Information Content in Organic Molecules: Brownian Processing at Low Levels

Daniel J. Graham*

Department of Chemistry, Loyola University Chicago, 6525 North Sheridan Road, Chicago, Illinois 60626

Received November 2, 2006

The informatic properties of organic molecules have been the subject of our research during the past several years. In the present study, we investigate the lower levels wherein information is expressed via Brownian processing. Organic molecules are like other electronic devices in that their informatic details depend on the operating level in question. The low and high levels are distinguished (among other ways) by the amount of work they require for processing. In this work, a Brownian model is developed by which the low-level content of a chemical system can be quantified. The model is demonstrated for diverse organic molecules. In so doing, several scaling properties of low-level information are illustrated. In addition, the correspondence traits regarding the different levels are examined. Molecular information represents a capacity for work control such as during chemical reactions. Thus, the information expressed at low levels is examined in connection with the reaction pathway selectivity of organic compounds.

I. INTRODUCTION

Probability distributions underlie the different types of information, its communication and reception.[1] Information is the hallmark of all input/output codes and instruction sets, and complex mixtures of both. As with thermodynamic quantities, the information expressed by a system is stable under isolated conditions.[2] By contrast, a system's information can be altered through the agency of other probability distributions—other sources of information.[3] Interestingly, the informatics of a system turn on *both* the source and processor attributes. The net cast in this area of science has certainly been wide during the past several decades.[4] For organic molecules, information theory has been applied extensively to structure/function problems surrounding natural products and pharmaceuticals.[5−8] An equivalent statement holds for materials chemistry, for example, toward the development of electronic devices.[9−12] Information theory has been integrated as well with chemical reaction dynamics[13] and molecular electronic structure theory.[14−16] Is there an area of chemistry where information theory contributes zero insight?

Information is an intricate quantity because it depends on the details of representation. The previous sentence, for example, can be expressed using 89 Roman alphabetic characters, counting spaces and punctuation. It can also be communicated using an ASC hexadecimal code:[17]

```
49H 6EH 66H 6FH 72H 6DH 61H 74H 69H 6FH 6EH
69H 73H
61H 6EH
69H 6EH 74H 72H 69H 63H 61H 74H 65H
71H 75H 61H 6EH 74H 69H 74H 79H
.
.
.
```

The above string can alternatively be expressed using octal and binary codes, namely,

```
111 156 146 157 162 155 141 164 151 157 156
151 163
141 156
151 156 164 162 151 143 141 164 145
161 165 141 156 164 151 164 171
.
.
.
```

and

```
001001001 001101110 001010110 001101111 001110010 001101101 001100001
001110100 001101001 001101111 001101110
001101001 001110011
001100001 001101110
001101001 001101110 001110100 001110010 001101001 001010011 001010001
001110100 001100101
001110001 001110101 001100001 001101110 001110100 001101001 001110100
001111001
.
.
.
```

Clearly, information admits different representations which are related to one another by nontrivial rules. Importantly, the representations evince a natural ordering dictated by work requirements. For instance, it costs more work to pen each sentence of text material using a binary code compared with an alphabetic one of equivalent-sized characters. The binary representation can be thus be categorized as a lower-level one compared with its alphabetic counterpart. In the expression and transfer of information, different levels/representations are appropriate to different applications. By and large, however, the lower ones reflect more closely the devices that effect the actual storage and processing. This paper has been written with the aid of a Pentium processor coupled to a flash drive. The higher-level (alphabetic) format is the front-runner choice for human activities such as reading. Yet the binary account is equally critical because it lies at the crux of all the storage and processing issues. These are important issues because they hinge on (among other things)

* Corresponding author phone: 1-773-508-3169; fax: 1-773-508-3086; e-mail: dgraha1@luc.edu.

10.1021/ci600488x CCC: $37.00
Published on Web 03/09/2007

timing electronics, energy availability, and processing error rates. In effect, the lower-level representations better connect with the internal workings of the devices that carry and transmit information. Such a connection is obscured or lost outright in the high-level formats.

An organic molecule expresses information beginning with the atoms in a covalent bond network. Its electronics are processed—registered, appended, and redistributed—during collisions involving other molecules, typically in a solution environment. As is well-known, every organic compound admits different representations of information as conveyed by elementary dot structures, two-dimensional (2D) graphs, and electron orbital combinations. These are distinguished (among other ways) by the measure of work needed to obtain and process. Accordingly, the information associated with a molecule demonstrates the level properties of electronic systems in general. A mix of alphanumeric characters and graphs offers a high-level code; this demands less processing work compared with a molecular orbital (MO) (and consequently lower-level) description. As with papers about chemistry, the different representations appeal to different applications. Combinatorial chemistry, for example, is best characterized using the high-level format of structure graphs and resin icons.[18] By contrast, high-resolution spectroscopy data go hand-in-hand with low-level representations.[19] By and large, the high levels of molecular information are geared toward synthetic design and database applications.[18,20,21] Low-level accounts center more on an individual compound's structure and function.

We have aimed for the past few years for a better understanding of organic molecules as carriers and processors of information in solution environments. We have reported the results in a series of papers in this journal.[22−25] Our most recent endeavor has focused on the different levels of information and their correspondence properties. Our motivation can be described using computational analogies. For example, to comprehend how a Pentium computer and flash drive store and retrieve binary data advances their operations beyond those of a black box. Likewise, to grasp how a source code such as Pascal is converted into an object file offers another dimension for problem assessment and solving. In essence, an understanding of *both* hardware and software requires the precise relations between the levels of information.

It would be simplistic to view molecules merely as small-scale computers. However, information theory contributes insight to disparate fields, computers and chemistry being only two examples. Interestingly, several technical facets of solid-state electronics are shared by organic compounds. All parties demonstrate their own architectural flavor, capacity for information, and fungibility characteristics. Molecular information of the high-level variety was the topic of our previous works.[22−25] The object of the present writing is (1) to quantify and examine the lower-level content for select organic compounds and (2) to illustrate the correspondence properties regarding different information levels.

This paper is organized as follows. In section II, we review the Brownian processing characteristics of organic compounds and the quantification of their high-level content. This sets the stage for measuring the information expressed at lower levels. In section III, a procedure is developed for quantifying the low-level content for molecular systems in general. In section IV, representative data are illustrated for diverse organic compounds under isolated conditions. The correspondence properties regarding high- and low-level information are addressed in section V. Section VI closes the work with discussion of the correspondence properties of molecules and their work-control characteristics.

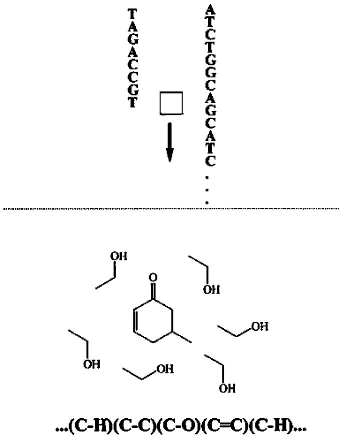


Figure 1. Fundamentals of Brownian processing. The upper half schematically illustrates Brownian processing by a polymerase enzyme (box). The information expressed by an input tape of organic base moieties (DNA, right-hand side) is registered during a forced random walk. The walk and registration processes catalyze the synthesis of an output tape (left-hand-side polymer) with equivalent information in a complementary format. The lower half depicts 5-methyl-3-cyclohexene-1-one dissolved in ethanol. The solvent and solute molecules function collectively as a Brownian processor in which information is registered via collisions. The tape recordings are encoded by atom−bond−atom electronic states and can be formulated using random walk methods.

II. MOLECULAR INFORMATION PROCESSING AND DIFFERENT LEVELS

The first chemical system viewed in information processing terms was the polymerase enzyme and deoxyribonucleic acid (DNA). In a seminal paper, Bennett detailed how the protein exercises a random walk which is biased along one direction of a polymer chain.[26] Bennett addressed specifically the thermodynamic aspects of the random walk in relation to its computational action. This system is represented schematically in the upper portion of Figure 1. As is well-known, the formation of a second polymeric chain is catalyzed during the walk; thus, the information expressed in the right-hand-side base sequence of Figure 1 is registered and duplicated as a complementary carbon copy (left-hand side). It is easy to see how the chemistry is *Brownian Computation* (Bennett's term) at its finest. The enzyme is able to discriminate one organic base moiety of a polymer from another and directs high-fidelity, nontrivial synthesis on that basis. The data processor, its input and output tapes, interface with one other in an aqueous solution. The conditions enable bond making and breaking as a reliable source of work. The liquid environment functions equally well as a sink for energy dissipation.

Yet even under mundane chemical circumstances, the ideas of Brownian computation—processing is a more general term—are no less relevant. One situation is illustrated in the lower half of Figure 1. In a room-temperature solution of 5-methyl-2-cyclohexene-1-one ($C_7H_{10}O$) and ethanol (C_2H_5O), each molecule experiences 10^{12}−10^{13} collisions per second via

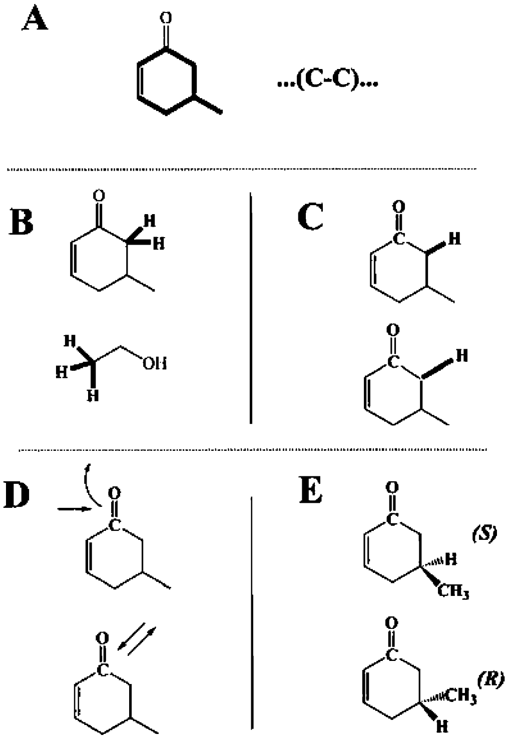


Figure 2. Several shortfalls of high-level coding. (A) Not all **(C−C)** units are alike within a given organic compound. (B) Not all **(C−H)** units are alike for different molecules. (C) The bond length associated with an individual code unit fluctuates over time. (D) The appearance of a code unit to a colliding party depends on the trajectory. (E) Information registered for a chiral center depends on the handedness of the processor.

translational motion and hindered rotations.[27] The collisions are not entirely random on account of the local structure imposed by the atoms and their geometric arrangement. In effect, the collisions are correlated over short—on the order of 10^{-12} s—but definitively nonzero time periods; the correlations extend over molecular dimensions of several angstroms.[28] Our previous work investigated collision strings which were formulated via Brownian (random-walk) methods.[24,25] These strings were encoded using the same atom−bond−atom units which compose the information sources. In a room-temperature solution, the colliding parties operate *collectively* as a Brownian processor in both serial and parallel modes, in a way analogous to polymerase and DNA. The difference is that atom−bond−atom units [**(C−H)**, **(C=O)**, etc.], instead of base moieties (**A**, **T**, **C**, **G**), encode the tape-recording data. Yet whether a source is registered as alphabetic text (e.g., this paper), organic base sequences **...ATATGCACG...** as in polymerase/DNA, or **...(C−H)(C−C)(C=C)(C−H)(C−C)...**, as in a transient collision scenario, the information can be quantified at multiple orders.

