A comparison of the chemical properties of drugs and FEMA/FDA notified GRAS chemical compounds used in the food industry

D.G. Sprous *, F.R. Salemme

Redpoint Bio, 7 Graphics Drive, Ewing, NJ 08628, USA

Abstract

The range of molecular properties of generally recognized as safe (GRAS) compounds that are typically used in food and beverage products is compared to marketed drugs. It is observed that GRAS compounds differ from marketed drugs with respect to several molecular descriptors, including molecular weight, H-bond acceptor count, H-bond donor count, aromatic ring count, basic group count, acidic group count, molecular flexibility, enhanced ether and ester bearing compound populations, and reduced nitrogen and halogen bearing compound populations. It was observed that log(P) and log(S), which provide computed estimates of compound solubility in organic and aqueous solvents, respectively, have significant overlap in the two populations. On the whole, GRAS compounds are seen to be more flexible, smaller, and composed of a more restricted set of elements than marketed drugs. In addition, a multivariable binary quantitative structure–activity relationship (QSAR) model incorporating the parameters listed above can distinguish GRAS and pharmaceutical compounds, further strengthening the distinction between the molecular spaces defining GRAS and pharmaceutical compounds. It is speculated that the difference between the GRAS and pharmaceutical property spaces is a result of the historical legacy of most GRAS compounds, which are primarily natural in origin. Compounds more recently added to the GRAS list appear more similar to pharmaceuticals with respect to their chemical properties.

Keywords: GRAS; Drug likeness; QSAR; Molecular properties

1. Introduction

"Generally Recognized as Safe" (GRAS) is the food ingredient category established by the 1958 Food Additive Amendment to the US Federal Food, Drug and Cosmetic Act which applies to natural or artificial chemicals that can be added to food and beverage products within a defined level of usage (Burdock, 2003; Hallagan and Hall, 1995; Smith et al., 2005; Woods and Doull, 1991). As will be discussed below, entry into the GRAS compound space is by four paths:

(1) compounds that entered by the FEMA GRAS approval process since 1958,
(2) compounds that entered by direct FDA notification by an individual corporation or group since 1958,
(3) compounds "grandfathered" into the regulation by virtue of widespread US use prior to 1958 and
(4) compounds determined GRAS by an individual group and entered into use without FDA notification.

The GRAS category was created to complement the slightly older "food additive" category, which requires an extensive safety assessment, FDA review and approval process. At the time, the "food additive" category was itself new and poised to be applied to all food ingredients then currently in use. This action would have created a severe disruption for the entire food industry. Consequently, it was

generally agreed that it would not be necessary to apply a full FDA-sanctioned safety study to ingredients already in use for decades. Instead, Congress created the generally recognized as safe (GRAS) exception to food additives. Ingredients were recognized as GRAS if they had been in use prior to Jan 1, 1958. ''Use prior to January 1, 1958'' had to be as a food ingredient, not as a drug or herbal remedy, and use must have been widespread enough to provide assurance that the compound did not induce any rare allergies or metabolic intolerances (Burdock, 2003). There is no favoritism in the law towards natural occurring substances.

New compounds can gain GRAS approval through a review of scientific material by an expert panel qualified by training and experience to determine food safety. Any corporation or group can convene an expert panel, determine GRAS status and notify the FDA through established published channels. FDA notification is not required by statute but is common for new GRAS materials. A survey of the FDA website shows that 198 GRAS notification letters were submitted since 1998. Ten submissions from 2006 are pending, as are five of twenty from 2005 and 1 from 2004. Seventeen submissions prior to 2005 have been met with a ''Notice does not provide a basis for a GRAS determination'' and details of the reason why are provided as is required by law. Twenty-six have been removed by the submitter's request. The remaining notifications have been met with the reply ''FDA has no questions''. Often, the FDA will make suggestions for package labeling in the ''no questions'' letter.

A significant fraction of entries to the GRAS list were reviewed by the Flavor and Extract Manufacturers Association (FEMA) Expert Panel which has operated since the early 1960s (Burdock, 2003; Hallagan and Hall, 1995; Munro et al., 1998; Smith et al., 2005; Woods and Doull, 1991). Data that the FEMA expert panel require to evaluate potential GRAS status are quite standardized and typically include data on metabolic fate, a repeat dose toxicology study (e.g., a 90-day rat study), and tests for reproductive effects for the a new compound. At some point, dosing of animals is performed at an elevated level to demonstrate a safety margin relative to the consumption level anticipated in any marketed food product. However, the FEMA panel 'eschews the rote approach of ''cookbook toxicology''' (Hallagan and Hall, 1995). In the case that a new compound has strong similarity to existent known safe compounds, this process can be reduced and heavy reliance is made on conservative decision trees (Munro et al., 1998). Ideally, the data upon which GRAS designation is based should be publicly known and made available in a published journal. Certification requires that the reviewing expert panel agree by consensus that the compound is GRAS and unanimously that the compound is safe. Although FDA notification is not formally required, the FDA together with interested scientific and professional public at large, is aware of the whole process. Indeed, Hallagan and Hall (1995) remark of the whole process that: ''Most importantly, all of this information is provided to the FDA so that the agency has the opportunity to challenge the GRAS status of flavour ingredients as determined by the FEMA expert panel''.

