathway. In practice, a pathway with more than 20 molecules is here considered to have an expected dissimilarity of 1.0. The pathway of synthesis and degradation of ketone bodies (hsa00072) is highlighted with a larger red circle. The figure was drawn using *R*. **b)** The symmetric matrix of between pathway similarities calculated using the MACCS fingerprints for the *human_all* data set (the equivalent matrix for the JChem fingerprint scores can be found in Figure S3: Metabolic pathway overlap using the JChem fingeprint similarity). The matrix of the original calculated similarities has been transformed for ease of visualization as follows: The diagonal is set by definition to 1.0, and every pairwise comparison is set to the same minimum value, if the similarity score is below the mean plus one standard deviation calculated for all pairwise comparisons of all 881 molecules in our data set. Each square is colored in a scale of gray according to the average similarity for the pair of pathways corresponding to that row and column. Higher similarities are darker in color. The pathways are ordered according to the class they belong to. The larger colored squares enclose the data for the comparisons within each class (colors are as follows: purple = CM, blue = EM, red = LM, green = NM, orange = AM, yellow = OAM, black = CVM.) Pathway numbers for the pathways involved in stronger interclass overlaps (squares that appear gray in the plot) are as follows: hsa00520 (8), hsa00530 (9), hsa00640 (13), hsa00062 (19), hsa00071 (20), hsa00072 (21), has00140 (24), hsa00230 (29), hsa00240 (30). The pair of CVM pathways with strong overlap are hsa00062 and hsa00071. The figure was drawn using the *plotrix* package in *R*. **c)** As (b), but each cell now represents the maximum similarity between two pathways among all their pairwise constituent molecule comparisons. In addition, the similarity values below the average plus two standard deviations (as calculated from the full data set of 881 molecules) are set to the same value to reduce the noise levels in the figure. The diagonal is set by definition to 1. Higher similarities are darker in color. The same matrix but with a less strict cutoff of mean plus one standard deviation for filtering the noise can be found in Figure S4: Matrix of pairwise maximum similarities between pathways. The figure was drawn using the *plotrix* package in *R*.

show significant overlap to only a small number of other pathways, and some do not overlap significantly with any other pathways. The most strongly overlapping pair is hsa00062 (fatty acid elongation in mitochondria) and hsa00071 (fatty acid metabolism), seen as the darkest squares in the lipid metabolism class in Figure 5b. Fatty acid metabolism essentially comprises a series of reaction steps that represent the reverse reactions to fatty acid elongation, as they break down long-chain fatty acids into shorter ones. Hence, these two pathways actually share many of their metabolites. Only this and two more pairs of pathways (hsa00500 (starch and sucrose metabolism)/hsa00052 (galactose metabolism) and hsa00790 (folate biosynthesis)/hsa00670 (one carbon pool by folate)) are significantly more related than one would expect by chance, i.e. they display a mean similarity higher than the average plus two standard deviations calculated for the whole data set of molecules (this number is higher (8 pairs) when the JChem fingerprint similarity score is used instead (see Figure S3: Metabolic pathway overlap using the JChem fingerprint similarity), but it is still very low overall). The plot in Figure 5b presents a picture of pathways consistent with the absence of any serious overlap that extends to the majority of their constituent molecules. The classes with the largest within-class overlap are the carbohydrate metabolism and the lipid metabolism classes, most likely reflecting the abundance of common highly hydrophilic and hydrophobic scaffolds among sugars and lipids respectively.

Another observation is that, using the mean dissimilarity criterion, there is no significant overlap within the amino acid and "other amino acid" classes and almost no overlap between these classes and other ones (except for the valine, leucine, and isoleucine degradation pathway that overlaps

with several lipid pathways, most likely due to the presence of compounds carrying the coenzyme A moiety). Interestingly, the isolation of amino acids in chemical space may have played a crucial role in their evolution of function, especially their selection as signaling molecules in higher organisms. Amino acid-derived neurotransmitters, for example, trigger fast nongenomic signaling pathways, such as the chemical communication within synapses, by tight and selective binding to membrane (metabotropic and ionotropic) receptors. Thus, the need for high specificity and selectivity may enforce tight rules for the selection of structures involved in signaling. Lipid metabolites acting as endocrine hormones and mostly mediating genomic signaling pathways through the activation of nuclear receptors would also be expected to be isolated in chemical space, although perhaps to a smaller extent, as there are relatively fewer nuclear hormone receptors (~2% of the human genome[33]) than there are membrane receptors (15%). Indeed Figure 5b supports this idea, although steroid hormone pathways show more in-class overlap than do the amino acid pathways.

Another interesting observation can be made from Figure 5c. Here, instead of comparing average similarity values between pathways, we examine the highest similarity between any molecules belonging to a pair of pathways. The figure is strikingly different than Figure 5b and shows that the vast majority of pathways share at least one pair of molecules with high similarity, even when, on average, they show little overlap in the chemical structure space. This may reflect the fact that pathways are ultimately part of the larger network, and sharing molecules is an important mechanism for integration of a pathway into that network. It may also reflect the way these pathways evolved from ancestor reaction networks, as, on some occasions, the depletion of one molecule could have triggered the evolution of a pathway that synthesizes it,[34] thus resulting in two pathways sharing that molecule as substrate and product. What is additionally interesting is that some rows in the matrix of Figure 5c appear to be almost totally white, i.e. there are pathways that are truly isolated from others in this space. The best example is retinol metabolism (hsa00830, pathway number 60 in the matrix), an animal pathway that has appeared relatively late in evolution. The thiamine metabolism (hsa00730, number 53), ubiquinone biosynthesis (hsa00130, number 51) and taurine and hypotaurine metabolism pathways (hsa00430, number 46) are also among the most isolated ones. Interestingly, the data for amino acid metabolism pathways seem to contradict that in Figure 5b. It is apparent that many of these pathways are connected to many others through the presence of pairs of highly similar metabolites, even though on average they appear to be very different structurally. On the other hand the two steroid pathways, androgen and estrogen metabolism (hsa00150, number 25) and the C21 steroid hormones metabolism (hsa00140, number 24), appear to have almost no connection to the network outside the lipid metabolism class.

