United States
Environmental Protection
Agency

Environmental Research
Laboratory
Athens GA 30613-7799

Research and Development

EPA/600/M-89/017 Aug.1989

EPA

ENVIRONMENTAL RESEARCH BRIEF

Computer Prediction of Chemical Reactivity—The Ultimate SAR

Samuel W. Karickhoff,[1] Lionel A. Carreira,[2] Clyde Melton,[3] Valeta K. McDaniel,[1]
Andre N. Vellino,[3] and Donald E. Nute[3]

Introduction

Approximately 70,000 industrial chemicals are listed by EPA's Office of Toxic Substances. As the evaluation and management of the environmental and human risk associated with the proliferation of these anthropogenic chemicals becomes an increasingly urgent social priority, there accrues a corresponding demand on physical and biological scientists and engineers to provide effective techniques for quantifying their release, fate, and potential environmental damage. Historically, Federal health and safety regulators (i.e., the U.S. Environmental Protection Agency and the Occupational Safety and Health Administration) have relied on field monitoring, toxicological test data, and expert scientific knowledge to condemn or vindicate a given chemical. Recent emphasis, however, on more quantitative and comprehensive risk/benefit analyses, coupled with the extension of the regulators' "umbrella" (via the Toxic Substances Control Act) over all manufactured chemicals (including those in the pre-manufacturing stages) has required the development of more sophisticated evaluation methods. These methods must be capable of forecasting pollutant behavior over wide ranges of environmental conditions, often without the benefit of measured data specific to the chemical or ecosystem in question.

Although a wide variety of approaches can be used in such judgmental exercises, a knowledge of the relevant chemistry of the compound in question is critical to any assessment scenario. For volatilization, sorption and other physical processes, considerable success has been achieved in not only phenomenological process modeling but also *a priori* estimation of requisite chemical parameters such as solubilities and Henry's Law constants.(1-3)

Although considerable progress has been made in process elucidation and modeling for chemical processes such as photolysis and hydrolysis(4-8), reliable estimates of the related fundamental chemical constants (i.e., rate and equilibrium constants) have been achieved for only a small number of molecular structures. The values of parameters, in most instances, must be derived from measurements or from the expert judgment of specialists in that particular area of chemistry. Parametric values

[1]Environmental Research Laboratory, USEPA, Athens, GA 30613-7799.
[2]Chemisry Department, University of Georgia, Athens, GA 30602.
[3]Advanced Computational Methods Center, University of Georgia, Athens, GA 30602

have actually been measured for, perhaps, only about one percent of the chemicals in the OTS inventory. Because these measurements may easily cost $20,000 to $100,000 per chemical, estimation techniques for these parametric values are very cost-effective. In any case, trained technicians and adequate facilities are not available for a measurement effort involving thousands of chemicals. We describe here a prototype system for the estimation of reactivity parameters that will cost the user only a few minutes of computer time.

This work seeks to develop methods for the computer estimation of fundamental reactivity parameters strictly from molecular structure. Although the prototype system, called SPARC (SPARC Performs Automated Reasoning in Chemistry), presently deals only with the prediction of UV-Visible light absorption spectra and pK_a, the techniques discussed below are being extended to other reactivity parameters such as hydrolysis rate constants. Any predictive method should be understood in terms of the purpose for which it is conceived and should be structured by appropriate operational constraints. The methods described herein are intended for what might be characterized as engineering applications in environmental assessments. More specifically they provide:

- an *a priori* estimate of reactivity parameters for chemical process models when measured data are unavailable,
- guidelines for ranking a large number of chemical parameters and processes in terms of relevance to the question at hand, thus establishing priorities for measurement or study,
- an evaluation or screening mechanism for existing data based on expected behavior, and guidelines for interpreting or understanding existing data and observed phenomena.

The primary operational constraint is that the data available for testing and calibrating predictive theories are limited in both quantity and quality. This lack of data recommends a less empirical--that is, more theoretical--approach moderated by the operational axiom that complexity not exceed need. In addition, predictive capability should extend to the entire world of organic chemicals. In particular, the theory should not be crippled by the computation and calibration requirements of large (molecular weights greater than 200) polyfunctional molecules.

