

Phytochemical Informatics of Traditional Chinese Medicine and Therapeutic Relevance

Thomas M. Ehrman,[†] David J. Barlow,^{*,†} and Peter J. Hylands[‡]

Pharmaceutical Sciences Division and Centre for Natural Medicines Research, King's College London, Franklin-Wilkins Building, 150 Stamford Street, London SE1 9NH, U.K.

Received May 17, 2007

Distribution patterns of 8411 compounds from 240 Chinese herbs were analyzed in relation to the herbal categories of traditional Chinese medicine (TCM), using Random Forest (RF) and self-organizing maps (SOM). RF was used first to construct TCM profiles of individual compounds, which describe their affinities for 28 major herbal categories, while simultaneously minimizing the level of noise associated with the complex array of diverse phytochemicals found in herbs from each category. Profiles were then reduced and visualized with SOM. The distribution of 10 major phytochemical classes, in relation to TCM profile, was delineated with SOM-Ward clustering. These classes comprised aliphatics, alkaloids, simple phenolics, lignans, quinones, polyphenols (flavonoids and tannins), and mono-, sesqui-, di-, and triterpenes (including sterols). Highly distinctive patterns of association between phytochemical class and TCM profile were revealed, suggesting that a strong phytochemical basis underlies the traditional language of Chinese medicine. Maps trained after random permutation of herbs assigned to each category were, by contrast, devoid of feature, providing additional evidence for the significance of these associations. Most classes were split into relatively few clusters, and further analysis revealed that simple descriptors, comprising skeletal type, molecular weight, and calculated log P, were in most cases able to readily discriminate within-class clusters. Relationships between TCM profile and predicted activities, relating to therapeutically important molecular targets, were explored and indicate that ethnopharmacological data could play an important role in pharmaceutical prospecting from Chinese herbs as well as identifying links between Chinese and Western medicine.

INTRODUCTION

In recent years a considerable body of information has accumulated on the chemical constituents of Chinese herbs and their pharmacological potential. This is reflected in the appearance of a number of new electronic databases^{1–5} which contain both structural details of several thousand herbal constituents and accompanying information on their uses in traditional Chinese medicine (TCM). A prerequisite for effective mining of the chemical information in these databases concerns understanding and interpretation of the language of TCM.⁶ Though obscure at first, many of the therapeutic categories found in Chinese herbal materia medica are interpretable in Western terminology, and a variety of texts now available in English, in particular the works of Bensky et al.,^{7,8} have helped to clarify these relationships.

Though traditional systems of herbal medicine are sometimes viewed as opaque to scientific analysis there is little evidence to date either to corroborate or refute this. Due to a shortage of information it has been difficult until recently to apply techniques of data analysis to the problem. This is illustrated by a consideration of publication dates for structures in our database⁵ and reveals that prior to the 1970s (toward the end of which period the first compendium of Chinese herbal constituents was published⁹) data of this sort were scarce. The volume of information has risen almost

exponentially, however, over the last few decades, and, though incomplete, current information therefore allows for considerably greater confidence in the application of exploratory data analysis than previously.

Depending on one's viewpoint, the efficacy of a herb may either be attributable to one or a few 'active ingredients' or else to the synergistic effects of all its constituents. There is evidence to support both arguments,¹⁰ and the truth may lie somewhere in between. A problematic issue in herbal pharmacology concerns the concentrations of different compounds within each herb. These vary considerably and are strongly influenced by factors such as genotype, season, and growing conditions. In the great majority of cases, however, those constituents with greatest pharmacological activity are present in relatively high concentration and are considerably 'elaborated', structurally speaking, in relation to minor constituents. In this respect the secondary metabolism of plants is strikingly similar to the activities of a combinatorial chemist. Thus, the number of ginsenosides reported from ginseng species number over 50 to date.¹¹ In each species constituents of the same class are furthermore very similar in terms of structure and tend to show characteristic *motifs* (substituent patterns) to the extent that an experienced phytochemist may be able to tell the botanical provenance of a particular compound, frequently to the genus if not the species level. Structural elaboration thus serves as the most reliable guide to assessing the importance of a related group of chemicals in any one herb and is of particular relevance to the type of analysis described here, whereby

* Corresponding author e-mail: dave.barlow@kcl.ac.uk.

[†] Pharmaceutical Sciences Division.

[‡] Centre for Natural Medicines Research.

this feature plays an important role in evaluating the likely significance of a particular phytochemical class within the classification scheme of TCM.

Most herbs which have been reasonably well characterized show a high degree of elaboration for more than one class. Thus, cinnamon not only contains many cinnamic acid derivatives (phenylpropanoids) but also is characterized by a large number of diterpene cinnasols. Likewise, liquorice contains many oleanane triterpene saponins and isoflavonoids, and ginseng, in addition to its ginsenosides, contains appreciable numbers of sterols, acetylenes, and sesquiterpenes.⁵ We might expect that these classes are differentially associated with distinct therapeutic categories, and, if so, structural analysis may reveal these associations more clearly.

In view of the fact that herbs constitute a complex array of chemicals it is essential to use methods which can (1) accurately identify compounds most characteristic of any particular TCM category and (2) assess the structural similarity of these to compounds found in *other* categories. In this way the chemical noise inevitably associated with each category is substantially reduced, and a clearer picture of relationships between individual compounds and herbal categories begins to emerge.

In this study we demonstrate that Random Forest (RF),^{12,13} based on consensus among multiple decision trees, may be used to construct a *profile* of each compound, showing its *affinity* for each herbal category (i.e., the probability that it is found in that category). The affinity of a compound for a particular category does *not* imply a direct therapeutic effect (though in many cases this may prove to be the case), the TCM profile being intended simply as a fingerprint which identifies the possible roles of a plant compound in Chinese medicine.

TCM profiles derived in this manner may be put to a number of uses. In the work reported here they are used to compare different phytochemical classes. Self-organizing maps, and allied methods such as counterpropagation networks, are versatile tools for exploring chemical diversity and structure–activity relationships¹⁴ and are here employed to visualize and cluster the distribution of these classes in relation to TCM categories. In addition, relationships between TCM profile and predicted activities against a number of therapeutically important targets are analyzed. The latter are taken from results which have, in part, been previously published¹⁵ and, despite the limited number of targets, illustrate how such an approach may be used to discover links between ethnopharmacological and molecular data.

This analysis can go some way toward a shift in emphasis from the level of the whole herb toward that of phytochemical constituents as a useful complementary focus in herbal practice as well as identifying reservoirs of plant compounds with potential in drug discovery. Though it might be supposed that this implies a reductionist approach, due to the fact that much work at the phytochemical level centers around experimental observation on isolated compounds, TCM profiles, by contrast, can be used to understand the properties of compounds in combination, thus shedding light on the chemical basis of traditional polypharmacy and its relationships with modern medicine.

MATERIALS AND METHODS

A. Materials. (1) *Phytochemical Databases.* The data used for this analysis come from our Chinese herbal constituents database (CHCD) which currently contains details, including structures, for 8411 compounds from 240 Chinese herbs.⁵ The emphasis is on herbs commonly used in the clinical practice of TCM, as opposed to the many herbs used in folk medicine throughout China, most of which are not part of official pharmacopoeias and whose TCM description is therefore rudimentary. Efforts have also been made to give as thorough a chemical characterization for each herb in the CHCD as possible, as opposed to providing a small set of representative structures.

Activities against molecular targets were predicted for each of these compounds, on the basis of RF models trained using sets of active compounds from our bioactive plant compounds database of phytochemicals with known target specificities (BPCD). This database currently holds 2597 compounds.⁵

(2) *Phytochemical Classification.* All compounds in the CHCD and BPCD have been classified according to natural product class (e.g., triterpene), skeletal type (e.g., oleanane triterpene), and glycoside status (whether the compound is a glycoside or not). The classification follows that in the *Dictionary of Natural Products*.¹¹ A ‘skeletal type’ is analogous to a congeneric series of compounds differing in their substituents but all based on the same ‘scaffold’.

(3) *TCM Herbal Classification.* For the purposes of this analysis, the classification of herbs follows the well-known categorization found in all Chinese materia medica, in which herbs are separated into just over 30 major categories, of which 28 are covered here (Table 1). Though all materia medica follow an almost identical scheme, the major reference used was *Chinese Herbal Medicine/Materia Medica*.⁷ In rare cases where the herb was not listed in this reference, other sources such as Hsu’s *Oriental Materia Medica*¹⁶ and Zhu’s *Chinese Materia Medica*¹⁷ were used instead.

The 28 categories covered were chosen both in terms of their importance in TCM and also by the quantity of chemical information available for each. Only these *major* therapeutic categories are used. Though this may appear simplistic given the variety of uses to which an herb may be put, there are several reasons for this. First, while there is almost universal agreement between different pharmacopoeias on the main category for each herb, when it comes to other uses agreement is less widespread. Second, the language employed to describe other uses of herbs does not lend itself, in many cases, to easy categorization. While this may have certain advantages, amenability to analysis is not among them. If analysis is restricted to major categories, then we can be sure, by contrast, that the associations so defined fall within a commonly agreed and recognized framework. Third, the therapeutic effects of herbal formulas are generally expressed in terms of these categories. Thus, for example, the well-known formula *Gui Pi Tang* (Restore the Spleen Decoction) is described in the traditional literature as a medicine which ‘tonifies the Qi (energy), nourishes the Blood, strengthens the Spleen, and calms the Shen (spirit)’.⁸ The actions of this formula thus combine three of the categories outlined in Table 1 (*Tonify Qi*, *Tonify Blood*, and *Shen*). Finally, as Table

Table 1. Descriptions of TCM Herbal Categories, Details of Numbers of Herbs and Compounds, and Representative Herbs for Each in the CHCD^a

no.	TCM category	Western equivalent (approximate)	signs and symptoms/conditions	NH	NC	representative herbs
1	Wind Cold	diaphoretic, antiviral, antibacterial	chills, headache, body and neck pain, no fever/mild fever	14	530	<i>Ephedra sinica</i> , <i>Cinnamomum cassia</i> , <i>Zingiber officinale</i> , <i>Magnolia</i> spp.
2	Wind Heat	diaphoretic, antiviral, antibacterial	fever, sore throat, mild chills, deep-seated infections; also used for rashes and eye problems	12	461	<i>Mentha haplocalyx</i> , <i>Arctium lappa</i> , <i>Chrysanthemum morifolium</i> , <i>Pueraria lobata</i> , <i>Vitex rotundifolia</i> , <i>Bupleurum chinense</i> , <i>Cimicifuga simplex</i>
3	Heat (Qi)	refrigerant, antipyretic, anti-inflammatory, antimicrobial	high fever, irritability, thirst, delirium, skin disease (certain types)	7	162	<i>Anemarrhena asphodeloides</i> , <i>Gardenia jasminoides</i> , <i>Lophatherum gracile</i> , <i>Phragmites communis</i>
4	Heat (Blood)	refrigerant, styptic, coagulant	fever, rash, nosebleed, vomiting blood, blood in stool and urine, skin disease (certain types)	6	174	<i>Rehmannia glutinosa</i> , <i>Scrophularia ningpoensis</i> , <i>Paeonia suffruticosa</i> , <i>Lithospermum erythrorhizon</i>
5	Damp Heat	antimicrobial, antipyretic, anti-inflammatory	dysentery, urinary difficulty, jaundice, skin disease (e.g., eczema)	5	298	<i>Scutellaria baicalensis</i> , <i>Coptis chinensis</i> , <i>Phellodendron amurense</i> , <i>Gentiana scabra</i> , <i>Sophora flavescens</i>
6	Toxic Heat	detoxicant, anti-inflammatory, antimicrobial, antiviral, diuretic	painful swellings, purulent infections, abscesses, dysentery, certain viral infections (mumps, encephalitis), skin disease	23	451	<i>Lonicera japonica</i> , <i>Forsythia suspensa</i> , <i>Isatis indigota</i> , <i>Taraxacum mongolicum</i> , <i>Viola yedoensis</i> , <i>Dictamnus dasycarpus</i> , <i>Andrographis paniculata</i> , <i>Smilax glabra</i>
7	Heat (Deficiency)	antipyretic, anti-inflammatory, antimicrobial, antiviral	fever, night fever, particularly in weakened patients	2	100	<i>Artemisia annua</i> , <i>Picrorhiza kurroa</i>
8	Laxative	laxative, purgative	constipation	3	168	<i>Rheum palmatum</i> , <i>Cassia angustifolia</i> , <i>Aloe</i> spp.
9	Cathartic	cathartic	oedema (severe), ascites, pleurisy	5	102	<i>Croton tiglium</i> , <i>Pharbitis nil</i> , <i>Euphorbia kansui</i> , <i>Daphne genkwa</i> , <i>Phytolacca acinosa</i>
10	Drain Dampness	diuretic	oedema, urinary dysfunction, damp sores, damp warm febrile disease, jaundice, heart disease	17	533	<i>Poria cocos</i> , <i>Coix lachryma jobi</i> , <i>Juncus effusus</i> , <i>Kochia scoparia</i> , <i>Plantago asiatica</i> , <i>Alisma plantago-aquatica</i> , <i>Tetrapanax papyriferum</i> , <i>Pyrrosia lingua</i>
11	Wind Damp	antirheumatic, analgesic, antipyretic, anti-inflammatory, anticoagulant	rheumatic conditions, pain and numbness (e.g., muscles, joints)	14	536	<i>Angelica pubescens</i> , <i>Gentiana macrophylla</i> , <i>Clematis chinensis</i> , <i>Acanthopanax gracilistylus</i> , <i>Siegesbeckia orientalis</i> , <i>Tripterygium</i> spp.
12	Phlegm Heat	expectorant, antitussive, anti-inflammatory, sedative	cough (dry), scrofula, goiter, convulsions, some psychiatric conditions	4	149	<i>Peucedanum decursivum</i> , <i>Fritillaria</i> spp., <i>Dioscorea bulbifera</i> , <i>Euphorbia helioscopia</i>
13	Phlegm Cold	expectorant, decongestant	cough (productive), phlegm	4	116	<i>Pinellia ternata</i> , <i>Arisaema amurense</i> , <i>Inula japonica</i>
14	Coughing & Wheezing	antitussive, expectorant, antibiotic, diuretic, laxative, antiasthmatic	cough (persistent), wheezing, asthma	8	340	<i>Aster tataricus</i> , <i>Tussilago farfara</i> , <i>Perilla frutescens</i> , <i>Eriobotrya japonica</i> , <i>Stemona sessilifolia</i> , <i>Morus alba</i> , <i>Ginkgo biloba</i> , <i>Nandina domestica</i>
15	Emetic	emetic	phlegm (severe), retained food, jaundice (certain types)	2	40	<i>Cucumis melo</i> , <i>Veratrum</i> spp.
16	Aromatic (Damp)	digestive stimulant	distention, nausea, vomiting, poor appetite, greasy tongue coating	5	207	<i>Agastache rugosa</i> , <i>Eupatorium fortunei</i> , <i>Atractylodes lancea</i> , <i>Amomum</i> spp., <i>Alpinia katsumadai</i>

