

Representation of Chemical Information in OASIS Centralized 3D Database for Existing Chemicals

Nikolai Nikolov,[†] Vanio Grancharov,[‡] Galya Stoyanova,[‡] Todor Pavlov,[‡] and Ovanes Mekenyan^{*,‡}

Centre of Biomedical Engineering “Ivan Daskalov”—Bulgarian Academy of Sciences, Bl. 105 Acad. G. Bonchev Str., 1113 Sofia, Bulgaria, and Laboratory of Mathematical Chemistry, University “Assen Zlatarov”, 8010 Bourgas, Bulgaria

Received April 18, 2006

The present inventory of existing chemicals in regulatory agencies in North America and Europe, encompassing the chemicals of the European Chemicals Bureau (EINECS, with 61 573 discrete chemicals); the Danish EPA (159 448 chemicals); the U.S. EPA (TSCA, 56 882 chemicals; HPVC, 10 546 chemicals) and pesticides' active and inactive ingredients of the U.S. EPA (1379 chemicals); the Organization for Economic Cooperation and Development (HPVC, 4750 chemicals); Environment Canada (DSL, 10851 chemicals); and the Japanese Ministry of Economy, Trade, and Industry (16811), was combined in a centralized 3D database for existing chemicals. The total number of unique chemicals from all of these databases exceeded 185 500. Defined and undefined chemical mixtures and polymers are handled, along with discrete (hydrolyzing and nonhydrolyzing) chemicals. The database manager provides the storage and retrieval of chemical structures with 2D and 3D data, accounting for molecular flexibility by using representative sets of conformers for each chemical. The electronic and geometric structures of all conformers are quantum-chemically optimized and evaluated. Hence, the database contains over 3.7 million 3D records with hundreds of millions of descriptor data items at the levels of structures, conformers, or atoms. The platform contains a highly developed search subsystem—a search is possible on Chemical Abstracts Service numbers; names; 2D and 3D fragment searches; structural, conformational, or atomic properties; affiliation in other chemical databases; structure similarity; logical combinations; saved queries; and search result exports. Models (collections of logically related descriptors) are supported, including information on a model's author, date, bioassay, organs/tissues, conditions, administration, and so forth. Fragments can be interactively constructed using a visual structure editor. A configurable database browser is designed for the inspection and editing of all types of data items. Database statistics are maintained on the number and quality of structures, conformers, and descriptors. Reports can be generated presenting any chosen subset of structures and descriptors into different formats suitable for inclusion into documents. In addition to fixed report formats, there is a powerful report template designer module with a visual report template editor to produce a customized page layout. The database is compatible at the import/export level with SDF, MOL, SMILES, and other known formats. The precalculated centralized 3D database could be useful for quantitative structure–activity relationship developers avoiding the time-consuming and cumbersome 3D calculation phase of model development.

1. INTRODUCTION

The use of a two-dimensional (2D) representation of chemical structures is an oversimplification which does not always produce adequate quantitative structure–activity models (QSARs). Such a description “shortcut”, however, allowed modelers to treat a large set of chemicals at a time when computing hardware demonstrated limited capabilities to quantum-chemically handle the steric and electronic structures of a large number of chemicals. On the other hand, it is common knowledge that the molecular interactions are three-dimensional (3D) in nature, and molecular models must treat chemicals as 3D entities. The computational barriers rather than empirical evidences shaped a present convention for use in the molecular modeling of single conformational

(3D) representations of chemicals derived from a computed minimum energy. However, once in the 3D description of a chemical structure, one should face another complication related to the conformational flexibility of molecules—often called the fourth dimension of the chemical structures. Thus, flexible molecules can exist as hundreds of different 3D geometrical conformations, and the electronic properties (hence, reactivity) of the different conformations of a single 2D structure can vary substantially. Minimum energy calculations often fail to identify so-called active conformers. The latter could be slightly less stable than the lowest energy structure; however, such a conformer could have the required shape and electronic properties to interact with biomacromolecules. This holds especially for enzyme-mediated reactions where enzyme-induced distortions in the direction of the transition state drive the molecules even out of the local potential energy minima. Our latest investigations showed that conformers of a chemical which have free energy in

* Corresponding author fax: ++35956880249; e-mail: omekenya@btu.bg.

[†] Bulgarian Academy of Sciences.

[‡] University “Assen Zlatarov”.

the neighborhood of 20 kcal/mol from the lowest energy structure (usually accepted threshold) often exhibited significant variation in potentially relevant electronic descriptors. For example, conformers of a relatively small molecule such as pyperonyl acetone [Chemical Abstracts Service (CAS) 003160370], being within a relatively narrow range of heat of formation, $-\Delta\Delta H_f^\circ = 3.99$ kcal/mol (MOPAC, AM1), had a range of 0.97 eV for E_{LUMO} , 0.37 eV for E_{HOMO} , 0.91 eV for $E_{\text{HOMO-LUMO}}$, and 2.27 Å for the maximum diameter. The observation that relatively small energy differences between conformers can result in significant variations in electronic structure highlighted the necessity of generating and analyzing all energetically reasonable conformers when defining the reactivity of chemicals. It appears that the selection of active conformers of chemicals is as important as the identification of structural descriptors of this chemical, which are juxtaposed with the modeled biological effect.

To address the issue of selection of the active conformers in QSAR studies with flexible molecules and complex biological interactions, the "dynamic" QSAR approach was introduced.^{1,2} The name "dynamic" was used to illustrate the attempts to mimic the infinite conformational space of a chemical by a set of static (discrete) conformers. The approach assumes that, in complex environments, such as biological tissues and fluids, chemicals can exist in conformations other than the lowest gas-phase energy state. The use of the latter in SAR studies is common, but inappropriate, because in complex systems, such as biological tissues and fluids, chemicals are likely to exist in a variety of conformational states. In fact, the lowest-energy, gas-phase conformations might be the least likely to interact with macromolecules, and solvation and binding interactions could more than compensate for energy differences among the conformers of a chemical.^{3,4}

The selection of the "most active" conformers required the development of complex structure-activity modeling algorithms that are based on the physical reality provided that the selection of active conformers is dependent on the specific interaction under investigation.⁴ Conformational flexibility appears to be a significant structural feature, especially when receptor-mediated toxic endpoints are modeled. Capabilities need to be developed to represent chemicals as a distribution of plausible conformations, quantify the molecular descriptors as a function of the conformation, and examine whether a chemical is flexible enough to conform to an "induced fit" by the receptor itself.⁶⁻⁸

Recently, we developed two approaches to conformer generation which were quite different from each other with respect to the algorithm used as well as their performance. The first approach, called 3DGEN, is based on a combinatorial procedure for the systematic search of conformational space.⁹ The systematic approach, however, was found to provide good performance for relatively small and rigid structures. A new approach for coverage of the conformational space of highly flexible chemicals by a limited number of conformers was developed—called the GAS algorithm.¹⁰ Instead of using the systematic search, whose time-complexity increases exponentially with degrees of freedom, a genetic algorithm (GA) was employed to minimize 3D similarity among the generated conformers. This makes the problem computationally feasible even for large, flexible molecules, at the cost of nondeterministic character of the algorithm.

In contrast to traditional GA, the fitness of a conformer is not quantified individually but only in conjunction with the population it belongs to. The approach handles the following stereochemical and conformational degrees of freedom: rotation around acyclic single and double bonds, inversion of the stereocenters, flip of the free corners in saturated rings, and reflection of the pyramids on the junction of two or three saturated rings. The latter two were particularly introduced to encompass the structural diversity of polycyclic structures.

A few insufficiencies in the GAS algorithms have been fixed in the latest modification of the approach.¹¹ The original formulation of the genetic algorithm was not applicable to rigid chemicals where the systematic algorithm was employed. Second, in some cases, the conformers produced by the original formulation of GA were not meeting the intuitive understanding for the coverage of conformational space. To fix this problem, the fitness function based on maximization of the root-mean-square distance between conformers only in the original formulation was combined with the Shannon function accounting for the evenness of the conformer distribution across conformational space. Finally, a new procedure has been developed for the automated determination of the number of conformers needed for reasonably good coverage of the conformational space.

When strained conformers are obtained by any of the algorithms, the possible violations of imposed geometric constraints are corrected with a strain-relief procedure (pseudomolecular mechanics; PMM) based on a truncated force field energy-like function, where the electrostatic terms are omitted.⁹ Geometry optimization is further completed by quantum-chemical methods. MOPAC 93^{12,13} is employed by making use of the AM1 Hamiltonian. Next, the conformers are screened to eliminate those whose heat of formation, ΔH_f° , is greater than the ΔH_f° associated with the conformer with an absolute energy minimum determined by a user-defined threshold—to be within the neighborhood of a 20 kcal/mol (or 15 kcal/mol) threshold from the low(est) energy conformers. Subsequently, conformational degeneracy, due to molecular symmetry and geometry convergence, is detected within a user-defined torsion angle resolution.