Scientific Computing

Until recently, scientific computing consisted almost entirely of mathematical calculations. These programs, although they often contain deep and involved mathematics, do not typically express a fully developed scientific theory. A scientific computer program, for example, might yield a solution to Schrodinger's equation for many particles using a self-consistent approximation method. Such scientific computer programs can be thought of as tools or instruments that extend the power of a theory, but they are not formulations of a theory.

The "new wave" computer technology of expert systems provides for imbedding theoretical knowledge as well as calculation algorithms into computer programs that, in principle, can shorten the distance between scientific theory and computer implementation. In most scientific applications to date, however, expert systems have fallen short of actual theory implementation, relying heavily on pattern matching or correlational inferencing in predictive strategy. Also, most of the current expert systems are oriented to a specific application and are targeted primarily to expert users within a particular scientific discipline. Interested readers should consult recent reviews by Gray[9] and Pierce and Hohne[10] in which existing expert systems for molecular structure elucidation, chemical synthesis, and molecular design are described.

In the field of organic chemistry, an extensively developed theoretical basis exists already for estimating chemical reactivity from molecular structure. This theoretical knowledge base refers to the body of facts, generalizations, models, laws, and theories that form the basis for mechanistic reasoning in physical organic chemistry. Mechanistic reasoning refers to the process of analyzing a chemical change in terms of more elementary component processes, such as critical motions of certain electrons or nuclei or sequential events through which the transformation proceeds. (For indepth descriptions, readers may consult physical organic chemistry texts such as that of Lowry and Richardson.[11])

The goal for our expert system is to capture the reasoning process that an organic chemist might undertake in reactivity analysis. The approach primarily involves deductive reasoning and is theory/mechanism oriented. Computational procedures are based on existing mathematical models of theoretical chemistry.

Chemical Modeling

Chemical properties describe molecules in transition, that is, the conversion of a reactant molecule to a different state or structure. For a given chemical property, the transition of interest may involve electron redistribution within a single molecule or bimolecular union to form a transition state or distinct product. The behavior of chemicals depends on the differences in electronic properties of the initial state of the system and the state of interest. For example, a light absorption spectrum reflects the differences in energy between the ground and excited electronic states of a given molecule. Moreover, chemical equilibrium—thus chemical equilibrium constants—depends on the energy differences between the reactants and products. Reaction rates, on the other hand, may depend on the energies of a transition state relative to either reactants or products.

For the reactions addressed in SPARC, these energy differences are extremely small compared to the total binding energies of the reactant involved. This presents a problem for *ab initio* computational procedures that calculate absolute energy values. Computing the relatively small energy differences needed for the analysis of chemical reactivity from absolute energies requires

extremely accurate calculations. Although achievable for small subclasses of molecules for certain reactivity parameters, these methods cannot provide the major computing thrust for SPARC considering the projected scope and aforementioned constraints.

There are, however, methods known under the general heading of "perturbation theories," that make it possible to calculate energy differences directly. These theories treat the final state as a perturbed initial state and the energy differences, then, are determined by quantifying the perturbation.

Perturbation methods can be used not only to predict reactivity of a given chemical but also to compute differences in reactivity for "similar" reactants. Although they provide potentially more accurate and simpler computations, the use of perturbation methods requires considerable circumspection. One must be careful to:

- define the boundary conditions of an algorithm's validity,
- choose appropriate reference states for calibration and extrapolation, and
- define, in terms of molecular structure, the meaning of terms like "chemical similarity."

These perturbation methods are ideally suited for expert system application due to their extreme flexibility and computational simplicity. The requisite conditions for applicability, as well as the selection of appropriate reference structures or reactions, can be easily built into the computation control portion of the expert system.

SPARC Computational Procedures

An indepth description of SPARC procedures is beyond the scope of this brief. The following, however, is a limited description of the logic of our approach to chemical parameter prediction and provides an overview of the reasoning and computational procedures currently incorporated in SPARC.

The basic philosophy is not to compute any chemical property from "first principles." Rather, it is to utilize directly the extensive knowledge base of organic chemistry. Organic chemists have established the types of structural groups or atomic arrays that impart certain types of reactivity and have described, in "mechanistic" terms, the effects on reactivity of other structural constituents appended to the site of reaction.

To encode this knowledge base, a classification scheme was developed that defines the role of structural constituents in effecting or modifying reactivity. Furthermore, models have been developed that quantify the various "mechanistic" descriptions commonly utilized in structure-reactivity analysis, such as induction, resonance, and field effects. SPARC execution (Figure 1) involves the classification of molecular structures (relative to a particular reactivity of interest) and the selection and execution of appropriate "mechanistic" models to quantify reactivity.