Table 1 (Continued)

no.	TCM category	Western equivalent (approximate)	signs and symptoms/conditions	NH	NC	representative herbs
17	Regulate Qi	digestive stimulant, circulatory stimulant, analgesic	pain (epigastric, hypochondriac and abdominal), diarrhea/constipation, irregular menstruation, stifling sensation in chest	10	527	<i>Citrus</i> spp., <i>Cyperus rotundus</i> , <i>Saussurea lappa</i> , <i>Lindera strychnifolia</i> , <i>Aquilaria agallocha</i> , <i>Santalum album</i> , <i>Allium chinensis</i> , <i>Melia toosendan</i> , <i>Diospyros kaki</i> , <i>Nardostachys jatamansi</i>
18	Stop bleeding	styptic	bleeding, vomiting and coughing blood, hematuria, excessive menstruation, trauma	9	368	<i>Typha</i> spp., <i>Agrimonia pilosa</i> , <i>Panax notoginseng</i> , <i>Bletilla striata</i> , <i>Cirsium japonicum</i> , <i>Sanguisorba officinalis</i>
19	Invigorate Blood	anticoagulant, circulatory stimulant	pain (severe, fixed), abscesses, ulcers, abdominal masses, thrombosis, ischemia	19	765	<i>Ligusticum chuanxiong</i> , <i>Salvia miltiorrhiza</i> , <i>Corydalis yanhusuo</i> , <i>Curcuma</i> spp., <i>Boswellia carteri</i> , <i>Commiphora</i> spp.
20	Interior Cold	circulatory stimulant, cardiogenic	cold extremities, lack of thirst, loose stool, nausea, diarrhea, chest and abdominal pain, slow pulse	13	490	<i>Aconitum carmichaeli</i> , <i>Evodia rutaecarpa</i> , <i>Zanthoxylum bungeanum</i> , <i>Eugenia caryophyllata</i> , <i>Piper</i> spp., <i>Myristica fragrans</i>
21	Tonify Qi	endocrine agent, immunostimulant	lethargy, weakness, poor appetite, weak voice, pale complexion, breathlessness, immunodeficiency	6	398	<i>Panax ginseng</i> , <i>Codonopsis pilosula</i> , <i>Astragalus membranaceus</i> , <i>Atractylodes macrocephala</i> , <i>Glycyrrhiza</i> spp., <i>Dioscorea opposita</i>
22	Tonify Blood	antianaemic	pallor, dizziness, vertigo, poor vision, lethargy, palpitations, amenorrhea, insomnia, pale tongue, fine pulse	4	69	<i>Angelica sinensis</i> , <i>Rehmannia glutinosa</i> , <i>Polygonum multiflorum</i> , <i>Paeonia lactiflora</i>
23	Tonify Yang	endocrine agent, stimulant	systemic exhaustion, fear of cold, cold extremities, withdrawal, sore and weak lower back, slow and deep pulse	10	357	<i>Cistanche deserticola</i> , <i>Epimedium grandiflorum</i> , <i>Psoralea corylifolia</i> , <i>Alpinia oxyphylla</i> , <i>Eucommia ulmoides</i> , <i>Dipsacus asper</i> , <i>Morinda citrifolia</i> , <i>Cnidium monnieri</i>
24	Tonify Yin	endocrine agent, antidiuretic, antihypertensive, anticholesterolaemic	dizziness, tinnitus, weak lower back and knees, low-grade fever, menopausal symptoms, scanty dark urine, red dry tongue, thin pulse	10	271	<i>Glehnia littoralis</i> , <i>Asparagus cochinchinensis</i> , <i>Ophiopogon japonicus</i> , <i>Dendrobium nobile</i> , <i>Polygonatum</i> spp., <i>Eclipta prostrata</i> , <i>Ligustrum lucidum</i> , <i>Lilium brownii</i>
25	Astringent	astringent, endocrine agent	diarrhea, polyuria, sweating, prolapse, discharge	8	304	<i>Schisandra chinensis</i> , <i>Cornus officinalis</i> , <i>Terminalia chebula</i> , <i>Nelumbo nucifera</i>
26	Shen	tranquillizer, sedative, nerve tonic	palpitations, anxiety, insomnia	5	292	<i>Zizyphus spinosa</i> , <i>Biota orientalis</i> , <i>Albizia julibrissin</i>
27	Phlegm (Heart)	resuscitant, tranquillizer, stimulant, nerve agent	delirium, seizure, coma, psychiatric conditions (e.g., bipolar)	3	128	<i>Polygala tenuifolia</i> , <i>Liquidambar orientalis</i> , <i>Acorus gramineus</i>
28	Internal Wind	antihypertensive, sedative, nerve tonic	tremor, spasm, hemiplegia, aphasia, paralysis, blurred vision, dizziness, paraesthesia, stroke	3	97	<i>Uncaria rhynchophylla</i> , <i>Gastrodia elata</i> , <i>Tribulus terrestris</i>

^a NH = number of herbs; NC = number of compounds.

1 may suggest, the principles of TCM are broad and encompassing, and the same basic categories recur repeatedly in different contexts. This implies that, if significant phytochemical associations can be identified, these may potentially have broad therapeutic applications.

B. Methods. (1) *Structural Characterization of Compounds.* For each entry in the CHCD, 110 descriptors were computed using MOE (Chemical Computing Group, Mon-

treau, Quebec), including constitutional descriptors (atom, bond and ring counts, H-bond donors/acceptors, molecular weight, etc.), topological indices (Balaban, Randic, Kier-Hall chi and kappa indices, etc.), and Labute's 'virtual surface area' (VSA) descriptors which characterize the molecular surface area (computed from a 2D representation, hence the term 'virtual') in terms of lipophilicity, molar refractivity, and partial charge.¹⁸ Finally, log P (octanol/water partition

coefficient), a measure of the relative solubility of a compound in aqueous and nonaqueous media, was calculated using an atomic contribution method.¹⁹

(2) *Compound Discrimination Using Random Forest*. RF predicts the category within the target variable to which a compound belongs on the basis of consensus among multiple trees. In constructing each of these trees, RF uses the CART algorithm and Gini splitting criterion to determine the appropriate value for each node in the tree.²⁰ Each tree is grown to its full extent, without the pruning characteristic of single trees.

Among the other features which characterize RF are the following: (1) *Random selection of descriptors*, whereby for each tree built a random subset of descriptors (often set, as here, to the square root of the total number) is selected at each step of the tree building process. (2) *'Out-of bag' (OOB) cross-validation*: In this procedure, for each tree generated, approximately 33% of compounds are randomly excluded and constitute an independent test sample. To measure the generalization error for each tree, the OOB sample is run through that tree and the error rate of prediction measured. The error rates for all trees are then averaged to give the overall generalization error for the entire forest. (3) *Balanced sampling*: If classes within the target variable contain different numbers of compounds, as is generally the case, then, because RF attempts to minimize overall error, it will concentrate on the majority class, resulting in a low degree of accuracy for the minority class. One solution is to draw equal samples from each class for each tree built before modeling, as has been used here. Further details may be found elsewhere.^{12,13}

In the present study 28 RF models were constructed, one for each of the TCM categories in Table 1. All compounds found in herbs for that category were given a value of 1, while others were scored as 0. A total of 500 trees was grown for each RF model.

The single decision trees used to discriminate SOM-Ward clusters within each phytochemical class (see below) followed an identical protocol to that above (CART algorithm with Gini splitting criterion). In this case, however, terminal nodes were constrained to have no fewer than five compounds.

RF models were trained using *Random Forests version 1.0* (Salford Systems, San Diego, CA). Single decision trees were constructed using *Statistica version 6.0* (StatSoft, Tulsa, OK).

(3) *Reduction, Visualization, and Clustering of TCM Profiles with Self-Organizing Maps*. For each TCM category, RF returns a probability for each compound of its association with that category. The full set of probabilities for all 28 categories constitutes the *TCM profile* of each compound. These profiles may be used to reveal the affinities of phytochemical classes (and subsets of these) for the full array of herbal categories found in TCM.

Self-organizing maps, also known as Kohonen maps after their inventor Teuvo Kohonen, are among the most popular methods for reducing and visualizing high dimensional data via a topologically ordered form of clustering.²¹ The training algorithm represents a type of unsupervised neural network in which the neurons (or nodes) of the map, arranged as a square or rectangular lattice, compete for input vectors (the set of descriptor values for each compound), generally via a

weighted Euclidean or Gaussian distance measure. Sequential updating of weights during training gives a final map in which (a) very similar compounds, in terms of TCM profile, are found in the same node, and (b) nodes are arranged such that increasing distance between two nodes reflects decreasing similarity.

In this example, a high resolution, 4000 node map was trained using only information on TCM profile (28 variables). These variables were all assigned an identical weight. Training was carried out using *Viscovery SOMine version 4.0* (Eudaptics Software, Vienna) which employs a modified, faster version of Kohonen's original algorithm.²²

Nodes were then clustered via SOM-Ward clustering, which simultaneously incorporates information on map topology with traditional Ward clustering.²³ Since this procedure is not widely known, details of the algorithm are given in the Supporting Information.

The variables used to cluster the nodes consisted of the 10 major phytochemical classes characterizing our data (flavonoids, monoterpenes, triterpenes, etc.). In this way phytochemical classes were split into subsets showing different TCM profiles. The final number of clusters used was determined by measuring the number of nodes occupied by the largest (and most heterogeneous) cluster. A significant drop in number was observed between 39 and 40 clusters (from 21% to 10% of the total map). Forty clusters were therefore used in subsequent analysis. Average quantization error was 0.1147, indicating that dimensionality reduction via SOM resulted in little loss of information on distances between neighboring compounds compared to unreduced data.

