LeadScope†: Software for Exploring Large Sets of Screening Data

Gulsevin Roberts, Glenn J. Myatt, Wayne P. Johnson, Kevin P. Cross, and Paul E. Blower, Jr.*

LeadScope, Inc., 1275 Kinnear Road, Columbus, Ohio 43212

Received June 17, 2000

Modern approaches to drug discovery have dramatically increased the speed and quantity of compounds that are made and tested for potential potency. The task of collecting, organizing, and assimilating this information is a major bottleneck in the discovery of new drugs. We have developed LeadScope a novel, interactive computer program for visualizing, browsing, and interpreting chemical and biological screening data that can assist pharmaceutical scientists in finding promising drug candidates. The software organizes the chemical data by structural features familiar to medicinal chemists. Graphs are used to summarize the data, and structural classes are highlighted that are statistically correlated with biological activity.

INTRODUCTION

In an effort to accelerate the drug discovery process pharmaceutical companies have dramatically increased the number of potential drugs that are made and tested. Through combinatorial chemistry, high throughput screening, and an ever increasing number of disease targets, pharmaceutical companies are now generating volumes of screening data that are orders of magnitude greater than they have experience handling.

To find and optimize a single compound that will become a drug requires information about biological targets, chemical structure and property data, and biological response data from multiple receptors—results from thousands of experiments. The task of collecting, organizing, and assimilating this information to effectively formulate hypotheses and design experiments is a major bottleneck in the discovery of new drugs. Software for analyzing and interpreting chemical and biological test results has not kept pace.

A wide variety of structural descriptors have been defined and used in QSAR,[1] similarity search,[2] clustering,[3] and diversity analysis programs[4] to relate a compound's molecular constitution to biological activity and other physical properties. These include generalized atom-pairs,[5,6] molecular fingerprints,[7] substructure search screens, two-dimensional and three-dimensional shape descriptors,[8−10] physiochemical property descriptors,[1,11] partial atomic charges, topological indices, and other graph theoretical descriptors (see lists in refs 2, 4, 12, and 13). As Cosgrove[14] and Willett point out, the molecular descriptors used for correlations, clustering, and the like are often abstract, theoretical constructs that are difficult for medicinal chemists to understand and visualize. More importantly, it is difficult to use the results from QSAR or clustering programs to decide what compounds to synthesize or test next.

There have been a number of recent reports of computer programs for analyzing compound sets in terms of molecular descriptors that are more meaningful to medicinal chemists. The SLASH program[14] analyzes compounds in terms of the functional groups they contain. Three classes of fragments are recognized: rings, groups, and chains. Each class is organized as a hierarchy which increases in atom-bond specificity as one descends the hierarchy. It uses an algorithmic approach based on a set of fragmentation rules rather than a database of predefined fragments. The program can calculate fragment weights or scores allowing the user to judge how fragments correlate with activity.

Boyd et al. developed HookSpace,[15] a program that analyzes the geometric relationship between pairs of functional groups in a molecule. The main purpose was to assess the structural diversity of a database and to highlight the differences between two datasets.

Bemis and Murcko[16] analyzed the Comprehensive Medicinal Chemistry database[17] in terms of four general shape descriptors: ring systems, linker groups, side chains, and framework. They performed the analysis at two levels: graph only and graph plus atomic properties.

Lewell et al. developed the RECAP program[18] to identify common building blocks or *privileged substructures* in biologically active molecules. The program fragments a molecule at 11 predefined bond types chosen because they can be formed by established combinatorial syntheses. They used RECAP to fragment the World Drug Index[19] and analyzed the high-frequency fragments by therapeutic class.

The Stigmata program, developed by Blankley et al.,[20] identifies common structural features in chemically diverse datasets using a *modal fingerprint*. This is a bit string, derived from Daylight fingerprints,[7] which encodes common bits found in molecular fingerprints of the input dataset. The bits encode atom paths and must be present in a certain percentage of compounds to be included in the modal fingerprint. A semiautomated procedure maps bits in the modal fingerprint back to the input structures and color-codes common features

Sheridan and Miller[21] have developed a method for identifying and visualizing common, high scoring, topological structures among a pairs of active compounds. The score is determined by comparing the size of the common

† LeadScope is a trademark of LeadScope Inc., Columbus, OH.

* Corresponding author phone: (614)442-8373; e-mail: pblower@leadscope.com.

10.1021/ci0000631 CCC: $19.00
Published on Web 09/13/2000

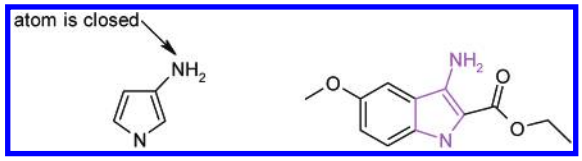


Figure 1. Query structure for *pyrrole, 3-amino(NH2),* and a matching substructure.

substructure with the expected size estimated for a pair of compounds selected at random from the MACCS−II Drug Data Report.[17]

Recently, Varmuza and Scsibrany[22a] described the *substructure isomorphism matrix*. This is a binary $n \times p$ matrix corresponding to n target molecules and p query substructures where the matrix entries indicate presence or absence of the query substructure in the target molecule. The authors describe a method for constructing a structural hierarchy[22b] by using the same set of compounds as both target and query structures.

We set out to develop a software package linking chemical and biological data and allowing medicinal chemist to visualize and interactively explore large sets of chemical compounds, their properties, and biological activities. Our intention was to provide a tool that would allow the medicinal chemist to get more information out of available screening data by focusing on the most statistically significant structural features.

LeadScope is novel computer software linking chemical and biological data that allows medicinal chemists to visualize and interactively explore large sets of chemical compounds, their properties, and biological activities. Chemical structures are organized in a large taxonomy of familiar structural features such as functional groups, aromatics, and heterocycles, each combined with common substituents— the common building blocks of medicinal chemistry. We begin by describing the structural feature hierarchy and software components.

THE STRUCTURAL FEATURE HIERARCHY

We have developed software for systematic substructural analysis of a compound set using predefined structural features stored in a template library. The structural features chosen for analysis are motivated by those typically found in small molecule drug candidates: aromatics, heterocycles, spacer groups, simple substituents.[1,11,16] At the present time, the feature library contains approximately 27,000 structural features.

All features in the hierarchy are substructures; they are defined as queries and recognized using substructure search techniques. Any open position in the query structure may be substituted by any atom in the matching structure unless specifically prohibited by an atom or bond restriction in the query. This is illustrated in Figure 1 for the query structure defining *pyrrole, 3-amino(NH2).* An atom restriction on the amine substituent prohibits further substitution of that atom. However, all ring positions may be substituted by any atom. In particular, the pyrrole may be part of a larger indole ring.

Each structural feature is assigned a chemical name generally based on the systematic nomenclature developed by Chemical Abstracts.[23] However, some differences and ambiguities arise because the feature templates represent substructures which may contain generic atoms or bonds.

Table 1. Major Structural Classes

amino acids	heterocycles
bases, nucleosides	naphthalenes
benzenes	natural products
carbocycles	peptidomimetics
carbohydrates	pharmacophores
elements	protective groups
functional groups	spacer groups

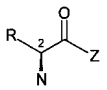
benzene, 1-,2-subst	(level 1)
benzene, 1-R-,2-alkoxy-	(level 2)
benzene, 1-acetamido-,2-alkoxy-	(level 3)
benzene, 1-acetoxy-,2-alkoxy-	
benzene, 1-acetyl-,2-alkoxy-	
benzene, 1-alkenyl-,2-alkoxy-	
benzene, 1-alkoxy-,2-alkyl-	
benzene, 1-alkoxy-,2-alkylamino-	
benzene, 1-alkoxy-,2-alkylthio-	
benzene, 1-alkoxy-,2-alkynyl-	
benzene, 1-alkoxy-,2-amino-	
benzene, 1-alkoxy-,2-amino(NH2)-	

Figure 2. Portion of the *benzene, 1-,2-subst* hierarchy.

Most structural features are named as a parent plus substituent prefixes (radicals) written inverted for sorting. Generally, positions with unspecified substituents are open and may be substituted with any atom. In cases where we want to restrict this to C or H attachments we specify this by an annotation in the name; see Figure 1.

The features are arranged in a hierarchy with the major structural classes listed in Table 1. Each of these classes is described in more detail below.

Amino Acids. This hierarchy contains the 20 naturally occurring amino acids plus homocysteine, homoserine, isovaline, and ornithine in D-, L-, and no stereo forms. The generalized substructure (L-form shown) is



Z matches N,O,S. The N has only single bonds and no bonds to heteroatoms. Atom 2 is closed to further substitution as are all carbons of the R-group. However, heteroatoms in the R-group, such as the side chain hydroxyl in serine, may be substituted

Bases, Nucleosides. This hierarchy contains the purine and pyrimidine bases *adenine*, *cytosine*, *guanine*, *thymine*, and *uracil*; their N-glycosyl derivatives; and the nucleosides *inosine* and *xanthosine*. For the nucleosides, both the D-*ribo* and *2′-deoxy*-D-*ribo* glycosides are included plus *No Stereo* analogues for each.

