CHAPTER 12

THE MONTE CARLO APPROACH TO LIBRARY DESIGN AND REDESIGN

MICHAEL W. DEEM
Rice University

12.1 INTRODUCTION

Once a high throughput laboratory has the ability to synthesize and screen for figure of merit large numbers of compounds, the question naturally arises as to what compounds should be made. This question exposes the basic feature of HTE, which is that only an infinitesimally small fraction of the space of possible chemical compounds can actually be explored. To search a system with five mole-fraction variables and two noncomposition variables at a resolution of 1% would take, for example, a library of 9×10^{10} compounds. Similarly, to search a small molecule system with 15 possible templates, each with 6 attachment sites, and 1000 possible ligands would take a library of 1.5×10^{19} compounds. Even more dramatically, it would take a library of $20^{100} \approx 10^{130}$ members to search exhaustively the space of all 100 amino-acid-long proteins.

The task of deciding what to make, that is, the task of deciding how to search the chemical and experimental variable spaces, is central to the design of high-throughput experiments. There is a striking analogy between searching this variable space and searching configuration space in Monte Carlo computer simulation. Searching configuration space using Monte Carlo methods is also an apparently challenging problem, as very often systems with many thousands of continuous degrees of freedom are examined, with effective complexities in excess of 10^{2000}. To date, HTE has not been successful at searching spaces of anywhere near this complexity. This chapter pursues the analogy between searching composition and noncomposition space in HTE and searching configuration space using Monte Carlo methods.

Monte Carlo methods are based upon the rigorous mathematical framework of Markov sampling. This framework ensures that points are sampled in a certain statistical way in a Monte Carlo protocol. In particular, the probability of visiting any given set of composition and noncomposition variables is given by $p = \exp(-\beta E)$, where E is the negative of the figure of merit, and β is a control parameter that de-

Experimental Design for Combinatorial and High Throughput Materials Development, Edited by James N. Cawse.
ISBN 0-471-20343-2 © 2003 John Wiley & Sons, Inc.

termines the degree to which favorable figures of merit are sampled. In some cases, E may be a function of the value returned by the experimental screen. If chemical binding energies were to be optimized, for example, E would typically be the logarithm of the measured binding constant. If a Monte Carlo protocol for library design is to be valid, it must possess several properties [1]. First, the protocol must be Markovian. That is, the library design for the next round can depend only on the current compositions, not on compositions in previous rounds. Second, the library design must lead to effective sampling of the entire variable space, or what is known as regular sampling. Third, the library design must satisfy what is known as the *balance condition*. This last condition ensures that the figure of merit will be properly sampled. An easier and sufficient condition of detailed balance is often imposed rather than balance. Detailed balance basically requires that the forward and backward fluxes to go from one state to another by a specific type of move are equal. This condition, which is overly strict and can be relaxed, is one way to ensure that equilibrium is reached. The Monte Carlo methods presented in this chapter satisfy these three conditions of validity.

In practice, a Monte Carlo method modifies the variables in the system in a random, yet statistical way. In the simplest possible method, the current value of the variables is modified by a small random amount. If the figure of merit increases by this change, the new value of the variables is used. If the figure of merit decreases, the new value of the variables is occasionally used. The probability with which the new value is used in this case is given by $\exp[-\beta(E_{\text{proposed}} - E_{\text{current}})]$. This simple procedure is called *Metropolis Monte Carlo.*

A significant advantage of Monte Carlo methods is that they are guaranteed to sample the figure-of-merit space. This is in contrast to most other optimization methods, in which a single global optimum is sought. Of course, Monte Carlo methods are themselves effective techniques for global optimization. However, the sampling that Monte Carlo provides is an additional, significant advantage. For example, simple global optimization may be misleading, since concerns such as ease of synthesis, patentability, and cost of materials are not usually included in the experimental figure of merit. Another way of understanding this issue is to realize that the primary screen, or the screen most easily performed in the laboratory, is usually correlated only roughly with the true figure of merit. That is, after materials are found that look promising based upon the primary screen, secondary and tertiary screens are usually performed to identify that material that is truly optimal. Finally, it may be that the figure of merit is only roughly known or that it has not been fully defined by the experimenters or management or that several figures of merit are of interest. In this case, sampling on the roughly defined figure of merit would lead to a set of possible compounds, some of which may be found to be suitable when later examined with a more refined figure of merit. Global optimization on the rough figure of merit, on the other hand, would produce a single molecule, which would have a high chance of being rejected when the more refined figure of merit is identified and applied.