The main goal of the present work is to apply our experience to an automated 2D-3D migration of large chemical inventories and an automated conformational generation procedure for building and managing a centralized 3D database where all chemical structures are precalculated, in terms of conformer generation and quantum-chemical treatment. All existing chemicals across the environmental state agencies are incorporated into the database, which consists of about 200 000 unique chemicals represented by 4 million quantum-chemically optimized conformers. The precalculation of steric and electronic (reactivity) parameters for all chemicals in the database allows for smart searching to explore a mechanistically sound hypothesis. One could extract chemicals for QSAR analysis without the tedious and time-consuming quantum-chemical calculations. Moreover, the active conformers of chemicals with respect to specific interactions could be searched. The representation of chemicals by a limited number of conformers covering conformational space makes the search of chemicals with steric and reactivity parameters more robust because, in general, not all conformers of the molecules meet the ranges of the 3D queries. This is especially important when modeling receptor-

mediated endpoints where the interaction specificity requires the investigation of all conformers of chemicals. Similarly, in the drug design, hits for effective drugs could be easily missed if the conformational space of the search chemicals is not investigated. The conformational flexibility of chemicals is also important in the similarity search where some of the conformers of the compared chemicals could be close with respect to specified molecular parameters, whereas others could be significantly different with respect to the same parameters. In the next section, the database platform is described along with the methods and principles of the specialized software capable of representing and processing the necessary amount and diversity of chemical information. Except for the discrete organic chemicals, the platform and software support defined and undefined mixtures and polymers which make the system unique as compared with existing databases. The content of the database and performance of the database manager are presented and discussed in the Results section.

2. METHOD. OASIS DATABASE PLATFORM

The OASIS Database platform is a software framework for building chemical databases. It contains a database schema and accompanying software for the management of chemical information. The system is named OASIS because the platform is part of the OASIS chemical software (<http://www.oasis-lmc.org>).

2.1. Overview. The platform provides an extensive list of features, such as the following: (i) the storage of chemical 2D and 3D structures; (ii) user-defined structural (2D), conformational (3D), and atomic descriptors and model parameters (different testing protocols); (iii) representation of discrete structures, defined and undefined mixtures, hydrolyzing chemicals, and polymers; (iv) an import/export interface with all known connectivity formats, such as SDF, MOL, and SMILES; (v) a search by CAS number, chemical names, and descriptors; an extensive 2D and 3D fragment search including atom modifiers, wildcard atoms, and distances; a similarity search; logical combinations; saved queries; and result exports as well as very fast dedicated search algorithms; (vi) a database browser for the inspection, insertion, modification, and deletion of all types of data items contained in the database; (vii) a visual structure editor for the defining and editing of structures and fragments; (viii) database statistics, descriptor distribution, and model correlation tools; (ix) a report generator with a visual template editor; (x) OASIS Web database (a software suite for implementation of the OASIS database functionality on WWW servers for public or restricted access to chemical information)

2.2. Data Types. The central notion in the representation is that of *chemical*. The basic data items for a chemical are its identification data (CAS numbers, chemical names, affiliation in other databases, etc.) and its 2D structure—data on atoms, bonds, and topology (Figure 1). These data are not obligatory; for example, chemicals without CAS numbers or names can also be represented. Similarly, chemicals such as undefined mixtures could have CAS numbers, names, affiliation, and so forth but not a 2D structure.

In addition to these basic data, chemicals can have the following:

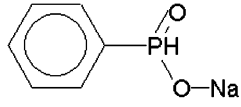
Chemicals may have	Example
CAS number	4297954
2D structure	
One or more names	Phosphinic_acid,_phenyl-,_sodium_salt Sodium phenylphosphinate
Database affiliations	The chemical databases this structure is known to belong to. Examples: Danish EPA, DSL, IUCLID, HPVC
Observed descriptors	Application-defined and specific to a database. Examples: RBA [%], LC50, LD50
Calculated descriptors	Application-defined and specific to a database. Examples: Molecular weight, logKow, logBCF, number of cycles, number of bonds, etc.
Sets of conformers	Generated according to specified criteria for optimum coverage of a conformational space.

Figure 1. Data groups maintained for a chemical.

A conformer may have	Examples
3D data (atom coordinates)	The 3D data of the conformer atoms and bonds
Conformational descriptors	Calc_Heat_Form, E(HOMO), E(LUMO), Egap, Volume polarizability, Dipole moment, Electronegativity
Atomic descriptors	Charge, atom polarizability, donor / acceptor delocalizability

Figure 2. Conformer data items.

(i) *Descriptors.* These are numeric, text, or logical values assigned to the chemical or its parts. Observed and calculated descriptors are supported. Descriptors can be assigned to the whole chemical or to a set of chemicals (*structural descriptors*), to a single conformer, or to a single atom. Each OASIS database can support its own set of descriptors; these can be defined by the user and altered at any time.

(ii) *3D Structure: Sets of Conformers.* Chemicals can include sets of conformers with 3D information and descriptors.

Every conformer has 3D coordinates of atoms and may optionally have atomic and conformational descriptors.

The sets of conformers are associated with the 2D structure, so chemicals with the same 2D structure share the same set of conformers (Figure 2).

The OASIS data scheme includes the representation of discrete chemicals (hydrolyzing and nonhydrolyzing), mixtures (defined and undefined), and polymers. The basic functionality used for all of these structures is the mechanism that allows chemicals to have other chemicals as components. Components can be things such as hydrolyzing structures, which in turn can have components. Of course, hydrolyzing structures can be registered as independent chemicals as well. This representation has the advantage of allowing CAS numbers, names, descriptors and so forth at all levels where these are appropriate. On the other hand, the presence of optional data items in the data model allows for an adequate representation of actual chemical information (e.g., mixtures could not have conformers but mixture components could have them).

When the component-based representation is used, the mixtures and hydrolyzing structures are stored as shown in

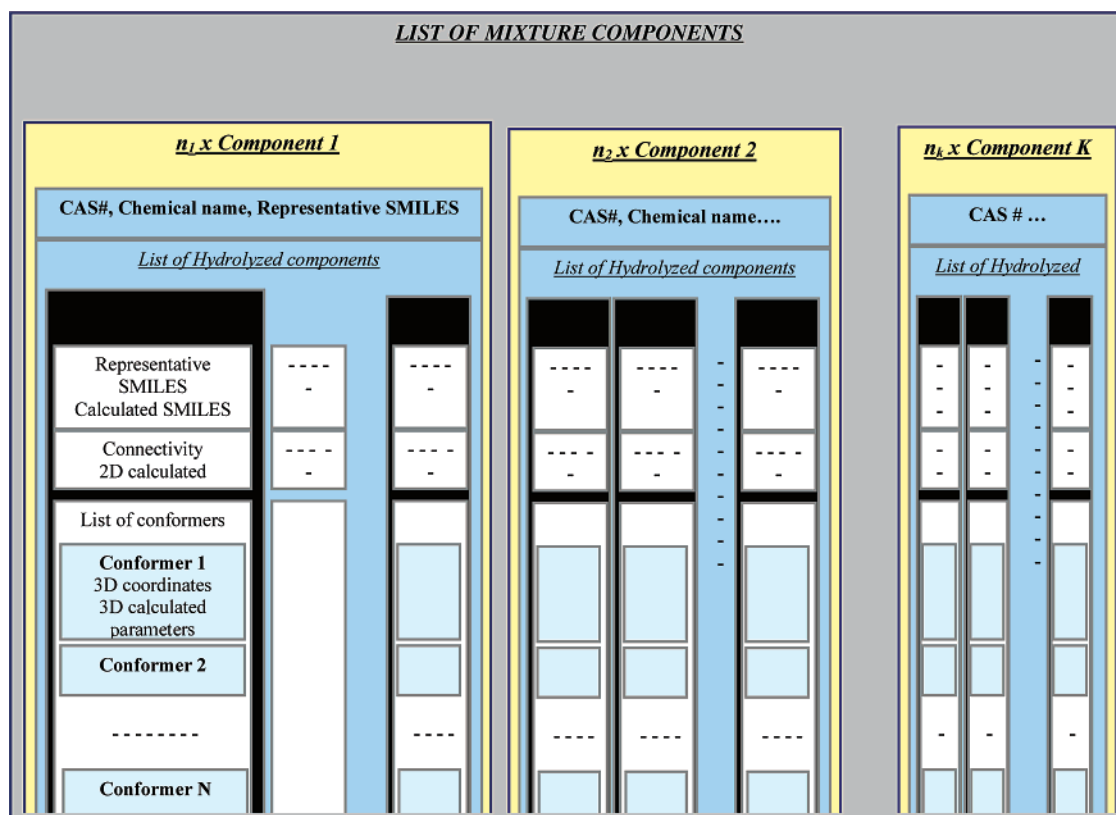


Figure 3. Principle for representing defined mixtures, hydrolyzed structures, and polymers.

Figure 3. Mixtures can have identification data, some set of descriptors, and a list of components. A mixture component can be either nonhydrolyzed or a hydrolyzed structure. In the case of hydrolyzed structures, their components form a third representation level. The bottom level, independently of its position, may have any kind of data specific to a discrete structure, including a calculated 2D structure, conformers, and so forth.