Benzenes. This hierarchy contains five major subcategories. The *Benzenes, 1,2-subst* class contains 1,2-disubstituted benzene rings with a wide variety of substituents. As with most features in the hierarchy, the benzene may be further substituted and embedded in a larger ring system such as a naphthalene. This class in the hierarchy is redundant. For example, *benzene, 1-amino, 3-chloro* can be found under *benzene, 1-R, 3-amino,* and *benzene, 1-R, 3-chloro.* A portion of the *Benzenes, 1,2-subst* class is shown in Figure 2. The *Benzenes, 1,3-subst* and *Benzenes, 1,4-subst* classes are completely analogous

Table 2. Main Functional Group Classes

acid anhydrides	guanidines	phosphorus groups
acid halides	halides	quinones
alcohols	hydrazines	silicon groups
aldehydes	hydroxylamines	sulfides
alkenes	imines	sulfonamides
alkynes	iminomethyls	sulfonates
allenes	isocyanates	sulfones
amidines	isonitriles	sulfonic acids
amines	ketones	sulfonyl groups
azides	mercaptans	sulfonyl halides
boron groups	misc nitrogen groups	sulfoxides
carbamates	misc oxygen groups	thiocarboxamides
carbonyls	misc sulfur groups	thiocarboxylates
carboxamides	nitriles	thiocarboxylic acids
carboxylates	nitro	thioxomethyls
carboxylic acids	nitroso	ureas
ethers	organometals	

The *Benzenes, substituents* class contains substituted benzene rings with an even wider variety of substituents. Open positions on the ring may bear other substituents. The *Benzenes, substitution patterns* class contains benzene rings substituted by two types of atoms: any acyclic atom and any fused ring atom. All possible combinations and positional variations with 1−4 substituents are included. Unlike most features in the hierarchy, open positions on the ring are substituted by hydrogen. Thus, any unsubstituted phenyl ring would be in the *benzene, monosubst* class.

Benzene-containing groups are found as substituents in the other major structural categories; for example, *ketone, phenyl-* under *Functional Groups:ketones* and *pyridine, 3-phenoxy-* under *Heterocycles:pyridine*.

Carbocycles. This hierarchy contains common all-carbon ring systems such as *adamantane*.

Carbohydrates. This hierarchy contains a large variety of 4-, 5-, and 6-carbon monosaccharides and has five major subbranches: *furanoses, pyranoses, pentoses, hexoses*, and *inositols*. In all cases, the OH substituents have been replaced by a Z (which matches {N,O,S}) to recognize the common amino and thio analogues of sugars. Each feature is represented in D-, L-, and no stereo forms.

Elements. This hierarchy contains features of the form ⟨*E*⟩-A, where A is any atom, ⟨*E*⟩ is taken from {As, B, M, P, Se, Si, Te}, and *M* represents any metal atom. The primary intent of this hierarchy is to provide a convenient way to eliminate compounds with undesirable elements.

Functional Groups. The main functional group classes are listed in Table 2. Some functional groups do not appear as class headings but are listed under other classes. For example, *phosphate, phosphine oxide, phosphite, phosphonamide,* etc. are listed under *phosphorus groups;* and *sulfenic acid and deriv*s*., sulfinate ion, sulfinic acid and deriv*s*., sulfuric acid ester,* etc. are listed under *misc sulfur groups.*

Heterocycles. The *Heterocycles* class contains nearly 60% of all features in the LeadScope feature hierarchy and is by far the largest class. Table 3 lists the major classes in the *Heterocycles* branch. These classes can be divided into two major categories: rings with substituents and other rings. All of the substituted ring classes are organized as illustrated for *pyridine* in Figure 3. At the top level of the class is the parent ring (e.g., *pyridine*). At level 2 are ring parents with a generic "R" group in the various substituent positions. At level 3 are rings with specific substituents at appropriate ring positions. Every open ring position corresponds to a branch in the hierarchy.

Table 3. Main Heterocycle Classes

azepine	indole	quinoline
azetidine	isoindole, 1,3-dioxo	quinoline, 2-oxo
aziridine	isoindole, 1-oxo	quinoline, 4-oxo
benzimidazole	isoquinoline	quinoxaline
benzimidazole, 2-oxo	isothiazole	rings size 3−4 N+Z
1,4-benzodiazepine	isothiazolidine	rings size 4−7 O+S
1,4-benzodioxin	isoxazole	z-rings size8−14
1,3-benzodioxole	isoxazolidine	3,4-ring systems
benzofuran	morpholine	spiro amines
benzopyran	1,2,3-oxadiazole	spiro aminoethers
benzopyran, 2-oxo	1,2,4-oxadiazole	spiro ethers
benzopyran, 4-oxo	1,2,5-oxadiazole	spiro lactams
benzopyrazole	1,3,4-oxadiazole	spiro lactones
5,1-benzothiazepine, 2-oxo	1,3-oxazepine	1,2,3,4-tetrazine
1,4-benzothiazine	1,4-oxazepine	1,2,3,5-tetrazine
1,2-benzothiazole	1,2-oxazine	1,2,4,5-tetrazine
1,3-benzothiazole	1,3-oxazine	tetrazole
1,2-benzothiazole, trioxo	1,4-oxazine	1,2,4-thiadiazine, dioxo
benzothiophene	oxazole	1,2,3-thiadiazole
1,4-benzoxazine	oxazolidine	1,2,4-thiadiazole
1,2-benzoxazole	oxepin	1,2,5-thiadiazole
1,3-benzoxazole	oxetane	1,3,4-thiadiazole
beta lactam	oxolane	thiane(H)
bicyclic amines	piperazine	1,3-thiazepine
1,2-diazepine	piperidine	1,4-thiazepine
1,3-diazepine	pteridine	1,2-thiazine
1,4-diazepine	purine	1,3-thiazine
1,3-diazepine, 2-oxo	purine, 2,6-dioxo	1,4-thiazine
1,2-diazine(H)	pyran(H)	thiazole
1,3-diazine(H)	pyrazine	thiazolidine
1,3-dioxane	pyrazine(H)	thiepin
1,4-dioxane	pyrazole	thietane
1,3-dioxolane	pyrazolidine	thiolane
1,4-dithiane	pyridazine	thiophene
1,3-dithiolane	pyridine	1,2,3-triazine
episulfide	pyridine(H)	1,2,4-triazine
epoxide	pyridine, 1,4-dihydro	1,2,4-triazine(H)
furan	pyrimidine	1,3,5-triazine
5,5-fused N-rings	pyrimidine, 2,4-dioxo	1,3,5-triazine(H)
5,6-fused N-rings	pyrrole	1,2,3-triazole
6,6-fused N-rings	pyrrolidine	1,2,4-triazole
imidazole	pyrrolidine, 2-oxo	1,3,4-triazole
imidazolidine	quinazoline	1,2,3-triazolidine
indazole	quinazoline, 4-oxo	1,2,4-triazolidine

For the most common 5- and 6-member heterocycles, the hierarchy contains more than one bond variant: an aromatic version (e.g., *pyrimidine, furan*) and a nonaromatic version (e.g., *1,3-diazine(H), oxolane*). The nonaromatic version will match any bond variation, saturated or partially saturated, except aromatic. The *(H)* in ring names like *1,3-diazine(H)* indicate that the ring is at least partially saturated.

For the 3-, 4-, and 7-membered heterocycles and less common 5- and 6-member heterocycles, the hierarchy contains only one bond variant. Bonds will match any bond variation, saturated, partially saturated, or aromatic. The hierarchy contains the most common 5,6-, and 6,6-fused bicyclic heterocycles. There is only one bond variation, usually aromatic.