The ultimate test of new, theoretically motivated protocols for HTE is, of course, experimental. So as to motivate such experimentation, the effectiveness of Monte

Carlo protocols will be demonstrated in this chapter by high throughput experiments where the experimental screening step is replaced by a figure of merit returned by a model. That is, to test the proposed protocols in an efficient fashion, a model will be used in lieu of the real experimental screening process. The model is not fundamental to the protocols; it is introduced as a simple way to test, parameterize, and validate the various library design methods. Nonetheless, the validity of the model is important if discrimination between the various searching protocols is to be achieved. The particular type of model used is known in the statistical physics literature as a random-energy model. This type of model cannot be used to design a compound on a computer, but it can be used to determine how well various protocols would work on a variety of possible compounds or a variety of possible figures of merit. A random-energy model, in particular, can mimic the generic features of an experimental figure of merit. For example, the NK model is used to model combinatorial chemistry experiments on peptides [2], the block NK [3] and generalized NK [4] models are used to model protein molecular evolution experiments, the Random Phase Volume Model is used to model materials discovery [5], and a random-energy model is used to model small molecule design [6]. These random-energy models are rather subtle aspects of statistical mechanics. The reader interested mainly in the Monte Carlo protocols themselves may skip the subsections that describe the details of these models, simply assuming that the figure of merit has been measured experimentally.

This chapter explores the Monte Carlo approach to library design and redesign in HTE of both materials and molecular systems, summarizing our recent work [5,6,7]. The task of library design for materials discovery is addressed first in Section 12.2. The space of variables is identified, and how to search this space by a Monte Carlo method is discussed. A Random Phase Volume Model is introduced as a surrogate for the experimental screen, and the effectiveness of the Monte Carlo protocols is judged. Attention is then turned to library design for small molecule discovery in Section 12.3. The space of variables is identified, and a random-energy model for the experimental screen is introduced. Several Monte Carlo strategies are described, and their performance on both figure of merit and diversity is compared with genetic algorithms. Finally, the interesting case of templated materials synthesis, where there are both molecular and solid-state variables, is considered in Section 12.4. Concepts from the rest of the chapter are brought together to show how Monte Carlo methods can be applied to this final case of library design and redesign. Some final thoughts are offered in Section 12.5.

12.2 MATERIALS DISCOVERY

The goal of high throughput materials discovery is to find compositions of matter that maximize a specific material property [8,9,10]. The property may be, for example, superconductivity [11], magnetoresistance [12], luminescence [13,14,15], ligand specificity [16], sensor response [17], or catalytic activity [9,18,19,20,21,22,23]. The task of finding desirable materials can be reformulated as one of searching a multidimensional space, with the material composition, impurity levels, and synthesis conditions as variables. The property to be optimized, the figure of merit, is generally an unknown function of the variables. Indeed, if how the figure of merit varied with composition were known, desirable materials could be determined purely by theoretical means. Generally this will not be the case, and the figure of merit will be measured only experimentally. It is under these conditions that HTE is useful.

Approaches to high throughput materials discovery to date almost always perform a grid search in composition space, followed by a gradient-type optimization of the figure of merit. This approach becomes inefficient in high-dimensional spaces or when the figure of merit does not vary smoothly with the variables, and use of the grid search has limited most current combinatorial chemistry experiments to ternary or quaternary compositions.

12.2.1 The Space of Variables

There is a large chemical and experimental space to search when seeking the material with the optimal figure of merit. Material composition is clearly a variable. In addition, there is a variety of noncomposition variables. Film thickness [24] and deposition method [25], for example, are variables for materials made in thin-film form. The processing history, such as pH, pressure, temperature, and atmospheric composition, is a variable. The impurity levels or guest compositions can greatly affect the figure of merit [23]. The crystal habit or "crystallinity" of the material can affect the observed figure of merit [24]. Finally, the method of synthesis or nucleation may affect the phase or morphology of the material and thereby affect the figure of merit [26].

There are several mechanisms by which these variables can affect the figure of merit, often in unexpected ways. First, a small impurity composition can cause a large change in the figure of merit, as seen by the rapid variation of activity with Cu content in a Cu/Rh oxidation catalyst [23]. Second, the phases formed in a thin-film synthesis are not necessarily the same as those formed in bulk, as seen in the case of a thin-film dielectric, where the optimal material was found outside the region where the bulk phase forms [24]. Third, the crystallinity of the material can affect the observed figure of merit, again as seen in the thin-film dielectric example [24].