Pharmaceuticals represent a second class of natural and novel molecules to which humans are exposed at significant levels. In contrast to GRAS compounds, a large, nearly decade-old literature genre exists that deals with the definition of ''classical'' pharmaceutical molecular properties (Blake, 2000; Blake, 2003; Ertle et al., 2000; Lipinski, 2003; Oprea and Gottfries, 2000; Walters et al., 1999) and defines ''druglike'' characteristics in terms of property ranges for specific molecular descriptors. These property distributions have subsequently found use both as guides for synthetic medicinal chemists and as the basis for designing commercially marketed drug screening libraries (see websites for Asinex,ChemDiv,ChemBridge for examples of how different companies approach this issue). However, owing to the vast size of chemical diversity space, by some estimates believed to encompass some 10^{85} molecules with classical ''drug-like'' properties, there is ample room for either fine-grained sub-structure in chemical structure space or excursions outside the property space represented by currently marketed drugs. This makes it unwise to assume that the present bounds of pharmaceutical property space are fixed. Indeed, it has been shown that compounds that inhibit specific drug target families, such as G-protein coupled receptors (GPCR) (Jimonet and Jager, 2004) and kinases (Briem and Gunther, 2005; Ford et al., 2004; Prien, 2005; Sprous et al., 2005) have property distributions that are distinct from presently marketed pharmaceuticals and in some cases represent distinct expansions in drug property space. A practical consequence of this observation is that focused libraries can be designed that specifically target these protein families (Jimonet and Jager, 2004; Prien, 2005). Screening these targeted libraries for biological activity produces a significantly higher fraction of active compounds than libraries that simply maximize chemical diversity.

There is substantial similarity in the safety studies required to obtain a FEMA GRAS designation and the pre-IND toxicology studies required by the FDA to qualify a new drug for Phase I human clinical trials. However, pharmaceuticals must not only be safe (or at least have an acceptable risk/benefit ratio), but be therapeutically effective as judged by the outcome of an extensive clinical trials process. For a new drug there is the perception of some level of acceptable risk that is largely absent for a GRAS compound added to food or beverage products. By contrast to the wealth of analysis and classification performed on pharmaceutical compounds, there has been no systematic evaluation of the GRAS list compounds as a chemical compound population. Given the similarities of the determination of safety, but differences in use and acceptance of risk, we felt it useful to explore potential differences in chemical properties between pharmaceuticals and GRAS compounds.

We were only able to study a restricted population of GRAS compounds, since only those compounds that

entered into GRAS status by routes 1 and 2 above (via FDA notification or FEMA GRAS expert panel GRAS affirmation) are readily defined as an obtainable group. There is no means of gathering a list of compounds given GRAS status by private concerns that skip the optional FDA notification protocol. Likewise, compounds in common use prior to 1958 are ipso facto GRAS but are difficult to define (although this might be interesting for a future study). Hence, the present study is based on a the use of a subset of GRAS chemistry, comprising a union of the FEMA GRAS list and list of compounds which passed through the FDA notification process. Although not complete, this union represents a coherent compound list that occupies a well defined regulatory niche. The present study was largely stimulated by two questions: "Are GRAS compounds distinct as a chemical compound population?" and "Assuming GRAS compounds are distinct, how do they compare as a population to pharmaceuticals?". As we performed the analysis, we developed a third question: "Is GRAS (specifically the FEMA GRAS/FDA notification GRAS union) chemical space changing with time or is it constant?". The present paper will present the datasets, descriptors and analysis techniques we applied which have permitted us to answer that GRAS compounds are a distinct population, easily recognized from pharmaceuticals. However, we believe that this is mainly an artifact reflecting the historical predominance of nature-identical compounds used in food products, and that more recently developed GRAS compounds are populating an expanding and changing chemical space.

2. Materials and methods

The materials and methods in this study are the molecular databases employed, the descriptors calculated for those datasets and the statistical analysis done to quantify the differences between the two datasets.

2.1. Databases preparation

2.1.1. GRAS1882

The GRAS database was obtained from the Flavor-Base software package (Flavor-Base, 2004) and includes all entries up to the GRAS21 release. The GRAS22 release was reserved for additional analysis. The database provided by Flavor-Base is a list of names, some chemical, some common and some origin focused, with associated data concerning natural origin, commercial suppliers and notes concerning flavour. The database contains both small molecules and proteins, with no distinctions between the two. Our first task in this project was to convert the chemical names into 2D SMILE strings (Weininger, 1988), in order to create computer-readable descriptions of the GRAS chemical structures. This was accomplished by use of the LexiChem software (LexiChem, 2006). The program had the desired quality of being "brittle" when confronted with syntax errors in compound names, which facilitated their manual correction and consequent correct structure generation. Some alternative programs that we tested would "successfully" return incorrect SMILEs strings from names with syntax errors. The largest sources of failure to return a 2D SMILE description were from syntax errors such as missing brackets or misspelling in the names themselves. After correcting these errors, we obtained a set of defined small molecules that could be more-or-less automatically derived from the GRAS list that contains 1882 compounds from a total Flavor-Base GRAS list of approximately 3000 compounds. Although additional compounds could potentially be recovered using essentially manual conversion methods and/or literature references, we restricted the present study to the 1882 compounds whose SMILE descriptions could be computationally derived. Henceforth, in this paper, the dataset for which we obtained structures will be called the GRAS1882 set and is inclusive of the FEMA/FDA GRAS set up to GRAS21.