Because they occur frequently in files of known drugs such as the World Drug Index[19], the hierarchy contains the some monocyclic and bicyclic heterocycles with *oxo* substituents. There is only one bond variation: constituent

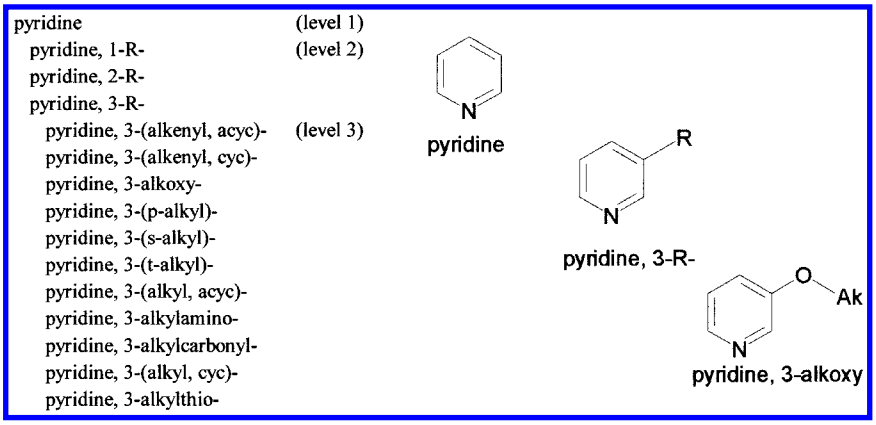


Figure 3. Portion of the pyridine hierarchy.

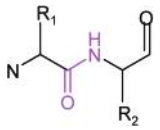


Figure 4. Amide Bond Ψ[CONH].

benzene rings are aromatic; the heterocycles match any bond variation.

Naphthalenes. The hierarchy contains 1- and 2-substituted naphthalene rings with a wide variety of substituents. As with most features in the hierarchy the naphthalene may be further substituted and embedded in a larger ring system.

Natural Products. In the current version of the feature hierarchy, the *Natural Products* class only contains one subclass *steroids*. Other natural product classes such as alkaloids, antibiotics will be added in the future releases.

Peptidomimetics. In the current version, the *Peptidomimetics* class only contains one subclass *amide bond mimetics*. The features are dipeptide structures containing the amide bond surrogates (the highlighted substructure in Figure 4) cited in Sawyer[24] and named using the Ψ-convention. Other *Peptidomimetics* classes will be added in the future releases.

Pharmacophores. The features grouped under the class *pharmacophores* are pairs of generalized physiochemical atom types joined by a path of atoms/bonds of indeterminant type with path length 3−8. There are five physiochemical atom types: aromatic, hydrogen bond donor/acceptor, and positive/negative charge. These features are very similar to the *binding property pairs* of Kearsey et al.[6] However, we used the atom type definitions of Greene and Wang,[9−10] modified as suggested by Dr. Jens Loesel.[25]

Protective Groups. Features contained under *Protective Groups* are the IUPAC recommended blocking groups for heteroatoms.[26] All protective groups have a Q atom at the point of attachment which will match any of {N, S, O, P, Cl, Br, F, I, Si, Se, B}.

Spacer Groups. Many "drug-like" molecules have chains of carbons separating important structural features such as heterocyclic rings and functional groups; for example, see ref 16. The class of *Spacer Groups* is based on this concept. Spacer group classes are organized by chain length and named as *1*,*n*-derivatives of normal alkanes, from methane through hexane. Atoms in the "spacer" are acyclic, unfunctionalized carbons; i.e., saturated carbons with no heteroatom attachments. The *Pharmacophores* class is based on a related concept but with more generic structural features.

THE SOFTWARE

LeadScope comprises an analysis subsystem that prepares data for exploration and a user interface that allows users to visualize and manipulate the compiled information.

User Interface. LeadScope provides powerful pictorial representations[27,28] and dynamic querying[29] capabilities making it easier to interpret the complex data in large structure−activity datasets. It presents a high level view of the whole data set organized in a large hierarchy of familiar structural features and graphically summarized using 2D histograms and scatter plots. The user can dynamically construct a series of detailed cross sections of the dataset by expanding more specific structural classes, creating subprojects, and using interactive controls for properties.

As illustrated in Figure 5a, the user interface comprises several coordinating panels on the computer screen. The left panel contains the feature hierarchy, a series of structural features arranged in a class hierarchy described above. The graphic panel in the middle shows a graph of the contents of the underlying compound set relative to the structural features in the left panel. Histogram bars give frequencies of each class in the data set (see Databases below) plotted on a log scale.

Sliders are used to focus on interesting property ranges. The filter panel on the right contains a series of two-ended sliders[28] each corresponding to a property such as molecular weight that allows the user to dynamically adjust the members of the underlying compound set. As the sliders are adjusted to remove compounds with undesirable property values, the graphs are dynamically redrawn to show the effect of the property changes on the whole data set as well as individual subsets. In Figure 5a, the property filters were adjusted to select compounds with "drug-like" properties.[30] The user can import any property with numerical values, and LeadScope will create an interactive slider for it.

The legend panel provides controls for color-coding the histogram to show how structural features are statistically correlated with property data such as biological activity. Histogram bars and scatterplot cells are color-coded based on the *feature z-score* which is the difference—expressed in standard deviations (NSD)[31a]—between the mean activity of the subset of compounds containing a structural feature and the mean activity of the full set. For example, the dark red bars in Figure 5a indicate that the mean activity[32] is more the 5 NSD ($z > 5.0$) from the mean activity of the full cancer screening set (the expected value). Features with gray bars

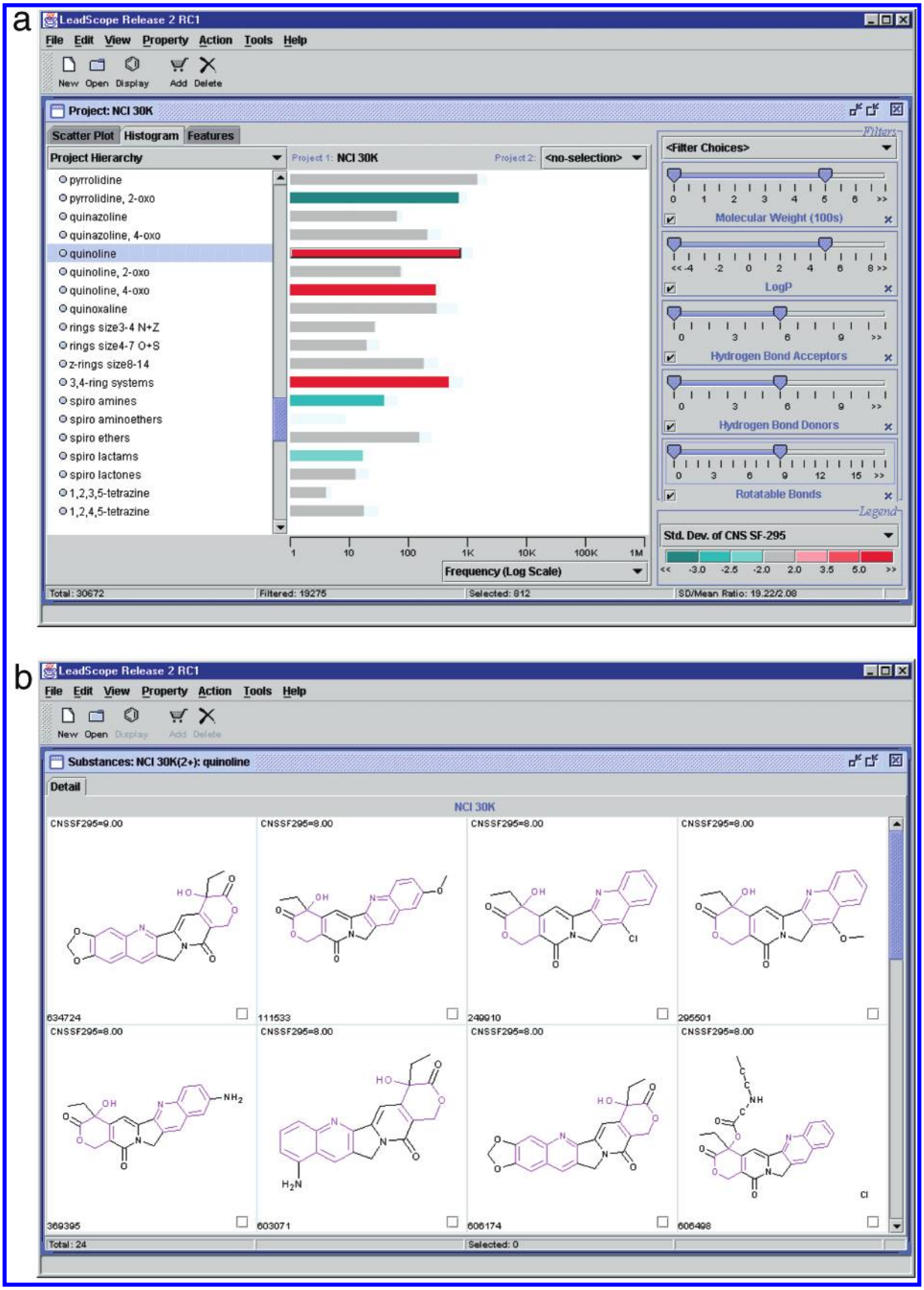


Figure 5. (a) The LeadScope the user interface showing the histogram view. The left panel shows the structural feature hierarchy open to reveal a portion of the *Heterocycles* branch with *quinoline* selected; the histogram in the middle panel gives frequencies of each feature class in the data set plotted on a log scale; the right panel contains a series of the property filters adjusted to select compounds with "drug-like" properties; and at the bottom, four information windows give useful statistics: 30,672 compounds in the full NCI data set, 19,275 compounds satisfy the all filter settings, and 812 quinolines with an activity ratio of 2.09 and a z-score of 19.22. (b) *Camptothecin* derivatives showing the *quinoline* and *3-oxy-pyran* substructures.

have z-scores between (−2.0) − 2.0 NSD. Negative z-scores indicate that the mean activity of the subset is less than the expected value. The user can adjust legend by setting cutoff values for standard deviations ranges and selecting the coloring scheme.