As organized, the system automatically partitions the mixtures into components, and a module is designed to automatically identify hydrolyzable structures and subsequently simulate the hydrolysis (Figure 3).

The OASIS database platform contains a special feature intended to store test results and related information for the test protocol. This information is called a model. A model contains a collection of logically related observed descriptors which represent test outcomes and other test conditions. Models are specified by a model's name, an author's name, the date of creation, bioassay, organ or tissue, test conditions, and administration, as shown in Figure 4.

Models can be defined, inspected, and edited, and descriptors can be registered to models using the model browser.

2.3. OASIS Database Functionality. 2.3.1. Search Capabilities. The OASIS database can apply a variety of search queries, single, combined, or result-based. A single search is defined by one search condition. A detailed list of the available types of search conditions is presented in section A below. Combined search queries contain one or more queries combined with the logical operators AND, OR, or NOT. Single as well as other combined queries can be used to obtain a search query of arbitrary complexity and level of nesting. Result-based queries are executed over the results of a previous search. All requested queries and their results

are stored during a software session. The queries can also be saved to files and used, viewed, or modified in subsequent sessions. The database browser windows maintain a list of the requested queries. Whenever a new search is started, one will be asked if it is going to be a result-based one. Result-based searches can use any of the previously completed searches. Result-based queries can be themselves either single or combined.

A. Single Search Queries. For examples of single search queries, see Chart 1.

B. Combined and Result-Based Search Queries. Single search queries could be organized in combined searches. Negation, conjunction, or disjunction are used to combine these single queries in complex queries. The logical operators can be applied to single as well as combined searches. For example, a query that looks for structures that have a CAS number between 200 000 and 1 000 000 *and* are in TSCA but not IUCLID *and* have a molecular weight of more than 300 *and* either have a phenolic fragment with a halogen attached or contain two oxygen atoms whose distance is between 10 and 12 Å can be formulated in a number of equivalent ways. One of them is combining all clauses in one query, as shown in Scheme 1, where Q0 stands for CAS BETWEEN 200 000 AND 1 000 000; Q1 belongs to TSCA; Q2 belongs to IUCLID; Q3 is MOL_WEIGHT > 300; Q5 is Fragment c1ccccc1RX1 where the wild atom RX1 stands for F,Cl,Br,I; and Q6 is O_O{10 < DISTANCE < 12}}; note that single queries are displayed green.

Another equivalent search alternative is to execute some queries first and subsequently the rest of the queries on the subset resulted after the first search. Such a searching scenario is called a result-based search. Result-based queries can in some cases be more efficient and have the additional

Figure 4. Model browser designed to view or edit the models available in the database. In the illustration, model 12 holds for measured RB affinity in the rat, at 25 °C; data produced by John Katzenellenbogen et al.^{14,15}

advantage of showing the intermediate results. Result sets of equivalent queries formulated either way coincide.

C. Similarity Search. Similarity search queries return sets of chemicals similar to a query chemical within a specified similarity threshold. A similarity search is based on the notion of *degree of similarity*, a real number between 0 and 100, calculated for all chemicals of interest, where 0 corresponds to total dissimilarity and 100 to total similarity. Because the basic data types that can take part in the similarity calculation are diverse, different cases are to be considered. For this reason, we use a special version of the degree of similarity, defined below, and based on known statistical distance measures.^{16,17} Given the user-defined similarity threshold, the search results include the chemicals exhibiting a degree of similarity more than or equal to the threshold.

There are two types of similarity search: descriptor-based and fragment-based.

A descriptor-based similarity search requires a chemical *S* to be compared with chemicals *C* with respect to structural (observed or calculated) or conformational descriptors from a list of descriptors d_1, \dots, d_n .¹⁶ The descriptors are used in the similarity calculation as follows.

The differences between structural descriptors are normalized against the variety of possible values over the database. In this respect, for each structural descriptor d_i , we calculate its range of values r_i over the whole database:

$$r_i = \frac{\max[d_i(C)] - \min[d_i(C)]}{\text{C is a chemical from the database}} \quad (1)$$

We define the contribution v_i of each structural descriptor d_i of the similarity degree as

$$v_i = \frac{|d_i(C) - d_i(S)|}{r_i} \quad (2)$$

where *C* is a chemical to compare with the query chemical *S*.

When using conformational descriptors for a similarity calculation, there are sets of values instead of just one value for each descriptor (because the chemicals can be represented by a number of conformers). In this case, conformational descriptor distributions (Figure 5) are compared by Hellinger metrics:¹⁷

$$v_i = \text{Hellinger Distance}\{\text{distribution}[d_i(C)], \text{distribution}[d_i(S)]\} \quad (3)$$

Next, the degree of similarity of a chemical *C* to the query chemical *S*, with respect to a list of descriptors d_1, \dots, d_n , is calculated as the mean value of descriptor contributions to the similarity:

$$\text{Similarity}(C, S, d_1, \dots, d_n) = 1 - \frac{\sum_{i=1}^n v_i}{n} \quad (4)$$

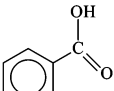
Two different algorithms are applied for estimating the fragment-based similarity: standard and advanced. The standard fragment-based similarity search counts the presence or absence of specified fragments as binary (0 or 1) contributions. The similarity is calculated by eq 4, where v_i is 1 if the corresponding fragment is found in the structure under consideration and is 0 otherwise.

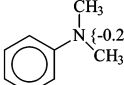
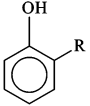
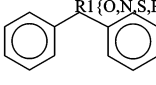


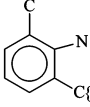
An advanced fragment similarity search is the average of two values, the first equal to the standard fragment-based similarity calculated as above and the second taking into account the number of atoms participating in matching atom-centered fragments.

To get the values as a percentage, the calculated similarity value (which is normalized between 0 and 1) is multiplied by 100.

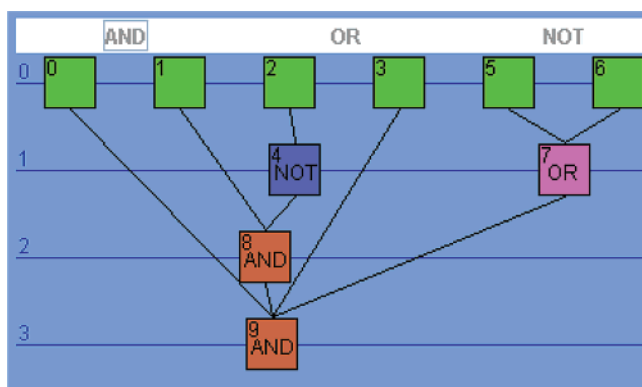
The results are presented in a descending order of their degree of similarity to the selected chemical, until the defined threshold is reached. The actual descriptors are selected by the user, and any type of descriptor combinations may be

Chart 1

Search type	Example
1. Structure identification search	
1.1. CAS search by CAS numbers equal to, or greater than (or equal to), or less than (or equal to), or different than a specified value, or within a specified range of values. Alternatively, more than one value can be provided for matching. Text files can be used to specify the lists of CAS numbers for searching.	<ul style="list-style-type: none"> > CAS=50033 > CAS > 100000 > CAS between 200,000 and 1,000,000 > CAS in { }
1.2. Name search for structures with names beginning with, or ending with, or containing, or exactly coinciding with a specified text string.	<ul style="list-style-type: none"> > Name contains 'acid' > Name begins with 'benz'
1.3. Database affiliation search for structures found in other specified chemical databases. Chemical database names could be selected from the list of available database affiliations.	<ul style="list-style-type: none"> > Structure in TSCA > Structure in TSCA and not in EUCLID
1.4. SMILES search for structures having a specific SMILES. This is not the fragment search also available in the OASIS Database platform. Instead, this is an additional tool for just literal matching of specific SMILES strings or substrings.	SMILES is "c1ccccc1"
2. Descriptor search: by calculated structural, observed, conformational, or atomic descriptors. Search is possible for structures having descriptor value equal to, or greater than (or equal to), or less than (or equal to), or different from a specified value, or within a specified range of values. For atomic and conformational descriptors, searches can be executed structure-wise or conformer-wise, depending on the selected option.	<ul style="list-style-type: none"> > MOL.WEIGHT > 300 > Log(Kow) < 0.5 > E_HOMO > -9 eV > Charge of O atom between -0.9 and -0.2 a.u.
3. Fixed fragment search with no varying characteristics or extra conditions. The possible characteristics and conditions that can vary are described in the next subsection. A special tool called Fragment Editor was designed to enable interactive building of fragments and structures to search for. A fragment can be constructed with the editor or just specified by typing the corresponding SMILES.	 <chem>C(=O)(O)c1ccccc1</chem>

Search type	Example
4. Extended fragment search queries can include additional conditions on the fragment to be found. The following types of additional conditions are implemented:	 <chem>c1(N(-0.28<Q<-0.38)(C)C)ccccc1</chem>
> Wildcard atom; one can use as many wildcard atoms into a fragment as one wishes	 <chem>Rc1c(O)ccccc1</chem>
> Enumerated wildcard atom. Create a list of alternatives for an atom and mark the desired site in the fragment.	 <chem>R1{O,N,S,P}</chem> <chem>c1([R1])c2ccccc2)ccccc1</chem>
> Atomic conditions. Conditions involving atomic descriptors can be requested for any atom in the fragment. A Local condition box is used to select an atomic descriptor and a range of its desired values.	 <chem>R1{O,N}</chem> <chem>c1([R1]){0.28<DONOR.DELOC.<0.40}(C)C)ccccc1</chem>
> Distance. Structures having a specified distance between specified atoms or fragments can be searched.	 <chem>O_O[5<DISTANCE<7]</chem>
> Atom qualifiers. Additional specifications could be required for any atom in the fragment to fulfill various conditions. A list of all available qualifiers and their meanings can be found in Appendix III.	 <chem>c1(C)c(N)c(C{acy})ccccc1</chem>

Scheme 1



used (observed, calculated, and conformational descriptors) in the same query and so forth.