The components of an everyday computer are powered by batteries or an external supply. By contrast, it is the thermal fluctuations of several quantities which drive the information processing for organic molecules. With every collision in solution, for example, involving an adenine (**A**) unit in DNA, or **(C=C)** of $C_7H_{10}O$, there result local alterations of the electric charge density, energy, and free volume. The changes are erratic and momentary in occurrence; they can be positive or negative depending on the details of molecular structure and interaction. Yet if new correlations are a consequence, then information is registered locally, with a concomitant decrease in the system entropy. With so many transactions (collisions, charge density, and free-volume alterations) per second, the entropy deficits are overwhelmingly fleeting. They are erased via ensuing collisions with the consequential loss of information. In effect, Brownian processing is associated with the birth and death of innumerable correlations, and the corresponding local gains and losses of information. High-temperature conditions enhance fluctuations and thus the number and variety of correlations. Low temperatures discourage fluctuations and indeed cause some correlations to be frozen out.

It can be shown that, for an equilibrium system, the mean square entropy fluctuations $<(\Delta S)^2>$ are tied to the heat capacity as follows:

$$<(\Delta S)^2> = k_B C_p \qquad (1)$$

where k_B is Boltzmann's constant and C_p is the heat capacity at constant pressure.[29] For liquid ethanol at room temperature, the specific heat is 2.43 j/g K.[30] This equates with a per-molecule heat capacity of $\sim 17k_B$. Thus, by eq 1, a typical entropy fluctuation of a Figure 1 solvent molecule is $(17k_B^2)^{1/2} \approx 4k_B$. Entropy can be viewed (among other ways) as missing correlations.[31] A local, reversible change of $\sim 4k_B$ thereby equates with a purchase or loss of $\sim 4k_B/k_B \log_e(2) \approx 6$ bits of information. To be sure, the fractional entropy fluctuations are vanishingly small at the macroscopic level. For a 100 g sample of ethanol, one estimates $<(\Delta S)^2>^{1/2}/<S>$ to be $\sim 10^{-10}$, assuming $<S>$ to be on the same order as C_p. Yet the fluctuations wield significant impact at the molecular scale. For inside a closed flask of $C_7H_{10}O$/ethanol at room temperature, the informatic transactions transpire indefinitely. In effect, bits and pieces of locally manufactured tape recordings **...(C−C)(C=C)(C−H)(C−C)...** are registered as quickly as other fragments are deleted.

Now the informatic perspective of Figure 1 is distinctly high-level. It costs comparatively little work to represent base moieties by single characters **A**, **T**, **C**, and **G**. An analogous statement applies to the structure graphs for $C_7H_{10}O$ and ethanol. It is then easy to see where substantial limitations are imposed by the high-level formats. The six **(C−C)** units, of $C_7H_{10}O$ (Figure 2A), for example, are not identical owing to molecular asymmetry and resonance structures. Along the same lines, a **(C−H)** in $C_7H_{10}O$ is not equivalent to a neighboring one in ethanol (Figure 2B). Likewise, a given **(C−H)** in $C_7H_{10}O$ is not the same at all times owing to bond length fluctuations (Figure 2C). Further, the cross-section presented by a unit such as **(C=O)** depends on the trajectory of the colliding party (Figure 2D). Last, $C_7H_{10}O$ contains a chiral carbon in the 5 position. A handed processor would thus register the *R* and *S* enantiomers differently (Figure 2E).

In high-level studies, many informatic subtleties can be fleshed out by the analysis of pair [**(C−H)(C=O)**, **(C−O)(C−C)**, etc.], triplet [**(C−C)(C−H)(C=C)**, **(C=C)(C−C)(C−H)**, etc.], and higher-order states.[24,25] The efficacy of this procedure is not guaranteed, however. For example, the chair and the boat forms of cyclohexane (C_6H_{12}) express the same atom−bond−atom code units; high-level analyses at multiple orders offer no distinctions between the two conformers.

Even for molecules not subject to significant conformational changes, for example, *cis*- and *trans*-dichloroethylene

($C_2H_2Cl_2$), high-level analyses offer no capital for discriminating the two isomers during Brownian processing. For these and other systems, equal quantities of high-level information are expressed at all orders; the disparities can only be identified via the lower-level representations. This is not an uncommon situation in the information sciences. In computer programming, for example, the statements "if x < y then procedure_1" and "if m < n then procedure_2" offer identical message lengths at the source level. Yet the binary code affiliated with each line in an execution file depends on the context, variable types, and function calls. By analogy, the information expressed by a system—chemical, computational, or otherwise—cannot always be gauged at the higher levels with minimal work expenditure. In our past studies, Brownian techniques were used to quantify information at the structure graph level. The next task concentrates on the lower levels.

III. THE QUANTIFICATION OF MOLECULAR INFORMATION AT LOWER LEVELS

Quantifying molecular information at the high levels utilizes Brownian methods coupled with mathematics originated by Shannon.[32] One identifies all of the $j = 1...N$ code units (atom−bond−atom states) for a system and their occurrence frequencies $p(1)...p(j)...p(N)$. The value of H_I is then obtained via the formula

$$H_I = -\sum_j p(j) \log_2 p(j) = -(1/\log_e 2) \sum_j p(j) \log_e p(j) \quad (2)$$

where H_I is measured in bits, and

$$\sum_j p(j) = 1 \quad (3)$$

is a necessary condition. There are several ways of interpreting the individual terms and summation result of eq 2. First is that each $-\log_2 p(j)$ term quantifies the *self-information* or *surprisal,* allied with the *j*th state.[33] An infrequently occurring state—rare sequence of code units—is thereby distinguished from a pedestrian one by its high suprisal value. Second is that H_I establishes the optimum number of bits required to encode random draws of atom−bond−atom sequences given the molecular structure attributes. Third, H_I measures the length of the typical message expressed by the molecule during Brownian processing at the covalent bond level, the atom−bond−atom units providing an effective code. For all compounds in solution, the variable draws transpire via collisions and thermal fluctuations. A given state is registered by its electric charge distribution in a way which is distinctly robust. Not only is **(C−C)** distinguishable from **(O−H)** where the atomic charge distributions are substantially different but also states wherein the disparities are subtle, for example, **(C−O)** and **(C=O)**, **(C−C)** and **(C=C)**.

Expanding a molecule's informatics beyond the source high levels, however, is nontrivial because the underlying probability distributions are not discrete. Thus, for a Brownian processor to differentiate the **(C−C)** units in $C_7H_{10}O$, *R* and *S* enantiomers, and so forth, requires a close reading of charge functions which are inherently continuous. Moreover, there is more than one distribution in play for a given source unit on account of bond length and multiple trajectories (cf. Figure 2). Where to begin?

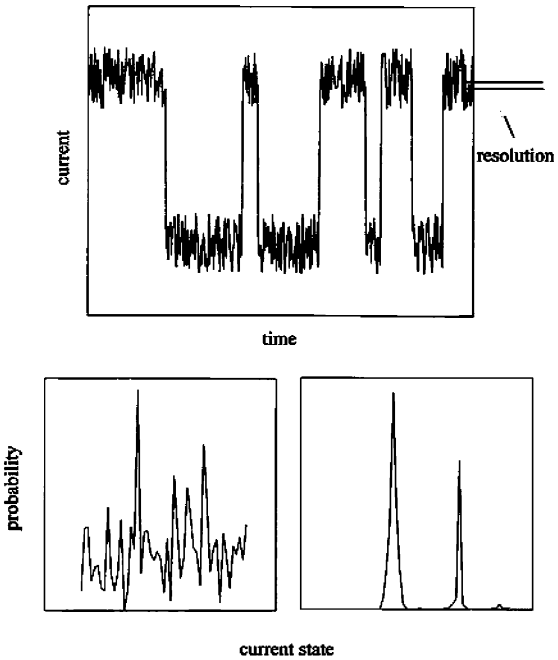


Figure 3. Information and continuous sources. The upper panel schematically illustrates an electric current signal as a function of a coordinate variable; the state is determined by the amplitude. The resolution of a noise-free measuring device (e.g., multimeter or oscilloscope) is indicated. The lower panels illustrate two distribution functions derived from the signal. The function appearing in the left-hand panel results from a higher-resolution measurement and predicates an information content of 5.37 bits. The right-hand-side panel results from a lower-resolution measurement and expresses 3.71 bits.

The procedure we have developed is grounded upon electronic devices in general, where a continuous source predicates a distribution approximated by digital means. This is the theme of Figure 3. The upper panel shows an electric current due to a circuit element such as a transistor with the fluctuations dependent on a suitable coordinate variable. Let the state of the current be defined by the amplitude. A noise-free multimeter or oscilloscope will register different states over some finite range and resolution. Repeated measurements enable the assembly of a state distribution function which depends on *both* the source and sensing devices operating in tandem. If the sensor operates at high resolution, the distribution will reflect the source as in the lower-left panel of Figure 3. If the sensor performs at low resolution, the distribution will resemble that illustrated in the lower-right panel. The point is that the information of the recorded distribution will not exceed that originated by the source. If a distribution corresponding to 5.37 bits is registered as in the lower-left panel, then the source could only have expressed *at least* 5.37 bits. Further, the distribution fluctuations will not exceed those of the source. If multiple versions of the lower-right plot show a spread of 0.45 bits—the average being 3.71 bits—this could only mean that the source variations were 0.45 bits or more.

As electronic devices go, organic compounds are among the leaders in small-sweepstakes contests. Their charge distributions are established experimentally via X-ray diffraction, magnetic resonance, and other spectroscopic methods[34] and computationally using a variety of algorithmic strategies.[35] Central to our approach is that *lower-limit* values of information can be realized on the basis of such distributions, both experimental and computational. These distributions underpin the high-level states such as those portrayed in Figures 1 and 2. Not surprisingly, the results bring new chemical properties—and questions—to the fore. The root concepts, however, do not stray from those applied routinely at all levels, the high ones included.

. If an adenine (**A**) moiety is registered by polymerase, for example, then *at least* $\log_2 2^2 = 2$ bits of information are involved. The value can only be construed as a lower limit given the other chemical details, for example, sugar–phosphate bonds and solvent molecules.