Any set of selected features can be used to create a subproject containing only the selected compounds. LeadScope provides the same capabilities within a subproject, including the ability to create subprojects, except that the statistics in the subproject reflect only the subpopulation. By creating subprojects, users can find combinations of structural features with mean activity significantly higher than any of the constituent features.

In summary, the interface allows users to both see and manipulate the underlying information on structural features and associated properties in order to focus on what is relevant and eliminate distracting information.

Project Creation. Before the user can begin exploring a set of structures and associated properties, the data must be analyzed and compiled as a LeadScope project. During this process, each structure is dissected into structural features defined in the template library using standard substructure search techniques.[34] To facilitate structure analysis, the software has sets of structural definitions[35] for (1) aromatic systems, (2) tautomeric systems, (3) generic groups, and (4) structural feature templates; these resources are maintained as indexed binary files. The templates correspond to the structural features available to the user during project exploration. Template atoms are elements, generic atoms such as Ak (alkyl) and Ar (aryl), or generic groups that have been defined by other templates. Additional restrictions can be placed on the entire template structure, on a set of atoms or bonds, and on individual atoms and bonds.

In addition to structural analysis, several properties are calculated for each compound as part of project creation: (a) molecular weight—the sum of atomic weights for all atoms including nonspecified hydrogens; (b) rotatable bonds—the number of single, nonterminal acyclic bonds; (c) the number of hydrogen bond donors—defined by a generic group pattern;[9,10] and (d) the number of hydrogen bond acceptors—also defined by a generic group pattern.[25]

The user can import any type of numerical property associated with the compounds for use in filtering compounds or for statistical correlations. LeadScope will create an interactive slider for the property and an entry in the legend panel for use with statistical correlations. Some example properties are cLogP and pK_a values, activity categories from HTS, IC_{50} data from dose−response studies, activity ratios for two receptor subtypes, and quantity of compound on hand.

LeadScope treats numeric property data as ordinal categorical data. If the input data is continuous values such as IC_{50} or cLogP data, the user can determine how values are assigned to categories: the number of categories and the cutoff values between categories. The system provides a number of predefined property types, and the user can define new property types.

Customizing the Structural Feature Hierarchy. The default view of the feature hierarchy shows those structural features found in one or more compounds in the database. For a large dataset, it will contain many low frequency features with near average activity. LeadScope provides several ways to customize the feature hierarchy. The *S*orted *F*eature *L*ist is a flat list of all features that satisfy the frequency requirements, sorted in user specified order. This provides a way to quickly find those feature most highly correlated with activity and that satisfy frequency requirements. LeadScope also provides a user defined features branch of the feature hierarchy. New features can be defined as substructure search queries created with a chemical structure drawing program like ISIS Draw or ChemDraw.[36] LeadScope will execute the substructure searches and add the resulting compound sets as new features in the hierarchy. User-defined features can be used in the same way as the system-defined features; in particular, (1) histogram bars are color-coded according to activity, (2) substructure matches are highlighted in substance displays, and (3) the features can be used to define subprojects.

Table 4. Activity Categories for IC_{50} Data for the SF-295 Tumor Cell Line from the CNS Panel

category	values	count
0	$x \leq 4.0$	15046
1	$4.0 < x \leq 5.0$	10351
2	$5.0 < x \leq 6.0$	2582
3	$6.0 < x \leq 7.0$	794
4	$7.0 < x \leq 8.0$	473
5	$8.0 < x \leq 9.0$	151
6	$9.0 < x$	33

EXAMPLES OF USE

This section presents two examples illustrating the use of **LeadScope** to find sets of active compounds with combinations of structural features in large sets of structures and biological test results.

Databases. Two databases, used in this work to illustrate LeadScope capabilities, were obtained from the National Cancer Institute's (NCI) Developmental Therapeutics Program.[37] Since 1990, the DTP Human Tumor Cell Line Screen has been testing compounds for growth inhibition against a panel of 60 human tumor cell lines.[38−41] The data provide three concentration parameters for each compound-cell line pair: the IC_{50} value is the concentration that causes 50% growth inhibition; the TGI value is the concentration needed for "total growth inhibition"; and the LC_{50} value is the lethal concentration at 50%. The example below uses IC_{50} data for the SF-295 cell line from the CNS panel; IC_{50} values, which are provided by the DTP as −log(1/C), were assigned to categories as shown in Table 4.

The other DTP database used as an example is the AIDS antiviral screen which is also available online.[37] These dataset contains 32.343 compounds, 230 compounds with NCI-assigned category CA (confirmed active), 444 with category CM (confirmed moderately active), and 31,436 with category CI (confirmed inactive). Activity data is missing for 233 compounds.

Example 1. The first example uses the NCI anticancer screening set containing 30,672 compounds and IC_{50} data for growth inhibition of the CNS SF-295 cell line. Table 4 shows a breakdown across activity categories for this assay; note that 1242 compounds are missing activity data. Since many anticancer compounds are large natural products such as actinomycins and taxols, we will seek smaller, more "drug-

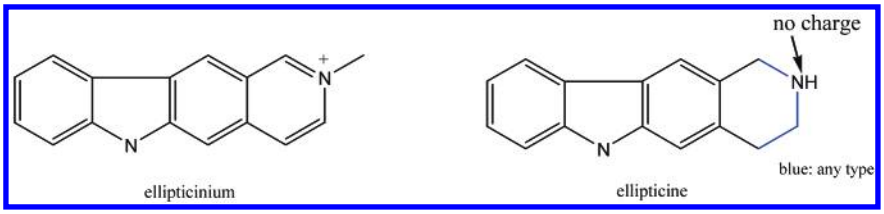


Figure 6. Query substructures for the *ellipticinium* and *ellipticine* classes.

like" molecules with acceptable bioavailability[30] properties. Our objective is to create a succession of subprojects of increasingly higher activity corresponding to combinations of structural features. The *Heterocycles* hierarchy is usually a good place to start because many known drugs contain heterocyclic motifs.[16,18,42] Substituted heterocycles have several desirable attributes as pharmacophore constituents including multiple binding sites, well-defined geometry, and some degree of rigidity.

Our strategy is to look for features that show a good balance of high activity and high frequency in the first stage and then put more emphasis on high activity in later stages. We will use color-coding and length of the histogram bars as quick visual clues to guide the search. If needed, more detailed information on activity ratios and frequencies is available in the information windows at the bottom of the screen.

Figure 5a shows a portion of the *Heterocycles* hierarchy with *quinolines* selected. The five property filters have been set to select smaller, semirigid molecules with acceptable bioavailability[30] properties. As explained above, histogram bars have been color coded for IC_{50} against the SF-295 CNS cancer cell line. Information windows at the bottom of the screen give frequency and activity data: 30,672 compounds in the full NCI data set, 19,275 compounds in the full data set that satisfy the all property filter settings, 812 quinolines with an activity ratio ($\mu_{\text{quinoline}}/\mu$) = 2.09 and a z-score of

$$z = (\mu_{\text{quinoline}} - \mu)/\sigma = 19.22$$

where σ is calculated from variance formula (1).

First we created a subproject comprised of all quinolines satisfying the property filter settings. There are several of members of the *quinoline* hierarchy with higher mean activity. However, there may be several highly potent subclasses of quinolines that could be overlooked by focusing on high activity too early. The general quinoline class offered a good balance of high activity and high frequency.

By browsing the *quinolines* subproject, we identified a 24-member subset containing the 3-oxy-pyran substructure (activity ratio = 2.76, relative to the subproject mean). The compounds in this set are displayed in Figure 5b, sorted in order of decreasing activity. This is a set of camptothecin derivatives, a well-known class of topoisomerase I inhibitors.[41]

Example 2. Any type of numerical property can be loaded for compounds. Then the techniques illustrated above can be used to identify classes enriched with compounds having high or low values of the property. In this example, we will describe a technique for finding compounds that are selectively more potent growth inhibitors of CNS cell lines than non-CNS lines. For each compound, we first calculated a CNS-selectivity (CNS−Sel) score which is the average compound pIC50 value over CNS cell lines minus the average compound pIC50 value over non-CNS cell lines. Since CNS−Sel is the difference of average log values, a CNS−Sel value of 1.0 means the compound is 10 times more potent against CNS compared with non-CNS cell lines.