Part II: Toward a Statistical Predictive Model of Metabolic Class and Pathway Membership. This part of the work attempts to answer the question: given the structure of a small molecule, can we predict the pathway (or at least the pathway class) that this molecule could either belong to or lie closest to in chemical space. This question is important, as new metabolomics experiments are likely to reveal the structures of new metabolites, whose presence was unknown, and which cannot be easily placed in an existing metabolic network of an organism. In addition, it could help reveal potential interference of the small molecule with a specific pathway.

Classifying Small Molecules Using Physicochemical and Topological Descriptors. Measuring the overlap of pathways in physicochemical and topological descriptor space is complicated by the large number of dimensions of that space and the fact that the descriptors are of different type (categorical and continuous variables) and scale, and many are highly correlated. An easier problem might be to train a classifier that given a small molecule can predict the "closest" metabolic pathway. Initial tests with a subset of physicochemical descriptors commonly used to describe the properties of drugs and the application of linear discriminant analysis gave poor results. A set of the 32 descriptors suggested by Labute[18] improves this classification, but the overall 10-fold cross-validated errors are still more than 20% for the majority of pathways (see Figure S5: classification errors from linear discriminant analysis).

As an alternative we used the *random forest* classifier.[13] Random forest classifiers have many advantages over the more traditional classification methods: they do not overfit, can handle a large number of mixed variable types of data, and can also cope with missing data. A random forest is an ensemble of individual classification trees trained on bootstrap samples of the data, with each tree voting for a class given a set of observations/features. The class with the majority of votes wins. The random forest incorporates a form of cross-validation, known as the out-of-bag (OOB) error. Approximately one-third of the data are left out of the training of each tree and form a test sample for that tree. Hence, each observation can be classified by the trees that have not used it during training, and the errors of all trees can be averaged to give an overall classification error estimate. In our analysis, the features describing each small molecule are the 32 Labute descriptors, and the classes to which molecules are assigned are either individual pathways from our data set or KEGG pathway classes (different random forests were built for the two different classifications). We present three ways of assessing how successful a classification was. The first is the classification error for each class derived from the confusion matrix, the second is the percentage of votes given to each observation that correspond to the class the observation actually belongs to (percentage of "correct" votes), and the third is the Matthews correlation coefficient for each class (see Methods).

Classification of Small Molecules to Pathway Classes. A random forest was built that classifies the human metabolites into their corresponding pathway classes, as defined in KEGG, i.e. carbohydrate, energy, lipid, nucleotide, amino acid, other amino acid, and cofactor and vitamins metabolism. The *human_unique* data set is used for training, as each observation should be associated with a single class. An overview of the results is depicted in Figure 6a where it is obvious that the carbohydrate, lipid, amino acid, and nucleotide metabolism classes stand out as the four major types in the metabolome, with the remaining classes overlapping (the amino acids class overlaps strongly in physicochemical space with the other aminoacids class). Moreover, no class forms an isolated cluster, but there are metabolites

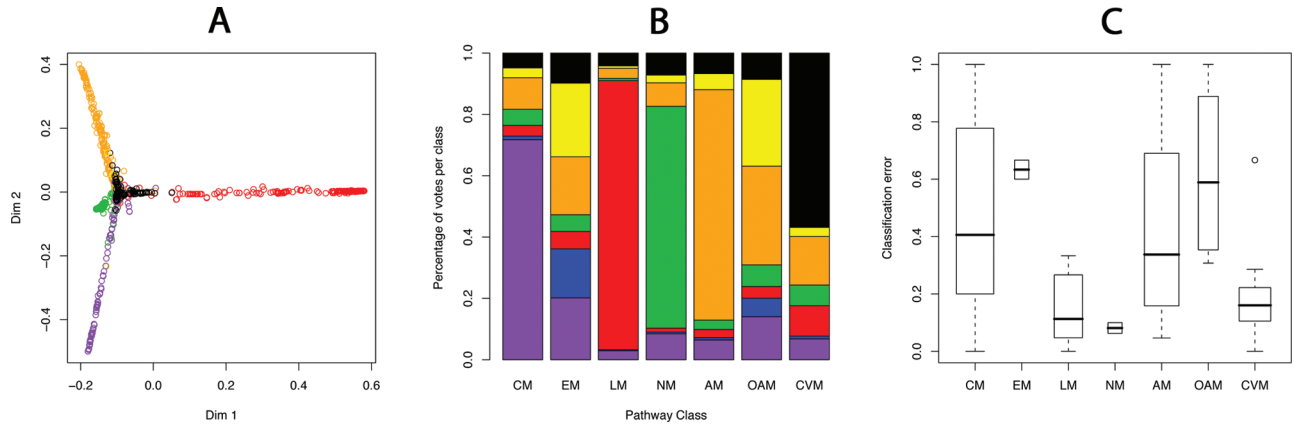

Figure 6. Pathway and pathway class classification of small molecules. **a)** A view of the human metabolites in the *human_unique* data set using multidimensional scaling on the proximity values provided by the random forest trained to assign pathway classes to small molecules. The proximity value for a pair of metabolites is an estimate of how often these two metabolites are found in the same terminal nodes of a tree. Each molecule is represented as a circle in this plot and colored according to its pathway class (purple = CM, blue = EM, red = LM, green = NM, orange = AM, yellow = OAM, black = CVM). This plot shows that pathways in KEGG fall into four major distinct classes (CM, LM, AM, and NM) and that the remaining CVM, EM, and OAM classes strongly overlap with them. **b)** Barplots of the distribution of votes for each pathway class (CM = carbohydrate metabolism, EM = energy met., LM = lipid met., NM = nucleotide met., AM = amino acid met., OAM = other amino acids, CVM = cofactors and vitamins metabolism). Each bar depicts the percentage of votes corresponding to each of the 7 pathway classes (colored using the same scheme as in Figure 6a), as a mean for all molecules in that class. For example, on average, molecules in the carbohydrate class are predicted to be in that class with a 72% probability. Molecules in the carbohydrate, lipid, amino acid, and nucleotide metabolism classes are the ones most often identified correctly as being in their true class. **c)** Classification errors from the confusion matrix of the random forest trained in the classification of individual pathways. Each box-and-whiskers plot groups together the classification errors corresponding to all the pathways within a given class.