In addition to GRAS1882 that spans up to GRAS21, the most recent compounds adopted into the FEMA/FDA GRAS regulatory mechanism are available as the GRAS22 set. The GRAS22 dataset comprised ~170 compounds. This set was reserved from the GRAS1882 dataset, allowing us to address questions about the differences between recent GRAS entries and historical GRAS entries. This subset of GRAS entries is henceforth referred to as GRAS22 in the present paper.

2.1.2. Marketed drug dataset [MDD1120]

The Prestwick Chemical Library (Prestwick, 2005) is comprised of 1120 off patent drugs formatted for screening. This dataset was used to define "drug space". Where GRAS1882 required several days to assemble, the MDD1120 set was provided in a simple SDF format.

2.2. Descriptor calculation

Operations associated with descriptor calculation were done within Molecular Operating Environment (MOE) (MOE, 2005). Prior to descriptor calculation, all compounds were standardized by the following steps: the MOE "Wash" utility was employed to remove co-ions and spurious matter; all compounds were provided explicit hydrogens; all acid groups were rendered as unprotonated and all basic groups were rendered as protonated. The descriptors calculated and employed in the QSAR model generation described below are listed in Table 1. In Fig. 1, the populations of the two datasets are presented for each of these descriptors and three additional parameters (nitrogen atom count, ether count, ester count). Most of the descriptors employed have been used frequently in modeling drug-like properties or oral bioavailability. Indeed, several of the major suppliers of chemical compounds for drug screening provide statistics for their libraries using these parameters and ship electronic catalogs that include these parameters for each compound (see websites listed in citations for Asinex, ChemBridge and ChemDiv for examples).

Several types and combinations of descriptors were evaluated in the discrimination models to establish a minimal set that could be employed that provided the best classification of compounds, resulting in the final set listed in Table 1.

2.3. Analysis

Population comparisons were performed from within Excel with VBA (visual basic) scripts calculating populations as a function of specific descriptors. To facilitate comparison between the two data sets which contain different total numbers of compounds, all populations were

Table 1
Descriptor definitions

Name	Description	MOE Tag	SMILES
MW	Molecular weight	Weight	
Flexibility	Rotable bonds normalized by total number of bonds	b_1rotN	
log(P)	Labute log(P) model	LogP (o/w)	
log(S)	Labute solubility model	logS	
Acceptors	H-bond acceptor count	a_acc	
Donors	H-bond donor count	a_don	
Acidic atoms	Acidic atom count	a_acid	
Basic atoms	Basic atom count	a_base	
Halogen	Halogen atom count		F, Cl, Br or I
Aromatic atoms	Aromatic atom count		a