By using CNS−Sel as an activity score, we can identify compound classes enriched with compounds showing CNS selectivity. For this, we used the LeadScope *Sorted Feature List*, with a frequency range of 20−100, to locate the structural classes with the highest *feature z-scores* for CNS−Sel. The top ranked structural feature was *isoquinolinium, 2-alkyl* (72 compounds, $z = 13.52$).

On examining the compounds in this class, we found a large number of ellipticinium compounds, which are known[43−45] to show selective growth inhibition activity against CNS tumors. To explore this class further and compare it with ellipticines, we constructed the substructural queries[33] in Figure 6 using ChemDraw[36] and executed the substructure searches in LeadScope. Figure 7 shows the results of the search as a new branch in the *User-Defined Features* branch of the hierarchy. Note the dramatic difference in CNS selectivity of the ellipticinium compounds (41 compounds, $z = 12.92$) compared with the ellipticines (41 compounds, $z = 0.17$).

DISCUSSION

These examples illustrate a couple of important points about using LeadScope that should be emphasized.

Statistical Correlations. Histogram bars and scatterplot cells are color-coded based on the difference—expressed in number of standard deviations—between the mean activity of the subset of compounds containing a structural feature from the mean activity of the full set. The standard deviation is calculated from the variance formula (1).[31b] Terms in (1) refer to entries in Table 5 where $i = 1,2$ and $j = 1,..., J$, and the + subscript in the marginal totals indicates the sum over the index (e.g. $n_{1+} = \sum_{j=1}^{J} n_{1j}$).

This tabulates the number of compounds in each of J activity categories, where J is the most active. Row 1 tabulates compounds that contain a given structural feature X, row 2 compounds that do not contain X.

$$variance = \left(\frac{1}{n_{1+}\left(1 + \frac{n_{1+}}{n_{2+}}\right)(n-1)}\right)\left(\sum_{j=1}^{J} j^2 n_{+j} - \frac{\left(\sum_{j=1}^{J} j n_{+j}\right)^2}{n}\right) \quad (1)$$

Note that the variance is completely determined by n_{1+}, the size of the subset containing feature X.

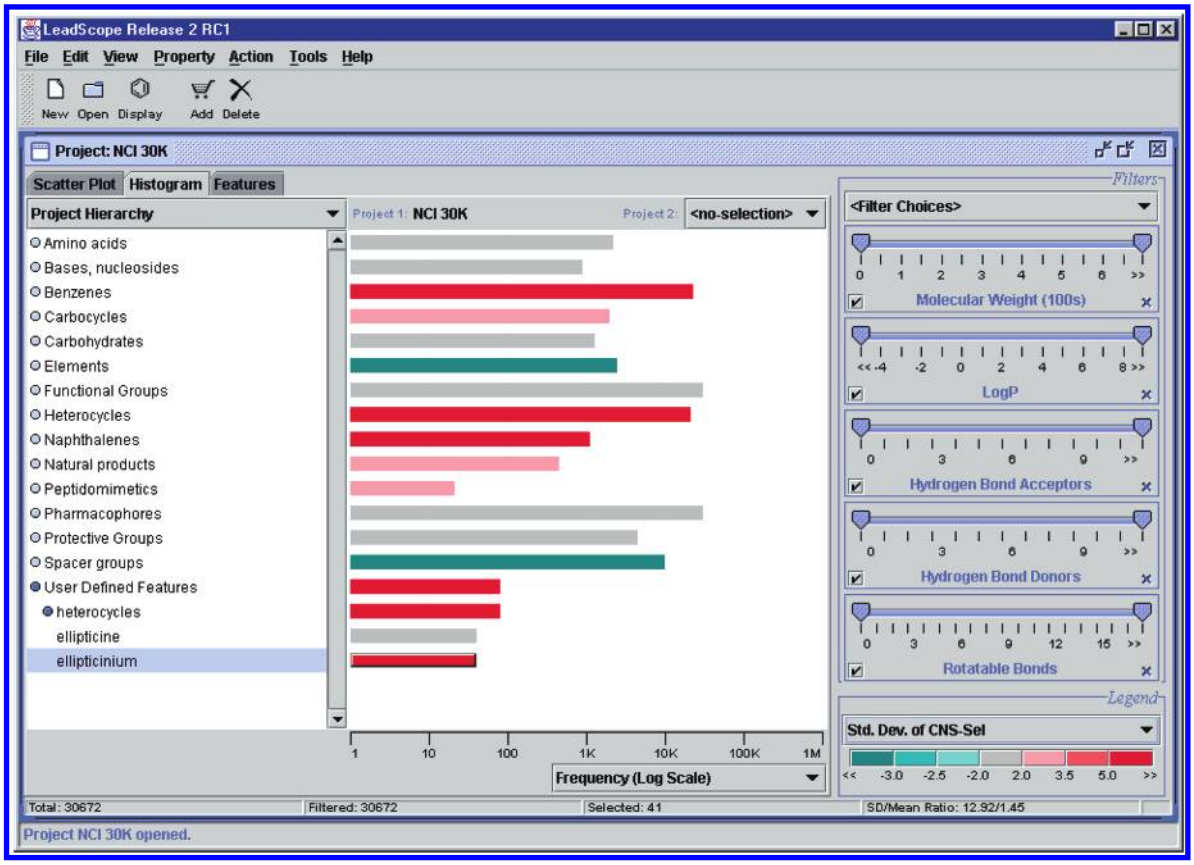


Figure 7. The LeadScope histogram view showing the *ellipticinium* and *ellipticine* classes in the *user-defined features* branch.

Table 5

compound/activity	1	2	···	J	total
have feature X	n_{11}	n_{12}	···	n_{1J}	n_{1+}
do not have X	n_{21}	n_{22}	···	n_{2J}	n_{2+}
all compounds	n_{+1}	n_{+2}	···	n_{+J}	n

The mean activity of the full NCI anticancer screening set is 0.7048 using the categorical activity values[50] of Table 4 for the SF-295 tumor cell line. Suppose we select a 100-membered subset from the NCI set with mean activity 1.0. A question we would like to ask is what is the p-value or the probability of selecting a 100-membered subset with mean activity ≥ 1.0.

Standard deviations give approximate information on p-values. To illustrate the information provided by the statistical color coding of histogram/scatterplot cells, we will use data from Table 4. Figure 8a shows the probability distribution for the mean activity of 100-member subsets selected from the 29,430 compounds with SF-295 categorical activity values from the Table 4 above. The total activity for any 100-member subset can vary from 0 (all 100 compounds in category 0) to 5.33 (= (33*6 + 67*5)/100). The *x*-axis is mean activity of the subset, and the *y*-axis is probability. The blue curve is 10 * p(x), the probability that the subset will have mean activity = x. The magenta curve is cumulative distribution function (*cdf*). Cdf(x) is the probability that the subset will have mean activity $\leq$x. Thus, the p-value at x is 1 − cdf(x). Only the portions of the functions between mean activity = 0.4 and 1.05 are shown. For subsets with activity outside this range, the probability is so low that point would appear to be on the *x*-axis.

The expected total activity for a 100-member subset is $\mu = 0.7048$, with a standard deviation of $\sigma = 0.0934$. The vertical purple line is μ, and the vertical teal-colored lines are 1σ, 2σ, 3σ from μ.

Figure 8b shows a blow-up of the region around 3σ. The colored vertical line is $\mu + 3\sigma = 0.9849$, and the circled point gives the p-value at 3σ:

$$\text{p-value} = 1.0 - \text{cdf}(0.99)$$

$$= 0.0019$$

Thus, the probability of selecting a 100-member subset with total activity units more than 3σ from the mean is approximately 2 out of 1000.

The size of the subset of compounds containing a given structural feature has an important effect on the probability distribution. Suppose X is an *m*-membered subset selected from a set of compounds tested for some activity. Let the mean activity of the full compound set be μ with standard deviation σ, and let μ_X be the mean activity of X. The *central limit theorem* from statistics says that the probability distribution of μ_X, over all subsets X of size *m*, is approximately normal for sufficiently large values of *m*, generally $m > 29$. Further, the standard deviation is approximately $\sigma/\sqrt{(m)}$. Thus, as *m* increases, the standard deviation of the distribution of μ_X decreases. These results hold regardless of the distribution of activity values in the full compound set.