Let $\rho(x,y,z)$ describe the charge distribution for a chemical system of interest—atom, molecule, or complex. In turn, $\rho(x,y,z)\ \Delta x\Delta y\Delta z$ quantifies the fraction of charge in the volume element $\Delta x\Delta y\Delta z$ with

$$\sum\nolimits_{xyz} \rho(x,y,z)\ \Delta x\Delta y\Delta z = 1 \tag{4}$$

By analogy with eqs 2 and 3, an informatic functional *could be* introduced as follows:

$$\begin{aligned} I(\rho) &= -\sum\nolimits_{xyz} \rho(x,y,z)\ \Delta x\Delta y\Delta z \log_2 \rho(x,y,z) \\ &= -[1/\log_e(2)] \sum\nolimits_{xyz} \rho(x,y,z)\ \Delta x\Delta y\Delta z \log_e \rho(x,y,z) \end{aligned} \tag{5}$$

The result, $I(\rho)$, however, would overlook important dimensional issues. Since ρ has the units of inverse volume, its singular placement in a logarithm argument is dubious. One addresses this concern by targeting not I per se but, rather, its family relation K, the *relative* information as defined by Kullback.[36] By this tack, a reference distribution φ also having reciprocal volume dimensions is called upon for assistance. The functional $K(\rho{:}\varphi)$ is then obtained via

$$\begin{aligned} K(\rho{:}\varphi) &= +\sum\nolimits_{xyz} \rho(x,y,z)\ \Delta x\Delta y\Delta z \log_2(\rho/\varphi) \\ &= +[1/\ln(2)] \sum\nolimits_{xyz} \rho(x,y,z)\ \Delta x\Delta y\Delta z \log_e(\rho/\varphi) \end{aligned} \tag{6}$$

$K(\rho{:}\varphi)$ can be regarded as the number of *extra* bits required for coding the messages allied with ρ, all the while taking the distribution to be described by φ. Clearly, $K(\rho{:}\varphi)$ equals zero if ρ and φ turn out to be identical. Much more commonly, the summation terms in eq 6 are positive or negative; the summation total $K(\rho{:}\varphi)$ exceeds zero in accordance with the Gibbs inequality.[37] One notes practical restrictions regarding reference functions, namely, continuity and convergence; as with levels of information, different φ's suit different applications. For an in-depth discussion of information and continuous distributions, the reader is directed to chapter 8 of ref 33. In addition, Dinur and Levine have explored in detail the entropy associated with continuous distribution functions with applications to chemical kinetics.[38]

Finally, the information mathematics of the continuous wavefunctions of molecules has been presented in depth by Parr, Nalewajski, and their co-workers.[14−16]

Note here that $K(\rho{:}\varphi)$ is closely related to another functional, $D(\rho{:}\varphi)$. This has been termed by Kullback as the *divergence*:[36]

$$\begin{aligned} D(\rho{:}\varphi) &= \sum\nolimits_{xyz} (\rho - \varphi)\ \Delta x\Delta y\Delta z \log_2(\rho/\varphi) \\ &= [1/\ln(2)] \sum\nolimits_{xyz} (\rho - \varphi)\ \Delta x\Delta y\Delta z \log_e(\rho/\varphi) \end{aligned} \tag{7}$$

$D(\rho{:}\varphi)$ weighs the disparities of two distribution functions in a way which is alternative to $K(\rho{:}\varphi)$. It can be shown that

$$D(\rho{:}\varphi) = K(\rho{:}\varphi) + K(\varphi{:}\rho) \tag{8}$$

Thus, in targeting the divergence, each of the two distributions, φ and ρ, serves as a reference for the other. The sum total quantifies the ease (as measured in bits) of discriminating between the two probability functions.

In our studies, K and D have served as the low-level quantifiers, given chemically appropriate choices for φ, $\Delta x\Delta y\Delta z$ (volume partitions), and summation cutoffs. Regarding φ, we have considered the informatic role played by molecules as 2-fold. First is that they effect charge distributions in patterned, spatially biased ways. Accordingly, when assessing the low-level content for a system, we have taken φ to equate with a uniform distribution of an equivalent number of charges. For example, in applying this rule to the electron distribution of 5-methyl-2-cyclohexene-1-one, a calculation of $K(\rho{:}\varphi)$ gauges the information *relative* to that of a diffuse charge cloud of 60 electrons. In effect, the uniform distribution provides a reference information value of zero because it lacks a chemically useful role.

The second role is no less critical. Organic molecules implement charge distributions which are subject to comparison. Thus, the distribution native to 5-methyl-2-cyclohexene-1-one has currency not only because it is spatially biased but also because it can be compared to the distributions of other compounds, even replicas. Accordingly, when comparing the informatics for two molecules or fragments thereof, we have taken φ to equate with ρ for one of the systems. Such compare-and-contrast calculations result in $D(\rho{:}\varphi)$.

In a liquid solution, molecular information is processed at a high but necessarily finite resolution. Thus, **(C=C)** is distinguished from **(C≡C)**, **(C=O)** from **(C−O)**, and so forth. Yet precluded is the differentiation of isotopic species, for example, **(^{13}C=^{12}C)** versus **(^{12}C=^{12}C)** and **(C−H)** versus **(C−D)**. Accordingly, the chemical responses are equivalent for all versions of 5-methyl-2-cyclohexene-1-one in solution: $^{12}C_6{}^{13}CH_{10}O$, $^{12}C_7CH_9DO$, and so forth. This limitation has been kept in mind throughout the investigation. Thus, in both $K(\rho{:}\varphi)$ and $D(\rho{:}\varphi)$ calculations, the partition $\Delta x\Delta y\Delta z$ has been equated with a cube of typical dimensions (0.1 Å^3). At such a resolution, the informatic effects of different isotopes have no bearing. The same statement applies to bond length fluctuations at this resolution.

The logistics of $K(\rho{:}\varphi)$ and $D(\rho{:}\varphi)$ calculations are illustrated via Figure 4. When targeting K, φ is first constructed as a homogeneous distribution inside a virtual container as depicted in panel A. The electron distribution for an atomic or molecular system of interest is then positioned at the container center (panel B) followed by the application of eq 6. The container size dictates the summation cutoff with the choice dependent on the compound(s) being

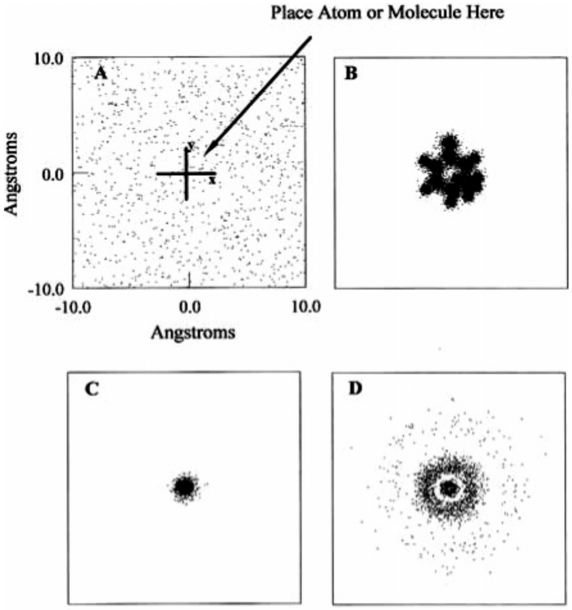


Figure 4. Logistics of quantifying low-level molecular information. Panel A illustrates a cubic container with a uniform charge density. Panel B illustrates the charge density function of 5-methyl-3-cyclohexene-1-one, having been situated at the container center. The $K(\rho{:}\varphi)$ calculations target the information relative to that of a uniform distribution effected by an equivalent number of electrons. Panels C and D illustrate two of the charge density functions, 1s and 3s, for the hydrogen atom. In the divergence calculations, the distributions of interest are probed under maximum-alignment conditions so as to yield minimum $D(\rho{:}\varphi)$.

investigated. A cube of volume 20 Å^3 accommodates $C_7H_{10}O$ and like-size molecules. It is ill-suited for larger molecules such as chlorophyll-a.

The divergence calculations proceed in the same way with one important qualification. For D to measure the true ease of distinguishing two distributions, φ and ρ must be examined under maximum-alignment conditions. If this condition is not met, otherwise-identical distributions such as for two ethanol molecules would at once be distinguishable—applying eq 7 would result in a divergence near infinity! This point is illustrated via panels C and D in Figure 4. Represented are the familiar electron distributions for the 1s and 3s states of the hydrogen atom.[38] Both distributions have been centered in the container such that the maximum registration of the charge clouds is demonstrated. As is known in beginning chemistry classes, the 1s and 3s distributions are disparate in their electron density and nodal structure. By applying eq 7, one quantifies the lower limit informatic difference via minimum $D(\rho{:}\varphi)$. Analogous procedures can be exercised quite generally so as to quantify the divergence for two atom—bond—atom states such as **(C—O)** and **(C=O)** and molecules such as 5-methyl-2-cyclo*hexene*-1-one and 5-methyl-3-cyclo*hexane*-1-one.

IV. LOW-LEVEL INFORMATION ASPECTS OF DIVERSE SYSTEMS

The charge distributions for organic compounds are predicated by their constituent atoms C, H, O, N, and so forth. Along the same lines, the structure and reactivity of molecules are determined by diatomic assemblies such as **(C—H)**, **(C—C)**, **(C=O)**, and so forth. With this in mind, ground work for this paper included extensive calculations for atomic and diatomic systems. Not surprisingly, several of the trends demonstrated by isolated atoms and diatomics are borne out in the much larger systems. For simplicity, we confine our attention to electron distributions; the effects of nuclear charges are addressed only indirectly by the electron clouds they sustain.

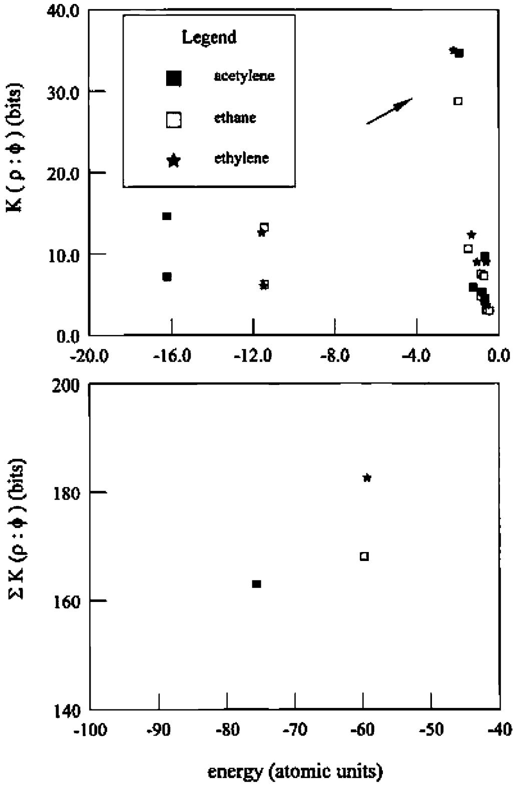


Figure 5. Low-level information data for acetylene, ethylene, and ethane. The upper panel illustrates $K(\rho{:}\varphi)$ versus molecular orbital energy. The arrow points to maximum $K(\rho{:}\varphi)$ for orbitals near the valence level and which are important to carbon—carbon bonds. The charge distributions derive from a limited-basis set, energy-minimized MO algorithm developed by the author. The lower panel shows ΣK of the two-carbon systems versus energy.