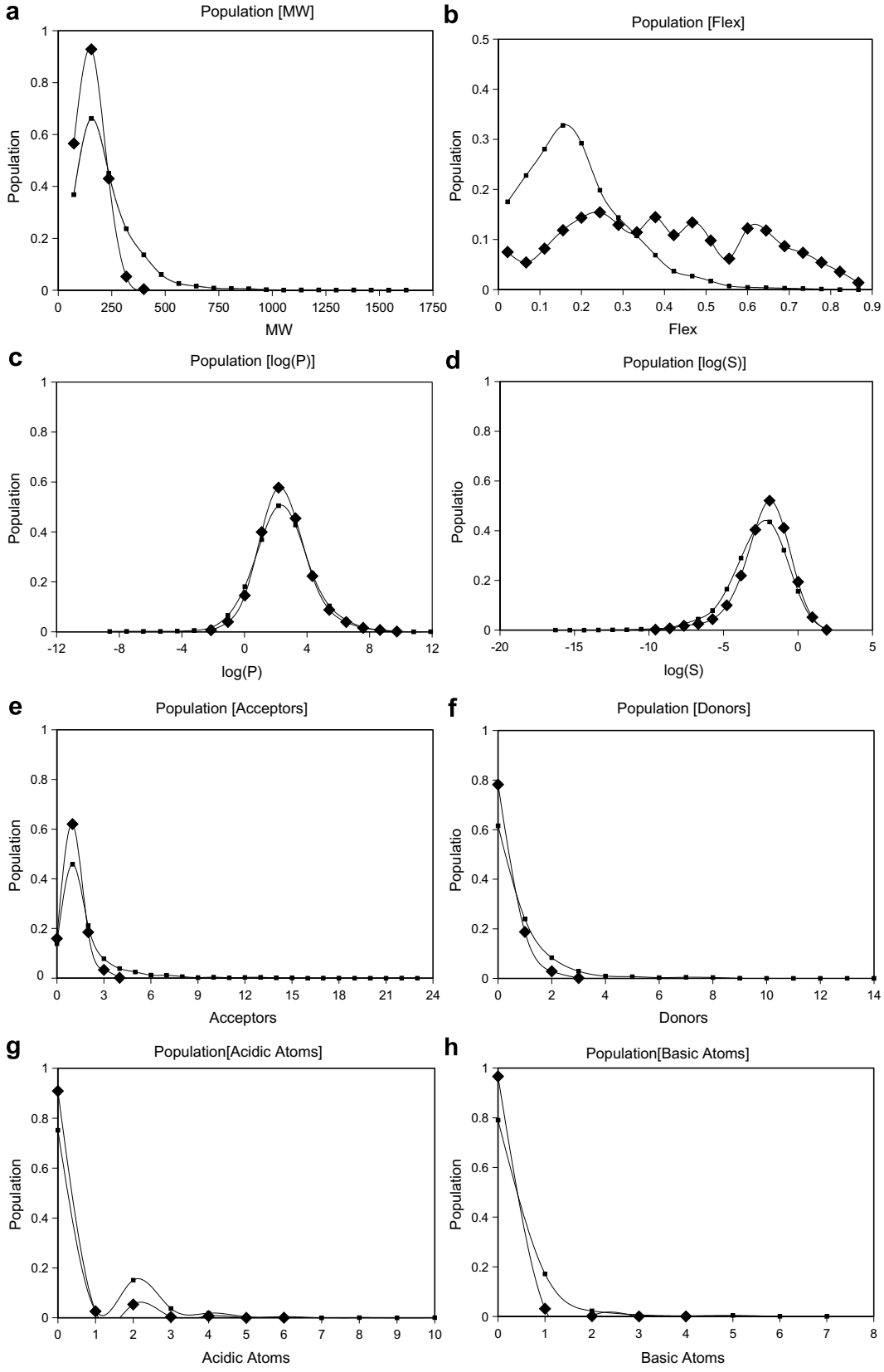

Fig. 1. Population as a function of various metrics for the GRAS dataset (large diamonds) and the Prestwick marketed drug dataset (small squares). (a) Population as a function of molecular weight. (b) Population as a function of Flex. (c) Population as a function of log(P). (d) Population as a function of log(S). (e) Population as a function of H-bond acceptor count. (f) Population as a function of H-bond donor count. (g) Population as a function of acidic group count. (h) Population as a function of basic group count. (i) Population as a function of halogen count. (j) Population as a function of aromatic atom count. (k) Population as a function of nitrogen atom count. (l) Population as a function of ester. (m) Population as a function of ethers.

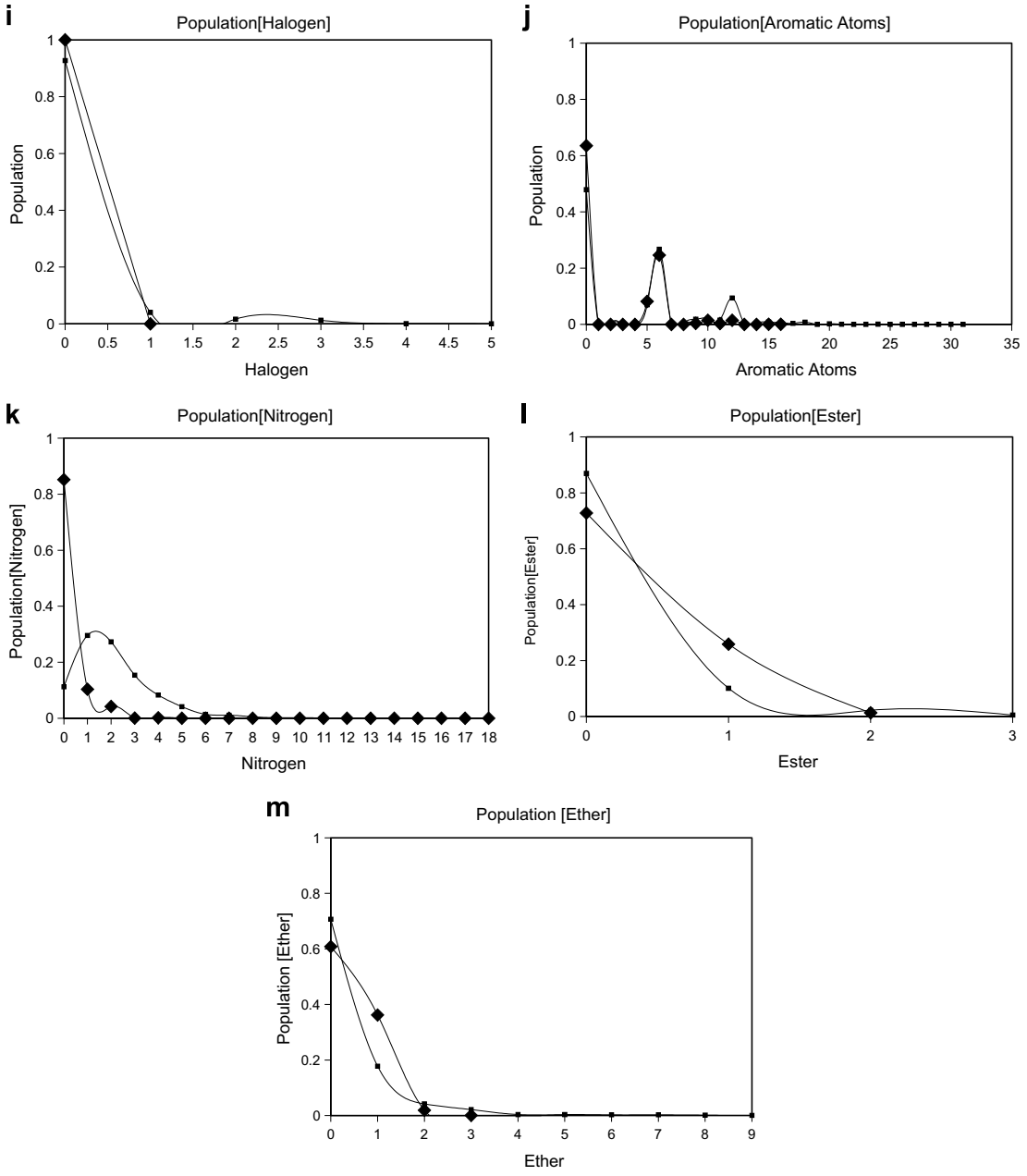

Fig. 1 (*continued*)