Table 2. True and False Positives and Negatives for the Random Forest Classification of Small Molecules into Pathway Classes and Corresponding Matthews Correlation Coefficient for Each Class

class	true positives	false positives	true negatives	false negatives	Matthews correlation coefficient
carbohydrate metabolism	101	19	550	11	0.84
energy metabolism	1	4	669	7	0.15
lipid metabolism	159	2	512	8	0.96
nucleotide metabolism	72	7	596	6	0.91
amino acid metabolism	183	30	458	10	0.86
other amino acid metabolism	14	7	644	16	0.54
cofactors and vitamins metabolism	78	4	584	15	0.88

connecting all classes in a stringlike fashion, reinforcing the view of the metabolome as a continuum.[2] The energy metabolism class is a badly defined class, since many of the substrates in the pathways of this class do not have unique annotations and have been removed from this set (none of the compounds of this class receives more than 50% of the votes from any single class, and so the classification is highly ambiguous). The compounds corresponding to less well-defined classes receive also a smaller percentage of votes from their "true" class, as is obvious from the distributions of class votes in Figure 6b. In particular, the other amino acids class molecules are most of the time assigned to the class of amino acids. These results are also reflected in the Matthews correlation coefficients calculated for the seven classes and presented in Table 2. We have also trained the random forest in unsupervised mode and, unsurprisingly, less structure was obvious in the data, but generally the major types of clusters are still observed (see Figure S6: Multidimensional scaling on the proximity values from an unsupervised run of the random forest classifier). As the Labute descriptors are harder to interpret, we also used all available descriptors in MOE to build a separate random forest classifier. In this case the single most important variable for discrimination of the classes is logP(octanol/water). Related descriptors that capture hydrophobicity/hydrophilicity are also relatively important (SlogP_VSA$_k$), as are variables intended to capture direct electrostatic interactions (PEOE_VSA$_k$), and those intended to capture polarizability (SMR_VSA$_k$). Chirality (descriptor chiral_u) and the number of nitrogens in the molecule also appear to be important for pathway class discrimination using a random forest.

Classification of Small Molecules to Individual Pathways. There are many more pathways than there are classes, and this means that it is more difficult to train a classifier to distinguish between individual pathways, as many comprise relatively few metabolites. The classification errors associated with each pathway (rather than class) in our *human_unique* data set using the 32 Labute descriptors as independent variables are shown in Figure 6c. We observe again that the worst classification errors are for pathways in the other amino acids and energy metabolism classes. However, all classes, except the lipid metabolism and the nucleotide metabolism have at least one pathway with an overall classification error higher than 60%.

Why are some pathways so prone to misclassification? One possible answer is that they simply do not contain enough data for the trees to train on. Some pathways with the worst classification errors are actually ancient pathways at the center of metabolism (pyruvate metabolism, TCA cycle), and these contain many molecules that are also part of other pathways and have thus been removed during the classification step (e.g., the TCA cycle would normally contain 20

molecules in our data set but contains only three in the *human_unique* data set - all three are misclassified). To test whether the small class size is causing the misclassification, we have tested additionally a binary classification. In this case, we consider a single pathway each time and include all metabolites from this pathway (i.e., without removing those with multiple annotations). These form one class, with the second class being formed by all other metabolites not belonging to this pathway. In the TCA cycle example, one class contains the 20 metabolites of the TCA cycle and the other class contains all remaining metabolites. A random forest trained on these data results in poor classification (the class error is 75%), even though class weights are used to ensure that the large imbalance in the data (one class being much smaller than the other) is not affecting the classifier. We conclude that, although the small class size is generally a problem, it is unlikely to be the only reason why some pathways are misclassified.

Another problem is the assumption that metabolites are connected by stepwise metabolic transformations, which is not always the case. For example, in the energy metabolism class, we find the oxidative phosphorylation pathway (hsa00190) comprising molecules as diverse as NADH and succinate, as the reactions involving these molecules are not connected in the linear way the reactions of a molecule's biosynthetic pathway would be connected.

A closer examination of the pathways with the highest metabolite classification errors also reveals a slight tendency toward a larger number of connections to other pathways (as shown in KEGG reference maps) compared with the number of connections for pathways whose metabolites are rarely misclassified. The number of connections tends to vary considerably between pathways; for example, in *E. coli* catabolic pathways tend to be more interconnected whereas biosynthetic pathways tend to be more linear.[35] In our data set, the mean number of connections for the pathways with more than 50% of misclassifications is 7.3, whereas the corresponding mean for the pathways with few (less than 10%) misclassifications is only 3.6 (a Welch two sample *t* test for the two distributions gives a *p* value of 0.02 for the null hypothesis that the two sets of numbers come from distributions with the same mean). There is a simple explanation for this observation: A large number of connections between pathways show that these pathways are well integrated into a network where the product of a reaction of one pathway is used as substrate in another. This means that such pathways also share some molecular scaffolds and are thus more likely to comprise molecules with similar physicochemical properties leading to a higher rate of misclassification of these metabolites. Indeed many of the misclassified molecules are actually classified as members of a pathway that share a connection with the pathway these molecules belong to.

Classifying the Molecules Belonging to Multiple Pathways. As an application of our classifier, we used it to classify the molecules in our data set that have multiple annotations (and which were left out of the training of the classifier). We observe the following: a) As expected, few molecules (24 of 190, 12.6%) with multiple annotations are assigned to a class with a high percentage (more than 80%) of votes. This contrasts with 207 out of 681 (30.4%) molecules being assigned to a class with high confidence in the uniquely annotated data set. b) For 31% of the molecules, the predicted *N* pathways with highest probabilities correspond exactly to the *N* pathways to which this molecule belongs to, and for 77% of these molecules, at least one of the top *N* predicted pathways corresponds to a pathway among the *N* "true" ones. c) If we calculate the difference between the highest and next highest fraction of votes for each compound, we find a pronounced difference between the two distributions for the uniquely and nonuniquely annotated data sets. In about 1/3 of all cases (34%), the difference in the vote percentages for the top two pathways is less than 10% for molecules with multiple annotations. This is true for only 15% of the compounds in the uniquely annotated set. Hence, it may be possible to distinguish between nonuniquely annotated and uniquely annotated compounds based on their physicochemical and topological properties.