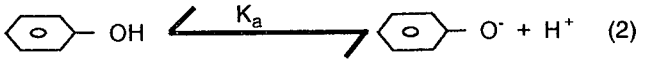
The computational approaches in SPARC are a blending of conventional linear free energy theory (LFET), structure activity relationships (SAR), and perturbed molecular orbital (PMO) methods. In general, SPARC utilizes LFET to compute thermal properties and PMO theory to describe quantum effects such as delocalization energies or polarizabilities of π electrons. In reality, every chemical property involves both quantum and thermal contributions and necessarily requires the use of both perturbation methods for prediction. These approaches have been extensively developed and utilized in physical organic chemistry. For detailed descriptions, readers should consult texts by Leffler and Grunwald[12] and Hammett[13] on LFET and SAR, Dewar[14] and Dewar and Dougherty[15] on PMO theory, and the reviews of Taft *et al.*[16] on SAR applications.

Structure Classification

Reactivity assessment in SPARC begins with locating potential sites within the molecule for a particular reaction of interest. These reaction sites, which are termed reaction centers (C), are in general the smallest subunit(s) to which the reactivity of interest can be ascribed. Any molecular structure appended to C is viewed as a "perturber," denoted P. All reactions to be addressed in SPARC (from light absorption to hydrolysis) are analyzed in terms of some critical equilibrium component:

$$PC_o \xrightarrow{f(\Delta E)} PC_f \qquad (1)$$

where C_o and C_f denote initial and "final" states of the reaction center C, P is the "perturber"—the structure that is presumed unchanged by the reaction, $f\Delta(E)$ denotes some reaction parameter of interest that is a function of the energy change (ΔE) of the reaction. For example, the ionization of phenol is described by:



(2)

where C_o, C_f and P are -OH, -O— and the phenyl group, respectively.

For light absorption, C_o and C_f are ground and excited states of the reaction center and $f(\Delta E)$ is the intensity of light absorption as a function of incident light energy (that is, the absorption spectrum). For reaction kinetics, C_o and C_f may denote the initial and transition states of the reaction center, and $f(\Delta E)$ is the rate constant expressed as a function of the energy required for achieving the transition state.

The energy change is expressed in terms of contributions of factored structural components. For thermal properties,

$$\Delta E = \Delta E_c + \delta_p (\Delta E_c) \qquad (3)$$

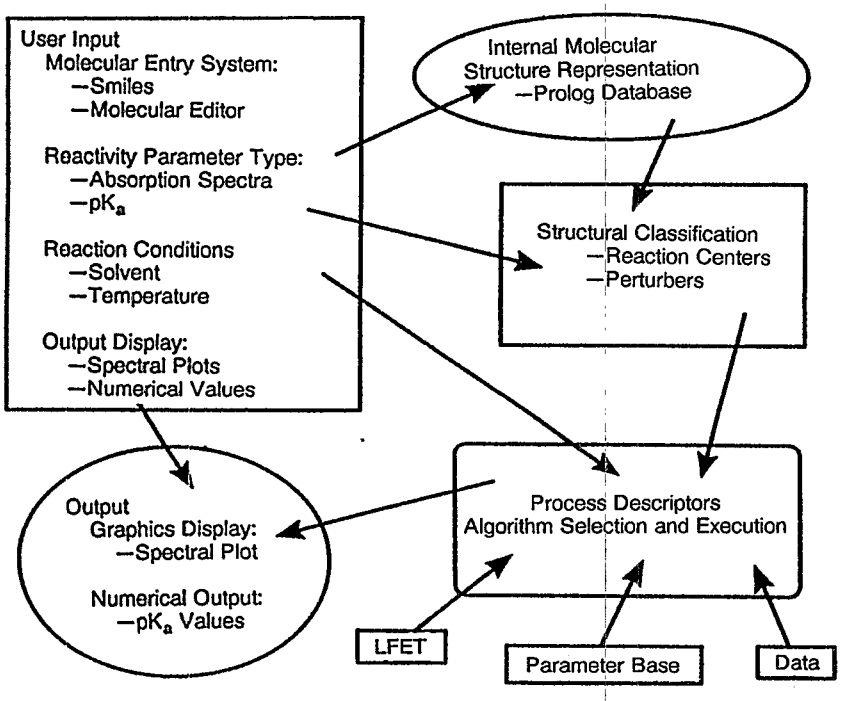


Figure 1. SPARC overview.