expressed as a fraction of the total population, such that the maximum population for either GRAS1882 or MDD1120 is 1.0. For graphs involving integers as the independent variable (e.g., number of H-Bond acceptors), the increment was simply one. For the property graphs that involved continuously variable properties [e.g., molecular weight or log(P)], a common range was found for both the GRAS1882 and MDD1120 datasets and a common increment i of 1/20 of said range was employed. The population reported at a specific point p is the population found between $p - 0.5 * \mathrm{i}$ to $p + 0.5 * i$ along the axis defined by the specific descriptor. Distribution visualization compliments the binary QSAR method as will be discussed below.

2.4. Binary QSAR (quantitative structure–activity relationship)

In recent years, drug researchers developing methods to recognize "GPCR-inhibitor-like" and "Kinase-inhibitor-like" compounds from more general HTS-screening libraries have had a discrimination question that is similar to the present case; namely, how do we develop a computational model that distinguishes between two sets of molecules with different activities, or an active set of molecules from an inactive set? Models have been developed to recognize both kinase inhibitors and GPCR inhibitors using neural nets, machine learning and partial least squares (PLS) (Briem and Gunther, 2005; Ford et al., 2004). All of these methods perform well, presumably reflecting genuine chemical differences between these different kinds of protein binding ligands. In a previous study by Sprous et al. (2005) of protein kinase inhibitors, it was seen that PLS can be applied to binary data provided some checks for boundary conditions are employed. However, the general use of PLS is to models where dependent variables (e.g., inhibitory activity) span 3–5 log units (see chapter 12.12 in Leach, 1996 for an introduction and reference to these issues). Binary QSAR models, in contrast, are better suited for basic classification questions such as the discrimination of GRAS compounds from drugs.

The Binary QSAR method of LaBute (1999) is available as a component of the MOE package. Binary QSAR was initially developed in the MOE context for modeling high throughput screening (HTS) data that is generally produces a binary pass/fail classification of tested compounds whose data structure is essentially identical to the present GRAS/drug discrimination problem. The fundamental equation for binary QSAR was presented by LaBute (1999) and is shown below:

$$p(x) = 1 \Big/ \left[1 + \frac{m(I)+1}{m(A)+1} \prod_{j=1}^{\text{Nbins}} \frac{F(x,j,I)}{F(x,j,A)}\right] \quad (1)$$

$p(x)$ is the final reported probability that molecule x is active. This is dependent on the distributions for active ($F(x,j,A)$) and inactive ($F(x,j,I)$) molecules with respect to specific descriptors. The distribution function employed is:

$$F = 0.5 * \sum_{k=1}^{\text{Nbins}} \frac{P(k)+1/c}{(c+\text{Nbins}/c)} [g(k,) - g(k-1)] \quad (2)$$

which is dependent on the following terms:

$$g(k) = \text{erf}\left(\frac{P(k)-z}{\sigma\sqrt{2}}\right) \quad (3)$$

$P(k)$ is the population at some specific bin k, σ is the variance of the distribution across the specific descriptor of interest, z is the total population across all Nbins and c is a constant that is solved as part of the regression process. In Eq. (1), the multiplication of the product term by the ratio of actives ($m(A)+1$) and inactives ($m(I)+1$) in the training set over which the model is developed ensures that the result ranges between 0 and 1. Being a product, one distribution about a descriptor that shows no population at that value of the descriptor is captured easily by virtue that multiplication by zero is zero. In contrast, a method based on sums of the product of solved coefficients and descriptors would have trouble capturing this situation. Typically in linear regression models (whether PLS or LLS), a final equation is presented showing the solved coefficients. For Eq. (1), the equivalent would be a large table of individual populations per bin per specific descriptor, with the number of bins themselves changing from descriptor to descriptor. This particular report would be difficult to digest but the visualization is easily accomplished by use of the distribution figures discussed previously in Section 2.4.

Regardless of the technique employed, any modeling based on training needs to be challenged using a cross-validation process, where the accuracy of the model is determined for points that *were not used* to develop the model in the first place. Leave-one-out (LOO) cross-validation is typically applied to small *nonredundant* datasets. In LOO cross validation, each compound is removed from a training set, a model is developed, the properties of the excluded compound are predicted and recorded, and the process is iterated for each dataset member. However, with datasets of more than 50 or 100 members, each entered compound has a reasonable chance of having one or more compounds present in the dataset that are essentially duplicates, making LOO cross validation unsuitable for large dataset (see Clark et al., 2001 and references therein). In the present study, to demonstrate prediction accuracy for compounds exterior to the training set, the authors employed systematically larger ratios of reserve test sets to training sets based on random partitions of the combined database. Specifically, each compound was assigned a random number from 0 to 1.0 and the training set was defined as being those compounds with random numbers above some specified threshold. The binary QSAR is capable of predicting set membership accurately (90%) for the reserve test sets even when the test set is more than 95% of the possible total. While this is a product of binary QSAR technical strength and the appropriateness of the descriptors employed, it must be stressed that it is best described as a signature of the two sets being so different that devising a recognition function is trivial.