Part III: The Relationship between Drugs and Human Metabolites in Chemical Space. 4464 drugs from the DrugBank database flagged as "small molecules" were downloaded and are referred to here as the *all_drugs* data set. After preparation of the structures (see Methods), all 2D descriptors from MOE were calculated, and Principal Component Analysis was applied to the matrix of scaled descriptors for the joint data set of all drugs and all human metabolites (from the *human_all* data set). Figure 7 shows the distribution of these molecules in the space of the first few principal components. Clearly, the two data sets of molecules overlap, even if we separate out drugs with associated KEGG ids in DrugBank (which are likely to be in their majority metabolites themselves) - see Figure 7b,c. This is perhaps not unexpected, given the fact that many old drugs were designed to mimic endogenous metabolites (such as the amino acid - derived neurotransmitters), and many of the newer drugs were designed as analogues of the old drugs.[19] However, these results appear to clash with the results from Karakoc et al.[23] (the only study we are aware of that directly compares human metabolites to a data set of drugs) who showed that human metabolites occupy a distinct part of chemical space and show little overlap with drugs, druglike molecules, antibacterial compounds, and bacterial metabolites (all of which overlap significantly in the same space). There are two possible explanations for the differences observed in the two studies. One is that the data set of human metabolites is significantly different in the two studies, although this is unlikely. The second explanation would be that the descriptors used in the Karakoc study are essentially different than ours. Indeed, although approximately half of their descriptors are identical to our own (both calculated using the MOE software), the other half comprises their own "inductive" descriptors. Using only the conventional 2D descriptors on the Karakoc data set, we can reproduce the extent of the overlap between human metabolites and drugs that we are observing in this study.

Drugs Classified Using the Metabolite Pathway Classifier. The 32 Labute descriptors for the molecules from DrugBank were calculated, and then the random forest trained on the *human_unique* data set was used to predict the pathway most closely resembling each of these drugs. Our first observation is that the majority of the drugs in this data set (3676 of 4464, 82%) are not assigned to any single pathway with more than 50% of the votes. Of the remaining 788, 333 appear to be themselves metabolites (have a KEGG

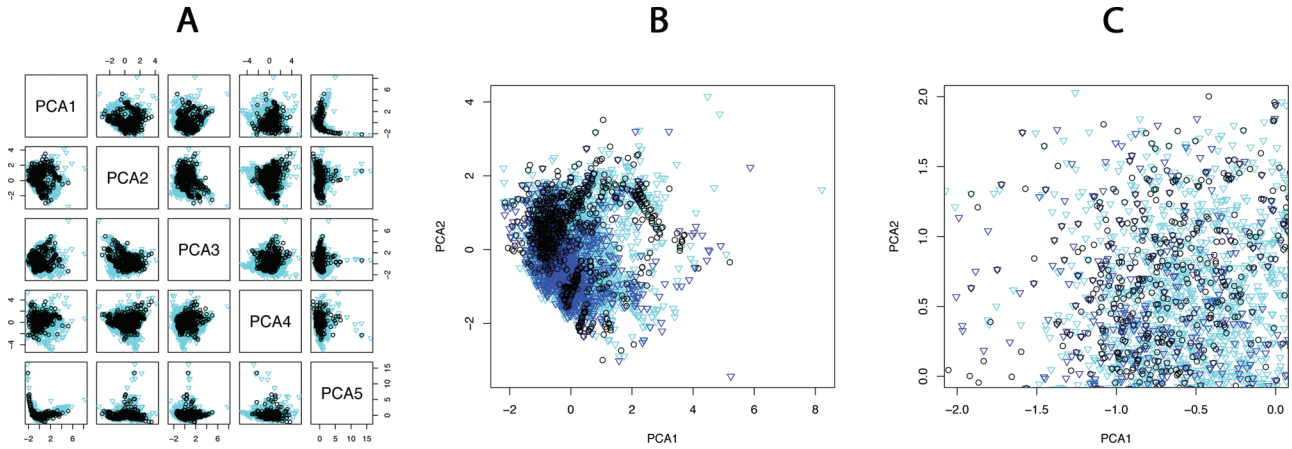

Figure 7. Drugs and human metabolites in chemical space. **a)** Matrix of scatterplots for the first five principal components from PCA on the joint data set of 183 physicochemical descriptors for the *human_all* (black circles) and *all_drugs* (cyan triangles) data sets. The cumulative proportion of variance explained by these five principal components is 0.68, with the first and second components contributing 0.386 and 0.132, respectively. Inulin (drugbank code DB00638) has been excluded from this plot, as it is a major outlier (a polysaccharide of over 6000 g/mol molecular weight). The figure was drawn with *R*. **b)** The PC1 vs PC2 plot of a, but additionally all drugs associated with a KEGG id (1304 molecules) have been annotated as blue triangles (remaining drugs are cyan triangles, and human metabolites are black circles). **c)** Zoomed-in view of a dense area of the plot in b for a better view of the data.