The charge distributions for organic compounds are accessible by a variety of methods.[34,35] For conciseness, we limit our discussion to energy-minimized, minimal-basis-set combinations of atomic core and valence orbitals. The computational programs which arrived at these electron distributions were written and refined by the author. They are extensions of the semiempirical linear combination of atomic orbitals—molecular orbital classics originated by Wolfsberg and Helmholz and developed further by Hoffmann.[40,41] As with our survey studies of atoms and diatomics,[42] we focus solely on electronic ground states. The object is to illustrate in the simplest way possible the low-level information characteristics.

Figure 5 begins with data for familiar two-carbon systems: acetylene (C_2H_2), ethylene (C_2H_4), and ethane (C_2H_6). The upper panel illustrates $K(\rho{:}\varphi)$ versus MO energy. The data points are congested due to the number of MOs in question, especially near the valence level. There is a take-home point in that, for organic compounds, the largest $K(\rho{:}\varphi)$ values are not affiliated with the core orbitals which play minor roles in covalent bond formation. Rather the highest $K(\rho{:}\varphi)$ values are those expressed near the valence level; the highest information is thus tied to the MOs critical to

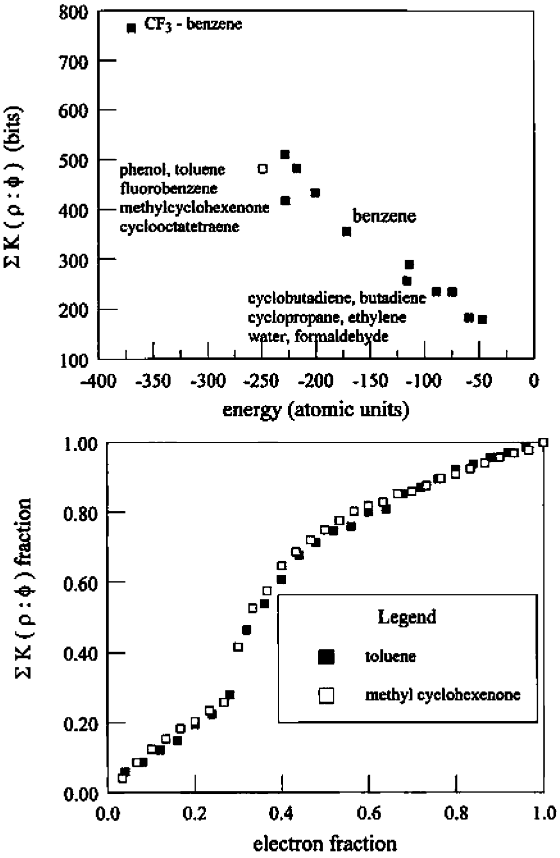


Figure 6. Low-level information for diverse carbon compounds. The upper panel illustrates $\sum K$ versus MO energy. The lower panel shows fractional $\sum K(\rho{:}\varphi)$ versus the electron fraction for 5-methyl-3-cyclohexenene-1-one and toluene. The ordering of the data points is determined by the MO energies: left-to-right ↔ lowest energy to highest. The container parameters are the same as exercised for acetylene, ethylene, and ethane in Figure 5.

covalent binding. In addition, one observes the $K(\rho{:}\varphi)$ values to be dispersed over a range of 3 −12 bits near the valence levels. We note that all of the Figure 5 data were obtained via the same container and summation parameters: volumes of partitions of 20 and 0.1 Å,[3] respectively.

Summed information data—$\sum K$ equals the sum of K values over occupied orbitals[43]—are illustrated for C_2H_2, C_2H_4, and C_2H_6 in the lower panel of Figure 5. Not surprisingly, $\sum K$ is found to be comparable for the three molecules. Interestingly, $\sum K$ for C_2H_6 falls between that measured for C_2H_2 and C_2H_4. C_2H_6 carries the greatest number of electrons (i.e., 18), yet its propensity for concentrating charge is compromised by the C−C bond length, that is, 1.54 Å versus 1.20 and 1.32 Å for acetylene and ethylene, respectively. By and large, $\sum K$ increases with the number of charges, as observed for atoms and diatomics.[42] Figure 5 points out, however, that information is acutely sensitive to high-level details such as the degree of unsaturation.

$\sum K$ is illustrated for diverse organic compounds in Figure 6: α,α,α-trifluorotoluene, butadiene, cyclopropane, and so forth. Included are data for 5(*S*)-methyl-2-cyclohexene-1-one ($C_7H_{10}O$, open square) which was central to the section II discussion. The $\sum K$ scaling proves fairly linear with the best-fit slope, 1.82 bits/au and R^2 = 0.970.

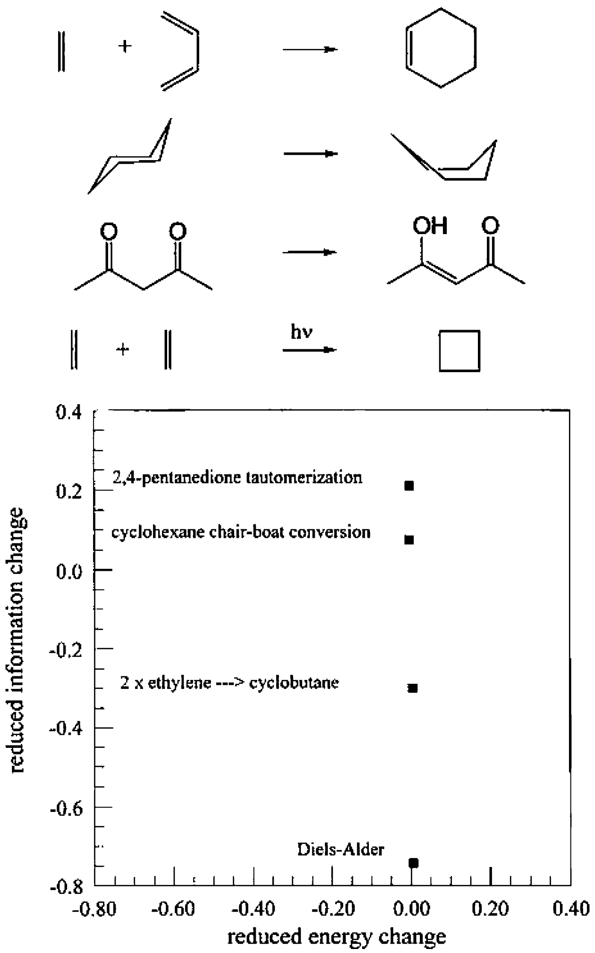


Figure 7. Changes of low-level information via chemical reactions. The upper half depicts four well-known reactions: Diels−Alder, chair-boat conversion of cyclohexane, tautomerization of 2,4-pentane-dione, and photochemically induced condensation of ethylene. The lower panel illustrates the reduced information and energy changes for each reaction.

A second type of scaling is illustrated in the lower panel. As is well-known, basis set orbitals can approximate the charge distributions of an infinite number of molecules. This trait has ramifications on how information is parceled within a given compound. Thus, if one plots fractional $\sum K$ versus the electron fraction, an unusual scaling behavior is observed for organic compounds. The data in the lower panel of Figure 6 pertain to toluene and 5(*S*)-methyl-2-cyclohexene-1-one, the ordering of the points dictated by the MO energies: lowest-to-highest ↔ left-to-right in the plot. The chemistry of toluene differs radically from that of a cyclohexenone derivative. However, the way that information is packaged internally for the two molecules is strikingly similar. In our experience, this point can be made using virtually any pair of organic compounds. The results show that ~40% of the low-level information content is affiliated with ~30% of the electrons. In turn, the MOs which lie near the valence level are under-represented, contributing <20% of the information content.

A molecule's structure and function change during a chemical reaction. What is striking is that the energy changes are typically minor compared with changes in low-level $\sum K$. The reactions of Figure 7 along with reduced variables illustrate this point. One defines the energy and information differences involving the reactants and products (indicated by subscripts) as follows:

$$\Delta\sum\epsilon = \frac{\sum\epsilon_p - \sum\epsilon_r}{\sum\epsilon_r} \qquad (9)$$

The lower panel of Figure 7 then shows the contrast between

$$\Delta\sum K = \frac{\sum K_p - \sum K_r}{\sum K_r} \qquad (10)$$

the energy and information changes with $\Delta\sum\epsilon \approx 0$ and $\Delta\sum K$ in the range -0.30 to $+0.30$. One learns that molecular information is not a frozen quantity. Rather, it is acutely sensitive to tautomerizations, chair-boat conversions, and other types of chemical reactions.

V. THE LEVEL CORRESPONDENCE PROPERTIES OF MOLECULAR INFORMATION

Organic molecules process information via collisions, typically in liquid environments. Their operating methods are consequently poles apart from those of everyday computers. The two processor types nonetheless share questions of correspondence regarding the lower and higher information levels. The purpose of this section is to address the questions relevant to molecules, with digital computers providing ready analogies. As in previous sections, a handful of compounds such as 5-methyl-2-cyclohexene-1-one illustrate the properties of carbon systems in general.

A. The Correspondence between Source and Processing Information. The information of a source can be significantly more compressed than that entailed during real-time processing. For example, the Pascal programs developed for this research were a fraction of the code which implemented the calculations—loops, procedures, and function-calls effected multiple expansions of each source file upon compilation. Likewise, the quantity of information handled during the calculations exceeded the execution files by several orders of magnitude. A desktop computer can readily process $\sim 10^{14}$ bits during a 1 h calculation, in response to a 10^5-bit source. The energy supply coupled with the low-level software enables a billion-fold stepup of the source information.

In an analogous way, a molecule such as 5-methyl-2-cyclohexene-1-one serves as an information source albeit in a compressed format: $\sum K \approx 482$ bits as illustrated in the upper half of Figure 6. The molecule's processing action, however, occurs via extended collision strings **...(C=C)(C−H)(C−C)(C=O)...** in solution. The first question that arises is: how does the source information compare quantity-wise with that expressed during processing?