3. Results

3.1. Populations as functions of descriptors

Fig. 1a–m visually present the differences between the GRAS1882 and MDD1120 datasets. Fig. 1a presents Population [MW] for both GRAS1882 and MDD1120. The peak for both GRAS1882 and MDD1120 is at 150 au for both populations (increment was 75 au). However, the peak for GRAS1882 constitutes 90% of the population where for MDD1120 it is 60%. Thereafter, the remaining 10% population is found prior to reaching 325 au. By contrast, the MDD1120 set has significant population well past 400 au and numerous members above 750 au. Flexibility in Fig. 1b is the number of rotatable bonds divided by the number of bonds. This descriptor ranges from 0 for a compound with no degrees of freedom to 1.0 for a compound that has the maximum possible degrees of freedom for that many bonds. As can be seen in Fig. 1b, GRAS1882 shows a much greater difference in Flexibility ranges compared to MDD1120, and has more members that have Flexibility greater than 0.40. The values of log(P) (log of octanol/water partition coefficient) and log(S) (log of solubility) are frequently parameters of interest in evaluating bioavailability. As can be seen in Fig. 1c and d, the populations are quite close, but differ in that MDD1120 has a greater number of outliers outside where GRAS1882 has no outliers beyond two standard deviations for both log(P) and log(S). The quartet of MW, Flexibility, log(P) and log(S) (Fig. 1a–d) collectively suggest that GRAS1882 compounds are often noticeably smaller and more flexible than MDD1120 compounds, but not different in bulk solubility properties.

In Fig. 1e and f, populations as functions of hydrogen bond acceptors and donors are presented. In both cases, the range for acceptors and donors for MDD1120 extends considerably farther beyond the maximum encountered for the GRAS1882 dataset. GRAS1882 has noticeably higher single peaks for both acceptors and donors than those seen in MDD1120, although both datasets have their peaks at essentially the same place.

Population as a function of acidic atoms shows a major peak at 0 count and minor peak at 2 count (Fig. 1g). However, GRAS1882 is about 10% higher in the first peak and 10% lower in the second peak as compared to MDD1120. The population as a function of basic atoms is simpler (Fig. 1h): GRAS1882 has >95% of its population at 0 count and the remainder is found at 2 and 3 counts, respectively. In contrast, MDD1120 has about 79% at 0 count and the remainder spread from 0 to 7 counts.

Analysis of element, element per hybridization and functional group population differences proved to be problematic since sets can frequently be overlapping and correlated (e.g., oxygen atoms count versus alcohols). We are presenting data associated with halogens, aromatic atoms, total nitrogen counts, ester and ether counts. The populations as a function of halogens can be summarized as that

GRAS1882 essentially has no halogens while MDD1120 has halogens rarely (Fig. 1i). Population as a function of aromatic atom counts shows GRAS1882 to have as much population with a single aromatic ring (Fig. 1j, the hump at 5 or 6 atoms) but to be less aromatic in content overall and lacking any counts above 15 atoms (Fig. 1j, differences in peaks at 11 and 12 atoms and 15–18). We looked at total nitrogen atom counts, nitrogen counts per $N.sp^3$, $N.sp^2$, $N.sp^1$ and N.ar, primary, secondary and tertiary amines. At all variants of nitrogen we saw little present in GRAS but common occurrences in pharmaceuticals. For this reason, we present total nitrogen count in Fig. 1k to present the nitrogen rarity in GRAS and nitrogen ubiquity in pharmaceuticals. More than 90% of the GRAS population *has no* nitrogen where 90% *has* nitrogen in pharmaceuticals. Integration over counts at 1 and 2 nitrogen atoms reaches approximately 60% of the pharmaceuticals leaving a sizable population with even more nitrogen counts. We looked at total oxygen atom counts, oxygen counts per $O.sp^3$, $O.sp^2$ and O.ar, alchohol, aldehyde, ketone, ester and ether counts. After, careful review of structures and population analysis, we concluded that the most accurate presentation of the phenomena overall are done by presenting ester (Fig. 1l) and ether (Fig. 1m) counts. Overall, GRAS compounds have a noticeably larger portion of esters and ethers than pharmaceuticals and those functional groups occur *once* per molecule. The larger pharmaceuticals have few members with these groups overall but have *some* members with multiple occurrences per molecule. Alcohols, ketones and aldehydes are equally common to both GRAS and pharmaceutical compounds (data not shown).

3.2. Binary QSAR model

As was detailed in the Methods section, the critical question for a QSAR model is how well can it predict beyond its training set. The partitions between training and test sets are presented in Table 2 with the accuracies seen in testing the models. The training set members are all compounds with a random number above the specified "Rand" (threshold random number). Hence, the first entry in Table 2 has no members exterior to the training set while the second line has half the members in the training set. For the case where the threshold is such that the training set population is the entire set, the binary QSAR model shows a baseline accuracy of 94% for recognizing GRAS1882 members and an accuracy of 92% for recognizing MDD1120 members. As can be seen at 50% and 95% levels, there is little change of predictivity when applied to the test set. Only at the most aggressive partition where the test set constitutes 99% of the whole do we see loss of predictivity.