2.3.2. Browsing. The database browser is the subsystem designed for interactive work with the OASIS databases. Database browsers can display either a whole database or a set of structures resulting from a search.

In addition to the list of chemicals and their basic data, database browsers can display all available data for each chemical: conformers; CAS numbers and names; 2D and

3D pictures; structural, conformational, and atomic descriptors; database affiliations (the known chemical databases that contain data on the chemical); and quality assurance data.

The 2D box and parameters associated with the 2D molecular structure are illustrated in Figure 6.

The 3D box has the complete functionality known from other OASIS software (rotate, zoom, auto-rotate, local descriptors check, atom properties, calculation of distances, angles, and torsion angles). The differences in 3D descriptor values for two of the conformers of n4-pentylphenol (CAS 14938353) are shown in Figure 7. The list of all representative conformers of the chemicals with their global and local molecular descriptors is available.

Mixtures, hydrolyzed structures, and polymers can be viewed collapsed (without their components) or expanded (the components are listed along with the chemical). All available data on components are also displayed (Figure 8).

The browser window is configurable—the set of items to display can be selected manually or automatically.

2.3.3. Editing. The database editor provides the following functionalities (Figure 9): (i) view and modify all types of structural and conformational data; (ii) undo and redo functionality for all operations; (iii) append (also by structure

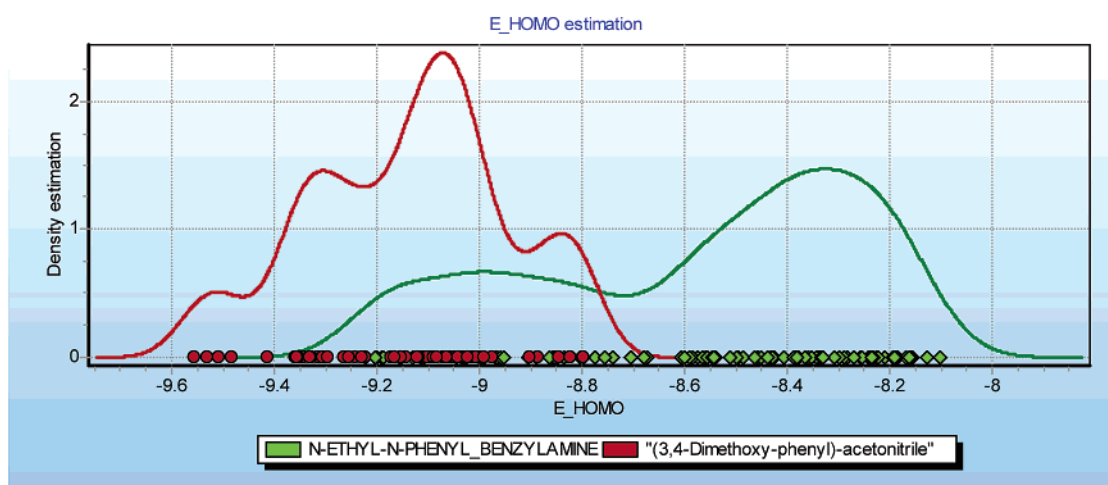


Figure 5. Conformational distribution of two chemicals—*N*-ethyl-*N*-phenyl benzylamine (CAS 92-59-1) and (3,4-dimethoxy-phenyl)-acetonitrile (CAS 93-17-4)—across *E*(LUMO).

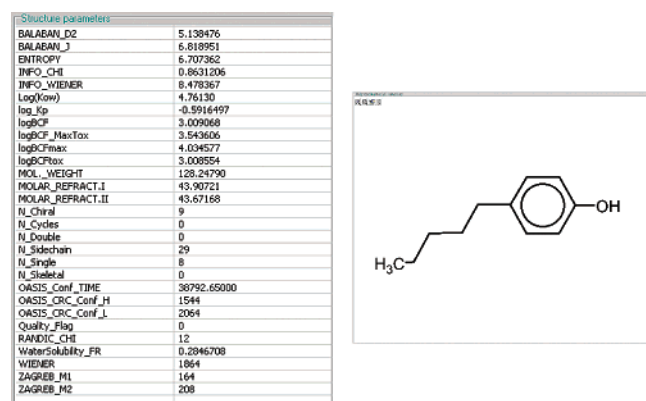


Figure 6. 2D box and list of parameters related to 2D molecular structure.

drawing); (iv) delete structures, fragments, descriptors; and (v) update structures, fragments, descriptor names, or values.

The structure editor is designed for the interactive building of fragments and structures. Drag and drop operations are used to build fragments from available palettes of atom and bond types and simple fragments. The adding, editing, and cutting of atoms, bonds, and fragments are available. Wildcard atoms (standing for any atom) and enumerated wildcard atoms (e.g., "F,Cl,Br,I") can be used.

2.3.4. Database Statistics. The database statistics provide charts about the database affiliation, quality of chemicals, and distribution of chemicals across selected physicochemical parameters. The following features are included, here: (i) the database affiliation (contributions of different inventories in the database), (ii) the database affiliation of unique chemicals (unique chemicals from different inventories in the centralized database), (iii) the quality of quantum chemical calculations (MOPAC) by records (the quality of quantum chemical optimization procedures), (iv) the record quality filter (the quality of quantum of records with respect to their 3D geometry), (v) conformer multiplication (the degree of conformer multiplication across the database), (vi) the distribution of metabolites across selected physicochemical parameters, (vii) the distribution of structures according to their chemical type (the number and percentage of discrete chemicals, hydrolyzing structures, inorganic structures, polymers, defined or undefined mixtures; Figure 10).

2.3.5. Data Import and Export. The OASIS database provides the import of structural (2D and 3D) and parameter data stored in the following types of files:

SMI. SMI files are text files where, except for SMILES strings, the system recognizes information on CAS numbers and chemical names.

SDF. SDF files can be used to import CAS numbers, chemical names, 2D and 3D structure data, and numeric descriptors.

MOL. MOL files are text files carrying data on chemical structures.

XYZ. XYZ files are text files carrying data on chemical structures.

CMP. A CMP file is a binary file in a legacy OASIS format that stores series of chemicals and all pertinent data describing them. Each logical record of the file describes a separate chemical or a particular conformer of a chemical.

DBF. The system uses Dbase database files to import CAS numbers of structures and various numeric and character string descriptors.

Text Files. The software can work with text files including columns (tab- or blank-delimited) that can store, for example, CAS numbers of structures and various numeric and string descriptors. This format could be used to input parameters associated with chemicals specified by their CAS or 2D structure.

The whole database or any subset of it can be exported in the following file formats:

Excel. A tab-delimited text file is generated that contains a user-selected combination of CAS number, 2D structure, name, or any collection of descriptors (calculated, observed, or conformational).

HTML. A HTML file is generated that contains a user-defined combination of CAS number, 2D structure, name, or any collection of descriptors (calculated, observed, or conformational). 2D structures can be exported as images. Image files are compressed, and appropriate links to them are created.

OASIS Database. Any selection of structures can be saved as a new OASIS database or merged into an existing one.