(4) *Two-Way Clustering of Information on TCM Category and Target Affinity*. It is often the case, as in microarray analysis, that relationships between row and column vectors need to be arranged in such a way that subsets characteristic of both are found 'clustered' together. In this study, relationships between TCM category and predicted target affinity were analyzed using Hartigan's original modal block algorithm.²⁴ This employs a simple iterative approach to rearrange row and column vectors until subsets of each, sharing identical or similar values, are optimally clustered. The analysis used *Statistica's* implementation of the modal block algorithm.

RESULTS

(1) Description of Herbal Categories and Their Role in TCM. The 28 herbal categories used in the analysis are described in Table 1, and the number of herbs, number of compounds, and representative herbs from the CHCD are given for each. The order of categories follows that found in standard materia medica, in which these are arranged, where possible, in terms of similarity to one another. In keeping with modern TCM practice, terms such as Qi, Blood, Yin, and Yang are all given an initial capital letter, both in Table 1 and the text, to distinguish them from such terms in non-TCM usage. Herbal categories (e.g., *Wind Heat*, *Tonify Qi*) are given in italics.

It will be noticed that some of these categories are relatively small in terms of the number of herbs they contain. Thus, for instance, *Damp Heat* herbs number only five in all. However, the importance of a herbal category bears little

relationship to its size. These five herbs are very widely used and occur in 43 out of 193 (22%) of the best known herbal formulas. Similarly the *Heat (Blood)* category has six herbs which occur in no less than 35% of these formulas.⁸ Another reason for the small size of some categories concerns the fact that medicinals other than plants make a substantial contribution, for which little chemical information is available and which are not covered in the CHCD. An example is the *Internal Wind* category where a variety of invertebrates are listed in traditional materia medica.⁷

Where possible, each category is described both in terms of the signs and symptoms traditionally used in diagnosis and also in terms of equivalent Western categories. Concerning the latter, however, it is sometimes difficult to find precise equivalents, and a number must be regarded as approximations.

The majority of categories, however, do not pose significant problems in this respect, and it is hoped that the information in Table 1 is, in most cases, reasonably self-explanatory. Space does not permit a detailed discussion of these categories here and their relationships to TCM theory. It should be noted, moreover, that Table 1 only gives an outline of the uses to which herbs in each category may be put, and more specialized disciplines, such as TCM dermatology, contain much additional information which is not included. For an overview of Chinese medicine, the reader is referred to one of the many introductory texts now available in English.²⁵

Some categories are difficult to translate into Western terms. *Shen*, for instance, refers to the 'spirit' in TCM. Herbs in this category are known for their tranquilizing and sedative effects. Similarly, the category named *Phlegm (Heart)* refers to those herbs which 'clear Phlegm from the Heart Orifices'. These are mainly resuscitants, used in fainting, delirium, coma, and so forth, though a number are also used in less acute circumstances where they have a role to play in the treatment of some psychiatric conditions (such as manic depression or bipolar disorder).

It will be noticed that Wind occurs in several categories. There are two types—external and internal. The former is found in *Wind Cold*, *Wind Heat*, and *Wind Damp* categories. Herbs from these categories are used in the treatment of infectious diseases, such as coryza and influenza, and also play a prominent role in rheumatic conditions. Herbs from the *Internal Wind* category, by contrast, are used in nerve-related conditions, such as tremor, spasm, aphasia, vertigo, and Parkinson's disease.⁷

A number of categories are also Heat related. In TCM, Qi-level Heat (*Heat (Qi)*) refers to those situations where there is a high fever, while Blood-level Heat (*Heat (Blood)*) encompasses a range of symptoms including bleeding and certain skin conditions. The relationship between Heat and bleeding is explained by the TCM concept that 'Heat in the Blood' gives rise to 'reckless movement of Blood' which in turn leads to hemorrhage.²⁶ *Damp Heat* herbs are frequently used in the treatment of dysentery, jaundice, and liver disease and are generally bitter in taste, while *Toxic Heat* herbs, which comprise a very large and varied category, are used to treat abscesses, certain infectious diseases, and skin conditions among others.⁷

Tonics are differentiated into four categories, as shown, and are almost always included where TCM diagnosis reveals

'deficiencies' of various sorts. Many herbs in these categories exert a known or suspected influence on endocrine function and tend, with some exceptions, to be mild in their effects.⁷

(2) Performance of Random Forest. RF provides a measure of performance in terms of misclassification rate. Compounds which RF recognizes as most characteristic of each category receive a high score, usually of 0.7 or greater (i.e., over 70% of trees 'vote' for those compounds to belong to that category), while compounds which have rare structural features for that category receive lower scores.

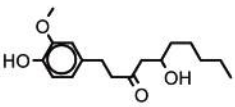
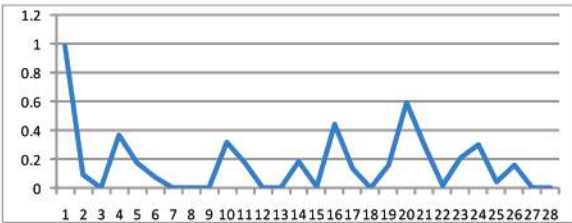
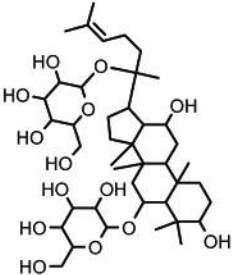
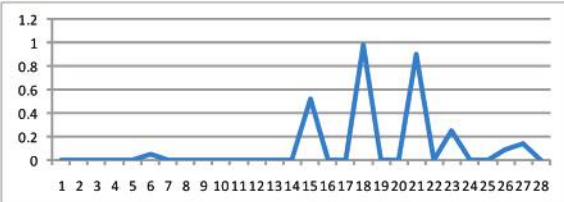
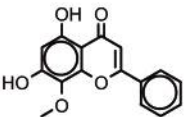
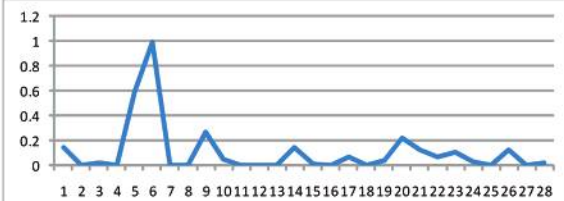
Similarly, compounds found in herbs from *other* categories may receive a high score if RF finds it difficult to distinguish them from compounds in the category in question. In the case of phytochemical classes which are both widely distributed and where compounds are structurally very similar to one another (e.g., flavonoids) this is likely to happen quite frequently. However, this will depend on the distribution of descriptor values for similar compounds within the categories in question. Apparent similarities between two compounds in different categories may not result in them achieving a comparable score for both, if RF can find a sufficient number of descriptors with significant discrepancies in this respect. RF thus provides a sensitive measure of similarity between related compounds, based on the statistical distribution of structural features in different categories.

RF classifies a compound as belonging to the class with the highest vote. In the case of binary classification, more than 50% of votes are required to assign the compound to the appropriate class. The misclassification rate is the discrepancy between actual and predicted class. In the case of two categories (*Tonify Blood* and *Phlegm (Heart)*) misclassification was above 30%, reflecting the fact that these categories have relatively few herbs and that their constituents are poorly differentiated, structurally, from compounds in other categories. The majority of categories had misclassification rates between 10% and 30%, while for six categories (*Heat (Blood)*, *Damp Heat*, *Laxative*, *Emetic*, *Phlegm Heat*, and *Phlegm Cold*) rates were less than 10%. The latter are largely composed of compounds with features which are rare in other categories.

This also suggests that, for the majority of herbs, RF classification will correctly predict the category to which each individual herb belongs. For 146 of the 223 herbs included in the analysis (65%), the average score (affinity) of all constituents of any one herb for its parent category was >0.7, indicating that these are predicted to belong to the correct category with a high level of precision. Still correctly predicted, but with lower average scores (0.5–0.7), were the constituents of a further 69 herbs (31%). Only in the remaining 8 herbs (4%) may prediction be considered to have failed. These comprise one *Drain Dampness* herb (*Artemisia scoparia*), one *Toxic Heat* herb (*Chrysanthemum indicum*), one *Heat (Qi)* herb (*Prunella vulgaris*), two *Invigorate Blood* herbs (*Liquidambar formosana* and *Leonurus heterophyllus*), one *Tonify Blood* herb (*Paeonia albiflora*), one *Tonify Yin* herb (*Polygonatum odoratum*), and one *Wind Damp* herb (*Gentiana macrophylla*). For four of these herbs, however, chemical information is limited (less than 15 constituents per herb listed in the CHCD).

Due to the fact that, like neural networks, RF constitutes a form of 'black box' whereby only a crude picture of the influence of different predictor variables can be established,

Table 2. TCM Profiles of Three Phytochemicals from Chinese Herbs^a

COMPOUND	TCM PROFILE
 <p>[6]-Gingerol</p>	
 <p>Ginsenoside Rg1</p>	
 <p>Wogonin</p>	

^a The X-axis indicates the 28 herbal categories listed in Table 1. The Y-axis indicates affinity for each category (from 0 to 1).

discussion of the relative importance of individual descriptors in determining membership of each category is omitted, in addition to the fact that relationships between simple descriptors and SOM-Ward clusters are explored in greater detail below.

(3) Illustrations of TCM Profiles. The structures and associated TCM profiles of three phytochemicals from well-known herbs are shown in Table 2.

The first, [6]-gingerol, is predominant among the long-chain aromatics found in *Zingiber officinale* (ginger). This herb belongs to the *Wind Cold* category and is found in a number of formulas used in the treatment of conditions such as the common cold. The TCM profile for [6]-gingerol also tells us that it has an affinity for the *Internal Cold* category [20], and it is of note in this respect that the *dried* form of ginger is classified in this category. However, in this analysis, the constituents of ginger were simply classified in the *Wind Cold* category which implies that [6]-gingerol must share close structural similarities with compounds from *other* herbs, apart from ginger, in the latter category.

The ginsenosides from *Panax ginseng* (ginseng) are among the most widely studied of all plant constituents. From a

biomedical perspective they appear to have a bewildering array of potential (though rather vaguely defined) activities,²⁷ so it is informative to discover their associations in TCM. Ginsenoside Rg1, a major constituent of the Chinese species of ginseng, is strongly associated with *Stop bleeding* [18] and *Tonify Qi* [21] categories. Ginseng is classified among the *Qi* tonic herbs, and a number of other herbs in this category contain compounds which are structurally very similar to ginsenosides, such as the cycloartane triterpene glycosides found in *Astragalus membranaceus*. Ginsenosides are also found in the closely related *Panax notoginseng* which is classified in the *Stop bleeding* category, hence the high score. In this respect it is perhaps significant that ginsenosides do appear to have mild coagulant properties and are known to counteract the effects of warfarin.²⁸

In the case of wogonin, a flavone from *Scutellaria baicalensis* (skullcap), there are two associations, with the *Damp Heat* [5] and *Toxic Heat* [6] categories, respectively, both of which are closely related in TCM. Herbs from these categories are frequently combined, particularly in the treatment of skin conditions, abscesses, certain infectious diseases, dysentery, and jaundice.⁷

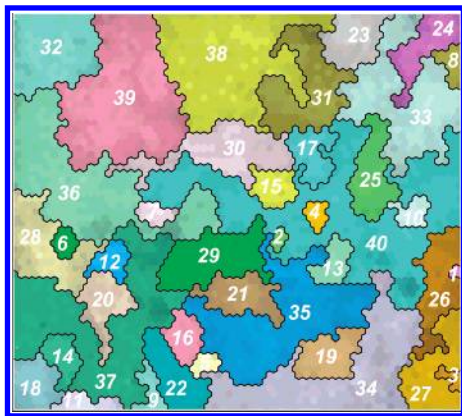


Figure 1. SOM-Ward clustering showing 40 clusters of compounds belonging to different phytochemical classes with distinct TCM profiles. [Details for each cluster are given in Table 4. Clusters are numbered in order of increasing size. Dark background shading indicates nodes containing more compounds].

(4) Distribution of Phytochemical Classes in Relation to TCM Profile. Figure 1 shows the distribution of 40 clusters representing subsets of 10 major phytochemical classes with distinct TCM profiles. These 10 classes comprise the following: aliphatics (including acetylenes) [574 compounds], alkaloids [861], simple phenolics (phenols, phenylpropanoids, coumarins, and other small phenolics) [1203], quinones (and related classes such as anthracenes) [276], lignans [327], polyphenols (flavonoids and tannins) [1162], monoterpenes [801], sesquiterpenes [881], diterpenes [388], and triterpenes (including sterols) [1644]. The clusters are numbered in order of increasing size (regarding number of nodes not number of compounds, though the two are strongly correlated with an $r^2 = 0.972$). The gray background shading of individual nodes reflects the number of compounds per node, the darker nodes containing more compounds.