where ΔE_c describes the intrinsic behavior of the reaction center, and $\delta_p(\Delta E_c)$ denotes some perturbation derived from the appended structure, P. Changes in localized bonding energies are incorporated in ΔE_c and are assumed unchanged by P. For each reaction type, SPARC catalogues reaction centers and appropriate characterization data, $f(\Delta E_c)$. These reference data are not calculated *a priori*, but are inferred directly from measured data. Figure 2 contains sample reaction centers for acid or base ionization constants (pKa's) and UV-Visible light absorption. SPARC computes reactivity perturbations, $\delta_p(Reac)$ that are then used to "correct" the intrinsic behavior of the reaction center for the compound in question.

To facilitate quantitative modeling, $\delta_p(Reac)$ is expressed in terms of potential "mechanisms" for interaction of P and C.

$$\delta_p\,(Reac) = \delta_e\,(Reac) + \delta_r(Reac) + ... \qquad (4)$$

where subscripts e and r denote electrostatic and resonance interactions, respectively. Electrostatic interactions derive from local electric dipoles or charges in P (fixed or induced) interacting with charges or dipoles in C. Because $\delta_p(Reac)$ describes changes in the reaction $C_o \rightarrow C_f$ effected by P, $\delta_e(Reac)$ represents the difference in the electrostatic interactions of P with the two states, C_o versus C_f. Resonance interactions involve the delocalization of π electrons into or out of C, but again, $\delta_r(Reac)$ describes the change in electron delocalization accompanying the reaction. Additional perturbations include specialized interactions of structural elements of P contiguous to the reaction center such as H-bonding or steric blockage of access to C for another molecule (solvating or reacting).

The following examples are indicative of the extrapolation capability of SPARC. Figure 3 shows calculated and measured UV-Visible absorption spectra of sexiphenyl isomers, each of which was calculated as a "perturbed" benzene. The *para* isomer shows extended π conjugation throughout the six-membered ring, whereas the *meta* isomer shows only pair-wise ring coupling. In the *ortho* isomer, steric twisting of the essential single bonds removes, to a large extent, ring coupling. These spectra reflect the ability of SPARC to consider both electrometric and steric factors in resonance perturbations.

Figure 4 shows predicted and measured pK_a's for the reaction center, -OH, in a variety of molecular structures. These compounds demonstrate the extendability of SPARC procedures to inorganic compounds and to both aliphatic and aromatic organic compounds. Figure 5 shows similar information for the phosphono reaction center.

Model Verification and Testing

In chemistry, as with all physical sciences, one can never determine the "validity" of any predictive model with absolute certainty. This is a direct consequence of the empirical nature of the science. The closest one can get to "truth" in chemistry is to make use of the established laws

Light Absorption		pK_a	
Π-Π^*	- rigid structures - ethylene - benzene - napthalene	Acids	- OH - SH - CO_2H - PO_2H | - $B(OH)_2$ - SO_2H - SeO_3H - AsO_2H |
η - Π^*	- NR_2 - OH - SH - N (In ring) - S (In ring - C=O - C=S - C-NH	Bases	- NR_2 - N (In ring) - C=N - - C=O

Figure 2. Reaction centers for light absorption and acid/base equilibrium.

and theories, whose validity derives not from logic but from experience, having withstood exhaustive challenges or attempts to falsify. This established theoretical knowledge provides our best description of what the molecular world is really like. Because SPARC is expected to predict reaction parameters for processes for which little or no relative data exist for corroboration, "validity" must derive from the efficacy of the model constructs in "capturing" or reflecting the existing knowledge base of chemical reactivity.

In every aspect of SPARC development, from choosing the programming environment to building model algorithms or rule bases, testability was an important criterion. As discussed earlier, the capability to extrapolate and/or to avoid situation-specific or *ad hoc* descriptions was a primary goal for SPARC. The basic mechanistic models were designed and parameterized so as to be portable, in principle, to any type of chemistry or chemical structure. This extrapolatability enhances testability in several important ways. First, as the diversity of structures and the chemistry that is addressable increases, so does the opportunity for failure. More importantly, however, in testing against the theoretical knowledge of reactivity, specific situations can be chosen that offer specific challenges. This is particularly important when testing performance in areas where existing data are limited or where additional data collection may be required. Finally, this expanded prediction capability allows one to choose, for exhaustive testing, the reaction parameters for which large and reliable data sets do exist to test against. Test data sets are presently being encoded, including UV-Visible absorption spectra (about 5000 compounds) and ionization pK_a's (about 18,000 compounds). These unbiased and unscreened data sets will provide an exhaustive test of performance covering a broad domain of chemistry.