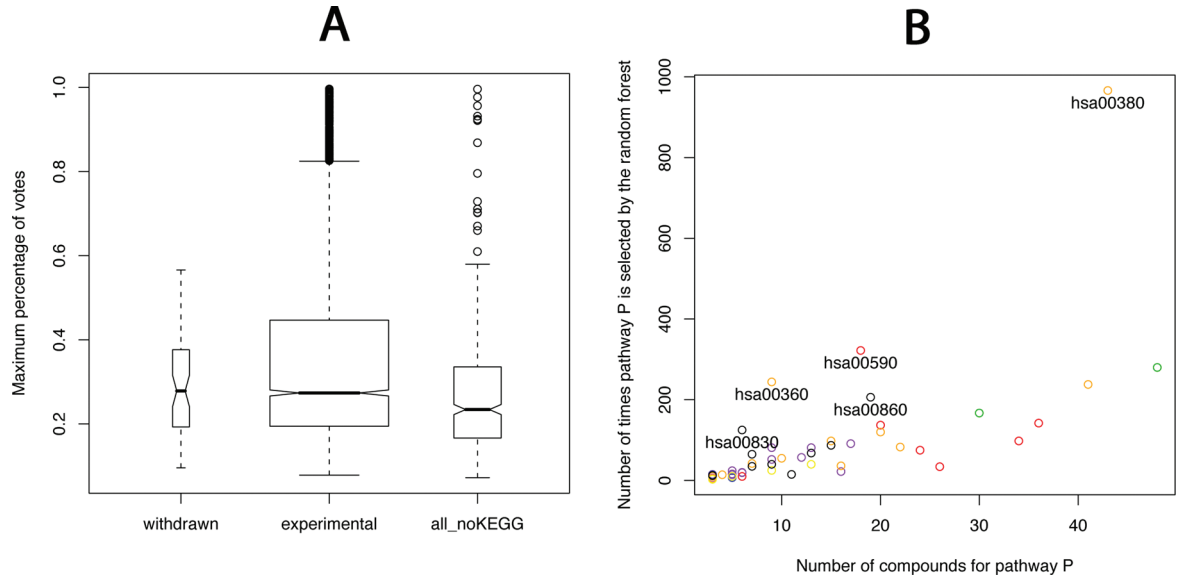

Figure 8. Molecules from DrugBank classified using the random forest trained on human metabolic pathways. **a)** Box-and-whisker-plots of the distribution of the maximum percentage of votes from the random forest classifier for the following subsets of DrugBank molecules: 1) 64 drugs tagged as being withdrawn, 2) 3103 drugs tagged as being experimental, and 3) 1328 drugs with no KEGG id associated with them, and no experimental or withdrawn tags. Boxplots drawn as described in caption of Figure 1. **b)** The relationship between the size of a pathway and how often it is selected by the random forest as being the target of a drug (data shown for all 4464 small molecule drugs). Outliers in this plot are annotated with their KEGG codes: tryptophan metabolism (hsa00380), arachidonic acid metabolism (hsa00590), phenylalanine metabolism (hsa00360), retinol metabolism (hsa00830), and porphyrin/chlorophyll metabolism (hsa00860). The most common targets of the 62 withdrawn drugs (according to our random forest classification) are the tryptophan metabolism (selected 28 times), the arachidonic acid metabolism (selected 7 times), and the phenylalanine and retinol metabolism (each selected 5 times).

id associated with them), and many of the ones with no KEGG id are actually metabolites, but no connection between the databases has been established yet.

A small proportion of drugs in DrugBank are tagged as "withdrawn" from the market (62 of the 4464). One might think that the side effects of at least some of these drugs may be due to serious interference with an off-target metabolic pathway. Figure 8a shows that when classified with our random forest, these drugs do not generally resemble greatly the metabolites in our pathways, and hence their effects are more likely to be due to interference with proteins outside the central metabolic pathways. However, both drugs tagged as withdrawn and those tagged as experimental by DrugBank appear to get larger proportions of the total number of votes for a single pathway, when compared with drugs with no such tags and no KEGG compound ids associated with them. Hence, there is a weak indication that successful marketed drugs tend to overlap less with human metabolic pathways. The individual cases where a strong overlap is observed (more than 60% of the votes toward a single pathway) are, in the majority, cases where the drug either belongs to our data set (despite the fact that DrugBank does not report a KEGG id for it), or where it obviously resembles the metabolites of a pathway (often that pathway being the target pathway of the drug).

Of the 2002 DrugBank molecules with at least one target pathway listed, 1073 have a target pathway in our training data set of 52 and have no associated KEGG compound id. In about 20% (218) of these cases, the pathway with maximum number of votes is the actual target pathway (in about half of these cases (102) the maximum number of votes exceeds 50%). Results are similar for drugs that are associated with KEGG but do not belong to our data set: 21% of these (77 of 361) are predicted to overlap more strongly with their actual target pathways. Predictably, results are much better for drugs belonging to our training data set and having a target pathway also in our data set. In this case, 74% of the time the target pathway is the one with the predicted largest overlap.

As expected, there is a correlation between pathway size and the number of times a pathway is selected by the random forest as being the likely target of a drug (see Figure 8b). However, a few pathways are selected disproportionately often, the most obvious outlier being the tryptophan metabolism (hsa00380), followed by the arachidonic acid (hsa00590), phenylalanine (hsa00360), and retinol metabolism (hsa00830) pathways. These comprise metabolites important in signaling, inflammation, growth, vision, and the immune system, so it is not surprising that drugs are often found in the chemical space of these pathways. As withdrawn drugs are also assigned primarily to these pathways, this is clearly both a risky and rewarding place in chemical space from a pharmaceutical point of view.

Clearly, the pathways that receive a lot of votes may indeed be possible candidates for interference from the action of the drug in question, even if they are not listed as the drug's targets. There is no easy way to prove or disprove such links, and the listed side effects of drugs are too general to be traced back to an individual metabolic pathway. However, other interesting links are sometimes observed between a DrugBank molecule and the pathway it is predicted to overlap with. For example, fumarate (DB01677 in DrugBank, but also a known cognate metabolite, C00122, in KEGG) is classified by our random forest as a member of the nicotinate/nicotinamide metabolism (hsa00760). This molecule belongs indeed to this metabolic pathway, but the enzymes associated with the reactions involving fumarate are at the moment not found in human genes.

DISCUSSION

In this work we have pursued three goals, all based on the idea that projecting pathways onto the chemical space defined by their constituent small molecules can offer an insight into the organization of the metabolic network and its likely interactions with exogenous molecules. The first goal was to visualize and quantify how human metabolic pathways may overlap or be distributed in chemical space, given the structures and physicochemical/topological properties of their constituent molecules. In other words, we looked for ways to project a set of pathways onto chemical space, and we found information from structure and properties to be complementary. Our visual exploration of the human metabolic pathways in chemical space suggests that metabolism covers a diverse range of physicochemical properties and structures, reflected in the relatively nonoverlapping (albeit continuous) parts of space occupied by the different pathway classes. Interestingly, chemical biologists who aim to produce small molecule inhibitors for any known gene are also opting for the methods of diversity-oriented synthesis,[36] a fact that reflects the need to spread in chemical space in order to target an equally diverse protein sequence space. We also found that pathways show very little overlap when average similarities are calculated for their respective constituent molecules, but that most are connected by a few highly similar or identical molecules. It is interesting that this latter observation agrees with earlier observations on the connectivity of molecules in the network (i.e., few molecules are very well connected, whereas the remaining molecules have very few connections[37]), despite the fact that we have not made use of any reaction information but used only physicochemical and topological descriptor information.