Figure 9a shows probability distributions for 20- (magenta) and 100-member subsets (blue) with mean activities calculated using categorical activity values from the Table 4. The mean activity for a 20-member subset can vary from 0 to 6

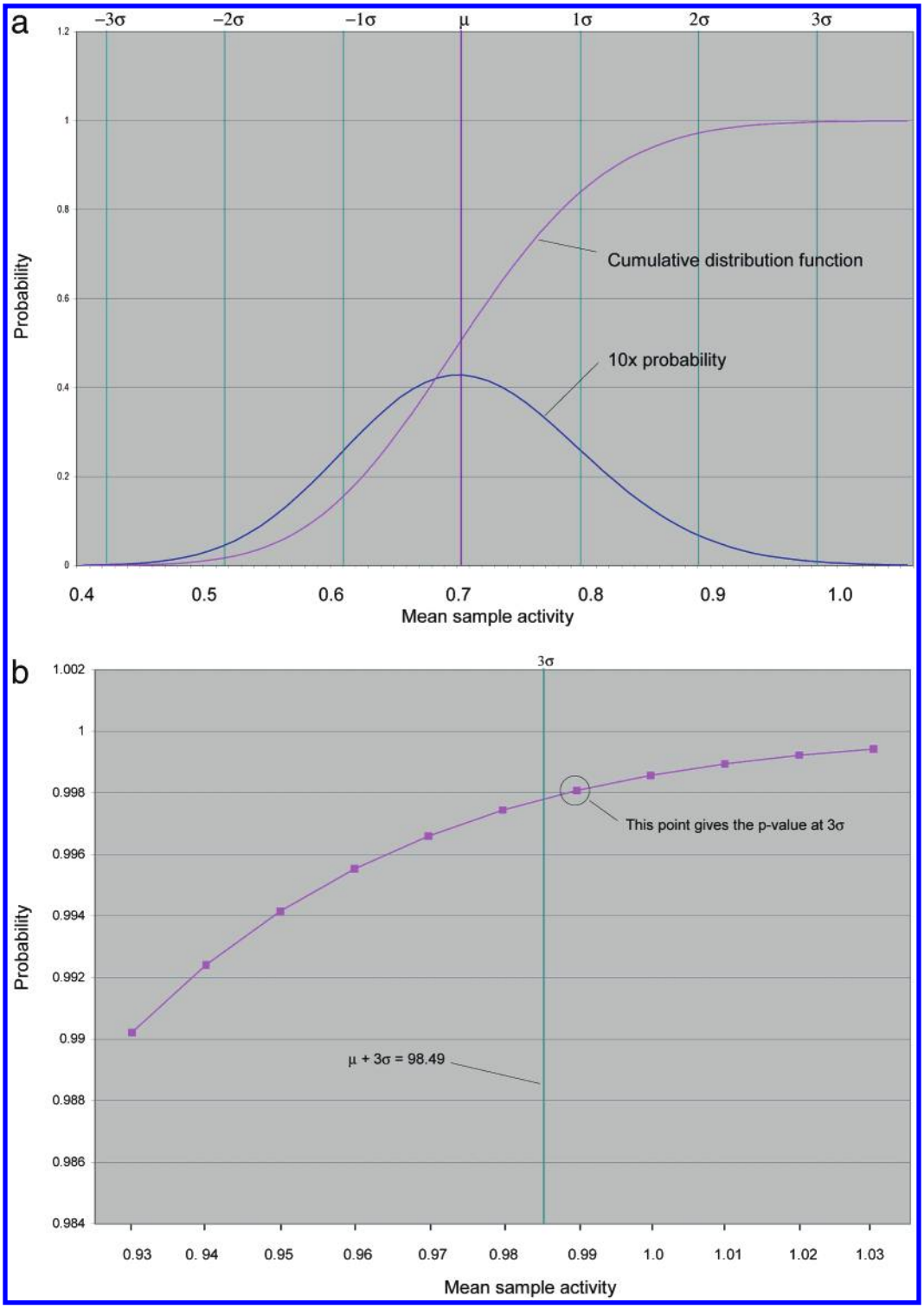


Figure 8. (a) The probability distribution for the mean activity of 100-member subsets using categorical activity values from the Table 4. (b) A blow-up of the region around 3□ for the cumulative distribution, the colored vertical line is $\mu + 3\sigma$.

(all 20 compounds in category 0 or category 6, respectively); the mean activity for a 100-member subset can vary from 0 to 5.33. Both distributions are somewhat skewed with a long right tail. This is more evident with 20-member subsets (magenta). Note that as *m* increases, almost all *m*-membered subsets occupy an increasingly narrow band centered on μ

Recursively Exploring Subsets. An approach to structure−activity relations that is gaining popularity is known as *recursive partitioning*.[46−48] This procedure partitions a set of chemical structures into subsets which contain from 0 to 3 instances of a predefined structural feature. For each feature, a statistics is computed comparing the mean or variance of the feature-count subsets. The feature with the lowest p-value is selected for the split. Then the procedure is recursively reapplied to the newly created subsets until a statistical threshold is exceeded. The procedure produces a dendrogram where the nodes are compound sets. The root node is the full compound set or parent set, and the offspring of any node is a partitioning of the parent set. The structural features that have been used are similar to those used for

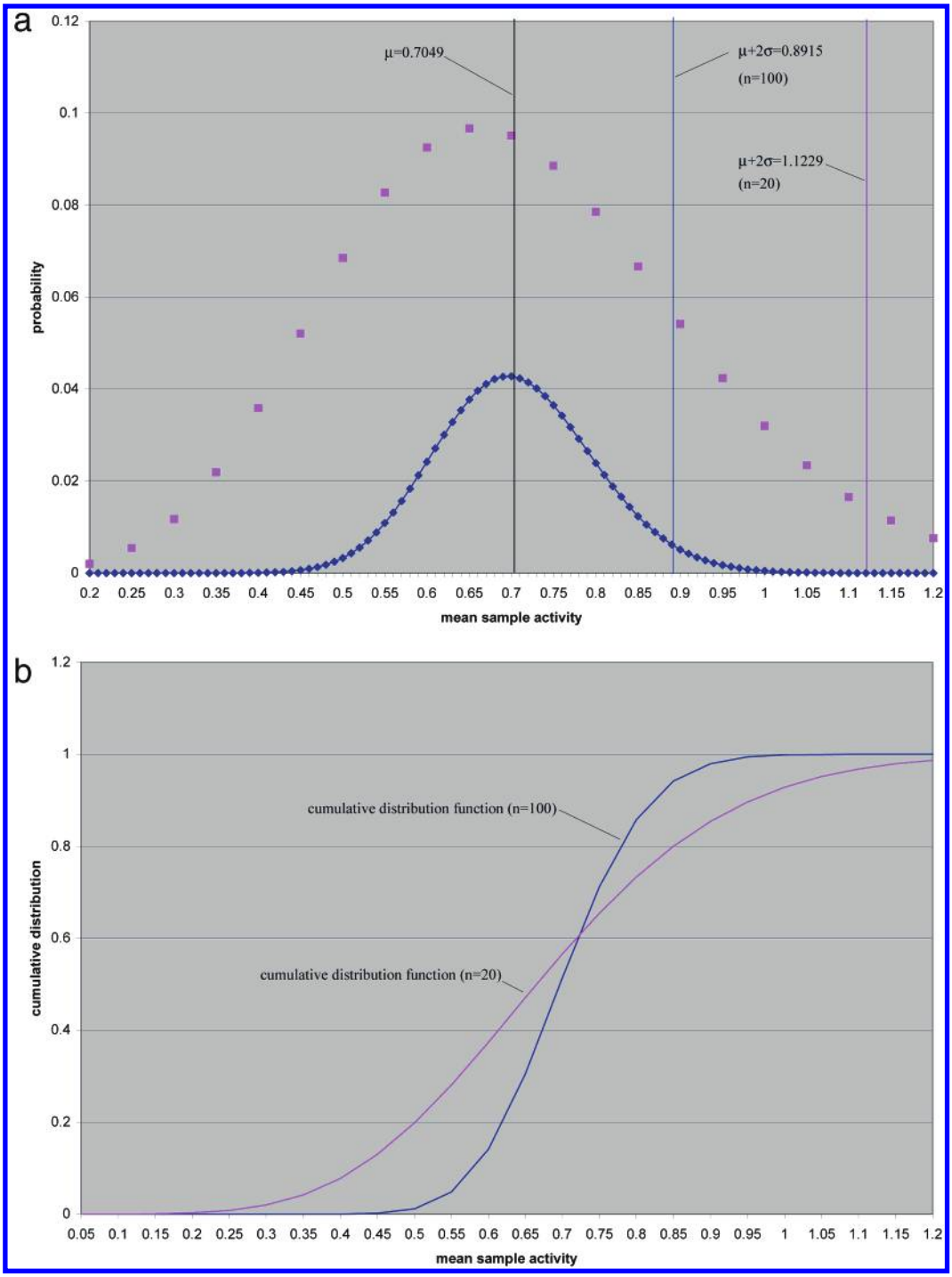


Figure 9. (a) Probability distributions for the mean activity of 20- and 100-member subsets using categorical activity values from the Table 4. (b) The corresponding cumulative probability distributions.

clustering and conventional SAR. Recursive partitioning is a fully automated procedure; there is no provision for the chemist to participate in or guide the partitioning process.