The GRAS22 dataset, the list of the most recent compounds accepted into the GRAS regulatory mechanism, was not incorporated into the GRAS1882 dataset. Rather, GRAS22 was kept independent to address questions of how GRAS chemical space may be changing with time. The binary QSAR models discussed above was applied to score each member of the GRAS22 dataset. This revealed a false negative rate of 13%. By comparison, the GRAS1882 dataset had false negative rates of no greater than 6% for the training sets and of ~7% within the reserve test sets at partitions of 0.50 and 0.95, respectively (see Table 2). Visual inspection showed that the compounds failing to be recognized as GRAS to be less flexible, richer in aromatic groups and larger than expected from a strict interpretation of Fig. 1's distributions. Typically, the false negatives had 11 or 12 aromatic atoms, which is more common in pharmaceuticals than GRAS compounds (see Fig. 1j). In the GRAS22 dataset, approximately 76 of ~170 compounds are not marked in the Flavor-Base dataset as to whether they are "natural", "natural identical" or "artificial". In contrast, the GRAS1882 set is predominately defined by 75% as "nature identical", ~600 "artificial", 8 "natural" and only 8 unmarked. Many of the compounds in the GRAS22 unmarked group have a more complex multifunctional chemical nature characteristic of synthetic products. All of the GRAS22 false negatives fell in this unmarked compound class, which appears to reflect a drift towards more purely synthetic entries now gaining GRAS certification.

4. Discussion

With regard to the questions "Are GRAS compounds distinct as a population?" and "Assuming GRAS compounds are distinct, how do they compare as a population to pharmaceuticals?", it is now possible to state that while GRAS compounds and pharmaceuticals are similar in some specific ways (log(P) and log(S)), they are distinct and easily recognized when compared to representative

Table 2
QSAR model performance

Rand	Population				Accuracy			
	Training		Test		Training		Test	
	GRAS	MDD	GRAS	MDD	GRAS	MDD	GRAS	MDD
0	1882	1120	0	0	0.94	0.92		
0.5	908	571	973	548	0.92	0.94	0.93	0.93
0.95	81	53	1800	1067	0.94	0.92	0.93	0.9
0.99	17	10	1864	1109	0.96	0.94	0.69	0.34

pharmaceuticals in several specific chemical metrics. Some conclusions concerning the differences between the two groups and the nature of GRAS chemical space include:

1. GRAS space is more compact than pharmaceutical space. Despite the fact that the GRAS1882 database is ~70% larger than the MDD1120 on which this study was performed, the GRAS dataset was seen to have a significantly more compact range of values. This can be seen as the long black line for MDD1120 extends in one or both directions beyond what is seen for the GRAS1882 line and that the peaks for the GRAS1882 line are higher and fewer or both as compared to MDD1120. The populations are directly comparable since they are both represented as a fraction.
2. GRAS compounds are smaller than pharmaceuticals (Fig. 1a).
3. GRAS compounds are more flexible than pharmaceuticals (Fig. 1b).
4. The difference between GRAS and pharmaceuticals populations is frequently the difference of "never" (for GRAS) versus "infrequent" (for pharmaceuticals). This is seen in profiles for hydrogen bond acceptors and donors.
5. Charge groups (acidic and especially basic atoms) are noticeably rare in GRAS compounds as compared to pharmaceuticals.
6. As per aromatic groups:
   (a) A larger percentage of GRAS compounds (~ 60%) has no aromatic groups.
   (b) GRAS compounds have fewer fused ring systems than pharmaceuticals.
   (c) GRAS and pharmaceuticals are identical in terms of populations of single aromatic rings.
   (d) Pharmaceuticals are more likely to have multiple aromatic rings than GRAS compounds.
7. GRAS compounds and pharmaceuticals differ at a functional group level (Fig. 1h–m):
   (a) GRAS compounds lack halogens.
   (b) GRAS compounds are rich in single occurrences for esters and ethers.
   (c) 90% of pharmaceuticals have nitrogen bearing functional groups.
   (d) 90% of GRAS compounds do not have nitrogen functional groups.
8. GRAS chemical space is most likely changing, as can be seen by the number, chemical qualities and origin of the false negatives in the GRAS22 dataset.

Do these results make intuitive sense for GRAS compounds that are used predominantly for flavouring versus pharmaceutical compounds used predominantly to modulate intracellular machinery? First, taste and smell are highly intertwined and volatiles are favoured (McGee, 2004). Volatility is enhanced for compounds with lower molecular weights and generally impeded by charge and H-bonding. In contrast, pharmaceuticals predominantly need to be both soluble and able to penetrate cell membranes (Blake, 2003), a situation where volatility is a liability, leading to again the shift in molecular weight, charge and H-bonding elements. Any organic text book will mention the prevalence of ethers and esters in compounds associated with specific and desirable odors, which is clearly seen in Fig. 1l and m. Common texts (example: McGee, 2004) concerning food and food chemistry will present compounds with other oxygen bearing groups including alcohols, aldehydes and ketones that are important flavourings in food. We investigated the frequency of occurrence of these oxygen functional groups, but these distributions were found to be essentially the same for both GRAS compounds and pharmaceuticals. The absence of halogens in GRAS is possibly a simple result that natural cellular biochemistry does not use halogens to a significant degree.

The differences between GRAS compounds and pharmaceuticals are manifest in the predictive power of the binary QSAR model documented in this paper. Inspection of Table 2 presents that a binary model based on these 10 descriptors can effectively predict outside of its own training set. This behavior suggests that the two populations are at present so distinct that a relatively small subset is adequate to span the pertinent chemical space and to devise a trained recognition function (i.e., Binary QSAR model) capable of accurately partitioning the two sets. The selection of descriptors used was based largely on the authors' intuition from visual inspection of the GRAS database versus past experience in pharmaceutical research. It is probable that other metrics will show distinctions as well for these two datasets.