One answers this question by examining Brownian collision strings (serial and/or parallel processing) such as those illustrated in Figure 8 and detailed in our previous works.[24,25] One then applies the section III methods in order to compute $\sum K$ for the individual atom−bond−atom states: **(C=C)**, **(C−H)**, **(C=O)**, and so forth. As with the high-level code, these quantities are averaged over several orders and compared to $\sum K$ of the original (compressed) source.

Data for several compounds are presented in the lower panel of Figure 8. The vertical axis refers to *processing* $\sum K$, having been averaged eight orders versus *source* $\sum K$ (horizontal axis). Not surprisingly, the greater the source amount, the more information is entailed during processing. This property is in effect regardless of the order of analysis, although the precise dependence differs from compound to compound. What is important is that every organic molecule demonstrates a unique processing capacity as measured by the average bits per code unit of the Brownian string. Of the Figure 8 compounds, benzoic acid demonstrates the greatest capacity at ~510/8 ≈ 63.8 bits/code unit realized via its Brownian collision strings. Note further that compounds such as 5-methyl-2-cyclohexene-1-one and phenol exhibit nearly identical *source* $\sum K$ values. By contrast, ~25% more information is expressed by the aromatic weak acid compound during processing. This means that, of the Figure 8 molecules, phenol demonstrates the greatest augmentation of its source content during processing. For phenol and derivative molecules, a comparatively small amount of information in a compressed format controls a larger quantity realized during processing.

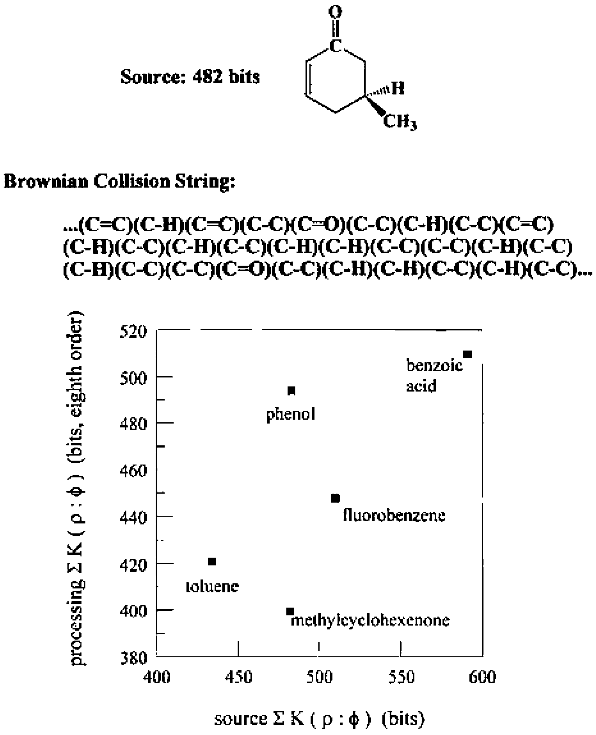
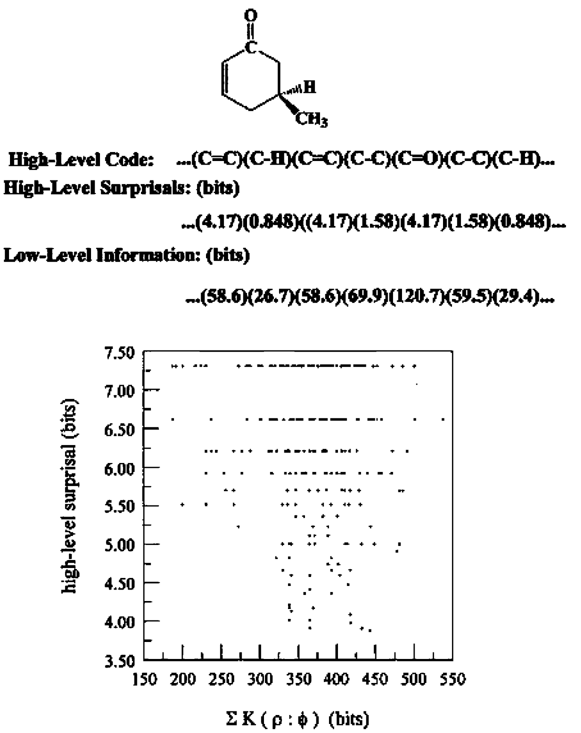


Figure 8. Correspondence between source and processing information. The upper portion illustrates 5-methyl-3-cyclohexenene-1-one and a collision string fragment expressed during Brownian serial processing. The lower panel illustrates *processing* $\sum K$ versus *source* $\sum K$ for several compounds. The information values follow from an eighth-order analysis of atom−bond−atom states.

B. The Correspondence between Low and High Information Levels. In everyday processing, there can be sparingly little correspondence between low and high information levels. For example, the symbols "e" and "z" appear with markedly different frequencies in written text materials; the self-information for these high-level code units is thus low and high, respectively. Matters are egalitarian, however, at the low levels of a computer word processor. In an ASC file, an equal number of bits is apportioned to each character regardless of occurrence frequency. This means that the information expressed at high levels is retained inefficiently at the lower levels. In turn, there is a notable absence of



High-Level Code: ...(C=C)(C-H)(C=C)(C-C)(C=O)(C-C)(C-H)...

High-Level Surprisals: (bits)

...(4.17)(0.848)((4.17)(1.58)(4.17)(1.58)(0.848)...

Low-Level Information: (bits)

...(58.6)(26.7)(58.6)(69.9)(120.7)(59.5)(29.4)...

Figure 9. Correspondence between high- and low-level information. The upper portion illustrates 5-methyl-3-cyclohexenene-1-one and a collision string expressed in Brownian serial processing. Included are the surprisal values (self-information) associated with the high-level coding and the low-level information quantities. The lower panel illustrates the high-level surprisals and $\sum K$ for all eighth-order fragments.

correlations between the different levels of a digital word processor.

The correspondence between the different information levels for organic compounds warrants analogous attention. The upper panel of Figure 9 shows a portion of a Brownian collision string for 5-methyl-2-cyclohexene-1-one. Immediately below appear the self-information (surprisal) values associated with each high-level code unit. Just beneath these are listed the $\sum K$ values calculated for the atom—bond—atom states. Hence, the second question is: what (if any) correspondence is demonstrated between the different information levels?

We have considered this issue via extensive comparisons of the high- and low-level information quantities for Brownian collision strings. Sample data are presented in Figures 9 and 10 for 5-methyl-2-cyclohexene-1-one. In the lower portion of Figure 9 are plotted the high-level surprisals associated with multiple order states of atom—bond—atom code units versus low-level $\sum K$. A peculiar correspondence is thereby observed. The *average* of the low-level content proves more-or-less independent of the high-level surprisal; this trait is detailed in the upper panel of Figure 10. One finds the $\sum K$ averages to vary less than 10% as a function of the high-level information content.

The spread of information at the low levels relates a different story as illustrated in the lower panel of Figure 10. One observes that the higher the low-level surprisal, the greater the standard deviation in $\sum K$. In other words, for an organic compound, the greater the self-information realized in the high-level code, the greater the *uncertainty* regarding the low-level content. For a given compound, certain collision sequences of atom—bond—atom states manifest only rarely during Brownian processing. It is these rare sequences—as opposed to the commonplace ones—which demonstrate the widest range in their low-level information content.

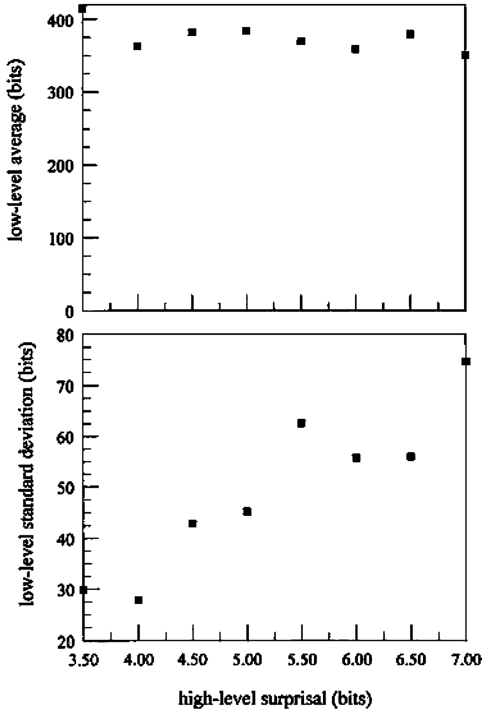


Figure 10. Averages and spreads of low-level information. The upper panel shows the dependence of the average low-level $\sum K$ versus high-level surprisals. The data follow from an eighth-order analysis of collision strings for 5-methyl-3-cyclohexenene-1-one. The lower panel shows the standard deviation of $\sum K$ versus high-level surprisal. The average $\sum K$ varies little with changes in the high-level information, whereas the standard deviation increases more-or-less linearly.

C. The Correspondence between Available and Accessible Information. The random access memory of a typical computer offers $32 \times 512 \times 10^6 \approx 1.6 \times 10^{10}$ bits of information. This amount is augmented by the media drives such as flash and CD-ROM. The interfaces and energy supply enable abundant queries and transfers, all wrapped in 32-bit (or larger) packages. The end effect is that 100% of the information—RAM, media, and otherwise—is processing-accessible.

Organic compounds handle information via collisions in solution. Figures 5 (two-carbon molecules) and 6 (diverse polyatomics) offer measures of low-level content which is *available* for processing. A third question thus arises; namely, what percentage of the content is operationally accessible?

Importantly, the amount of information that can be registered via Brownian processing is severely limited, at least via local, single-collision events. The upper panel of Figure 11 illustrates a typical Brownian sequence of the per-molecule information registered reversibly in liquid ethanol at room temperature. The magnitudes follow the entropy—heat capacity relation of eq 1 with the added assumption of Gaussian-weighted statistics. If the registered information is compared with a collision string for a solute such as 5-methyl-2-cyclohexene-1-one:

...(C—H)(C—C)(C—H)(C—C)(C=O)(C—C)(C—H)...