In Figure 2, the distributions of the 10 phytochemical classes are shown individually. The color spectrum, from blue at one extreme to red at the other, shows the proportion of compounds (using a scale of 0 to 1) from that class found in each node. Dark blue nodes contain no compounds from the class in question, while red nodes contain compounds solely of that class. Green, yellow, and orange identify those nodes containing a mixture of compounds from different classes (usually no more than two or three). Numbering has been suppressed to allow for easier visualization.

The results demonstrate that these classes have highly distinctive patterns of distribution in relation to TCM profile. With the exception of diterpenes, where admixture with other classes is predominant (notably the structurally related nortriterpenes), the high number of red nodes indicates that compounds with similar TCM profiles tend to belong very often to the same class. This is particularly pronounced in the case of quinones, polyphenols, monoterpenes, sesquiterpenes, and triterpenes.

It is also apparent that the majority of compounds in any one class fall within a limited number of clusters. If we ignore classes making a contribution of 20% or less to each cluster, these are as follows: (1) aliphatic - 16 and 36; (2) alkaloid - 27, 29, 31, and 37; (3) simple phenolic - 18, 28, 33, 36, and 38; (4) lignan - 6, 10, 12, and 21; (5) quinone - 8; (6) polyphenol - 1, 20, 21, 23, 24, 35, and 40; (7) monoterpene - 25 and 32; (8) sesquiterpene - 15 and 39; (9)

diterpene - 17, 18, and 30; and (10) triterpene - 2, 3, 4, 5, 7, 9, 11, 13, 14, 17, 19, 22, 26, 27, 30, 34, and 37.

Those clusters characteristic of a particular class are sometimes adjacent, notably in the case of polyphenols and triterpenes, indicating that these show a *continuum* of behavior in relation to herbal category.

In contrast, aliphatics, alkaloids, simple phenolics, lignans, monoterpenes, sesquiterpenes, and diterpenes show relatively sharp discontinuities. We might therefore suspect that these clusters can be discriminated using very simple descriptors. Accordingly, for each class, single decision tree models were run, using only skeletal type (within each class), glycoside status (whether a compound is a glycoside or not), molecular weight, and calculated log P as descriptors.

(5) Discrimination of Within-Class Clusters Using Single Decision Trees. The resulting decision tree for diterpenes is shown, as an example, in Figure 3. This is a simple tree with only 7 terminal nodes. For each pair of nodes in the tree the descriptor used to split the compounds found in the parent node and the corresponding value are shown. The distribution of compounds within each of the three diterpene clusters found in each terminal node are shown to the right of the tree, in addition to the number of compounds for that node.

At four of the seven terminal nodes, compound distribution is highly uneven between the three clusters, with over 95% of compounds associated with only one cluster. At the other three, discrimination is not as good with compounds associated with two, as opposed to only one, of the three clusters. Nevertheless, the former nodes account for 220 out of the 262 diterpene compounds found in these clusters, so that overall discrimination is fairly successful. It should also be noted that failure to assign compounds to two or fewer clusters did not occur.

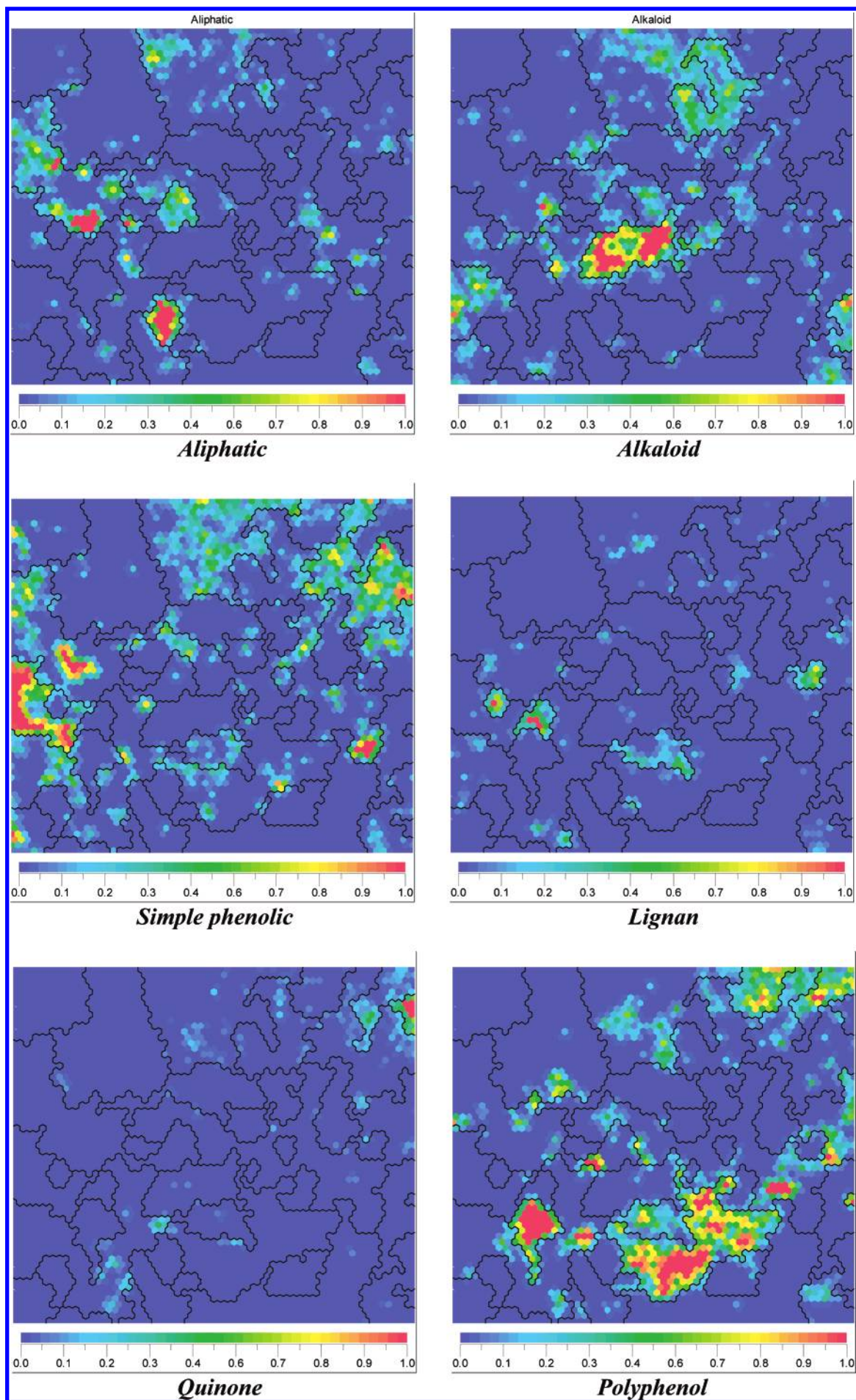
The results for all classes are shown in Table 3. Overall, these very simple trees, using only skeletal type, glycoside status, molecular weight, and calculated log P, appeared to successfully discriminate within-class clusters. In the majority (67%) of cases, terminal nodes contained only one dominant cluster, while over 70% of compounds were restricted to two clusters in a further 28% of cases.

The number of terminal nodes tells us how easily clusters were discriminated. In the case of monoterpenes, the simplest possible tree with only two terminal nodes (data not shown) managed to precisely discriminate the two monoterpene clusters, 25 and 32, using a criterion of $\log P > 0.987$. In the case of alkaloids, simple phenolics, polyphenols, and triterpenes, tree complexity was highest, though in the case of triterpenes, where the number of clusters was 17, the resulting tree was relatively less complex than for the other three classes.

Of the descriptors used, the least valuable was glycoside status, which was used on only two occasions. The use of the other descriptors was very evenly balanced, with skeletal type used on 30, molecular weight on 32, and log P on 31 occasions.

(6) Description of SOM-Ward Clusters. In Table 4, the 40 clusters shown in Figure 1 are explained both in terms of the main classes and main TCM categories involved.

For each cluster, only the main classes of compound are given. Similarly, only the major herbal categories shared in common between the TCM profiles found within that cluster



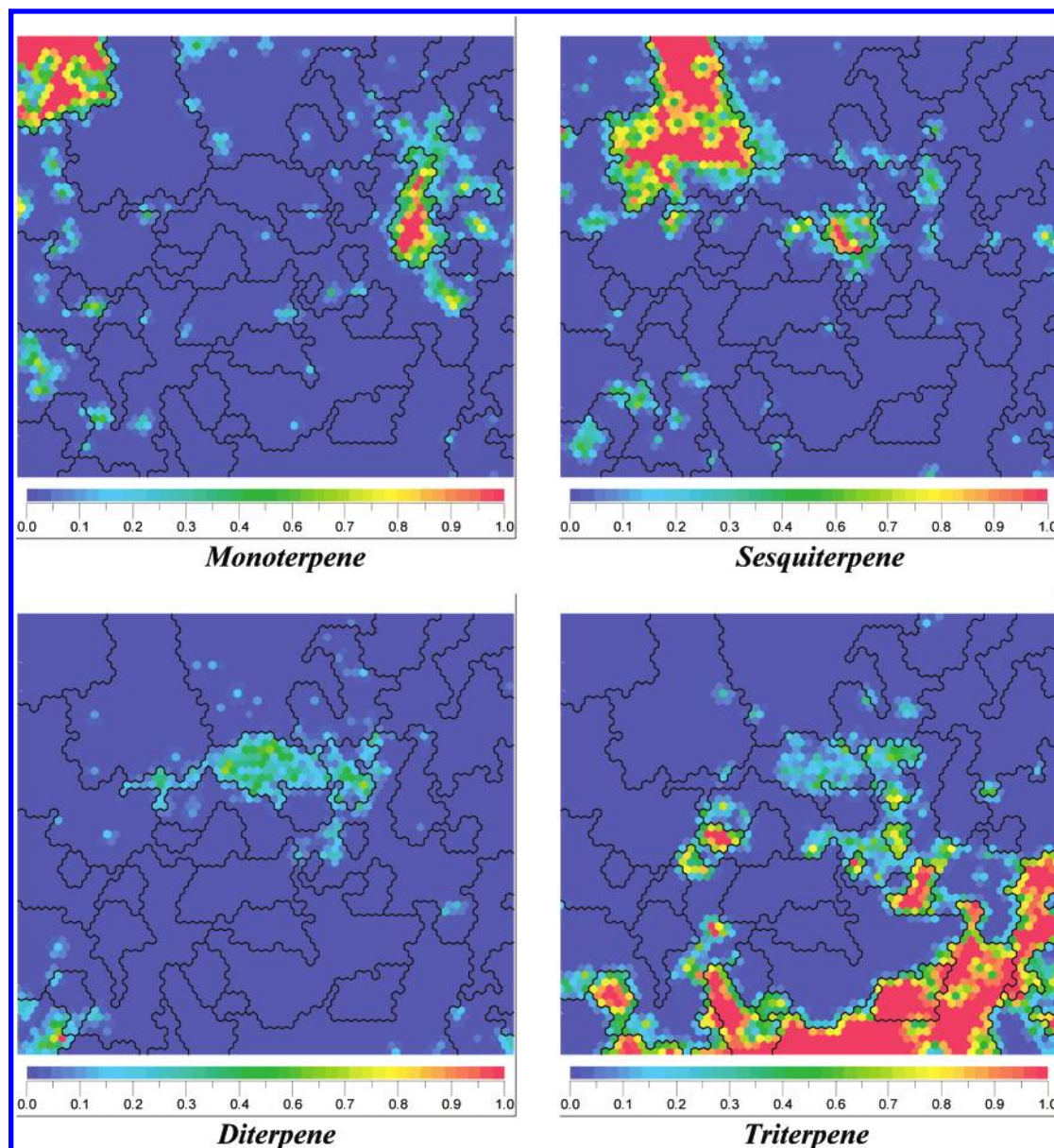


Figure 2. Distribution of major phytochemical classes in relation to clusters shown in Figure 1. Coloring indicates the proportion of compounds from that class found in each node, according to the scale shown (with colors spanning the spectral range from dark blue—indicating nodes where the given class is absent—through to bright red, indicating nodes that are populated exclusively by that class).

are given. These were identified by counting the number of profiles exhibiting a value of 0.7 or greater for each category.