5. Conclusions

GRAS compounds and pharmaceuticals are selected based on different performance requirements. The present paper has demonstrated that GRAS and pharmaceutical chemical space are distinct and easily recognized. However, we note that the compounds of the more recent GRAS22 subset, which are strikingly less populated by "nature identical" compounds than the entire previous accumulated set, are more diverse than the historical collection dominated by natural products, and suggests that the GRAS chemical property space is expanding. Despite the recognition that the present understanding for GRAS and pharmaceutical chemical space is imperfect and subject to periodic review, a better understanding of how the present GRAS and pharmaceutical chemical spaces differ, as well as the binary QSAR model that can recognize GRAS-like compounds from commercial chemical and natural product libraries, is useful for groups with a focus on nutrition, taste and cosmetics.

Dedication

The authors dedicate this paper to the recently deceased Karnail Atwal. A wonderful friend and medicinal chemist who left us and his family in December 2006.

Acknowledgements

The authors thank R. Bryant, R. Cortes, K. Atwal and R. McGregor for excellent reviews and discussions of this paper.

References

Asinex Inc. <www.asinex.com>, Moscow, Russia.

Blake, J.F., 2000. Chemoinformatics – predicting the physiochemical properties of druglike molecules. Cur. Opin. Biotechnol. 11, 104–107.

Blake, J.F., 2003. Examination of the computed molecular properties of 'drug-like' molecules. BioTechniques 34, 16–20.

Briem, H., Gunther, J., 2005. Classifying kinase inhibitor likeness using machine learning methods. Chem. Biol. Chem. 6, 558–566.

Burdock, G.A., 2003. The GRAS process. FoodTech. 57, 17.

ChemBridge Corp. <www.chembridge.com>, San Diego CA (USA).

ChemDiv Inc. <www.chemdiv.com>, San Diego CA (USA).

Clark, R.D., Sprous, D.G., Leonard, J.M., 2001. Progressive scrambling: Validating models based on large datasets. In: Holtje, H.-D., Sipple, W. (Eds.), Rational Approaches to Drug Design. Prous Science, S.A. Barcelona, Spain, pp. 475–486.

Ertle, P., Rohde, B., Selzer, 2000. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. J. Med. Chem. 43, 3717–4714.

Ford, M.G., Pitt, W.R., Whitley, D.C., 2004. Selecting compounds for focused screening using linear discriminant analysis and artificial neural networks. J. Mol. Mod. Graph. 22, 464–467.

Flavor-Base, 2004. Leffingwell and Associates, Canton Georgia, USA.

Hallagan, J.B., Hall, R.L., 1995. FEMA GRAS – A GRAS assessment program for flavor ingredients. Reg. Tox. and Pharma. 21, 422–430.

Jimonet, P., Jager, R., 2004. Strategies for designing GPCR-focus libraries and screening sets. Curr. Opin. Drug. Disc. Dev. 7, 325–333.

LaBute, 1999. Binary QSAR: A new method for the determination of QSARs. Pac. Sym. Biocomp. 4, 444-455.

Leach, 1996. Molecular Modelling: Principles and Applications. Pearson Education Limited. Essex, England.

LexiChem, 2006. OpenEyes. Santa Fe NM USA.

Lipinski, C.A., 2003. Chris Lipinski discusses life and chemistry after the Rule of Five. Drug. Disc. Today 8, 12–16.

McGee, H., 2004. Food and Cooking: The science and lore of the kitchen. Scribner, (Simon and Schuster), New York, NY, 10020.

MOE, 2005. Chemical Computing Group, Montreal, Quebec (Canada).

Munro, I.C., Shubik, P., Hall, R., 1998. Principles of the safety evaluation of flavoring substances. Food Chem. Tox. 36, 529–540.

Oprea, T.I., Gottfries, J., 2000. A one component model for oral bioavailability. J. Mol. Graph. Model. 5, 261–274.

Prestwick Chemical Library, 2005. Prestwick Chemical Inc. Illkirch, France.

Prien, O., 2005. Target family oriented focused libraries for kinases. Chem. Biol. Chem. 6, 500–505.

Smith, R.L., Cohen, S.M., Doull, J., Feron, V.J., Goodman, J.I., Marnett, L.J., Munro, I.C., Protoghese, P.S., Waddell, W.J., Wagner, B.M., Adams, T.B., 2005. Criteria for safety evaluation of flavoring substances. Food Chem. Tox. 43, 1141–1177.

Sprous, D.G., Zhang, J., Zhang, L., Wang, Z., Tepper, M.A., 2005. Kinase inhibitor recognition by use of a multivariable QSAR model. J. Mol. Graph. Mod. 24, 278–295.

Walters, W.P., Ajay, Murcko, M.A., 1999. Recognizing molecules with druglike properties. Curr. Opin. Chem. Biol. 3, 382–387.

Weininger, D., 1988. SMILES. 1. Introduction and encoding rules. J. Chem. Info. Comp. Sci. 28, 31–38.

Woods, L.A., Doull, J., 1991. GRAS evaluation of flavoring substances by the expert panel of FEMA. Regul. Toxicol. Pharmacol. 14, 48–58.