Our second goal was to build a statistical model that would allow us to place any molecule in the vicinity of one or more of the pathways in our data set with a given probability. We have employed random forests for this classification and a limited set of carefully chosen descriptors. The success of the classifier varies widely for the different classes of pathways, reflecting both the inherent diverse nature of these pathways as well as the current availability of data to train the classifier on. Carbohydrates, nucleotides, lipids, and amino acid pathways are among the best described, although some variation within each class was also observed. One of the most obvious applications of such a classifier would be in assisting with the identification of the biological role of identified small molecules in metabolomics experiments. A molecule lying in the vicinity of a known pathway may be likely to be part of that pathway, and this information may lead to the discovery of missing enzymes or new reactions in partially characterized pathways.

Another possible application for such a classifier would be in drug development. Indeed, our third goal was to understand the relationships of the collection of known drugs and the human metabolome as a whole and the use of our classifier in unravelling likely connections between drugs and metabolic pathways. Although the more conventional view of the effects of drugs may be that of a single target, the evidence of the polypharmacology of several drugs[38−40] suggests that a more likely scenario would be for a drug to be affecting several targets at once, albeit perhaps with a lesser effect on each one (with some accounting for unwanted side effects). One could argue that a synthetic compound with a weak association to multiple proteins of a pathway will be a lot less detrimental than one with a strong effect on a single enzyme. However, given the built-in redundancy of pathways and the many molecular mechanisms by which cells often seem to overcome the deletion of a single gene, it may be that a drug that affects a whole pathway by occupying a site in chemical space that is common to multiple metabolites of that pathway can have an equally important effect on that pathway, as if it were strongly inhibiting a single enzyme. Similarly, a drug occupying a site in chemical space that is common to metabolites endowed with genomic and nongenomic signaling properties can potentially have its pharmacological outcome explained and/or predicted. Although we have concentrated our study on the collection of known drugs, similar applications of our classifier should be possible for any other synthetic molecules

that may interfere with human health, such as agrochemicals and cosmetic products.

This study, like any other, has several caveats. Ultimately, our observations can only be as accurate as the data set we have constructed. We have already mentioned problems with using the KEGG metabolic pathways as the basis of constructing an organism-specific metabolome. The gene functional assignments are not perfect, and this results in missing or superfluous reactions among the pathways. As our knowledge improves, the reliability of the metabolic network should increase too, but until then a data set derived using our approach will most likely include at least a few metabolites absent from humans and will certainly not cover the whole of the human metabolome. However, we believe that the advent of metabolomics promises a vast improvement of metabolome databases, so that in the future data sets should be almost entirely experimentally verified. We have also mentioned in the Introduction that pathways are artificial constructs, not naturally defined modules, and that we necessarily ignore spatial and temporal constraints, which may interfere with the extent of the predicted pathway overlaps.

The calculation of the between and within-pathway similarity has its own caveats. Here, we took the simple route of calculating similarity as an average across all molecule pairs. Among many alternatives, we could have emphasized similarity or dissimilarity by putting extra weight on the most similar or dissimilar pairs between two pathways. Finding better measures of pathway overlap is a subject for future research.

Another serious caveat in projecting pathways onto the chemical space is a problem that has often hampered QSAR studies in the past and that has been termed by Maggiora[41] as the *lack of invariance of chemical space*. This problem arises from the fact that different descriptors are likely to provide different views of the molecules examined. In QSAR, as here, this would result in molecules being neighbors in one space but lying much further apart in another. Indeed this could be partly responsible for discrepancies we observed between our results and those of an earlier study that used different descriptors, in the extent of overlap between drugs and human metabolites. Using uncorrelated descriptors that cover well the fundamental properties of molecules (the approach we have adopted here) goes a little way toward alleviating this problem, but we believe that rather than seeing this as a problem, we should embrace it as an opportunity to help us better understand which properties might be responsible for bringing two pathways close to each other in chemical space and which might put them apart. This way we might be able to emphasize the specificity of synthetic compounds by choosing their properties as to avoid the regions in chemical space where the two pathways can overlap.

Our random forest classifier showed some promise toward discriminating between different pathways, although its success depended strongly on the type of pathway as well as, naturally, on the availability of data for the given pathway on which the classifier was trained. Alternative classifiers employing different statistical methods may be constructed in the future, but, in our opinion, they are unlikely to perform much better, given the same data. However, we think that there is scope for improvement given a different set of descriptors. Finding which descriptors will be optimal is a challenge in itself, but making use of 3D information and field-based descriptors may go some way toward incorporating information that we have ignored in this study. Such descriptors come with their own problems (e.g., how to pick the conformations that should be considered), but existing protein structure information for enzymes involved in these pathways could be used to guide the process of substrate conformation selection.

Understanding how drugs, and more generally synthetic compounds, may interact with the metabolism network was one of the aims of this study. More specifically, our classifier was an attempt to reveal likely drug-metabolic pathway interactions. The majority of drugs are not assigned with a great probability to a single pathway, as perhaps expected. Indeed, most drugs are not targeting the central metabolic pathways,[42] and hence a classifier covering solely these pathways is not likely to result in hits, as the side effects from such interactions would likely be very serious, an unlikely scenario for a drug that has survived cytotoxicity tests and rigorous clinical trials. Even among withdrawn drugs, few would actually interfere with the pathways in our study, as most would be withdrawn for their effect on receptors and their interference with gene expression regulation. In some cases drugs are meant to target a metabolic pathway that is present in our data set. Although we were able to predict the expected target pathway when the molecule was in our data set, we were only successful 20% of the time, when the molecule was not in our data set. This may sound disappointingly low, but in cases where a drug's targets are unknown, such a result would be of value. In addition, it is important to understand that the low prediction accuracy is not entirely the fault of the classifier. Other factors can play a role, for example, there are many cases where a target pathway is not listed in DrugBank, there is no mechanism of action, or side effects are not reported or not obviously related to a certain pathway. In these cases, even if we were successfully predicting the correct pathway, the prediction would not be counted in the 20% of successes. An example is the antidepressant mianserin (DB06148), which is known to act by blocking, among others, certain serotonin receptors. Our classifier assigns it closest to the tryptophan metabolism (hsa00380), the metabolic pathway comprising the reactions involving the conversion of tryptophan into serotonin. Given the fact that our data set of pathways covers only metabolism and not processes like the uptake of metabolites by receptors, this prediction was as successful as it could have been, yet it was not counted in our success rate. There are some further caveats of course: obviously this study does not take into account the accessibility of a metabolic pathway to a drug, i.e. whether the drug can actually reach the pathway, and there is also no account of metabolism of drugs by enzymes or drug−drug interactions. Additionally, and as noted by one of the reviewers, the "withdrawn" tag from DrugBank does not necessarily mean that the drug was withdrawn worldwide or that it has not been withdrawn simply for patent reasons.