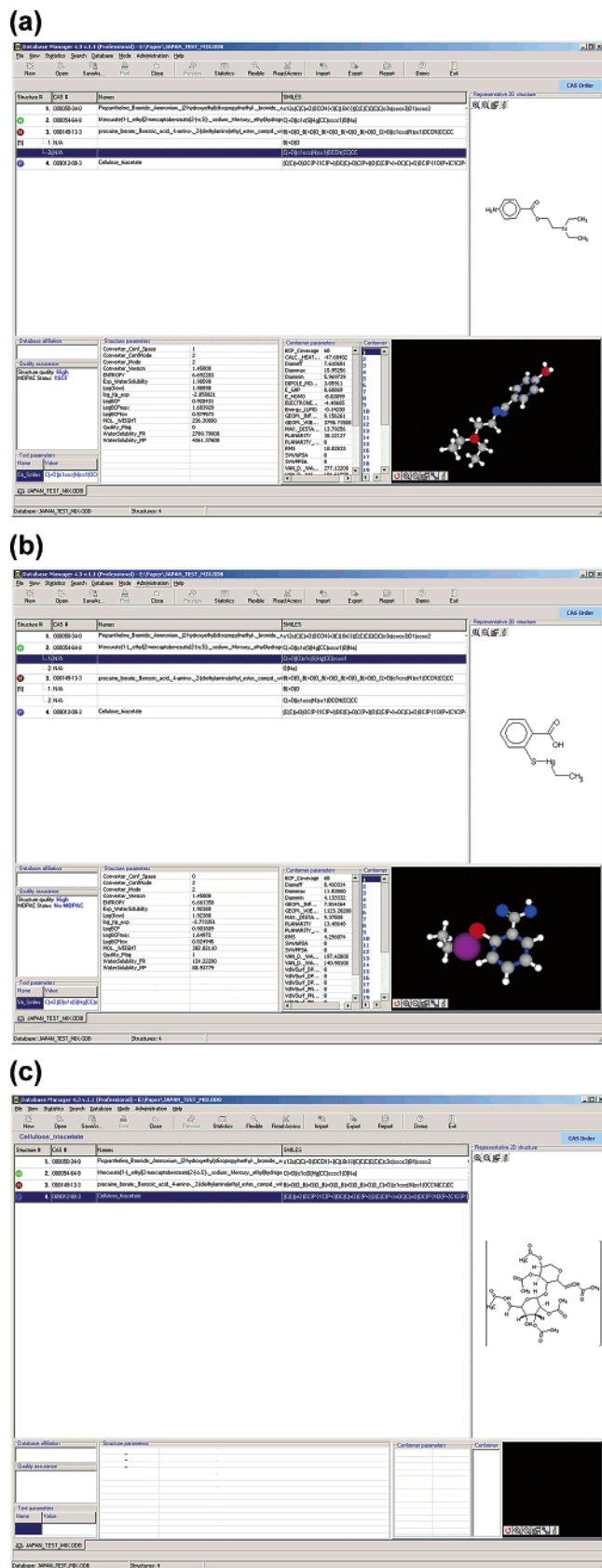


Figure 8. OASIS Database browser displaying discrete structures, mixtures, hydrolyzed structures, and polymers. (a) The “M” letter indicates a mixture, and the figure at the row of the components indicates the number of times it occurs in the mixture. One of the components of the mixture is presented here. (b) The “H” stands for a chemical which is hydrolyzed; the associated hydrolysis products could be seen, as shown. (c) The “P” letter indicates a polymer.

interpreter—there are (free) versions of it for all major operating systems; for some, it is even included by default.

Interface Subsystem. The interface subsystem collects users’ requests and transfers them on to the engine subsystem. A 2D editor provides an interface for fragment and structure drawing (Figure 14). The system could read structural information coded in commonly used connectivity formats, such as SDF. The interface subsystem is written in Java and HTML.¹⁸ It can be deployed on any Web server (no limitations on the server’s operating system).

Engine Subsystem. The engine subsystem performs requested search queries on the database and generates reports. These comprise search queries beyond standard SQL processing, for example, fragment queries or fragment or atomic distances. For these reasons, the engine subsystem includes executable modules of the OASIS database platform. It requires a Windows Web server with Internet Server Application Program Interface (ISAPI) capabilities. All connections from and to the interface subsystem are done through Transmission Control Protocol/Internet Protocol (TCP/IP), so the two subsystems may be on separate servers or on the same server.

All connections to the database are also done through TCP/IP, so the database may reside in turn on the same or a different server.

Database. The database consists of an Interbase/Firebird SQL server (third-party product for which both free and commercial versions exist) and OASIS database files.

The OASIS Web database has been implemented in the European Chemicals Bureau QSAR Web Site available at <http://ecbqsar.jrc.it>. The implementation contains QSAR data by the Danish EPA maintained in an OASIS database and is coupled with the OASIS Web database software.

3. RESULTS AND DISCUSSION

3.1. The Centralized 3D Database. Using the OASIS software framework for building databases and managing chemical information, we built a centralized 3D database. This is the largest OASIS database of existing chemicals to date. Chemicals which are under regulation by the respective government agencies are considered as “existing”. The individual databases of regulatory agencies in North America and Europe, including IUCLID of the European Chemicals Bureau (with 61 573 chemicals); the Danish EPA database (159 448 chemicals); TSCA (56 882 chemicals), HPVC_EU (4750 chemicals), HPVC_USEPA (10 546 chemicals), and pesticides’ active and inactive ingredients of the U.S. EPA (1379); DSL of Environment Canada (10 851 chemicals); and the Japanese Ministry of Economy, Trade, and Industry database (16 811), were combined in the database (Figure 15). The structural information for all chemicals is precalculated in terms of the conformer multiplication of all chemicals and the quantum-chemical optimization of each conformer. The 2D–3D migration, conformational multiplication, and quantum-chemical evaluation are described elsewhere (ref 19). Presently, the database contains approximately 185 500 structures and 3 700 000 conformers with hundreds of millions of descriptor data items. The platform maintains 2D and 3D data and molecular descriptors

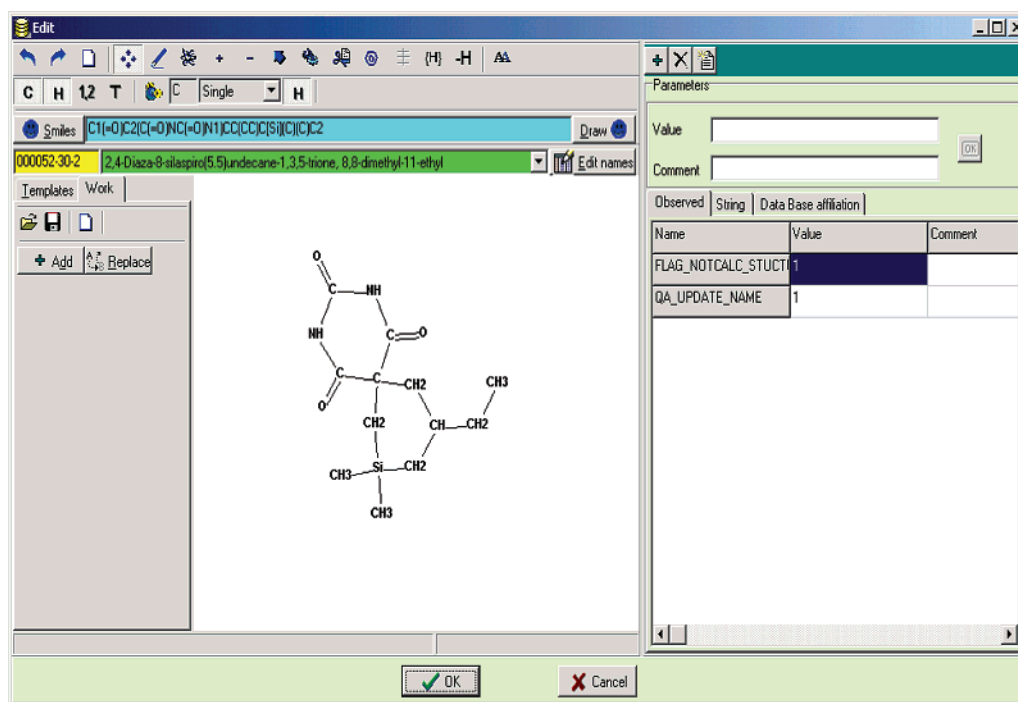


Figure 9. Interfacing window of the structure editor.

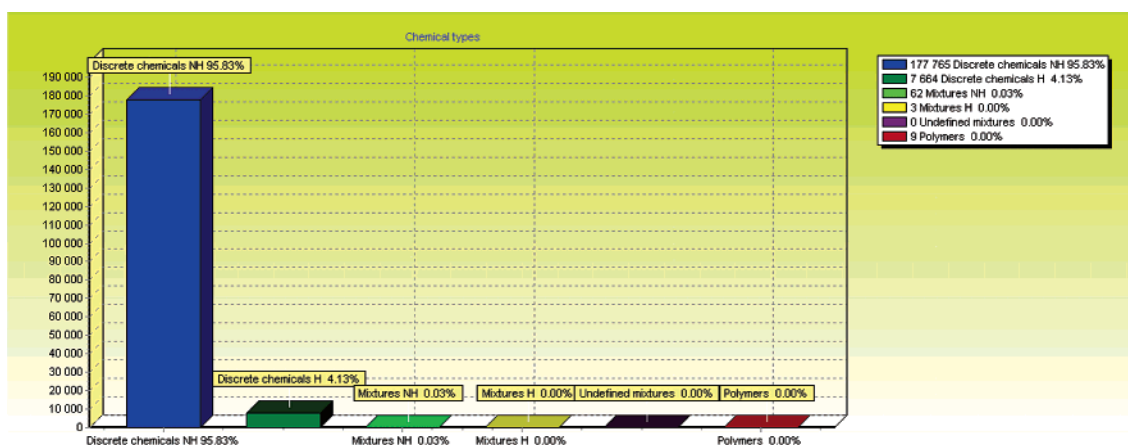


Figure 10. Distribution of structures of different chemical types in a data set.

for discrete chemical structures (defined and undefined), mixtures, hydrolyzing chemicals, and polymers; a search by CAS numbers, names, and descriptors; an extensive 2D and 3D fragment search including atom modifiers, wildcard atoms, and distances; a similarity search; logical combinations; saved queries; and result exports. A database browser is implemented. A 2D-structure editor was also developed for the defining and editing of structures and fragments. Database statistics, descriptor distribution (Figure 16), and model correlation tools are available. A report generator is included. The platform is compatible at the import/export level with a number of known formats.