The average values for molecular weight and log P are also shown for each cluster. Though glycoside status was not selected many times in the analyses above, and there is only a modest relationship between it and both molecular weight and log P ($r^2 = 0.58$ for multiple regression of % glycosides in each cluster on mean weight and mean log P), nevertheless in a number of clusters the great majority of compounds are glycosides, and this is likely to be significant, particularly in the case of monoterpenes and small phenolics. For this reason, it is included in Table 4. The skeletal types named in column 5 are shown in Table 5.

In general, with the exception of triterpenes and flavonoids, most classes are separated into clusters which are well differentiated both phytochemically and in terms of TCM profile. Monoterpenes and small phenolics share much in common as regards their distribution among herbal categories. In their simple nonglycosylated form they are conspicu-

ous for their association with the ‘warming’ categories of Chinese herbs such as *Wind Cold*, *Internal Cold*, and *Phlegm Cold* (cluster 32 in Figure 1 and Table 4). As glycosides, however, their TCM properties are entirely different, and they appear to be strongly associated with *Heat (Blood)* and *Tonify Yang* categories (25 and 33). It should also be noted that the mean log P values for the latter two clusters are among the lowest in Table 4.

The sesquiterpenes, which are slightly larger than both monoterpenes and the smallest phenolics, are closely associated with *Regulate Qi*, *Invigorate Blood*, and *Aromatic (Damp)* categories (39), though a smaller number also appear to have associations with *Wind Damp* and *Phlegm Cold* categories (15). It is intriguing to find parallels again in this respect among small phenolics, in that a fairly large group of these are also associated with *Aromatic (Damp)*, *Regulate Qi*, and *Invigorate Blood* categories (38), and these are on average slightly larger compounds than those phenolics which share TCM characteristics with the monoterpenes.

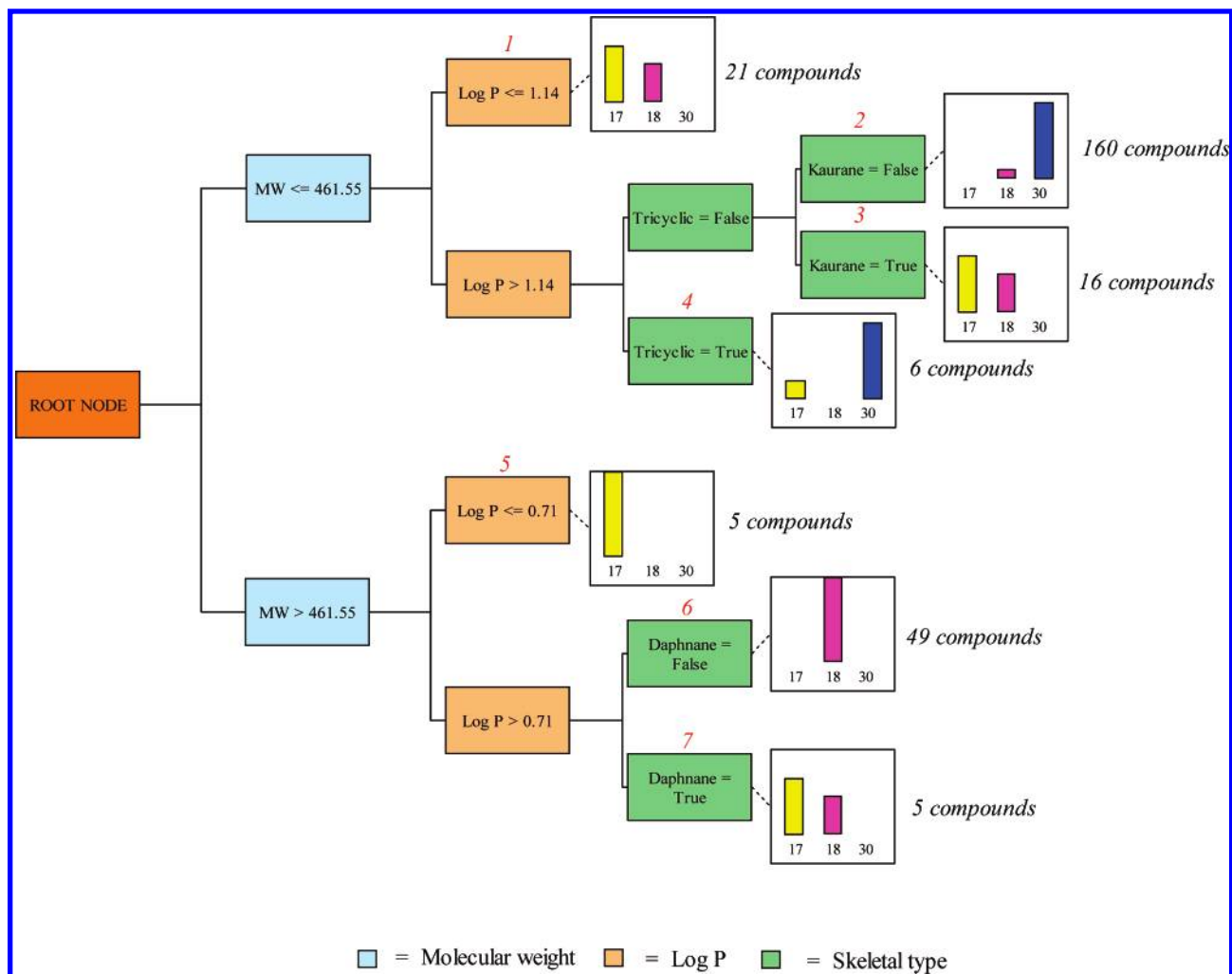


Figure 3. Decision tree discriminating 3 diterpene clusters using simple descriptors.

Table 3. Statistics for within-Class Decision Trees

							terpene			
	aliphatic	alkaloid	simple phenolic	lignan	quinone	polyphenol	mono-	sesqui-	di-	tri-
number of clusters	2	4	5	4	1	7	2	2	3	17
number of terminal nodes	5	13	18	5	0	28	2	4	7	22
terminal nodes with > 70% cases found in one cluster	5	10	12	5	0	16	2	3	5	12
terminal nodes with > 70% cases found in two clusters	0	2	6	0	0	10	0	1	2	8
terminal nodes with > 70% cases found in three clusters	0	1	0	0	0	2	0	1	0	2
skeletal type (splitting frequency)	1	5	3	1	0	11	0	0	3	6
glycoside (splitting frequency)	0	0	0	0	0	1	0	0	0	1
molecular weight (splitting frequency)	2	3	6	1	0	9	0	2	1	8
log P (splitting frequency)	1	4	8	2	0	6	1	1	2	6

Glycoside status also appears to play a role among lignans, where glycosylated forms are predominantly found among *Tonify Yang* and *Wind Damp* herbs (10), while aglycones are associated with *Wind Cold*, *Internal Cold*, and *Astringent* categories (6 and 12). Notable in the last category are the dibenzocyclooctadiene lignans from *Schisandra chinensis*.

Aliphatics are separated into two main clusters, one comprising low molecular weight, volatile compounds, which like the monoterpenes and small phenolics are found in *Wind Cold* herbs, as well as *Invigorate Blood* and *Drain Dampness* categories (36). The acetylenes, by contrast, are strongly

associated with two tonic categories, *Tonify Qi* and *Tonify Blood*, as well as *Stop Bleeding* and *Wind Heat* (16).

Alkaloids are found in four main clusters. The steroidal alkaloids are principally associated with *Emetic* and *Phlegm Heat* categories (27)—the latter category contains expectorant herbs, the mode of action of which shares much in common with the stronger emetics. Diterpene alkaloids, meanwhile, are the major constituents of *Aconitum carmichaeli* (aconite), which is the principal herb used to ‘rescue devastated Yang’, marked by symptoms of extreme cold. These and several other types (such as isobutylamide) are associated with

Table 4. Descriptions of SOM-Ward Clusters^a

phytochemical class(es)	CN	NC	TCM profile	skeletal type	glycosides (%)	MW (mean)	log P (mean)
aliphatic	16	64	<i>Tonify Qi, Stop Bleeding, Tonify Blood, Wind Heat</i>	acetylene	0	286	5.30
aliphatic, simple phenolic	36	589	<i>Wind Cold, Invigorate Blood, Drain Dampness</i>	low MW, volatile	5	250	3.47
alkaloid	27	263	<i>Emetic, Phlegm Heat</i>	steroidal	14	475	3.40
	29	186	<i>Astringent, Damp Heat, Cough & Wheezing, Drain Dampness</i>	protoberberine, erythrina, aporphine	4	345	2.82
	31	226	<i>Astringent, Internal Wind, Cough & Wheezing, Phlegm Cold, Toxic Heat</i>	mixed	10	355	1.87
alkaloid, ¹ simple phenolic	37	591	<i>Internal Cold, Stop Bleeding, Astringent</i>	diterpene, ¹ indole, ¹ isoquinoline, ¹ isobutylamide ¹	12	427	2.77
simple phenolic, ¹ diterpene ²	18	166	<i>Phlegm Heat, Wind Damp</i>	coumarin, ¹ jatrophane, ² abietane ²	6	487	3.94
simple phenolic	28	273	<i>Wind Cold, Internal Cold, Phlegm Cold</i>	phenol, phenylpropanoid, long-chain aromatic	2	234	2.44
	33	426	<i>Laxative, Tonify Yang, Heat (Blood)</i>	stilbene, xanthone, phenylpropanoid, phenol	66	415	0.05
	38	651	<i>Aromatic (Damp), Invigorate Blood, Regulate Qi</i>	coumarin, long-chain aromatic, phenylpropanoid, phenol, dibenzyl	6	287	2.78
lignan	6	31	<i>Wind Cold, Internal Cold, Astringent</i>	neolignan, furo- and epoxy-tetrahydrofuranoid	0	371	3.15
	10	44	<i>Tonify Yang, Wind Damp</i>	neolignan, furofuranoid	87	561	0.29
	12	45	<i>Astringent</i>	dibenzocyclooctadiene	0	434	4.10
quinone	8	70	<i>Laxative, Tonify Yang, Stop Bleeding</i>	anthraquinone	16	304	2.13
polyphenol	1	11	<i>Cough & Wheezing, Tonify Yin</i>	flavonol	0	299	2.30
	20	91	<i>Astringent, Stop Bleeding</i>	tannin, biflavonoid	27	753	2.71
	23	157	<i>Wind Heat, Tonify Blood, Damp Heat</i>	flavonoid, isoflavonoid, chalcone	55	478	0.82
	35	480	<i>Damp Heat, Tonify Qi, Tonify Yang</i>	flavonoid, isoflavonoid, chalcone	25	400	2.44
	40	831	<i>Toxic Heat, Tonify Yang, Tonify Yin, Cough & Wheezing</i>	flavone, flavonol	46	496	1.29
polyphenol, ¹ lignan ²	21	105	<i>Damp Heat, Tonify Yang</i>	isoflavonoid, ¹ pterocarpan, ¹ neolignan ²	21	392	2.40
polyphenol, ¹ quinone ²	24	167	<i>Laxative, Astringent</i>	flavan-3-ol, ¹ gallate ester, ¹ proanthocyanidin, ¹ anthraquinone ²	61	555	1.32
monoterpene	25	130	<i>Heat (Blood), Tonify Yang, Heat (Qi)</i>	iridoid, menthane, pinane	60	348	-1.04
	32	403	<i>Internal Cold, Wind Cold, Phlegm (Heart)</i>	menthane, thujane, camphane, pinane, fenchane, acyclic	0	148	2.97
sesquiterpene	15	63	<i>Phlegm Cold, Wind Damp</i>	lactones, xanthane, pseudoguaiane, eudesmane	0	326	1.47
	39	613	<i>Regulate Qi, Invigorate Blood, Aromatic (Damp)</i>	many types	0	232	3.49
diterpene, ¹ nortriterpene ²	17	75	<i>Cathartic, Toxic Heat, Astringent</i>	tricyclic, ¹ quassinoid ²	15	427	0.21
diterpene, ¹ tri-terpene ²	30	229	<i>Invigorate Blood, Wind Damp, Phlegm Heat</i>	abietane, ¹ clerodane, ¹ labdane, ¹ dammarane ²	2	340	3.08
triterpene	2	11	<i>Damp Heat, Toxic Heat</i>	quassinoid	0	430	2.19
	3	28	<i>Shen</i>	tetracyclic	3	518	3.14
	4	22	<i>Toxic Heat</i>	sterol	82	614	0.42
	5	29	<i>Tonify Qi</i>	sterol	76	643	2.06
	7	23	<i>Toxic Heat, Invigorate Blood, Tonify Yang</i>	pentacyclic	0	458	6.56
	9	61	<i>Wind Heat</i>	pentacyclic	54	618	3.35
	11	86	<i>Shen, Drain Dampness</i>	tetracyclic	0	543	3.43
	13	60	<i>Toxic Heat, Wind Heat, Phlegm Cold</i>	pentacyclic	56	827	1.36
	14	85	<i>Regulate Qi</i>	limonoid	4	588	2.81
	19	130	<i>Drain Dampness, Shen</i>	tetracyclic	2	484	4.87
	22	142	<i>Tonify Qi, Stop Bleeding, Wind Heat</i>	tetracyclic	82	723	2.01
	26	193	<i>Tonify Yin, Tonify Yang, Internal Wind, Heat (Qi)</i>	sterol	61	695	2.74
	34	561	<i>Drain Dampness, Wind Damp, Wind Heat, Toxic Heat</i>	pentacyclic	45	636	3.30

^a CN = cluster number (as in Figure 1); NC = number of compounds; glycosides (%) = % glycosidic compounds in cluster; MW = molecular weight.