CONCLUSIONS

In this work, we have explored visually and quantitatively the tightness and overlap of human metabolic

pathways in chemical space and have found that metabolism covers a diverse range of physicochemical properties and structures, reflected in the relatively nonoverlapping, but continuous, parts of space occupied by the different pathway classes. Consequently, the success of our random forest pathway classifier for small molecules varies widely for the different classes of pathways, reflecting both the inherent diverse nature of these pathways, as well as the current availability of data to train the classifier on. Finally, we have also found that, using the same classifier, the majority of drugs do not overlap significantly with any individual human metabolic pathway. However, where some overlap is observed, there is clearly a basis for investigating the possible interference of these drugs with the human metabolic network.

We believe the future of this work lies in exploring different possibilities for describing both the chemical space and the networks of reactions. For example, alternative ways of projecting pathways in the chemical space might lead to a better understanding of their cross-reactivity that could be missed, if only a single definition of the chemical space is used. Moreover, the way pathways are defined should be taken into account in the context of the application of the classifier. In the case of supporting metabolomics experiments, for example, it might be much more beneficial to define pathway templates that include all known enzymatic reactions across all species, as this is likely to lead to a more comprehensive coverage of reactions (which would be needed if the metabolites detected are not the outcome of reactions for which organism-specific genes are already known). This approach is likely to maximize the success of placing molecules in the context of pathways, when their origin and role in metabolism is unknown. Perhaps the most important improvement would be to extend the network to include not only metabolic pathways but also any other pathways involving small molecules. This would allow the broadest possible coverage of biological functions of metabolites. One could also consider how small molecules would map into the reaction space, i.e. avoid the use of traditional pathways in favor of groups of reactions that might cluster more naturally in chemical space. This space could additionally be defined using the substructures of the molecules corresponding to the reactive centers. We are currently exploring these possibilities.

ACKNOWLEDGMENT

We are grateful to Professor David Moss for his insightful comments on the calculation of the within-pathway dissimilarity measure.

Financial contribution from the Royal Society in the form of a research grant to I.N. (2005/R2) and a short visit grant to I.N. and A.M. (2006/R1) are gratefully acknowledged.

Supporting Information Available: Network of metabolites from the *human_unique* data set drawn with the Cytoscape software (Figure S1), inositol phosphate metabolism pathway from two different KEGG releases (Figure S2), metabolic pathway overlap using the JCHEM fingerprint similarity (Figure S3), matrix of pairwise maximum similarities between pathways (Figure S4), classification errors from the confusion matrix of the linear discriminant analysis (Figure S5), multidimensional scaling on the proximity values from an unsupervised run of the random forest classifier (Figure S6), and classification errors from using all 2D MOE descriptors and the 32 Labute descriptors (Figure S7). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Hattori, M.; Okuno, Y.; Goto, S.; Kanehisa, M. Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J. Am. Chem. Soc.* **2003**, *125*, 11853–11865.
(2) Nobeli, I.; Ponstingl, H.; Krissinel, E. B.; Thornton, J. M. A structure-based anatomy of the E.coli metabolome. *J. Mol. Biol.* **2003**, *334*, 697–719.
(3) Nobeli, I.; Thornton, J. M. A bioinformatician's view of the metabolome. *Bioessays* **2006**, *28*, 534–545.
(4) Koch, M. A.; Schuffenhauer, A.; Scheck, M.; Wetzel, S.; Casaulta, M.; Odermatt, A.; Ertl, P.; Waldmann, H. Charting biologically relevant chemical space: a structural classification of natural products (SCONP). *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 17272–17277.
(5) Goto, S.; Okuno, Y.; Hattori, M.; Nishioka, T.; Kanehisa, M. LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res.* **2002**, *30*, 402–404.
(6) Karp, P. D.; Arnaud, M.; Collado-Vides, J.; Ingraham, J.; Paulsen, I. T.; Saier, M. H. J.; The, E. coli EcoCyc database: No longer just a metabolic pathway database. *ASM News* **2004**, *70*, 25–30.
(7) Wishart, D. S.; Tzur, D.; Knox, C.; Eisner, R.; Guo, A. C.; Young, N.; Cheng, D.; Jewell, K.; Arndt, D.; Sawhney, S.; et al. : HMDB: the Human Metabolome Database. *Nucleic Acids Res.* **2007**, *35* (Database issue), D521–526.
(8) Brooksbank, C.; Cameron, G.; Thornton, J. M. The European Bioinformatics Institute's data resources: towards systems biology. *Nucleic Acids Res.* **2005**, *33* (Database issue), D46–53.
(9) Oprea, T. I. Chemical space navigation in lead discovery. *Curr. Opin. Chem. Biol.* **2002**, *6*, 384–389.
(10) Dobson, C. M. Chemical space and biology. *Nature* **2004**, *432*, 824–828.
(11) Lipinski, C.; Hopkins, A. Navigating chemical space for biology and medicine. *Nature* **2004**, *432*, 855–861.
(12) Villas-Bôas, S. G. In *Metabolome analysis: an introduction*; Villas-Bôas, S. G., Roessner, U., Hansen, M. A. E., Smedsgaard, J., Nielsen, J. Eds.; Wiley-Interscience: Hoboken, New Jersey, U.S.A., 2007.
(13) Breiman, L. Random forests. *Machine Learn.* **2001**, *45*, 5–32.
(14) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947–1958.
(15) Cannon, E. O.; Bender, A.; Palmer, D. S.; Mitchell, J. B. Chemoinformatics-based classification of prohibited substances employed for doping in sport. *J. Chem. Inf. Model.* **2006**, *46*, 2369–2380.
(16) Ehrman, T. M.; Barlow, D. J.; Hylands, P. J. Virtual screening of Chinese herbs with Random Forest. *J. Chem. Inf. Model.* **2007**, *47*, 264–278.
(17) Gupta, S.; Aires-de-Sousa, J. Comparing the chemical spaces of metabolites and available chemicals: models of metabolite-likeness. *Mol. Diversity* **2007**, *11*, 23–36.
(18) Labute, P. A widely applicable set of descriptors. *J. Mol. Graphics Modell.* **2000**, *18*, 464–477.
(19) Newman, D. J.; Cragg, G. M. Natural products as sources of new drugs over the last 25 years. *J. Nat. Prod.* **2007**, *70*, 461–477.
(20) Stahura, F. L.; Godden, J. W.; Xue, L.; Bajorath, J. Distinguishing between natural products and synthetic molecules by descriptor Shannon entropy analysis and binary QSAR calculations. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1245–1252.
(21) Ertl, P.; Schuffenhauer, A. Cheminformatics analysis of natural products: lessons from nature inspiring the design of new drugs. *Prog. Drug. Res.* **2008**, *66*, 219–235.
(22) Feher, M.; Schmidt, J. M. Property distributions: differences between drugs, natural products, and molecules from combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 218–227.
(23) Karakoc, E.; Sahinalp, S. C.; Cherkasov, A. Comparative QSAR- and Fragments Distribution Analysis of Drugs, Druglikes, Metabolic Substances, and Antimicrobial Compounds. *J. Chem. Inf. Model.* **2006**, *46*, 2167–2182.
(24) Lee, M. L.; Schneider, G. Scaffold architecture and pharmacophoric properties of natural products and trade drugs: application in the design of natural product-based combinatorial libraries. *J. Comb. Chem.* **2001**, *3*, 284–289.