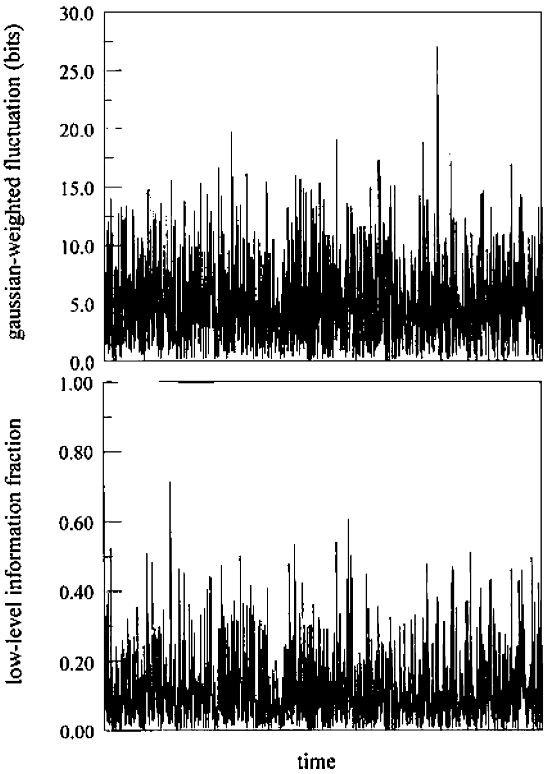


Figure 11. Correspondence between available and accessible information. The upper portion illustrates Gaussian-weighted information registered for a Brownian collision string of 5-methyl-3-cyclohexenene-1-one as allowed by thermal fluctuations in liquid ethanol. The lower panel shows the fraction of information registered. The highest fractions are associated with **(C−H)** states. The lowest fractions are observed for **(C═O)** and **(C═C)** states.

one finds that only a fraction of low-level content can be accessed. That is to say,

...(C-H) (C-C) **(C-H)** (C-C) (C=O) (C-C) **(C-H)...**

portrays in a more realistic way the amount of information logged in Brownian processing. The percentages for atom−bond−atom states range from 1 to 60% as shown in the lower panel of Figure 11. Naturally, each state poses a different available content dependent on the charge distribution. What is important is that the largest fractions of registered information involve the **(C−H)** units for an organic compound. Functional groups such as **(C═C)** and **(C═O)** are the active players in chemical reactions; however, their fractional role in Brownian processing is minor compared to that of the framework **(C−H)** units. Note that the section III methods offer lower-limit quantification of molecular information. Thus, the fractional values of Figure 11 represent upper limit values.

D. Molecular Information and Processing Errors. A digital computer demonstrates a nonzero—but fortunately low—processing error frequency. This quantity depends on several variables such as the material composition and temperature and the rate at which data are transferred. That the error frequency exceeds zero requires that detection and correction facilities be incorporated into the hardware and software.[44]

Organic molecules process information via collisions in liquid solution. It is reasonable to expect that a **(C═C)** unit

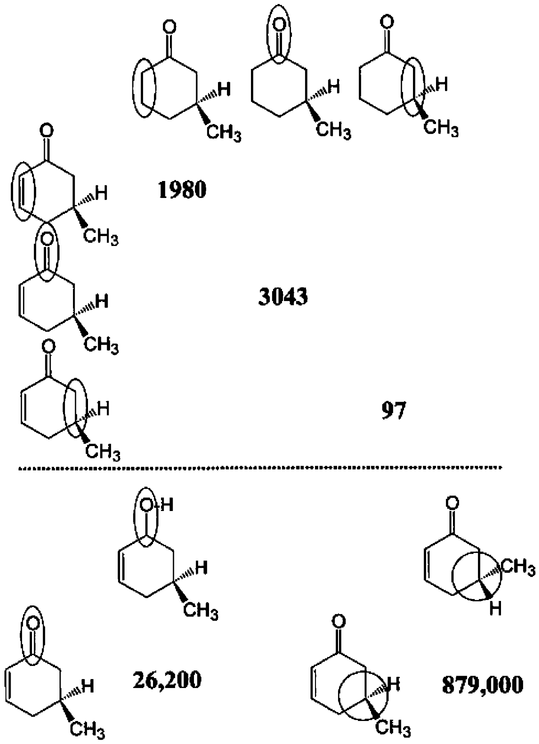
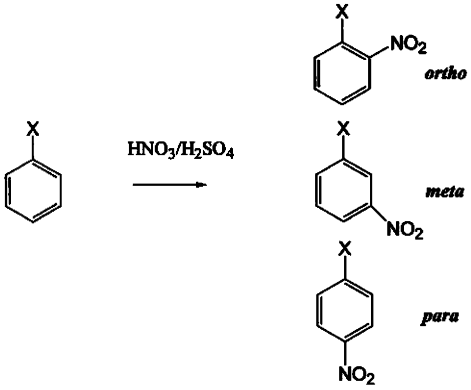


Figure 12. Minimum $D(\rho:\varphi)$ for atom−bond−atom states. The ovals indicate the states subject to divergence calculations under maximum alignment conditions. The summations have been restricted to a single valence molecular orbital electron for each compound. Numerical quantities are expressed in bits.

of a given compound will be registered occasionally as **(C−C)**, **(C═O)** as **(C−O)**, and so forth. Thus, the fourth question becomes: what is the correspondence between the information posed by a molecule and that which is registered accurately during processing? Alternatively, what fraction of the registered information for a molecule is expected to be error-tarnished?

We have considered this question by means of divergence calculations applied to a variety of molecular recognition scenarios. We consider here the Brownian processing of 5(*S*)-methyl-2-cyclohexene-1-one in an ethanol solution. A collision in which **(C═C)** is misregistered as **(C−C)** is equivalent to the accurate registration of the 2,3 site of 5(*S*)-methyl-2-cyclo*hexane*-1-one. Such an error could occur if the charge distributions for **(C═C)** and **(C−C)** in the two compounds resembled one another in a twinlike way, for example, in the manner of atom−bond−atom states involving different isotopes. A registration error could also arise if the **(C═C)** charge distribution was distorted on account of thermal fluctuations.

$D(\rho:\varphi)$ measures the ability to discriminate two distribution functions φ and ρ. For organic compounds, $D(\rho:\varphi) \approx 0$ bits in comparisons of **(C−H)** and **(C−D)**, **(^{13}C−^{12}C)** and **(^{14}C−^{12}C)**, and so forth. Figure 12 illustrates sample data for *minimum* $D(\rho:\varphi)$ for a variety of atom−bond−atom states. The upper portion shows $D(\rho:\varphi)$ associated with 5(*S*)-methyl-2-cyclohexene-1-one and its saturated ring counterpart, also having an *S* stereocenter. The charge states being compared are indicated by ovals superimposed on the structure graphs. For each case, comparison was made under maximum alignment conditions as for the H-atom distributions of Figure 4. Further, the eq 7 summations were restricted to a single



X	*o*	*p*	*m*	λ
OH	40	58	< 2	0.308
F	12	87	< 0.2	0.649
CH_3	58	38	4	0.261
CH_2F	28	54	17	0.0985
CO_2H	18	2	80	0.485
NO_2	5	2	93	0.731
CF_3	0	0	100	1.00
$C(CH_3)_3$	12	79.5	8.5	0.412

Figure 13. Electrophilic substitution of benzene derivatives. The upper portion shows the general form of the nitration reaction with ortho, meta, and para products. The lower portion features data compiled in ref 45 and includes values of the selectivity parameter λ.

electron affiliated with the highest-occupied MO of each molecule. The latter condition was imposed given that only a fraction of a given molecule's information is accessible in Brownian processing.

The results are compelling. One learns that $D(\rho{:}\varphi) \approx 1980$ bits for equivalent-site comparisons of **(C═C)** and **(C−C)** in 5(*S*)-methyl-2-cyclohexene-1-one and 5(*S*)-methyl-2-cyclo*hexane*-1-one. Equally noteworthy is that substantial $D(\rho{:}\varphi)$ values (ca. 3043 bits) are the rule when the carbonyl groups of the two compounds are compared. Likewise, $D(\rho{:}\varphi) \approx 97$ bits is obtained in comparisons of equivalent-site alkane bonds. These calculations show that there is ample basis for discriminating the atom−bond−atom states of a given compound, even when the electronic disparities are subtle. Such is true even when minimalist fractions of the total information content are registered.

The lower portion of Figure 12 amplifies the above points. One finds $D(\rho{:}\varphi) \approx 26\,200$ bits in comparisons of **(C═O)** and **(C−O)** of the closely related ring compounds solely on the basis of valence electrons. Moreover, $D(\rho{:}\varphi) \approx 879\,000$ bits is found in comparisons of the chiral centers of *R* and *S* enantiomers. The over-riding message is that typical minimum $D(\rho{:}\varphi)$ far exceeds the information which is operationally accessible by typical fluctuations.

The informatic errors associated with Brownian processing are clearly infrequent. For a sampling error to occur, the distance between the true and erroneously registered states being ~100 bits, requires local thermal fluctuations with a probability on the order of $\exp[-100 \log_e(2)] \approx 10^{-30}$. Thus, for a molecule subject to 10^{13} collisions per second, each serving as a registration event, an error transpires about once every 3 billion years.

E. The Correspondence between Molecular Information and Reaction Selectivity. The circuitry in a computer performs electrical work directed by the resident software. The energy supply quantifies the capacity for work. The information stored provides the instructions for how and what work is performed. The familiar result is that computers direct voltages and currents (*V*,*I*) in time- and spatially selective ways. A total of 32 bits of information offers double the *V*,*I*-instructional capacity of a 16-bit packet.

Organic molecules carry energy and information. The former enables the performance of work via chemical reactions; the latter functions as the on-site instruction set. Instructions furnish the criteria and enable the making of choices during work performance, for example, concerning reaction pathways. A large quantity of bits can specify greater pathway criteria compared with a small quantity. The question that arises then is: what is the correspondence between the information stored expressed by a molecule and the capacity for selecting reaction paths? Alternatively, how does molecular information correlate with the control of work performance?

We consider this question via a reaction genre that has received attention for over a century. Figure 13 illustrates the key elements of electrophilic aromatic substitution via nitration.[45] In the archetypical case, a single substituent of a benzene ring determines the percentages of ortho-, meta-, and para-substituted products. The lower portion of Figure 13 details the product distributions for eight compounds and includes for each an informatic selectivity parameter λ computed as follows. If a substituent X were to exert *no effect* on the choice of reaction pathways, then the *o*-, *m*-, and *p*-isomers would appear in equal numbers. Accordingly, the information H_I associated with the product composition in solution would be given by (cf. eq 2):

$$\begin{aligned} H_I &= (-1/\log_e 2) \textstyle\sum_j p(j) \log_e p(j) \\ &\approx (-1/0.693)3 \log_e(1/3) \\ &\approx 1.580 \text{ bits} \end{aligned} \tag{11}$$

Such would correspond to the maximum information possible given three states. One identifies λ by comparing H_I for each isomer distribution realized experimentally to the maximum information scenario, namely

$$\lambda = \frac{H_I^{(\max)} - H_I^{(\text{obs})}}{H_I^{(\max)}} \tag{12}$$

As is well-known, different substituents exert different selectivities; for example, $\lambda_{\text{toluene}} \approx 0.261$ whereas $\lambda_{\text{phenol}} \approx 0.308$. The disparities are traditionally explained in terms of the electron-withdrawing power of the substituent and the transition states involved in the nitration process.[45,46] The chemistry is not straightforward quantitatively, however. For example, $\lambda_{CH_2F-\text{benzene}} \approx 0.0985$ whereas $\lambda_{CF_3-\text{benzene}} \approx 1$. Thus, a single (electronegative) fluorine atom at the α-carbon position of toluene diminishes the reaction selectivity, whereas three α-fluorines maximize it!