Is SPARC verifiable? SPARC was designed to optimize falsifiability. A more pertinent question from a pragmatic viewpoint is what happens when SPARC fails? In general, failures to predict derive not by happenstance, but from errors or inadequacies in conceptualization or theory implementation. Again, one must resort to the theoretical knowledge base of chemical reactivity to determine the source(s) of failure. The "mechanistic" output of SPARC should aid in the process. In addition, the modular design

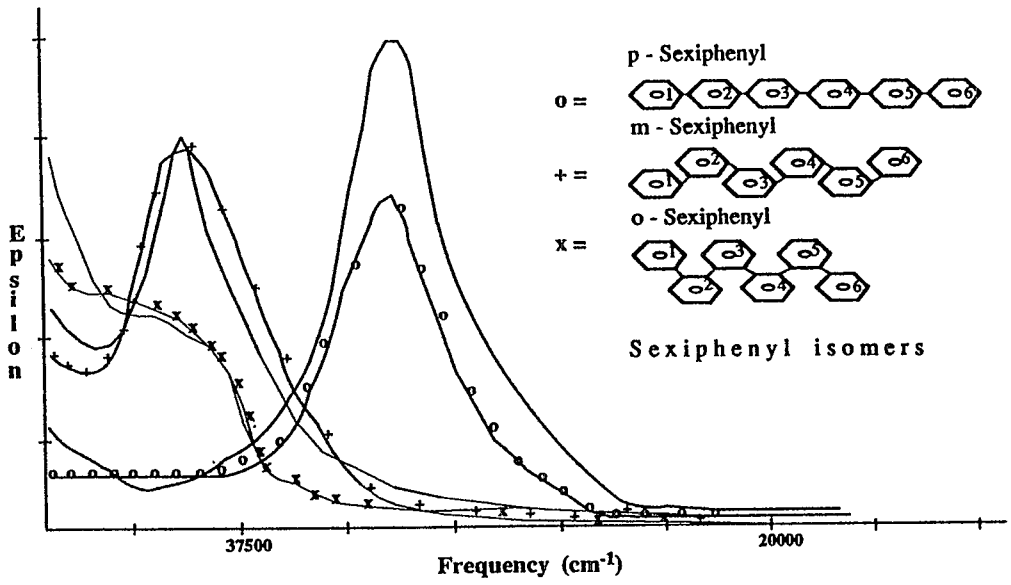


Figure 3. Predicted (solid lines) and measured spectra of sexiphenyl isomers.