(25) Henkel, T.; Brunne, R. M.; Muller, H.; Reichel, F. Statistical investigation into the structural complementarity of natural products and synthetic compounds. *Angew. Chem., Int. Ed. Engl.* **1999**, *38*, 643–647.
(26) Hata, K.; Koseki, K.; Yamaguchi, K.; Moriya, S.; Suzuki, Y.; Yingsakmongkon, S.; Hirai, G.; Sodeoka, M.; von Itzstein, M.; Miyagi, T. Limited inhibitory effects of oseltamivir and zanamivir on human sialidases. *Antimicrob. Agents Chemother.* **2008**, *52*, 3484–3491.
(27) Marcoff, L.; Thompson, P. D. The role of coenzyme Q10 in statin-associated myopathy: a systematic review. *J. Am. Coll. Cardiol.* **2007**, *49*, 2231–2237.
(28) Andrade, M. A.; Ouzounis, C.; Sander, C.; Tamames, J.; Valencia, A. Functional classes in the three domains of life. *J. Mol. Evol.* **1999**, *49*, 551–557.
(29) Cases, I.; de Lorenzo, V.; Ouzounis, C. A. Transcription regulation and environmental adaptation in bacteria. *Trends Microbiol.* **2003**, *11*, 248–253.
(30) Kanehisa, M.; Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **2000**, *28*, 27–30.
(31) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **2003**, *13*, 2498–2504.
(32) Gutteridge, A.; Kanehisa, M.; Goto, S. Regulation of metabolic networks by small molecule metabolites. *BMC Bioinf.* **2007**, *8*, 88.
(33) Hopkins, A. L.; Groom, C. R. The druggable genome. *Nat. Rev. Drug Discovery* **2002**, *1*, 727–730.
(34) Horovitz, N. H. On the evolution of biochemical synthesis. *Proc. Natl. Acad. Sci. U.S.A.* **1945**, *31*, 153–157.
(35) Kim, J.; Copley, S. D. Why metabolic enzymes are essential or nonessential for growth of Escherichia coli K12 on glucose. *Biochemistry* **2007**, *46*, 12501–12511.
(36) Schreiber, S. L. Target-oriented and diversity-oriented organic synthesis in drug discovery. *Science* **2000**, *287*, 1964–1969.
(37) Jeong, H.; Tombor, B.; Albert, R.; Oltvai, Z. N.; Barabasi, A. L. The large-scale organization of metabolic networks. *Nature* **2000**, *407*, 651–654.
(38) Hopkins, A. L.; Mason, J. S.; Overington, J. P. Can we rationally design promiscuous drugs. *Curr. Opin. Struct. Biol.* **2006**, *16*, 127–136.
(39) Paolini, G. V.; Shapland, R. H.; van Hoorn, W. P.; Mason, J. S.; Hopkins, A. L. Global mapping of pharmacological space. *Nat. Biotechnol.* **2006**, *24*, 805–815.
(40) Yildirim, M. A.; Goh, K. I.; Cusick, M. E.; Barabasi, A. L.; Vidal, M. Drug-target network. *Nat. Biotechnol.* **2007**, *25*, 1119–1126.
(41) Maggiora, G. M. On outliers and activity cliffs--why QSAR often disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535.
(42) Robertson, J. G. Mechanistic basis of enzyme-targeted drugs. *Biochemistry* **2005**, *44*, 5561–5571.
(43) Ihlenfeldt, W. D.; Takahashi, H.; Abe, S.; Sasaki, J. Computation and management of chemical properties in CACTVS: An extensible networked approach toward modularity and flexibility. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 109–116.
(44) Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* **2008**, *36*, D901–906.
(45) R Development Core Team. *R: A language and environment for statistical computing*; R Development Core Team: Vienna, Austria, 2005.
(46) Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493–500.
(47) Kochev, N.; Monev, V.; Bangov, I. Searching Chemical Structures. In *Chemoinformatics*; Gasteiger, J., Engel, T., Eds.; Wiley-VCH: Weinheim, 2003; pp 291−318.

CI900196U