LeadScope also provides facilities that allow the user to create a succession of subsets of increasingly higher activity which correspond to combinations of structural features. The color-coding of histogram bars as well as information on frequencies and activity ratios can guide the user in selecting subsets to explore, but the user is free to explore any subset. Instead of choosing the most statistically significant feature, the user may select a feature in which he or she has a special interest or to achieve a better balance of high frequency and high activity. In addition, the user can dynamically adjust the property filters to constrain the compound set to the subset for which all property values are in desired ranges. For example, the user could adjust the filter settings to select only "drug-like" molecules.[30]

Value of Inactive Compounds. With a dataset comprising both active and inactive compounds, the statistical color-coding in LeadScope can help to locate features or combinations of features that are highly correlated with activity. If inactive compounds are eliminated, most of the visual clues the statistics can provide are also eliminated. This is illustrated in Figure 10.

The example in Figure 10 was taken from the HIV project which use compounds and activty data from the DTP AIDS

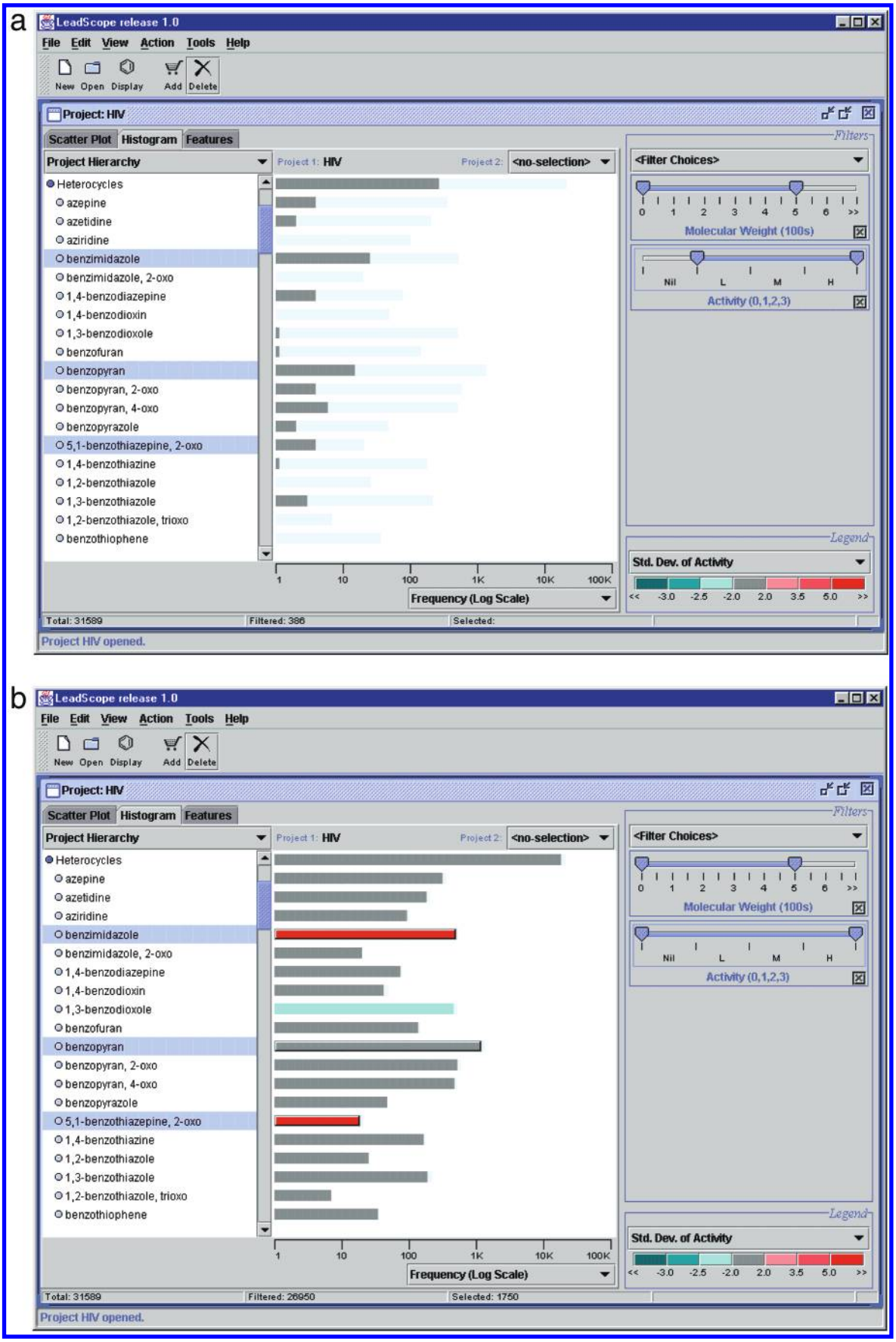


Figure 10. (a) A portion of the hierarchy for the HIV dataset with inactive compounds eliminated. (b) The same view with both active and inactive compounds.

antiviral screen. In Figure 10a, inactive compounds have been eliminated using the activity slider.[49] The statistical color-coding evidently gives no clue to which of the highlighted sets is more promising. As seen in Figure 10b, the situation

changes dramatically when the inactive compounds are restored.

CONCLUSION

The ability to link and effectively interpret the volumes of chemical and biological screening data being generated has become a critical bottleneck in making full use of these data. Because medicinal chemists have no means for effectively dealing with the torrent of data being produced, inactive compounds (negative results) are largely ignored. This can be as much as 99% of the structure−activity data collected.

LeadScope is a decision support solution for pharmaceutical researchers, to help draw knowledge from the large amounts of data now being generated from combinatorial chemistry and high throughput screening. It provides sophisticated data visualizations and tools for interactively exploring structures by properties and common structural features, helping to quickly find correlations between biological activity and structural features.

Instead of focusing on individual compounds and their activities, these techniques allow users to broaden their focus to sets of compounds, correlated structural features, and statistics over the full dataset. As a result pharmaceutical scientists can get more information out of available screening data and concentrate on the most statistically significant structural features in formulating a hypothesis to explain observed biological activity. This visual exploration amplifies the chemist's experience and expertise, enabling them to investigate many more compounds.

ACKNOWLEDGMENT

We gratefully acknowledge financial support from Pfizer Ltd, Sandwich, Kent CT13 9NJ, UK, and the assistance of our sponsors Mark Lord and Nick Terrett. We thank the Pfizer chemists who participated in the collaboration Paul Edwards, Jens Loesel, Mike Snarey, and Tony Wood for many changes, improvements, and additional features which were incorporated into the program. Finally we would like to thank an anonymous reviewer for several useful suggestions that improved the paper.

REFERENCES AND NOTES

(1) Kubinyi, H. *QSAR: Hansch Analysis and Related Approaches*; VCH Publishers: New York, 1993.

(2) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−96.

(3) Willett, P. *Similarity and Clustering in Chemical Information Systems*; Research Studies Press: Letchworth, 1987.

(4) Brown, R. Descriptors for Diversity Analysis. *Perspect. Drug Discovery Des*. **1997**, *7/8*, 31−49.

(5) Carhart, R. E.; Smith, D. H.; Venkataraghavan. R. Atom pairs as molecular features in structure−activity stidies: definitions and applications. *J. Chem. Inf. Comput. Sci.* **1985**, 25, 64−73.

(6) Kearsley, S. K.; Sallamack, S.; Fluder; E. M., Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical Similarity Using Physiochemical Property Descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, 36, 118−127.

(7) Software for generating molecular fingerprint is available from Daylight Chemical Information Systems, 27401 Los Altos, Suite #370, Mission Viejo, CA 92691 and Tripos, Inc., 1699 South Hanley Road, St. Louis, MO 63144.

(8) Brown, R. D.; Martin, Y. C. Use of Structure Activity Data to Compare Structure Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−84.

(9) Greene, J.; Kahn, S.; Savoj, H.; Sprague, P.; Teig, S. Chemical Function Queries for 3D Database Search. *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 1297−1308.

(10) Wang, T.; Zhou, J. 3DFS: A new 3D Flexible Searching System for Use in Drug Design. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 71−77.

(11) Hansch, C.; Leo, A. *Substituent Constants for Correlation Analysis in Chemistry and Biology*; John Wiley & Sons: New York, 1979.