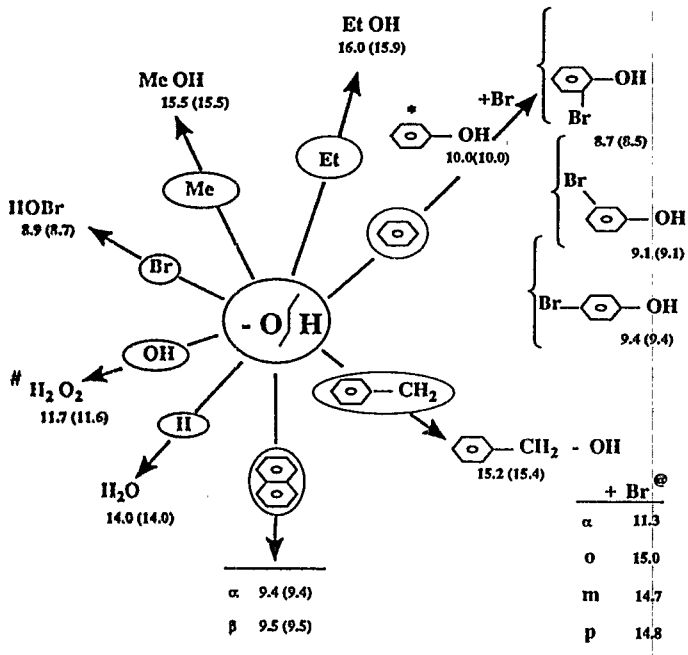


Figure 4. Predicted and measured (values in parentheses) pK_a's for compounds containing the reaction center, -OH. For example, the phenyl group and -OH react to produce phenol (indicated by *) with a predicted value of 10.0, which compares with the measured value of 10.0. Further reaction with bromine produces o-, m- and p- bromophenol to give predicted and measured values as shown. A typical inorganic addendum is OH with the reaction center to produce H_2O_2 (indicated by #) for a predicted pKa of 11.7 as compared to the measured pK_a of 11.6. Note: there are no measured pK_a values for the various bromine-substituted benzylalcohols (indicated by @).

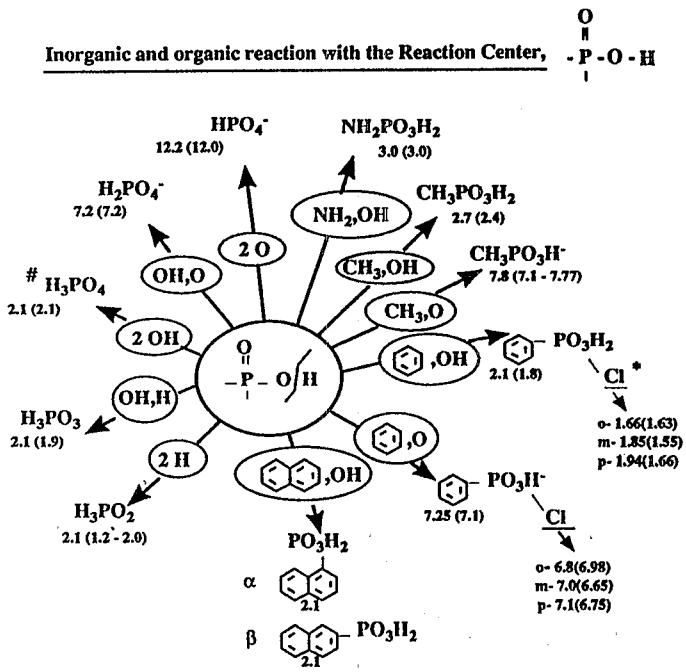


Figure 5. Predicted and measured (values in parentheses) pK_a's for compounds containing the phosphono reaction center. For example, the phenyl group, OH and chlorine react with phosphono to produce the 3 isomers (o-, m-, p-) of chlorophenylphosphonic acid (indicated by *) to give comparable values. Reaction of phosphono center with two OH groups gives the inorganic acid, phosphoric (H_2PO_4) (indicated by #) with a comparable value.

and programming environment of SPARC facilitates modifying or adding to model algorithms or the rule bases. This provides, we hope, for coherent growth or advancement in predictive capability. In this capacity, the SPARC approach also can serve as a research tool for resolving conflicting viewpoints or perhaps ultimately advancing the field of reactivity description.

Projection

The methods described above predict UV-Visible light absorption spectra, ionization pK_a's, and hydrolysis reaction rates. The prototype computer program currently runs on a widely used minicomputer and uses commercially available Prolog and Fortran compilers. Plans are being developed for a PC version. Future development in the modeling will include:

For Light Absorption Spectra—

- expand the reaction center database to include all commonly encountered "rigid" π structures,
- develop electrostatic effects models for spectra,
- begin exhaustive testing.

For Ionization pK_a—

- continue parameter optimization for electrostatic effects models,
- develop H-bonding effects and steric models,
- develop parameters for solvation effects,
- begin exhaustive testing.

For Hydrolysis—

- assemble reaction center database for hydrolysis reaction rates,
- begin exhaustive testing.

Application

The necessary software for trial use of SPARC is expected to be available for EPA use on minicomputers by late 1990. Fully documented software and a user's manual are planned for 1991. During this final development period, preliminary applications of SPARC to several field problems will be demonstrated; for example, estimation of hydrolysis rate constants necessary for exposure modeling by an EPA Regional Office analyst responsible for a hazardous waste site assessment.

This expert system should be a boon to any Agency Office or Region, state or other environmental group needing to predict the concentration of a pollutant in a particular environment for exposure assessment. It should be valuable to Agency regulators in their modeling efforts relative to the assessment of land disposal of wastes, compliance monitoring, development of remedial action plans, implementation of premanufacturing notice regulations, or in any area where the chemical reactivity of pollutants is important in exposure prediction. SPARC can estimate reactivity parameters at less cost, with greater accuracy, and with a broader scope than any conventional technology.