12.2.2 Library Design and Redesign

HTE differs from usual Monte Carlo simulation in that multiple searches of the variable space are simultaneously carried out. That is, in a typical high throughput experiment, many samples, e.g., 10,000, are synthesized and screened for figure of merit at one time. With the results of this first round of experimentation, a new set of samples may then be synthesized and screened. This procedure may be repeated for several rounds, although current materials discovery experiments have not systematically made use of this feature.

Using the analogy with Monte Carlo, each round of combinatorial chemistry corresponds to a move in a Monte Carlo algorithm. Instead of tracking a single system with many configurational degrees of freedom, however, many samples are synthesized and screened, each with several composition and noncomposition variables. Modern robotic technology is what allows for the cost-effective synthesis and screening of these multiple-sample compositions.

The development of robotic technology for materials discovery continues, and future progress can and should be influenced by theoretical considerations. In this spirit, the composition and noncomposition variables of each sample are assumed to be independently adjustable, as in spatially addressable libraries [8,21]. This flexibility is significant, because it allows great latitude in how the variable space can be searched with a limited experimental budget. In fact, the term *high throughput experimentation* is to be preferred to *combinatorial chemistry* when spatially addressable libraries are used. Any constraints that exist on how the variables can be changed in an experiment can easily be accommodated in a Monte Carlo protocol, so the approach is generally applicable.

Current materials discovery experiments uniformly tend to perform a grid search on the composition and noncomposition variables. As will be shown in this chapter, however, it is preferable to choose the variables randomly and statistically from the allowed values. It is also possible to consider choosing the variables so as to maximize the amount of information gained from the limited number of samples screened, via a quasi-random, low-discrepancy sequence [27,28]. These sequences attempt to eliminate the gaps and the redundancy that naturally occur when a space is searched randomly, and they have several desirable theoretical properties. Figure 12.1 depicts these three approaches to materials discovery library design.

As a high throughput experiment proceeds, information is gathered about the figure-of-merit landscape, and this information can be incorporated by multiple rounds of screening. The Monte Carlo protocol provides one convenient method to

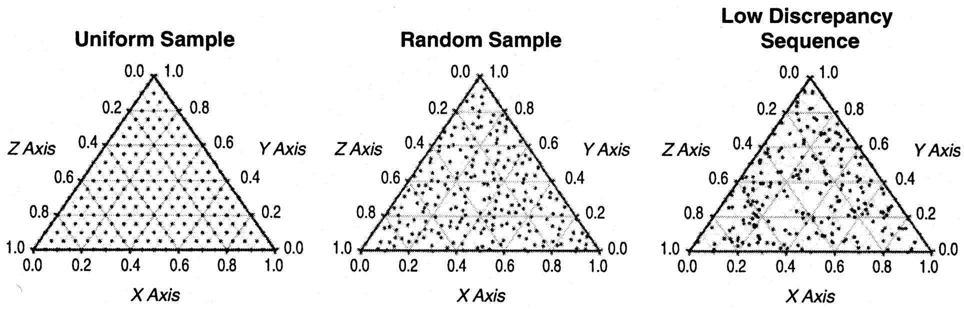

Figure 12.1 The grid, random, and low-discrepancy sequence protocols for designing the first library in a materials discovery experiment with three compositional variables. The random approach is freed from the regular pattern of the grid search , and the low-discrepancy sequence approach avoids gaps and overlapping points that may arise in the random approach. (Adapted, with permission from Deem, M. W., *Advances in Chemical Engineering,* Vol. 28, A. Chakraborty, Ed., 81–121; copyright © 2001. Academic Press, San Diego.)

incorporate this feedback in multiple rounds of experimentation. This approach leads to sampling the experimental figure of merit, E, proportional to $\exp(-\beta E)$. If the control parameter β is large, the Monte Carlo procedure will seek out values of the composition and noncomposition variables that optimize the figure of merit. If β is too large, however, the procedure will become trapped in relatively low-lying local optima. The first round is initiated by choosing the composition and noncomposition variables either from a grid search or statistically from the allowed values. The variables are changed in succeeding rounds as dictated by the Monte Carlo procedure.