(12) Cummins, D. J.; Andrews, C. W.; Bentley, J. A.; Cory, M. Molecular Diversity in Chemical Databases: Comparison of Medicinal Chemistry Knowledge Bases and Databases of Commercially Available Chemicals. *J. Chem. Inf. Comput. Sci*. **1996**, *36*, 750−63.

(13) Matter, H. Selecting Optimally Diverse Compounds from a Structure Database: A Validation Study of Two-Dimensional and Three-Dimensional Molecular Descriptors. *J. Med. Chem*. **1997**, *40*, 1219−29.

(14) Cosgrove, D. A.; Willett, P. SLASH: A program for analyzing the functional groups in molecules. *J. Mol. Graphics Model*. **1998**, *16*, 19−32.

(15) Boyd, S. M.; Beverley, M.; Norskov, L.; Hubbard, R. E. Characterizing the geometric diversity of functional groups in chemical databases. *J. Comput.-Aided Mol. Des*. **1995**, *9*, 417−24.

(16) Bemis, G. W. and Murcko, M. A. The Properties of Known Drugs, 1. Molecular Frameworks. *J. Med. Chem*. **1996**, *39*, 2887−2893

(17) Comprehensive Medicinal Chemistry and the MACCS−II Drug Data Report are available from MDL Information Systems Inc., San Leandro, CA.

(18) Lewell, Q. L.; Judd, D. B.; Watson; S. P.; Hann, M. M. RECAP − Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 511−22.

(19) The World Drug Index is available from Derwent Information Ltd., Derwent House, 14 Great Queen Street, London WC2B 5DF, UK.

(20) Shemetilskis; N. E.; Weininger; D.; Blankley, C. J.; Yang, J. J.; Humblet, C. Stigmata: An Algorithm to Determine Structural Commonalties in Diverse Datasets. *J. Chem. Inf. Comput. Sci*. **1996**, *36*, 862−71.

(21) Sheridan, R. P.; Miller, M. D. A Method for Visualizing Recurrent Topological Substructures in Sets of Active Molecules. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 915−24.

(22) Varmuza, K.; Scsibrany, H. Substructure Isomorphism Matrix. *J. Chem. Inf. Comput. Sci*. **1998**, *40*, 308−13. (b) The LeadScope structural feature Hierarchy described in this work is generated algorithmically from tables of parent and substituent substructures.

(23) *Chemical Abstracts Index Guide*, Appendix IV; Chemical Substance Index Names, American Chemical Society: 1997.

(24) Sawyer, T. K. In *Structure-Based Drug Design*; Veerapandian, P., Ed.; Marcel Dekker: 1997; pp 559−634.

(25) Jens Loesel, private communication. See, also: Jens Loesel, IAFs − Empirical Descriptors for non-Covalent Interactions, UK QSAR & Chemoinformatics Group, Spring 1999 Meeting, April 27, Pfizer Ltd, Sandwich, UK, http://www.iainm.demon.co.uk/.

(26) IUPAC Nomenclature and Symbolism for Amino Acids and Peptides, section 3AA-18.1; http://www.chem.qmw.ac.uk/iupac/AminoAcid/, web version prepared by G. P. Moss.

(27) Shneiderman, B. *Designing the User Interface: Strategies for Effective Human-Computer Interaction*, 3rd ed.; Addison-Wesley: 1998.

(28) Ahlberg, C.; Shneiderman, B. Visual information seeking: Tight coupling of dynamic query filters with starfield displays. *Proceedings of the CHI'94 Conference on Human Factors in Computing Systems*; ACM: New York, 1994; pp 313−317.

(29) Chuah, M. C.; Kerpedjiev, S.; Kolojejchick, J.; Lucas, P.; Roth, S. F. Toward an Information Visualization Workspace: Combining Multiple Means of Expression. *Human-Computer Interaction* **1997**, *12*(1 & 2), pp 131−186.

(30) Lipinski, C. A.; Lombardro, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and developmental settings. *Adv. Drug Del. Res*. **1997**, *23*, 3−25.

(31) Agresti, A. *Categorical Data Analysis*; John Wiley & Sons Inc.: New York, 1990; pp 286−287. (b) Dr. Mark Farmen, personal communication.

(32) As indicated in the legend title, the activity data shown is for the SF-295 tumor cell line from the CNS panel.

(33) Bonds that are colored in structure diagrams are generic and may match any bond or some combination of specific bonds such as "single or aromatic". The user can review the complete definition of the corresponding substructural query if desired.

(34) Stobaugh, R. E. Chemical Structure Searching. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 271−275.

(35) Myatt, G. J. Computer Aided Estimation of Synthetic Accessibility, Ph.D. Thesis, Leeds University, 1994.

(36) ISIS Draw is a trademark of MDL Information Systems Inc., San Leandro, CA. ChemDraw is a trademark of CambridgeSoft Corp., Cambridge, MA.

(37) Nonconfidential screening results and chemical structural data from the NCI's Developmental Therapeutics Program are available online: (a) Human Tumor Cell Line Screen: *http://dtp.nci.nih.gov/docs/cancer/cancer_data.html.* (b) AIDS Antiviral Screen: *http://dtp.nci.nih.gov/docs/aids/aids_data.html.*

(38) Weinstein, J. N.; Myers, T. G.; O'Connor, P. M.; Friend, S. H.; Fornace, A. J.; Kohn, K. W.; Fojo, T.; Bates, S. E.; Rubinstein, L. V.; Anderson, N. L.; Buolamwini J. K.; van Osdol, W. W.; Monks, A. P.; Scudiero, D'Monks, A. P.; Scudiero, D. A.; Johnson, G. S.; Paull, K. D.; Sausville, E. A. The NCI anti-cancer drug screen: a smart screen to identify effectors of novel targets. *Anti-Cancer Drug Des*. **1997**, *12*, 533−41.

(39) Boyd, M. R.; Paull, K. D. Some Practical Consideration and Applications of the National Cancer Institute In Vitro Anticancer Drug Discovery Screen. *Drug Dev. Des*. **1995**, *34*, 91−109.

(40) Weinstein, J. N.; Kohn, K. W.; Grever, M. R.; Viswanadham, V. N.; Rubinstein, L. V.; Monks, A. P.; Scudiero, D. A.; Welch, L.; Koutsoukos, A. D.; Chiausa, A. J.; Paull, K. D. Neural Computing in Cancer Drug Development: Predicting Mechanism of Action. *Science* **1992**, *258*, 447−51.

(41) Paull, K. D.; Hamel, E.; Maispeis, L. Prediction of biochemical mechanism of action from the in vitro anticancer screen of the National Cancer Institute. In *Cancer Chemotherapeutic Agents*; Foye, W., Ed.; ACS professional Reference Books, ACS: Washington, DC, 1992; pp 9−45.

(42) Gibson, S.; McGuire, R.; Rees, D. C. Principal Components Describing Biological Activities and Molecular Diversity of Heterocyclic Aromatic Ring Fragments. *J. Med. Chem.* **1996**, *39*, 4065−72.

(43) Shi, L. M.; Fan, Y., Myers, T. G.; O'Connor, P. M.; Paull, K. D.; Friend, S. H.; Weinstein, J. N. Mining the NCI anticancer drug discovery databases: genetic function approximation for the QSAR study of anticancer ellipticine analogues. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 189−99.

(44) Shi, L. M.; Myers, T. G.; Fan, Y.; O'Connor, P. M.; Paull, K. D.; Friend, S. H.; Weinstein, J. N. Mining the National Cancer Institute Anticancer Drug Discovery Database: cluster analysis of ellipticine analogues with p53-inverse and central nervous system-selective patterns of activity. *Mol. Pharmacol*. **1998**, *53*, 241−51.

(45) Ohashi, M.; Oki, T. Ellipticine and related anticancer agents. *Expert Opin. Ther. Pat*. **1996**, *6*, 1285−1294.

(46) Chen, X.; Rusinko, A., III; Young, S. S. Recursive Partitioning Analysis of a Large Structure−Activity Data Set Three-Dimensional Descriptors. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1054−62.

(47) Hawkins, D. M.; Young, S. S.; Rusinko, A., III. Analysis of a Large Structure−Activity Data Set Using Recursive Partitioning. *Quant. Struct.-Act. Relat*. **1997**, *16*, 1−7.

(48) The Cerius2.CSAR (Classification SAR) program from Molecular Simulations, Inc., San Diego, is a commercial implementation of this technique.

(49) Light gray shadows, particularly prominent in Figure 10a, show the original length of histogram bars before filters were applied.

(50) For statistical correlations, LeadScope treats activity data as ordinal categorical data. If the input data is continuous values such as IC_{50} data, the user can determine how values are assigned to categories: the number of categories and the cutoff values between categories.

CI0000631