Interior Cold (37). The other two clusters (29 and 31) are more heterogeneous, perhaps reflecting the structural and

pharmacological diversity of alkaloids as a whole. Of interest, however, are the protoberberine alkaloids associated, in

Table 5. Skeletal Types of Phytochemical Classes Listed in Column 5 of Table 4^a


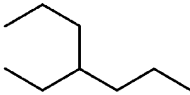
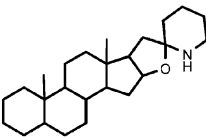
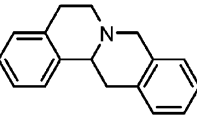
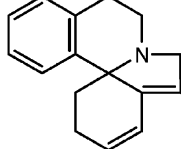
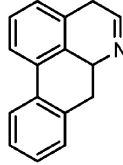
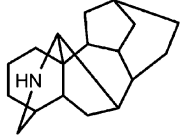
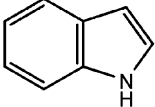
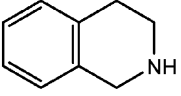
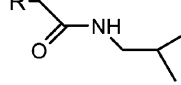
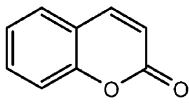
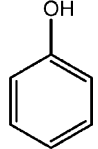
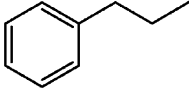
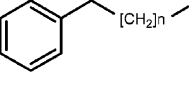
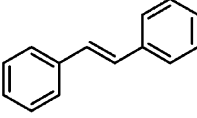
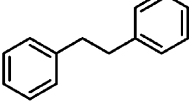
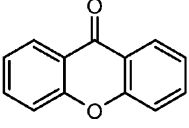
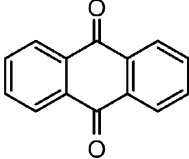
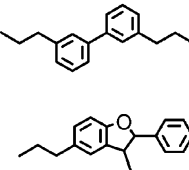
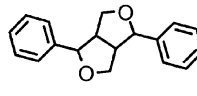
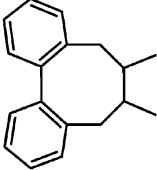
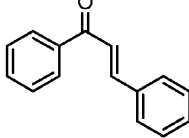
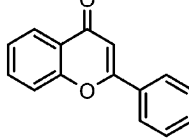
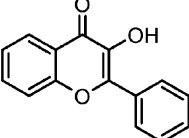
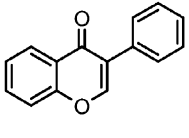
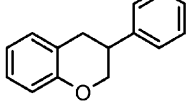
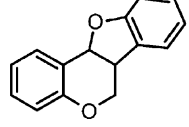
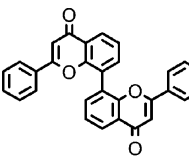
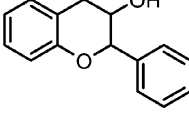
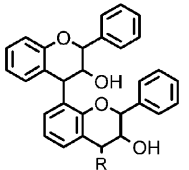
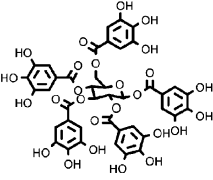
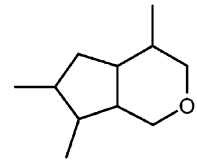
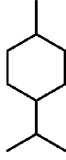
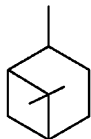
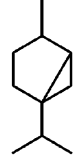
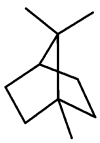
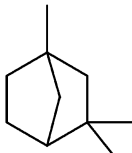
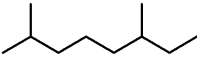
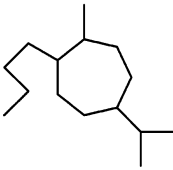
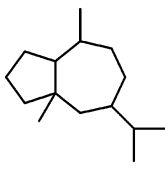
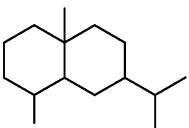
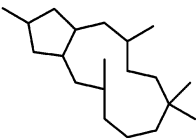
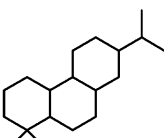
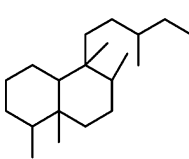
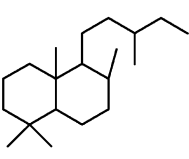
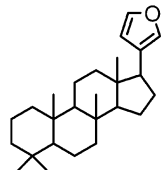
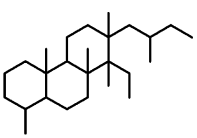
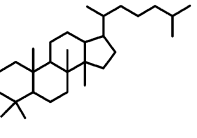
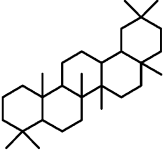
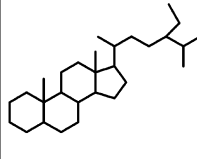
				
Acetylene (example)	Low MW aliphatic (example)	Steroidal alkaloid	Protoberberine alkaloid	Erythrina alkaloid
				
Aporphine alkaloid	Diterpene alkaloid	Indole alkaloid	Isoquinoline alkaloid	Isobutylamide (variable chain)
				
Coumarin	Phenol	Phenylpropanoid	Long chain aromatic (variable chain)	Stilbene
				
Dibenzyl	Xanthone	Anthraquinone	Neolignans (two types)	Furofuranoid lignan
				
Dibenzocyclooctadiene lignan	Chalcone	Flavone	Flavonol	Isoflavone
				
Isoflavan	Pterocarpan	Biflavonoid	Flavan-3-ol	Proanthocyanidin (variable units)
				
Hydrolyzable tannin (example)	Iridoid monoterpene	p-Menthane monoterpene	Pinane monoterpene	Thujane monoterpene

Table 5 (Continued)

 <i>Camphane monoterpene</i>	 <i>Fenchane monoterpene</i>	 <i>Acyclic monoterpene</i>	 <i>Xanthane sesquiterpene</i>	 <i>Pseudoguaiane sesquiterpene</i>
 <i>Eudesmane sesquiterpene</i>	 <i>Jatrophone diterpene</i>	 <i>Abietane diterpene</i>	 <i>Clerodane diterpene</i>	 <i>Labdane diterpene</i>
 <i>Limonoid nortriterpene</i>	 <i>Quassinoid nortriterpene</i>	 <i>Tetracyclic (dammarane) triterpene</i>	 <i>Pentacyclic (oleanane) triterpene</i>	 <i>Stigmastane sterol</i>

^a In cases where a particular class is not based on a common scaffold (e.g., acetylenes) an example is given instead.

particular, with the *Damp Heat* category. Herbs in this category are frequently used in the treatment of jaundice, dysentery, and skin disease. As far as the latter is concerned, a parallel can be found in the use of *Berberis* spp. in the West, which also contain these alkaloids, to treat similar conditions.²⁹

Antraquinones are largely restricted to one cluster (8), which is associated with *Laxative*, *Stop Bleeding*, and *Tonify Yang* categories. The laxative effects of anthraquinones are well-known,³⁰ these being found in herbs such as *Rheum* spp. (rhubarb) and *Cassia angustifolia* (senna) which are used in this way both in Chinese and Western herbal medicine.

Diterpenes are represented by two main clusters, the larger of the two (30) being strongly associated with *Invigorate Blood* and *Wind Damp* categories. Both of these categories contain herbs which are traditionally used in the treatment of pain and other symptoms, believed in TCM to result from 'stagnant Blood', and it is therefore no surprise that a number of these herbs are rich in compounds, such as diterpenes and coumarins, which show anticoagulant properties.³⁰ The other cluster (17) is made up of tricyclic diterpenes and closely related quassinoids, which show an affinity for *Cathartic*, *Toxic Heat*, and *Astringent* categories.

The remaining two major classes, polyphenols and triterpenes, are spread over a larger number of clusters, some of them virtually indistinguishable, indicating that in these instances SOM-Ward clustering has resulted in a certain degree of overprecision. Nevertheless, in most cases, clear and consistent patterns of association are evident.

As far as polyphenols are concerned, tannins, biflavonoids, and the closely related gallate esters and proanthocyanidins are primarily associated with *Astringent* and *Stop Bleeding* categories (20 and 24). The remaining clusters (1, 21, 23,

35 and 40), comprising largely flavonoids and isoflavonoids, are by contrast strongly associated with the tonic categories and also with *Wind Heat* and *Damp Heat*.

In the case of triterpenes, the sterols (4, 5, and 26) again show an affinity for the tonic categories (with the exception of *Tonify Blood*) and also for *Internal Wind*, *Heat (Qi)*, and *Toxic Heat*. In case of the tetracyclic (e.g., lanostane and dammarane) triterpenes, we can distinguish between the glycosides (22), which are associated with *Tonify Qi*, *Stop Bleeding*, and *Wind Heat*, and aglycones (3, 11, and 19), which have a greater affinity for *Drain Dampness* and *Shen* categories. The pentacyclic (e.g., oleanane and ursane) triterpenes (7, 9, 13, and 34) are meanwhile found in *Wind Heat*, *Wind Damp*, *Toxic Heat*, and *Drain Dampness* categories.

(7) SOM Training without Prior Use of Random Forest. In the Introduction, we stated that training maps using TCM profiles derived from RF as opposed to raw data is important for two reasons: (1) chemical noise associated with each herbal category is reduced, with commonly found structures rewarded and rare compounds penalized, and (2) associations between compounds found in different categories are quantified, in that compounds found in category A may still achieve a high score for category B if they are sufficiently similar to other compounds found in that category.

To investigate the differences which RF makes to subsequent training of maps, another map was constructed, using identical settings to those used in the analysis described above, except that in this case simple membership of each herbal category (either 1 if a member or 0 if not) was used instead of TCM profile.

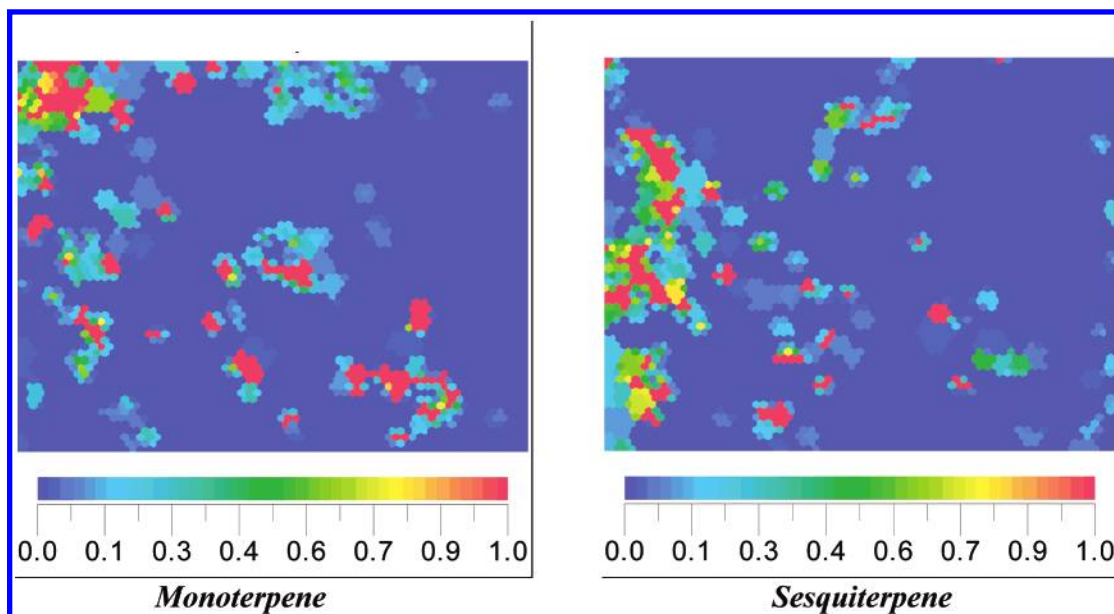


Figure 4. Self-organizing maps showing distribution of mono- and sesquiterpenes without prior transformation by RF. [For their RF transformed equivalents see Figure 2].