A software suite for the implementation of OASIS database functionality on WWW servers is also developed.

An information technology solution based on distributed computing was developed to overcome the time complexity of the OASIS 2D–3D migration of the centralized 3D database chemicals—including their conformer multiplication and quantum chemicals optimization. All PCs of the Laboratory of Mathematical Chemistry (by that time, about 25 using

P4 CPUs at 2 GHz—on average) are working on an intranet and are connected with a server where the database is located. Each station takes untreated chemicals from the database and works on them all the time when it is not occupied by other tasks; for this reason, the distributed mode of the OASIS 2D–3D migration procedure was called “idle calculation”. A report on the performance of all computers can be provided any time in a graph (the number of calculated structures, the currently processed structure, etc.). The distributed computing approach significantly accelerates the computing job. The enhanced computation effectiveness allowed the task of conformational multiplication and quantum chemical assessment (in PRECISE mode) of all existing chemicals to be accomplished in several months.

The precalculation of the chemicals in the centralized 3D database combined with the flexible searching capabilities (on 2D and 3D levels) allows testing hypotheses on the structural conditioning of modeled endpoints. Thus, the search of the database for chemicals which could elicit

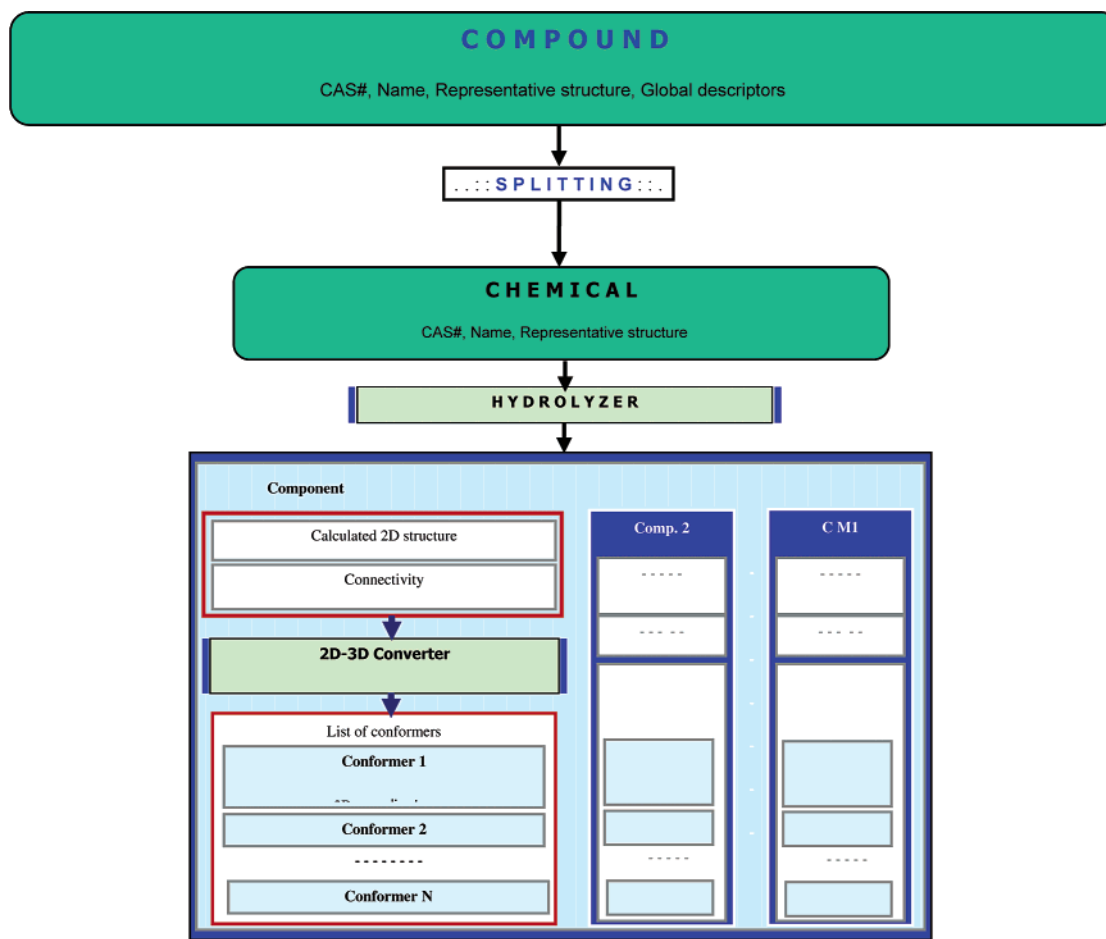
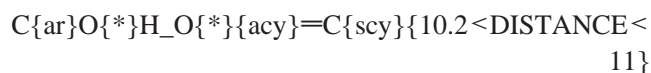


Figure 11. Processing chart for every chemical that is added to the database.

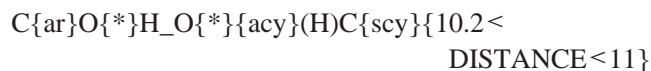
FixedText	CAS	Calc. Smiles	Repres. Smiles	Name	Picture	Parameter	Table	All parameters	Line	New page
Selected item (10,209,724,5) : Line										
CAS		NAME								
Picture 2D presentation		Molecular weight		MOL_WEIGHT						
		Log(Kow)		LOG(KOW)						
		E_HOMO		E_HOMO						

Figure 12. A report template.

significant estrogen receptor (ER) binding affinity with the earlier-defined 3D structural pattern⁷



or



and VAN_D_WAALS_SUR in the range of 284–365 Å² shows that 201 out of around 185 000 chemicals could be potential ER ligands. In the above expression of the search query, C{ar}O{*}H_O{*}\{acy\}=C{scy} holds for a distance range of 10.2–11.0 Å between the O atom of a hydroxyl group attached to an aromatic carbon (C{ar}O-{*}H) and an acyclic O atom bound with a cyclic C atom

by a double bond (O{*}\{acy\}=C{scy}) or a single bond (O{*}\{acy\}{H}C{scy}) (* indicates the atoms between which the distance is specified).

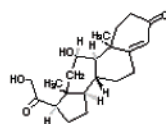
3.2. Search Capabilities and Efficiency of the OASIS Centralized QSAR Database. As an OASIS database, the centralized QSAR database is able to perform all types of searches featured by the OASIS database platform. This subsection presents some tests of the efficiency of search queries executed on the centralized QSAR database.

The tests include a series of single and combined searches. All tests were performed on two different systems, denoted below by PC and Web, respectively. PC is a desktop computer using the OASIS database platform, while Web is a WWW server running the OASIS web database. The Web server was outside the local area network of the PC where testing was taking place.

Report for: D:\Downloads\q\ODB DemoFile.ODB created on 12/27/2005 4:21:52 PM

50226

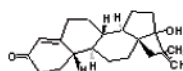
Corticosterone



Molecular weight	346.000
Log(Kow)	1.940
E_HOMO	min=-10.100 max=-9.950

797637

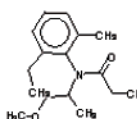
norgestrel



Molecular weight	312.000
Log(Kow)	3.480
E_HOMO	min=-10.000 max=-9.910

51218452

Metolachlor



Molecular weight	284.000
Log(Kow)	3.240
E_HOMO	min=-9.620 max=-9.190

Figure 13. Report produced for a set of three chemicals with the report template of Figure 12.