It would be nice and tidy to discover that organic molecules play by the same elementary rules as computers; hence, the greater their information content, the greater the selectivity exhibited during work performance. Nature is more complicated, however, because only a fraction of a molecule's information is operationally accessible. As a

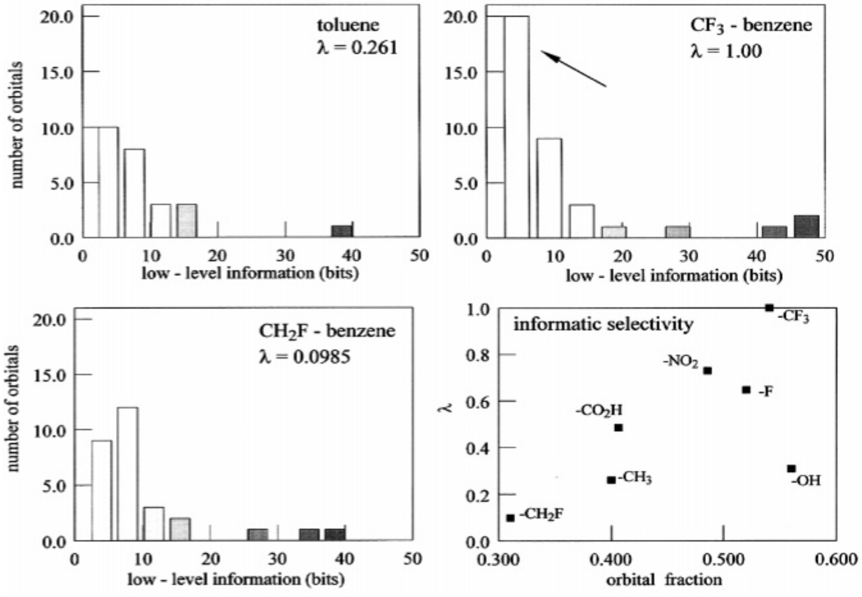


Figure 14. Correspondence between low-level information and reaction selectivity. The $K(\rho:\varphi)$ distributions are shown for toluene, α,α,α-trifluorotoluene, and α-monofluorotoluene. The dependence of λ on the fraction of first-bin information values is shown. The selectivity increases with this fraction as shown in the lower-right panel. The reaction selectivity demonstrated by phenol, however, does not adhere to this trend.

consequence, the control of work performance is governed not by the total content but rather by the finer points of the information distribution in a molecule.

Figure 14 shows how the low-level source information is distributed for three of the Figure 13 aromatics: toluene, α,α,α-trifluorotoluene, and α-monofluorotoluene. For each case, $K(\rho:\varphi)$ values allied with the molecular orbitals have been grouped among 10 equal bins. For these and related compounds, the ortho, meta, and para selectivity is determined by the left-most tails of the distributions. The reaction selectivity is enhanced when a substantial fraction of the MOs express low information content and is diminished otherwise.

The lower-right panel of Figure 14 shows the dependence of λ on the fraction of MOs which fall into the left-most (lowest information content) bin. One observes the pathway selectivity to increase linearly with this fraction. Phenol, interestingly, does not adhere to this trend. The aromatic molecule instead demonstrates a selectivity which is incommensurate with its low-level informatics.

VI. DISCUSSION

There are numerous ways to model an organic compound. One can appeal, for example, to ball-and-stick assemblies and structure graphs; these are the models with the deepest historical roots. Alternatively, one can utilize computer graphics based on electronic structure calculations. The former models are valuable for their immediacy and contact with valence principles. The latter are esteemed for their rigor and thoroughness.

Molecular models are numerous and varied; this paper has concentrated on the thread linking them. At heart, all models for organic compounds communicate information about angstrom-scale charge states. Accordingly, all models facilitate understanding of chemical stability, structure/activity relations, and synthetic routes. Our approach has been to view molecules first and foremost as information carriers and processors, their actions inextricably tied to probability distributions and level hierarchies. Our previous work focused on the informatic attributes of higher levels, as encoded by chemical formulas and structure graphs. The low-level aspects discussed in this writing were the natural next step of inquiry. Interestingly, digital computers are necessary to obtain the probability functions which underpin molecular information. In so doing, macroscopic devices—as with ball-and-stick models, 2D graphs, and so forth—offer useful analogies for the angstrom scale.

There were two aims of this study. First was to quantify the low-level information for chemical systems at a resolution insensitive to isotope effects. The methods for doing so were detailed in section III and demonstrated accordingly for diverse systems. The data were predicated on the electron distributions approximated via limited-basis-set orbitals. Importantly, the procedural aspects can be adapted readily to other electronic structure algorithms used to establish the charge distributions. The equivalent statement holds for the charge distributions accessed via experimentation. The second objective was to illustrate the level correspondence properties for organic compounds as in section V. Overall, one learns several things.

First, we were able to identify an appealing simplicity via the linear information-energy scaling. It is thus a straightforward matter to infer $\sum K$ for molecules not represented in Figure 6 on the basis of their MO energies. While the effects of nodal structure on $\sum K$ are nuanced and case-specific, the packaging architecture for information proves much the same for diverse compounds (cf. Figure 6, lower panel).

Note that, for organic molecules, the summed quantity of information does not reflect the stability—or the lack of it. The reactivities of cyclobutadiene, benzene, and cyclooctatetraene are radically different.[47] Yet there is no hint of such disparity in their information values (cf. Figure 6). Everyday computer processing offers an analogy. The binary words **01011101**, **1010101011101010**, and **10101001010101010111010101011011** offer 8, 16, and 32 bits of information, respectively. Yet in the absence of context, such do not reveal whether the messages are of a source, transient, or final-output nature. In a like manner, organic compounds express information which can be altered via chemical reactions. Their tendency to do so, however, is not disclosed by their message size alone.

The level-correspondence properties of organic compounds are equally noteworthy. In everyday computer processing, the expansion of information of a source file is limited only by the available energy supply. Matters are more complex for organic compounds given that thermal fluctuations drive all of the processing events. Section V.A thus emphasizes that every molecule in solution, as a source of information, demonstrates a unique expansion factor. This factor is governed by the particulars of the molecular structure and Brownian collision strings. The relation between *processing* $\sum K$ and *source* $\sum K$ is worth knowing given that information reflects a capacity for work control. Different compounds with a common function such as enzyme inhibition pose disparate *processing* $\sum K$/*source* $\sum K$ ratios. Importantly, compounds with the largest ratios offer the greatest control potential, that is, the greatest return on the source information investment. For a chemist investigating a library of compounds, the section III methods plus Brownian processing enable identification of the optimum investment strategies.

Section V.B addressed the correspondence between the low and high information levels for organic compounds. It was shown how the spread of low-level content increases with the high-level surprisal—the self-information affiliated with atom−bond−atom code units. This is a salient feature of organic compounds, one which connects closely with their action in solution. A molecule implements specific actions such as enzyme inhibition via the messages it transmits. The messages are encoded by collision sequences during Brownian processing. Section V.B establishes that the rarest of messages for a given compound are the ones furthest apart in terms of their low-level bit quantities. Such a trait reflects an effective code, one ideally suited to low error rates during communication. For organic molecules, the rare messages are those which can be distinguished not only by the code units themselves but also by their size in message space.

Section V.C emphasized that only part of the information posed by a molecule is accessible under typical solution conditions. This marks not a shortcoming but rather a vital trait of carbon-based systems. As is well-known, an information processor need not register all the bits posed by a source in order to elect a response. In computation, for example, the Boolean statement “if j mod 2 = 0 then procedure_1” turns solely on the right-most bit of the integer *j* address. In an analogous way, chemical action is governed by partial molecular information. The informatics are thereby consistent with the prominent roles played by valence electrons and functional groups in organic compounds; the core electrons and hydrocarbon frameworks play supporting roles by comparison.

Section II and Figure 2 identified several of the complexities of molecular information processing. At first glance, one would expect electronic scenarios driven by thermal fluctuations to be ill-suited to high-fidelity communication. The reality proves otherwise for organic compounds. By quantifying the low-level informatic divergence in several test cases, section V.D emphasizes the robustness of the transmission code. For molecules in solution, there is a paucity of registration errors in spite of the Brownian nature.

Section V.E considered how information connects with the selection of reaction pathways—the control of energy spending during chemical activity. Importantly, it is not the total information contained by a molecule which governs the pathway selection. Rather, it is the fraction of content which is accessible, that is, that information allied with the left-most tails of distributions such as in Figure 14. The section III methods enable quantification of the distributions for isolated compounds. A better understanding is needed, however, regarding solvent effects. For example, the behavior of phenol in pathway selectivity must be affected by solvent molecules and their information.

In summary, a study of Brownian processing and the low-level information content of organic molecules has been presented. The principal results were development of the quantification methods and illustration of the level correspondence properties for diverse systems. The subject does not terminate here given the followup critical questions that turn on practical applications: how do information level principles assist, for example, in the design and screening of enzyme inhibitors? Studies with a practical bent are currently being pursued with strategies furnished by this paper. In survey screening for “hits”, one is guided by the correspondence between the high and low levels (section V.B); one seeks molecules which evince the greatest range of low-level Brownian information. It is these compounds which offer the widest coverage of chemical message space. Along the same lines, one targets inhibitors which demonstrate the largest ratios between processing $\sum K$ and source $\sum K$ (section V.A); in these compounds, a small quantity of information is up-converted substantially in a Brownian environment. Other level principles offer guidelines of a fine-tuning nature. The greatest divergences are observed where asymmetric centers enter the picture (section V.D). Chiral (as opposed to achiral) molecules thus offer the greatest specificity and the lowest error rates during chemical message transmission. By the same token, molecular information of any flavor is useful only if it is accessible (section V.C). For organic compounds, the accessibility is tuned largely by the molecular orbitals near the valence level.

We close by adding that the level hierarchies are not governed solely by structure graphs and molecular orbitals. For proteins in particular, the information hierarchies entail the primary, secondary, tertiary, and quaternary structures which are all dictated by amino acid sequences. At present, we are also directing Brownian processing methods to globular proteins so as to establish structure−function relationships in an alternative way.

ACKNOWLEDGMENT

The author is grateful to William Sevchuk and Brant Carlson of the NSF-REU program for early-stage calculations. The author is similarly appreciative to Tim Thomas and Daniel Becker for helpful discussions, to the reviewers for comments and criticism, and to Jessica Greminger for technical assistance.