References

1. Lyman, W. J., W. E. Reehl, and D. H. Rosenblatt. 1982. *Handbook of Chemical Property Estimation Methods: Environmental Behavior of Organic Chemicals.* McGraw-Hill, New York, NY.
2. Yalkowski, S. H. and S. C. Valfani. 1980. Solubility and Partitioning I: Solubility of Non-electrolytes in Water. *Pharmaceutical Sci.,* 69:912-913.
3. Leo, A. J. 1975. Calculation of Partition Coefficients Useful in Evaluation of the Relative Hazards of Various Chemicals in the Environment. *I. J. C. Symposium on Structure Activity Correlations in Studies of Toxicity and Bio-concentrations with Aquatic Organisms.* (Veith, G. D., ed.), International Joint Commission, Windsor, Ontario.
4. Zepp, R. G. 1982. Experimental Approaches to Environmental Photochemistry. In: *Handbook of Environmental Chemistry* (Hutzinger, O., ed.), Vol. 2(B) Springer-Verlag, New York, NY, pp. 19-42.
5. MacKay, D., A. Bobra, W. Y. Shiu, and S. H. Yalkowski. 1980. Partition Coefficients. *Chemosphere,* 9:701-711.
6. Zepp, R. G. and D. M. Cline. 1977. Rates of Direct Photolysis in the Aquatic Environment. *Environmental Science and Technology,* 11:359-366.
7. Wolfe, N. L., R. G. Zepp, J. A. Gordon, G. L. Baughman, and D. M. Cline. 1977. Kinetics of Chemical Degradation of Malathion in Water. *Environmental Science and Technology,* 11:88-100.
8. Smith, J. L., W. R. Mabey, N. Bohanes, B. B. Hold, S. S. Lee, T. W. Chou, D. C. Bomberger, and T. Mill. 1978. *Environmental Pathways of Selected Chemicals in Freshwater Systems: Part II,* U. S. Environmental Protection Agency, Athens, GA, EPA/600/7-78/074.
9. Gray, Neil A. B. 1986. *Computer Assisted Elucidation.* Wiley and Sons, New York, NY.
10. Pierce, T. H. and B. A. Hohne (eds.). 1986. *Artificial Intelligence Applications in Chemistry,* American Chemical Society, Washington, DC, Symposium Series, No. 306.
11. Lowry, T. H. and K. S. Richardson. 1987. *Mechanisms and Theory in Organic Chemistry,* 3rd ed., Harper & Row, New York, NY.
12. Leffler, J. E. and E. Grunwald. 1969. *Rates of Equilibria of Organic Reactions.* Wiley and Sons, New York, NY.
13. Hammett, L. P. 1970. 2nd ed. *Physical Organic Chemistry,* McGraw Hill, New York, NY.
14. Dewar, M. J. S. 1969. *The Molecular Orbital Theory of Organic Chemistry,* McGraw Hill, New York, NY.
15. Dewar, M. J. S. and R. C. Dougherty. 1975. *The PMO Theory of Organic Chemistry.* Plenum Press, New York, NY.
16. Taft, R. W. (ed.). 1987. *Progress in Organic Chemistry,* Vol. 16. Wiley and Sons, New York, NY.
17. Karickhoff, S. W., Lionel A. Carreira, Clyde Melton, Valeta K. McDaniel, Andre N. Vellino, and Donald E. Nute. Predicting Chemical Reactivity By Computer, Part I: Approach. Submitted for publication. 1989.
18. Karickhoff, S. W., Lionel A. Carreira, Clyde Melton, Valeta K. McDaniel, Andre N. Vellino and Donald E. Nute. Predicting Chemical Reactivity by Computer, Part II: UV-Visible Light Absorption. Submitted for publication. 1989.
19. Karickhoff, S. W., Lionel A. Carreira, Clyde Melton, Valeta K. McDaniel, Andre N. Vellino, and Donald E. Nute. Predicting Chemical Reactivity By Computer, Part III: Ionization pK_a. Submitted for publication. 1989.

United States
Environmental Protection
Agency

Center for Environmental Research
Information
Cincinnati OH 45268

Official Business
Penalty for Private Use $300

EPA/600/M-89/017