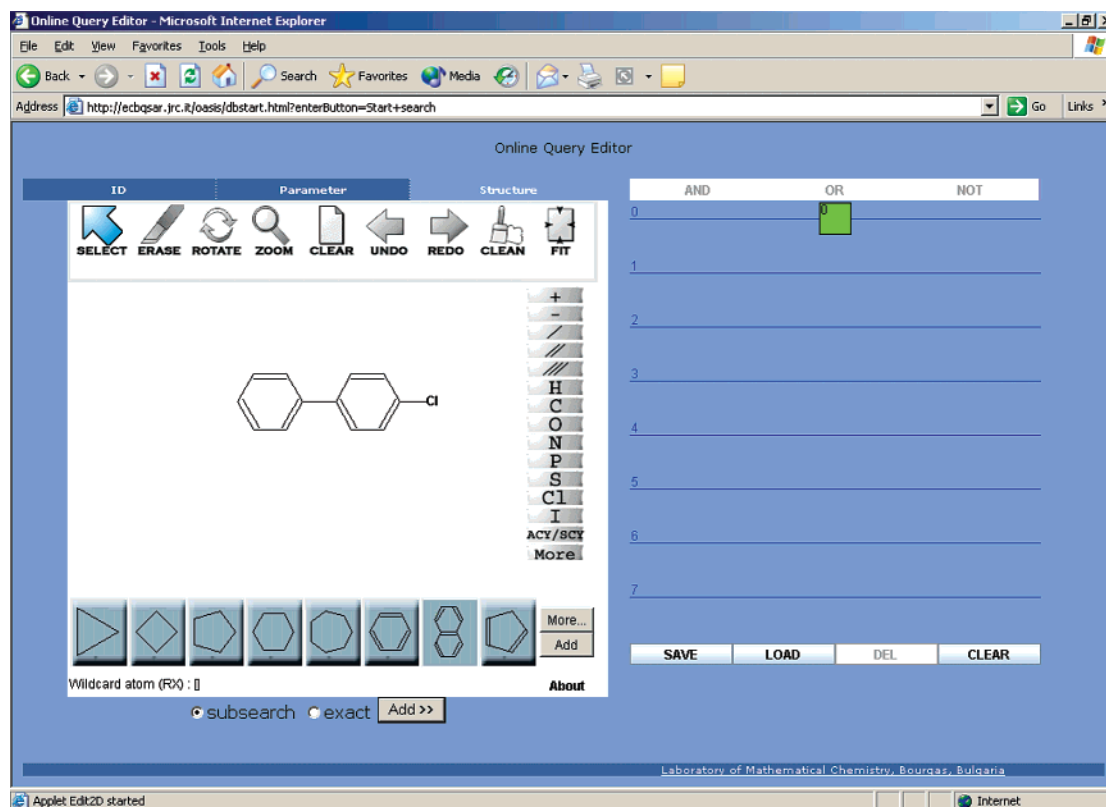


Figure 14. Interface for fragment and structure drawing of the OASIS Web database.

PC. This system consisted of the following: an Intel Pentium D dual-core 820 CPU at 2.8 GHz, 2 GB of RAM, ATA100 HDD, and the Windows 2003 Standard

Edition operating system. The New Technology File System (NTFS) was used on both the system and the testing partition.

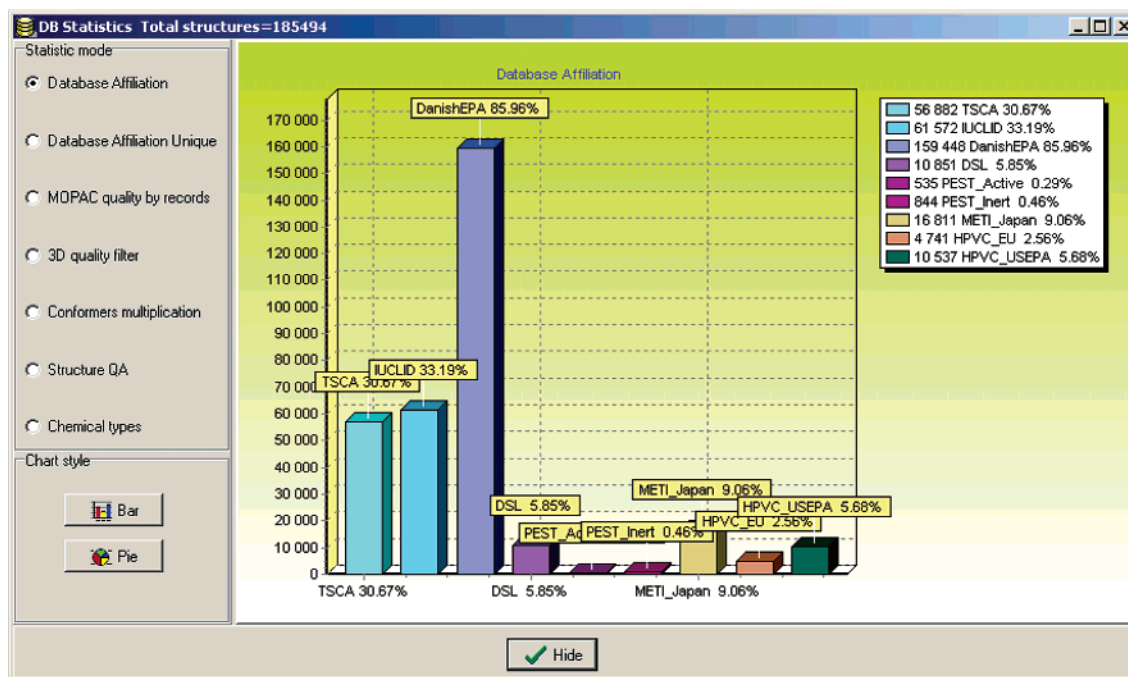


Figure 15. Rough data illustrating the database statistics. The affiliation of chemicals in the centralized 3D database across the different inventories.

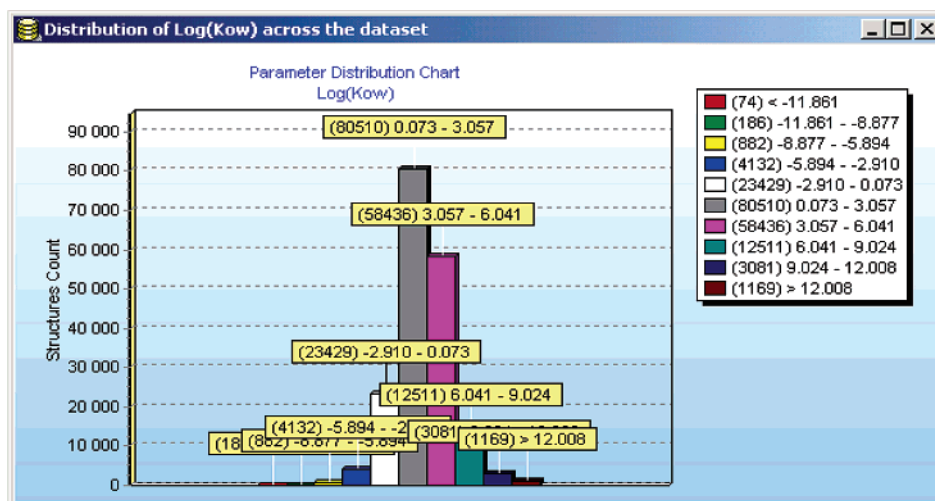


Figure 16. Database statistics: distribution of chemicals in centralized 3D database with respect to log Kow.²¹

Web. This system consisted of the following: an Intel Xeon x2 CPU at 2.8 GHz, 2 GB of RAM, SCSI HDD, the Windows 2003 Standard Edition operating system, and the NTFS file system.

The reference database is the OASIS Centralized QSAR Database with 185 500 structures.

Each test result was measured three times, and the average value was taken. The timings include not only the database search but also data retrieval and transfer to the client application.

The results of individual search tests, that is, each test was performed on the entire database, are presented in Figure 17.

The efficiency of the Web system is superior to the desktop one because, in addition to running on a server CPU and hard disk drives, the Web system contains modified search algorithms. These are optimized for better performance in the 2D search, especially the fragment search. However, the

version of the Web system used for testing currently lacks 3D functionality and atomic descriptors.

The tests in Figure 18 demonstrate the efficiency of combined queries for both the desktop and the Web versions of the OASIS database platform.

The tests in Figure 19 demonstrate the use of result-based queries—each query was based on the results of the previous one. This statistics is only available for the desktop version.

The results show a good efficiency of the search subsystem of the OASIS database platform on a large data set such as the OASIS Centralized QSAR database. Generally, the efficiency depends on the complexity of the query, on the amount of data items that should be checked, and on the number of its results. Some types of complexity issues can be handled, for example, by improving the relational database design. However, there are always types of searches that are not influenced by database considerations only, like the fragment search which reduces to a graph-theoretical prob-

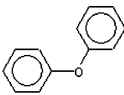
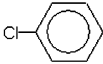
Query	Structures found	PC; Time[s]	Web; Time [s]
CAS = 19640370	1	<1	<1
CAS < 19640370	92556	4	4
CAS >= 19640370	92938	2	2
Name contains 'acid'	28101	4	3
Structure in TSCA	56882	3	4
LogBCFTox > -5	173838	8	4
MOL.Weight >= 600	4913	3	3
E_HOMO > -8	19255	9	N/A
Cl-atom: Charge > 0.5	369	21	N/A
Cl-atom: Charge > 0	12782	29	N/A
2D structure is C(C)(=O)OC(C)OC	1	<1	<1
2D structure contains 	1179	7	3
2D structure contains 	14415	10	3

Figure 17. Efficiency of search queries over the entire database. Each search query is performed independently. "PC" and "Web" show the corresponding timings in seconds.

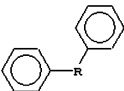
Query	Structures found	PC; Time[s]	Web; Time[s]
CAS >= 19640370 and structure in TSCA and MOL.Weight >= 600 and 2D structure contains  (R stands for any atom)	136	20	12

Figure 18. Efficiency of combined search queries. Each query is performed over the entire database.