REFERENCES AND NOTES

(1) Garrett, P. B. *The Mathematics of Coding Theory: Information Compression, Error Correction, and Finite Fields*; Pearson/Prentice Hall: Upper Saddle River, New Jersey, 2004.

(2) Lavenda, B. H. *Statistical Physics*; Wiley: New York, 1991.

(3) Tribus, M.; McIrvine, E. C. Energy and Information. *Sci. Am.* **1971**, *225*, 179.

(4) For a popular account, please see: Loewenstein, W. R. *The Touchstone of Life. Molecular Information, Cell Communication, and the Foundations of Life*; Oxford University Press: New York, 1999.

(5) Bonchev, D. *Information Theoretic Indices for Characterization of Chemical Structure*; Research Studies Press-Wiley: Chichester, 1993. See also: Bonchev, D.; Trinajstic, N. Information Theory, Distance Matrix, and Molecular Branching. *J. Chem. Phys.* **1978**, *67*, 4517.

(6) Basak, S. C.; Grunwald, G. D.; Niemi, G. J. Use of Graph-Theoretic and Geometrical Molecular Descriptors in Structure−Activity Relationships. In *From Chemical Topology to Three-Dimensional Geometry*; Balaban, A. T., Ed.; Plenum Press: New York, 1997; Chapter IV, p 73.

(7) González-Díaz, H.; Molina, R. R.; Uriarte, E. Stochastic Molecular Descriptors for Polymers I. Modeling the Properties of Icosahedral Viruses with 3D-Markovian Negentropies. *Polymer* **2004**, *45*, 3845. González-Díaz, H.; Molina, R. R.; Uriarte, E. Markov Entropy Backbone Electrostatic Descriptors for Predicting Proteins Biological Activity. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 4691. Gonzalez-Diaz, H.; Aguero-Chapin, G.; Varona-Santos, J.; Molina, R. R.; de la Riva, G.; Uriarte, E. 2D RNA-QSAR: Assigning ACC Oxidase Family Membership with Stochastic Molecular Descriptors. *Biorg. Med. Chem. Lett.* **2005**, *15*, 2932. de Ramos, A. R.; González-Díaz, H.; Molina, R. R.; Uriarte, E. Markovian Backbone Negentropies: Molecular Descriptors for Protein Research. I. Predicting Protein Stability in Arc Repressor Mutants. *Proteins: Struct., Funct., Bioinf.* **2004**, *56*, 715. González-Díaz, H.; de Ramos, A. R.; Molina, R. R. Markovian Negentropies in Bioinformatics. 1. A Picture of Footprints After the Interaction of the HIV Ψ-RNA Packaging Region with Drugs. *Bioinformatics* **2003**, 2079. González-Díaz, H.; Marrero, Y.; Hernandez, I.; Bastida, I.; Tenorio, E.; Nasco, O.; Uriarte, E.; Castanedo, N. C.; Cabrera-Perez, M. A.; Aguila, E.; Marrero, O.; Morales, A.; González, M. P. An Alternative in Silico Technique for Chemical Resarch in Toxicology. 1. Prediction of Chemically-Induced Agranulocytosis. *Chem. Res. Toxicol.* **2003**, *16*, 1318.

(8) Batista, J.; Godden, J. W.; Bajorath, J. Assessment of Molecular Similarity from the Analysis of Randomly Generated Structural Fragment Populations. *J. Chem. Inf. Model.* **2006**, *46*, 1937. See also: Stahura, F. L.; Godden, J. W.; Bajorath, J. Distinguishing between Natural Products and Synthetic Molecules by Shannon Descriptor Entropy Analysis and Binary QSAR Calculations. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1245.

(9) Remacle, F.; Levine, R. D. Towards Molecular Logic Machines. *J. Chem. Phys.* **2001**, *114*, 10239. Remacle, F.; Weinkauf, R.; Steinitz, D.; Kompa, K. L.; Levine, R. D. Molecular Logic Machines by Optical Spectroscopy and Charge Migration Along a Molecular Wire Realized as a Peptide. *Chem. Phys.* **2002**, *282*, 363. Remacle, F.; Schlag, E. W.; Selzle, H.; Kompa, K. L.; Even, U.; Levine, R. D. Logic Gates Using High Rydberg States. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 2973. Steinitz, D.; Remacle, F.; Levine, R. D. On Spectroscopy, Control, and Molecular Information Processing. *Chem. Phys. Chem.* **2002**, *3*, 43.

(10) Rambidi, N. G. Lure of Molecular Electronics—from Molecular Switches to Distributed Information Processing. *Microelectron. Eng.* **2003**, *69*, 485.

(11) Pease, A. R.; Jepessen, J. O.; Stoddart, J. F.; Yi, L.; Collier, C. P.; Heath, J. R. Switching Devices Based on Interlocked Molecules. *Acc. Chem. Res.* **2001**, *34*, 433.

(12) Bourret, R. B.; Stock, A. M. Lessons from Bacterial Chemotaxis. *J. Biol. Chem.* **2002**, *277*, 9625.

(13) Levine, R. D. Information Theory Approach to Molecular Reaction Dynamics. *Ann. Rev. Phys. Chem.* **1978**, *29*, 59.

(14) Nagy, A.; Parr, R. G. Information Entropy as a Measure of the Quality of an Approximate Electronic Wave Function. *Int. J. Quantum Chem.* **1996**, *58*, 323. Nalewajski, R. F.; Parr, R. G. Information Theory, Atoms in Molecules, and Molecular Similarity. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 8879. Nalewajski, R. F.; Parr, R. G. Information Theory Thermodynamics of Molecules and their Hirshfeld Fragments. *J. Phys. Chem. A* **2001**, *105*, 7391.

(15) Nalewajski, R. F. Entropic Measures of Bond Multiplicity from the Information Theory. *J. Phys. Chem A* **2000**, *104*, 11940. Nalewajski, R. F. Information Theoretic Approach to Fluctuations and Electron Flows Between Molecular Fragments. *J. Phys. Chem A* **2003**, *107*, 3792. Nalewajski, R. F. Information Principles in the Theory of Electronic Structure. *Chem. Phys. Lett.* **2003**, *375*, 196. Nalewajski, R. F.; Parr, R. G. Information Theory, Atoms in Molecules, and Molecular Similarity. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 8879.

(16) Nalewajski, R. F. *Information Theory of Molecular Systems*; Elsevier: Amsterdam, 2006.

(17) Cooper, J. W. *The Minicomputer in the Laboratory*; Wiley: New York, 1983; Chapter 14.

(18) See, for example: Terrett, N. K. *Combinatorial Chemistry*; Oxford University Press: Oxford, U. K., 1998.

(19) Herzberg, G. *Molecular Spectra and Molecular Structure*; New York, 1966.

(20) See, for example: Maizell, R. E. *How to Find Chemical Information*; Wiley: New York, 1998.

(21) See, for example: Fleming, I. *Selected Organic Syntheses*; Wiley: London, 1973.

(22) Graham, D. J.; Schacht, D. Base Information Content in Organic Molecular Formulae. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 942.

(23) Graham, D. J. Information Content in Organic Molecules: Structure Considerations Based on Integer Statistics. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 215.

(24) Graham, D. J.; Malarkey, C.; Schulmerich, M. V. Information Content in Organic Molecules: Quantification and Statistical Structure via Brownian Processing. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1601. Graham, D. J.; Schulmerich, M. V. Information Content in Organic Molecules: Reaction Pathway Analysis via Brownian Processing. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1612.

(25) Graham, D. J. Information Content and Organic Molecules: Aggregation States and Solvent Effects. *J. Chem. Inf. Model.* **2005**, *45*, 1223.

(26) Bennett, C. H. Thermodynamics of Computation - A Review. *Int. J. Theor. Phys.* **1982**, *21*, 905.

(27) The collision frequency is estimated using standard kinetic theory formulae. See: Hecht, C. E. *Statistical Thermodynamics and Kinetic Theory*; Freeman: New York, 1990; Chapter 6.

(28) Kivelson, D.; Madden, P. A. Light Scattering Studies of Molecular Liquids. *Annu. Rev. Phys. Chem.* **1980**, *31*, 523.

(29) Landau, L. D.; Lifshitz, E. M. *Statistical Physics*; Pergamon Press: London, 1958; Chapter XII.

(30) Masterton, W. L.; Hurley, C. N. *Chemistry Principles and Reactions*; Harcourt Brace: Fort Worth, TX, 1997; Chapter 8.

(31) Lloyd, S. Use of Mutual Information to Decrease Entropy: Implications from the Second Law of Thermodynamics. *Phys. Rev. A: At., Mol., Opt. Phys.* **1989**, *39*, 5378.

(32) Shannon, C. E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379.

(33) Reza, F. M. *An Introduction to Information Theory*; Dover: New York, 1994.

(34) See, for example: Wheatley, P. J. *The Determination of Molecular Structure*; Dover: New York, 1968.

(35) Cramer, C. J. *Essentials of Computational Chemistry*, 2nd ed.; Wiley: West Sussex, England, 2004.

(36) Kullback, S. *Information Theory and Statistics*; Dover: New York, 1997, Chapter 1.

(37) Caves, C. M. Quantum Limits on Bosonic Communication Rates. *Rev. Mod. Phys.* **1994**, *66*, 481.

(38) Dinur, U.; Levine, R. D. On the Entropy of a Continuous Distribution. *Chem. Phys.* **1975**, *8*, 17.

(39) Pauling, L.; Wilson, E. B., Jr. *Introduction to Quantum Mechanics with Applications to Chemistry*; Dover: New York, 1985.

(40) Wolfsberg, M.; Helmholz, L. The Spectra of the Tetrahedral Ions MnO_4^-, CrO_4^{2-}, and ClO_4. *J. Chem. Phys*. **1952**, *20*, 837.

(41) Hoffmann, R. An Extended Huckel Theory. I. Hydrocarbons. *J. Chem. Phys.* **1963**, *39*, 1397.

(42) Graham, D. J. Unpublished calculations.

(43) For example, one arrives at ΣK for the ground-state carbon atom via $2K_{1s} + 2K_{2s} + 2K_{2p}$. ΣK is obtained for N_2 via $2K_{1\sigma g} + 2K_{1\sigma u} + 2K_{2\sigma g} + 2K_{2\sigma u} + 2K_{3\sigma g} + 4K_{\pi u}$.

(44) Parity checks and Hamming codes are commonly used in computational error detection and correction. See: Ash, R. B. *Information Theory*; Dover Publications: New York, 1990; Chapter 4.

(45) Ferguson, L. N. Orientation of Substitution in the Benzene Nucleus. *Chem. Rev.* **1952**, *50*, 47.

(46) March, J. *Advanced Organic Chemistry*, 2nd ed.; McGraw-Hill: New York, 1977.

(47) le Noble, W. J. *Highlights of Organic Chemistry*; Dekker: New York, 1974; Chapter 9.

CI600488X