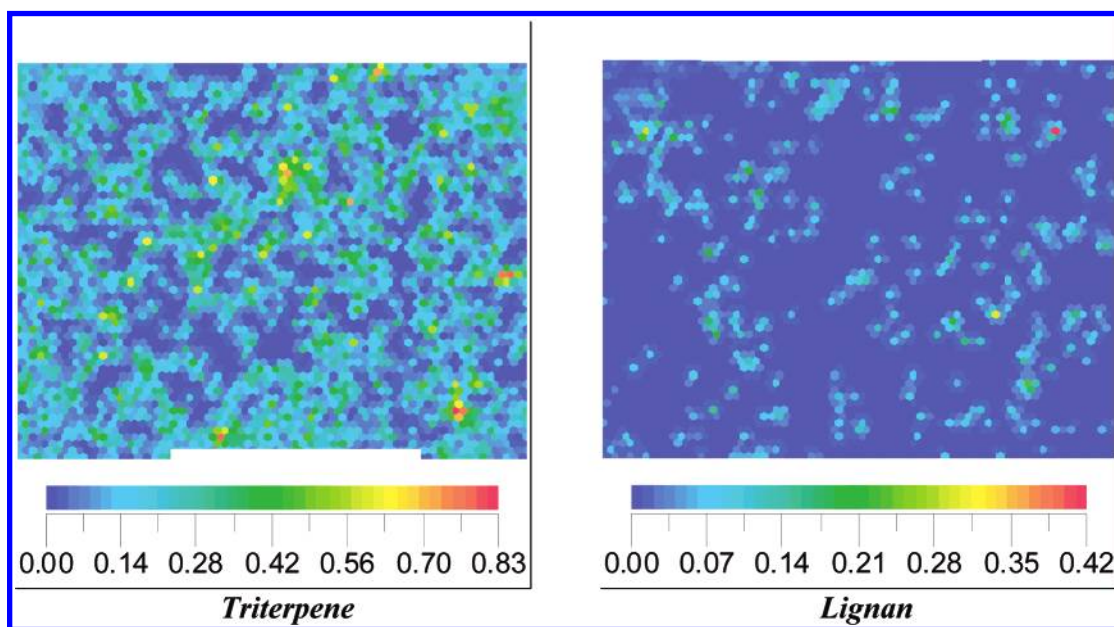


Figure 5. Self-organizing maps showing distribution of triterpenes and lignans after random scrambling of herbal categories (see text for details).

Maps for monoterpenes and sesquiterpenes are shown in Figure 4. It is clear that distribution of each class is more fragmented than in maps trained using TCM profile (Figure 2). Furthermore, each class covers a larger part of the map, and there is greater admixture with other classes, indicating that clustering and subsequent interpretation would prove difficult in many instances. This suggests that processing of data by a method such as RF is important prior to reduction by SOM and gives a clearer picture of phytochemical distribution in relation to TCM category than would otherwise be the case. Maps for the full set of classes may be found in the Supporting Information.

(8) Effects of Random Permutation of Herbal Categories on Phytochemical Distribution. In order to test the significance of the patterns revealed in Figure 2, an identical protocol was followed (SOM training on TCM profiles derived from RF), except that the herbal categories to which

each herb is assigned were randomly permuted or ‘scrambled’ prior to analysis. This results in random assignment of each herb to one of the 28 herbal categories, with the same number of herbs found in each category as before.

Scrambling results in entirely featureless maps for all classes, showing a stark contrast with the detail revealed in Figure 2. This provides striking visual evidence to reinforce the argument that a strong phytochemical basis underlies the language of TCM and demonstrates the extent to which TCM may use different phytochemical classes to achieve distinct therapeutic effects. Two maps, one for triterpenes, the other for the less common lignans, are shown in Figure 5. The others, which are all very similar, may be found in the Supporting Information.

(9) TCM Profile and Predicted Activity against Molecular Targets. In a previous paper we described the methods whereby activities of Chinese herbal constituents

	Voltage-gated Na ⁺ channel	Phospholipase A2	Estrogen aromatase	Testosterone 5 α reductase	Protein kinase C	cAMP phosphodiesterases	Protein kinase A	NO prod ⁿ in vivo	COX	LOX	Monamine oxidase	Aldose reductase	Topoisomerase II	Estrogen receptor	nACh receptor	Acetylcholinesterase	α 1-Adrenergic receptor	Dopamine receptor	HIV-1 reverse transcriptase	GABA receptor	Protein kinases (general)	HIV-1 integrase	HIV-1 protease	iNOS expression
Wind Cold																								
Heat (Def.)																								
Cathartic																								
Phlegm (Heart)																								
Heat (Qi)						S																		
Wind Damp																							T5	
Phlegm Heat	SA																							
Emetic	SA																							
Interior Cold																			L					
Toxic Heat					T5						MA									MA			2	
Phlegm Cold					T5						MA									MA				
Internal Wind						S					MA									MA				
Tonify Yin						S	F							F										
Shen													T4	T4									2	
Aromatic (Damp)									SP	SP	SP			SP										
Regulate Qi									SP	SP	SP			SP										
Invigorate Blood									SP	SP	SP			SP									T5	
Heat (Blood)											Q	SP	Q	Q							Q	Q		
Tonify Yang				F		S					Q	SP	Q	2							Q	2	T5	
Laxative											Q	2	Q	Q					T		Q	2		
Cough & Wheezing							F				MA		PA		PA	PA	PA	PA	PA	2				
Wind Heat					T5			T4					F	F						T4	F		T5	
Damp Heat		F	F	F		F		F				F	PA	F	PA	PA	PA	PA	PA	PA	F	F		
Drain Dampness												2	T4		PA	PA	PA	PA	PA	PA			2	
Tonify Qi				F				T4	A	A										T4		F		A
Stop Bleeding								T4	A	A				F			T	T	T	T4	T	T	T	2
Astringent											MA	T	PA		PA	PA	2	2	4	2	T	2	T	T
Tonify Blood									A	2				F							F	Q	F	A

A = Acetylene
 SA = Alkaloid (steroidal)
 MA = Alkaloid (Cluster 29)
 PA = Alkaloid (Cluster 31)

SP = Simple phenolic
 L = Lignan
 Q = Quinone
 F = Flavonoid
 T = Tannin

S = Sterol
 T4 = Triterpene (tetracyclic)
 T5 = Triterpene (pentacyclic)

Figure 6. Relationships between TCM and molecular target profiles for 12 phytochemical classes from Chinese herbs. [Each cell indicates the phytochemical class within that TCM category most likely to provide inhibitors against the relevant molecular target. In cases where more than one phytochemical class is involved, the number of classes is given instead].

against 10 therapeutically important molecular targets were predicted. Briefly, Random Forest was used to identify compounds expected to show activity on the basis of information contained in another database of bioactive plant compounds. The screening criteria were stringent, with over 80% consensus among trees required for a compound to be classified as active. Furthermore, it had to achieve this score for two out of three different analyses. Literature search subsequently provided evidence to support 29% of these predictions. The reader is referred to these papers for further details.^{5,15}

In this instance, we interpolated these, and additional scores for 17 other targets, onto the map and identified the major clusters into which they fell. Targets were chosen for which a reasonable number of active compounds are known (over 40). For an activity to be associated with a particular cluster, over 10% of its compounds had to be predicted as active against the target in question.

Details of the phytochemical class(es), cluster number, associated targets, potential therapeutic significance,³¹ and representative herbs from each cluster may be found in the Supporting Information. Unless otherwise stated, all compounds are assumed to be potential inhibitors.

Relationships between TCM categories and target inhibitors, taken from this table, are shown in Figure 6. Rows and columns were reordered by modal blocks so that subsets characterized by particular chemical classes are adjacent to each other.

Though the range of molecular targets indicated in Figure 6 is limited, it is noticeable that much of the chemical space in Chinese herbs is virtual 'terra incognita', in terms both of TCM categories and phytochemical classes. Four categories do not contain any predicted inhibitors of these targets, while a further five contain only class of compound.

In addition, many of the most important phytochemical classes are not featured at all. These include mono-, sesqui-, and diterpenes, small aliphatics, and a number of alkaloids, while others such as lignans and many of the smaller phenolics are poorly represented. It is therefore clear that, even for this small number of targets, exploration of the molecular basis of TCM is at a preliminary stage.

Those classes which dominate Figure 6 are, by contrast, flavonoids, tannins, anthraquinones, and some alkaloids, with a smaller contribution, in terms of predicted targets, from triterpenes and sterols.

For each molecular target there are a number of classes implicated, and these are preferentially associated with different TCM categories. Taking cyclooxygenases (COX) and lipoxygenases (LOX) as examples, predicted inhibitors are likely to be found among small phenolics in three categories, *Aromatic (Damp)*, *Regulate Qi*, and *Invigorate Blood*, whereas acetylenes are more likely to be involved in herbs from *Tonify Qi*, *Tonify Blood*, and *Stop Bleeding* categories. Similarly in the case of α 1-adrenergic and dopamine receptors, alkaloids are predicted to serve as inhibitors in herbs from *Cough & Wheezing*, *Damp Heat*, and *Drain Dampness* categories, while tannins may serve the same purpose in *Stop Bleeding* herbs.

If we consider all targets together, then it is clear that inhibitors from certain categories are more likely to belong to the same class than those from others. Thus, anthraquinones are likely to be active against several targets in herbs from *Laxative*, *Stop Bleeding*, and *Tonify Yang* categories. In the case of *Damp Heat* and *Drain Dampness* herbs, flavonoids and alkaloids are frequently implicated, while alkaloids and tannins are likely in the case of *Astringent* and *Stop Bleeding* categories.

In terms of the number of predicted inhibitors from different classes, we see that herbs from the *Astringent* category, in particular, are more likely to provide structurally varied inhibitors against some of these targets than other categories, with up to four different classes predicted to be active against HIV-1 reverse transcriptase.

Predicted activities for lignans and triterpenes are patchy in terms of TCM category. Lignans from *Interior Cold* herbs may be active against HIV-1 reverse transcriptase, while sterols associated with *Heat (Qi)*, *Interior Wind*, *Tonify Yang*, and *Tonify Yin* categories are likely to show some degree of activity against cAMP phosphodiesterases. The tetracyclic and pentacyclic triterpenes meanwhile show little overlap, either in terms of targets or TCM categories (with the exception of *Wind Heat*). The former are predicted to inhibit the GABA and estrogen receptors and to modulate nitric oxide production, with those compounds associated with *Shen*, *Wind Heat*, *Tonify Qi*, and *Stop Bleeding* preferentially involved. Pentacyclic triterpenes, by contrast, are predicted to inhibit protein kinase C and HIV-1 protease, though in each case different categories seem to be involved. In the case of the former, compounds associated with *Toxic Heat*, *Phlegm Cold*, and *Wind Heat* are indicated, while for the latter it is *Wind Damp*, *Invigorate Blood*, *Tonify Yang*, and *Wind Heat* categories which are involved.

DISCUSSION

(1) Random Forest and Noise Reduction. The results presented above indicate that RF is well suited to reducing noise in chemical data and to identifying significant signals in complex mixtures. Whereas interpretation of untransformed data in relation to TCM categories proved difficult and confusing, the use of RF to establish TCM profiles of individual compounds, prior to reduction with SOM, resulted in distribution patterns where individual phytochemical classes were well delineated and, for the most part, separable into meaningful clusters. Furthermore, subsequent analysis showed that these clusters could be discriminated using very simple descriptors.