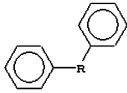
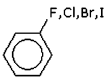
Query	Structures found	PC; Time [s]
CAS >= 19640370	92938	2
and structure in TSCA	26778	3
and MOL.Weight >= 600	834	1
2D structure contains  (R stands for any atom)	136	6
3D query: there are 2 oxygen atoms at a distance between 3 and 9 Å (for at least one conformer)	122	3
2D structure contains 	27	6

Figure 19. Efficiency of result-based search queries. Each query is searching within the results of the previous one.

lem. However, because of specially developed intelligent search algorithms, even in these cases, the amount of data items that should be checked can sometimes be drastically

Table 1.

item	maximum number
structures	unlimited
structure sets	unlimited
conformers	unlimited
conformers belonging to one structure set	unlimited
calculated descriptors (numeric)	65 535
calculated descriptors (character string)	65 535
observed descriptors (numeric)	65 535
observed descriptors (character string)	65 535
conformational descriptors	65 535
atomic descriptors	65 535
SMILES string length, characters	3000
descriptor name length, characters (all kinds of descriptors)	3000
chemical name length, characters	3200
chemical names of a structure	unlimited
components of a mixture/hydrolyzed structure	unlimited
models	unlimited

Table 2.

atom qualifier	meaning
{ACY}	atom not participating in rings
{SCY}	atom participating in one ring
{DCY}	atom participating in two or more rings
{SP1}	atom in sp ¹ hybridization
{SP2}	atom in sp ² hybridization
{SP3}	atom in sp ³ hybridization
{AR}	atom participating in aromatic rings
{SK}	skeletal atom
{SC}	nonskeletal atom
{2+}	dication
{2-}	dianion
{P+}	stereocenter of positive parity
{P-}	stereocenter of negative parity
{H0}	no hydrogen attached
{H4}	four hydrogens attached
{H3}	three hydrogens attached
{H2}	two hydrogens attached
{H}	one hydrogen attached
{+}	monocation
{-}	monoanion
{.}	monoradical
{*}	
BOND_ORDER	bond order descriptor

reduced, which leads to a significant improvement of 2D fragment search algorithms in the current version of the OASIS database platform (see Appendices I and II). When this is possible, the reduced resulting amount of data becomes a crucial factor in the search and data retrieval speed and can even lead to higher performance of the fragment search queries compared to some simpler search types (Figure 17).

ACKNOWLEDGMENT

Research associated with this paper was funded in part through an EPA cooperative research agreement (CR 83199501-0), EU VI Framework Project ReProTest, and research agreements with ExxonMobil and Unilever. Gratitude is expressed to Dr. Sabcho Dimitrov for the discussions improving the quality of the paper.

APPENDIX 1. OASIS DATABASE PLATFORM TECHNICAL DATA

Software Requirements. The following are the software requirements for this project:

(i) *MS Windows 98/ME/NT/2000/XP/2003.* The OASIS database runs on all 32-bit MS Windows platforms. On Windows NT/2000/XP/2003, NTFS is recommended for performance.

(ii) *Interbase/Firebird Server and Client Libraries.*

(iii) *OASIS Web Database Only*. The client side requires the Sun Java Runtime Environment and a Java-enabled HTTP browser, and the server side requires an ISAPI-capable HTTP server.

Limitations. Table 1 lists the maximum possible numbers of data items. For "unlimited" values, the theoretical limitations are so high that actual limits are usually imposed by hardware considerations.

APPENDIX 2. OASIS CENTRALIZED DATABASE DATASHEET

Hardware Requirements. Generally, CPU/RAM requirements depend on the database size, with Professional mode requiring more memory.

For OASIS QSAR Centralized Database, 256 MB of RAM are required for the Basic mode and 512 MB RAM for the Professional mode.

The current version of OASIS QSAR Centralized Database requires 1 GB of disk space in Basic mode and 17 GB in Professional mode.

APPENDIX 3. ATOM QUALIFIERS IMPLEMENTED IN THE OASIS FRAGMENT SEARCH

Table 2 gives a list of the atom qualifiers used in the OASIS fragment search.

REFERENCES AND NOTES

- (1) Mekenyan, O. G.; Veith, G.; Bradbury, S.; Russom, C. A QSAR Approach to Estimating the Toxicity of Unsaturated Alcohols to the Fathead Minnow (*Pimephales promelas*). *Quant. Struct.-Act. Relat.* **1993**, *12*, 132–136.
- (2) Mekenyan, O. G.; Nikolova, N.; Schmieder, P. Dynamic 3D QSAR Techniques: Application in Toxicology. *THEOCHEM* **2003**, *622*, 147–165.
- (3) Eliel, E. L. Chemistry in Three Dimensions. In *Chemical Structures*; Warr, W. A., Ed.; Springer: Berlin, 1993; p 1.
- (4) Ivanov, J. M.; Mekenyan, O. G.; Bradbury, S. P.; Schuurmann, G. A Kinetic Analysis of the Conformational Flexibility of Steroids. *Quant. Struct.-Act. Relat.* **1998**, *17*, 437–449.
- (5) Mekenyan, O. G.; Nikolova, N.; Schmieder, P.; Veith, G. D. COREPA-M: A Multi-Dimensional Formulation of COREPA. *QSAR Comb. Sci.* **2004**, *23*, 5–18.
- (6) Dimitrov, S.; Schmieder, P.; Veith, G. In Silico Modelling of Hazard Endpoints: Current Problems and Perspectives. *SAR QSAR Environ. Res.* **2003**, *14* (5–6), 361–371.
- (7) Mekenyan, O. G.; Kamenska, V.; Serafimova, R.; Poellinger, R. L.; Brower, A.; Walker, J. Development and Validation of an Average Mammalian Estrogen Receptor-Based QSAR Model. —*SAR QSAR Environ. Res.* **2002**, *13* (6), 579–595.
- (8) Ankley, G. T.; Mekenyan, O. G.; Kamenska, V.; Schmieder, P.; Bradbury, S. Reactivity Profiles of Ligands of Mammalian Retinoic Acid Receptors: A Preliminary Corepa Analysis.— *SAR QSAR Environ. Res.* **2002**, *13* (2), 365–377.
- (9) Ivanov, J.; Karabunarliev, St.; Mekenyan, O. G. 3DGEN: A System For an Exhaustive 3D Molecular Design. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 234–243.
- (10) Mekenyan, O. G.; Dimitrov, D.; Nikolova, N.; Karabunarliev, St. Conformational Coverage by a Genetic Algorithm. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (6), 997–1016.
- (11) Mekenyan, O. G.; Pavlov, T.; Grancharov, V.; Todorov, M.; Schmieder, P.; Veith, G. 2D-3D Migration of Large Chemical Inventories with Conformational Multiplication. Application of the Genetic Algorithm. *J. Chem. Inf. Model.* **2005**, *45* (2), 283–292.
- (12) Stewart, J. J. P. MOPAC: A Semiempirical Molecular Orbital Program. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1–105.
- (13) Stewart, J. J. P. *MOPAC 93*; Fujitsu Limited: Chiba 261, Japan; Stewart Computational Chemistry: Colorado Springs, CO, 1993.
- (14) Muthyala, R. S.; Sheng, S.; Carlson, K. E.; Katzenellenbogen, B. S.; Katzenellenbogen, J. A. Bridged Bicyclic Cores Containing a 1,1-Diarylethylene Motif Are High-Affinity Subtype-Selective Ligands for the Estrogen Receptor. *J. Med. Chem.* **2003**, *46*, 1589–1602.
- (15) Harris, H.; Katzenellenbogen, J. A.; Katzenellenbogen, B. S. Characterization of the Biological Roles of the Estrogen Receptors, ERalpha and ERbeta, in Estrogen Target Tissues in Vivo through the Use of an ERalpha-Selective Ligand. *Endocrinology* **2002**, *143*, 4172–4177.
- (16) Gibbs, A. L.; Su, F. E. On Choosing and Bounding Probability Metrics. *Int. Stat. Rev.* **2002**, *70* (3), 419–435.
- (17) Duda, R. O.; Hart, P. E.; Stork, D. *Pattern Classification*, 2nd ed.; John Wiley & Sons: New York, 2000; pp 538–542.
- (18) Developed in cooperation with the Syracuse Research Corporation.
- (19) Mekenyan, O. G.; Pavlov, T.; Grancharov, V.; Todorov, M.; Schmieder, P.; Veith, G. 2D-3D Migration of Large Chemical Inventories with Conformational Multiplication. Application of the Genetic Algorithm. *J. Chem. Inf. Model.* **2005**, *45* (2), 283–292.
- (20) Meylan, W. M.; Howard, P. H.; Boethling, R. S.; Aronson, D.; Printup, H.; Gouchie, S. Improved Method for Estimating Bioconcentration/Bioaccumulation Factor from Octanol/Water Partition Coefficient. *Environ. Toxicol. Chem.* **1999**, *18*, 664–672.

CI060142Y