Noisy, complex, and incomplete data are of considerable significance in many data mining applications, and RF may thus offer a simple and versatile tool for signal enhancement prior to other forms of analysis. In this instance, such analysis relies heavily on two features: (1) a sufficient volume of data to identify large-scale pattern and (2) robust methods with the ability to discriminate between large numbers of structurally diverse compounds. For each TCM category, RF managed to distinguish approximately 80% of the constituents, while simultaneously discriminating against those compounds atypical for that category and in favor of constituents in other categories with strong structural similarities. This resulted in TCM profiles where similarities between different categories were highlighted in a way difficult or impossible to achieve without such prior transformation.

(2) The Phytochemical Basis of TCM. TCM profiles can serve as a suitable starting point for many forms of investigation, including more detailed QSAR studies on particular groups of compounds and further work on the relationships between different phytochemical classes. With regard to the latter, our analysis, though preliminary, suggests that many classes of compounds are distributed in a highly significant fashion which is furthermore relatively easy to interpret in terms of the essential components of TCM herbal therapy. This should, in turn, lead to a better understanding of the ways in which such classes are combined within a traditional medical paradigm and provides a 'virtual' approach to mapping the landscape of potential synergistic interactions.

As is well-known, Chinese medicine seldom treats a particular condition (such as a migraine, for example) in an identical fashion irrespective of a patient's individual 'constitution', and one of the skills of TCM diagnosis is in finding the appropriate herbal formula to simultaneously treat both the symptom(s) and the underlying 'imbalance' in each patient.²⁵ As a result, a single formula may be used to treat a wide variety of conditions, while conversely, any one condition may be treated in several different ways, or as TCM more succinctly expresses it, 'Tong Bing Yi Zhi, Yi Bing Tong Zhi' (same disease different treatment, different disease same treatment). Distributional analysis of the type attempted here offers a way of understanding the *variety* of pharmacological strategies which may be used to treat a particular condition and in customizing treatment on an individual basis.

(3) TCM and Target-Based Therapy. Another potential use concerns the relationships between information derived from experimental or virtual screening of activity against particular molecular targets and ethnopharmacological data. Bringing the two together is often fraught with difficulty, though insights into such relationships are of potential benefit. TCM profiles of individual compounds provide a possible link between these two worlds, whereby the pharmacology of a compound can be more readily understood in terms of traditional usage.

Deficiencies in our knowledge are also more readily appreciated when such relationships are made apparent. In this instance it was demonstrated that little is known about expected specificities for those targets explored, both in terms of many TCM categories and also for some of the most important phytochemical classes found in Chinese medicine.

Nevertheless, despite these limitations, the information we have may still prove useful in understanding ways in which traditional knowledge of herbs can be combined with information on target specificities in the treatment of certain conditions. As an example, we might consider the herbal treatment of HIV infection. From Figure 6 we see that, on the basis of current information, we would expect to find inhibitors of integrase, protease, and reverse transcriptase in herbs from as many as 16 TCM categories. Furthermore the classes involved include lignans, alkaloids, triterpenes, quinones, and polyphenols (both flavonoids and tannins). This suggests that, *if* such compounds prove useful as prophylactics against immune system dysfunction, then Chinese herbs constitute a significant reservoir of potential inhibitors, and that a wide range of herbal formulas, customized according to individual presentation, might prove effective as a part of treatment.

(4) Ethnopharmacology and Pharmaceutical Prospecting. Mining of traditional herbal texts has sometimes been suggested as a means of discovering new drug leads.³² While promising, such an approach can suffer from a number of drawbacks. Among these are the following: (1) language which is difficult to translate into modern terminology; (2) imprecision and lack of consistency in describing the uses of herbs; (3) lack of information on chemical constitution, and (4) difficulties over identification and taxonomic status.

While the language of TCM may, at first sight, appear mysterious and arcane, closer acquaintance reveals that many categories are amenable to translation into Western terminology, as Table 1 illustrates. Furthermore, as already mentioned, Chinese herbs are among the best characterized of all plants chemically, and there are seldom difficulties over their taxonomic status, particularly for those in widespread use. Last, there is usually close agreement between different pharmacopoeias on the uses of particular herbs.

If, in addition, analysis of the relationships between plant chemistry and traditional usage reveals strong and consistent patterns, then this information may be used to detect regions of chemical space which are likely to be of value in searching for new leads and in suggesting new chemical classes for drug discovery.

A common objection to natural products as leads in drug discovery concerns their apparent complexity and intractability to chemical modification. This perception, however, may owe as much to historical factors as to natural product chemistry. Where phytochemicals have played a role in drug discovery, it is frequently the structurally more complex compounds that have been instrumental in developing new medicines, recent examples being artemisinin and taxol.³³ Yet, many of the most important classes in Chinese medicine are relatively simple structurally and represent as yet untapped potential in the search for new drugs. As shown above, little is known about target specificities of most of these, including mono-, sesqui-, and diterpenes, lignans, and many phenolics. Furthermore, as this analysis demonstrates, it is precisely these compounds which show strong associations with particular TCM categories, so that their possible roles in traditional herbal therapy may be readily inferred, and this information can then be used to explore their pharmacological potential in more depth. This is likely to be of benefit both in the search for new drugs and, equally significantly, in understanding how Chinese medicine works.

CONCLUSION

Regarded little more than 50 years ago as an archaic system of practice doomed to extinction, Chinese medicine has in recent times undergone a remarkable renaissance, first in China and the Far East, and more recently in the West, where its introduction may expand its potential applications into new and uncharted territory.

Given its long history, TCM is among the most developed and mature systems of alternative medicine, and closer scrutiny by Western science may prove important in finding new solutions to many contemporary health problems. One of its undoubted attractions lies in its apparently 'holistic' approach to human physiology. As shown by the rise of systems-based approaches in many areas of science, there is widespread interest in understanding the behavior of whole systems, though the challenges remain daunting,³⁴ and in this respect Chinese medicine offers a new and appealing perspective.

The language in which traditional systems of medicine are couched has often proved a stumbling block in the past and is frequently regarded as little more than a series of colorful and imaginative metaphors.³⁵ The results presented here suggest, on the contrary, that many of these 'metaphors' may have a well-defined basis in plant chemistry. It should be appreciated, however, that this type of distributional analysis is intended as a means of exploring patterns of association which merit further study. Informatics is primarily intended to build hypotheses, not test them.

Supporting Information Available: Maps for each phytochemical class trained using original data without prior construction of TCM profiles, and resulting from random permutation of herbal categories; a table containing information on predicted targets associated with each cluster shown in Figure 1, in addition to their possible therapeutic applications, and representative herbs; and details of the SOM-Ward clustering algorithm. This material is available free of charge via the Internet at <http://pubs.acs.org>.

REFERENCES AND NOTES

- (1) Qiao, X. B.; Hou, T. J.; Zhang, W.; Guo, S. L.; Xu, X. J. A 3D Structure Database of Components from Chinese Traditional Medicinal Herbs. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 481–489.
- (2) *Traditional Chinese Medicines 2.1*; CambridgeSoft Inc.: Cambridge, MA, U.S.A.
- (3) *China Natural Products Database*; NeoTrident Technology Ltd.: Beijing, P.R. China.
- (4) Chen, X.; Zhou, H.; Liu, Y. B.; Wang, J. F.; Li, H.; Ung, C. Y.; Han, L. Y.; Cao, Z. W.; Chen, Y. Z. Database of traditional Chinese medicine and its application to studies of mechanism and to prescription validation. *Br. J. Pharmacol.* **2006**, *149*, 1092–1103.
- (5) Ehrman, T. M.; Barlow, D. J.; Hylands, P. J. Phytochemical Databases of Chinese Herbal Constituents and Bioactive Plant Compounds with Known Target Specificities. *J. Chem. Inf. Model.* **2007**, *47*, 254–263.
- (6) Wiseman, N. Traditional Chinese Medicine: A Brief Outline. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 445–455.
- (7) Bensky, D.; Clavey, S.; Stöger, E. *Chinese Herbal Medicine: Materia Medica*; Eastland Press: Seattle, WA, 2004.
- (8) Bensky, D.; Barolet, R. *Chinese Herbal Medicine: Formulas & Strategies*; Eastland Press: Seattle, WA, 1990.
- (9) Hsu H.-Y.; Chen, Y.-P.; Hong M. *The Chemical Constituents of Oriental Herbs*; Oriental Healing Arts Institute: Taiwan and Los Angeles, 1982–1983; Vols I and II.
- (10) Williamson, E. M. Synergy and other interactions in phytomedicines. *Phytomedicine* **2001**, *8*, 401–409.
- (11) *Dictionary of Natural Products* (CD-ROM versions); Chapman & Hall/CRC: Boca Raton, FL, U.S.A.
- (12) Breiman, L. Decision Tree Forests. *Machine Learning* **2001**, *45*, 5–32.

- (13) Svetnik, V.; Liaw, A.; Tong, D.; 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.
- (14) Selzer, P.; Ertl, P. Applications of Self-Organizing Neural Networks in Virtual Screening and Diversity Selection. *J. Chem. Inf. Model.* **2006**, *46*, 2319–2323.
- (15) 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.
- (16) Hsu, H.-Y. *Oriental Materia Medica: A Concise Guide*; Keats Publishing Inc.: Connecticut, 1986.
- (17) Zhu, Y.-P. *Chinese Materia Medica: Chemistry, Pharmacology & Applications*; Harwood Academic: Amsterdam, The Netherlands, 1998.
- (18) Labute, P. A widely applicable set of descriptors. *J. Mol. Graphics Modell.* **2000**, *18*, 464–477.
- (19) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868–873.
- (20) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*; Chapman & Hall/CRC: Boca Raton, FL, 2000; Chapter 4, pp 103–104.
- (21) Kohonen, T. The self-organizing map. *Proc. IEEE* **1990**, *78*, 1464–1479.
- (22) Kastberger, G.; Kranner, G. Visualization of multiple influences on ocellar flight control in giant honeybees with the data-mining tool Viscovery SOMine. *Behav. Res. Methods Instrum. Comput.* **2000**, *32*, 157–168.
- (23) Kranner, G. The SOM-Ward Cluster Method. *PASE* **2000**, April, 2–5.
- (24) Hartigan, J. A. Modal Blocks in Dentition of West Coast Mammals. *Syst. Zool.* **1976**, *25*, 149–160.
- (25) Maciocia, G. *The Foundations of Chinese Medicine*; Churchill Livingstone: Edinburgh, 1989.
- (26) Maciocia, G. *The Practice of Chinese Medicine*; Churchill Livingstone: Edinburgh, 1994; p 710.
- (27) Duke, J. A. *Handbook of Biologically Active Phytochemicals and Their Activities*; CRC Press: Boca Raton, FL, 1992; p 70.
- (28) Plotnikoff, G. A.; McKenna, D.; Watanabe, K.; Blumenthal, M. Ginseng and Warfarin Interactions. *Ann. Intern. Med.* **2004**, *141*, 893–894.
- (29) Mills, S.; Bone, K. *Principles and Practice of Phytotherapy*; Churchill Livingstone: Edinburgh, 2000; Part 3, pp 286–295.
- (30) Pengelly, A. *The Constituents of Medicinal Plants*; CABI Publ.: Wallingford, 2004; Chapter 2, pp 21–23, Chapter 4, pp 48–50, Chapter 5, pp 64–66.
- (31) Chen, X.; Ji, Z. L.; Chen, Y. Z. TTD: Therapeutic Target Database. *Nucl. Acids Res.* **2002**, *30*, 412–415.
- (32) Holland, B. K. In *Prospecting for Drugs in Ancient and Mediaeval European Texts: A Scientific Approach*; Holland, B. K., Ed.; Harwood Academic: Amsterdam, The Netherlands, 1996; Chapter 1, pp 1–6.
- (33) Simmonds, M. S. J.; Grayer, R. J. In *Chemicals from Plants: Perspectives on Plant Secondary Products*; Walton, N. J., Brown, D. E., Eds.; World Scientific: Singapore, 1999; Chapter 5, pp 215–245.
- (34) Wolkenhauser, O.; Mesarovic, M.; Wellstead, P. A plea for more theory in molecular biology. *Ernst Schering Res. Found. Workshop* **2007**, *61*, 117–137.
- (35) Sneader, W. *Drug Discovery: A History*; Wiley: Chichester, 2005; Chapters 2–6, pp 8–40.

CI700155T