12.2.3 Searching the Variable Space by a Monte Carlo Protocol

The Monte Carlo approach to HTE redesigns the library for the next round based upon the results of the screen of the library for the current round. In particular, the composition and noncomposition variables of the current round are perturbed in some fashion to produce the samples to be examined in the next round. Two ways of changing the variables are considered, either a random change of the variables of a randomly chosen sample or a swap of a subset of the variables between two randomly chosen samples. Swapping is productive when there is a hierarchical structure to the variables. The swapping event allows for the combination of beneficial subsets of variables between different samples. Swapping is, in fact, the same as the crossover move from genetic algorithms. A swap move might, for example, combine a good set of composition variables with a particularly good impurity composition. Alternatively, a good set of processing variables might be combined with a good set of composition variables. These moves are performed on the samples until all the samples have been modified. The new samples are then screened for figure of merit. Whether the newly proposed samples will be accepted or the current samples will be kept for the next round is decided according to the detailed balance acceptance criterion. For a random change of one sample, the Metropolis acceptance probability is applied:

$$p_{\text{acc}}(c \to p) = \min\{1, \exp[-\beta(E_{\text{proposed}} - E_{\text{current}})]\}. \tag{12.1}$$

Proposed samples that increase the figure of merit are always accepted. Proposed samples that decrease the figure of merit are accepted with the Metropolis probability. By allowing the figure of merit to decrease occasionally, the protocol is able to escape from local optima. Periodic boundary conditions are used on the noncomposition variables. As discussed in Appendix 12.B, reflecting boundary conditions are used on the composition variables.

A similar acceptance criterion is applied when the swapping move is used. In particular, if a swap is made between samples i and j, the acceptance probability is

$$\begin{aligned} p_{\text{acc}}(c \to p) = \min\{1, \exp[-\beta(E^{i}_{\text{proposed}} + E^{j}_{\text{proposed}} \\ - E^{i}_{\text{current}} - E^{j}_{\text{current}})]\}. \end{aligned} \tag{12.2}$$

One round of the Monte Carlo procedure is shown in Figure 12.2a. The parameter β is not related to the real temperature of the system in the experiment and should be optimized for best efficiency. The characteristic sizes of the random changes in the composition and noncomposition variables are also parameters that should be optimized.

When the number of composition and noncomposition variables is too large, or when the figure of merit changes too roughly with the variables, normal Monte Carlo will not succeed in achieving effective sampling. Parallel tempering is a Monte Carlo method that is used to study statistical [29], spin glass [30,31], and molecular [32] systems with rough energy landscapes. The most powerful protocol for materials discovery incorporates parallel tempering for changing the system variables. In the parallel tempering approach, some of the samples are updated by Monte Carlo with parameter β_1, some by Monte Carlo with parameter β_2, and so on. After a round of synthesis and screening, samples are randomly exchanged between groups with different β's, as illustrated in Figure 12.2b. The acceptance probability for such an exchange of two samples is

$$p_{\mathrm{acc}}(c \rightarrow p) = \min\{1, \exp[\Delta\beta\Delta E]\}. \tag{12.3}$$

Here $\Delta\beta = \beta_i - \beta_j$, and ΔE is the difference, before the exchange is made, of the figures of merit between the sample in group i and the sample in group j. This parallel tempering exchange step involves no extra screening, and is, therefore, free in terms

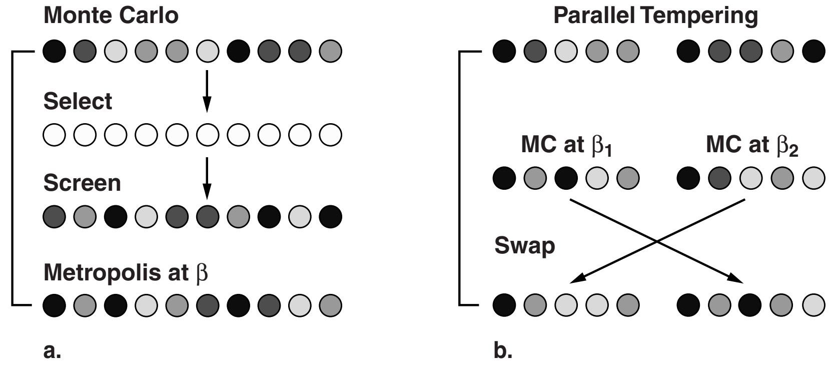

Figure 12.2 Schematic of the Monte Carlo library design and redesign strategy. (a) One Monte Carlo round with 10 samples. Shown are an initial set of samples, modification of the samples, and the Metropolis criterion for acceptance or rejection of the new samples. (b) One parallel tempering round with 5 samples at β_1 and 5 samples at β_2. In parallel tempering, several Monte Carlo simulations are performed at different temperatures, with the additional possibility of sample exchange between the simulations at different temperatures. (Reproduced, with permission, from Falcioni, M., and Deem, M.W., *Physics Review E,* 2000, 61: 5948–5952; copyright © 2000. American Physical Society, College Park, MD.)

of experimental costs. This step, however, can be, effective at helping the protocol to escape from local maxima. The number of different systems and the temperatures of each system are parameters that must be optimized for best experimental performance.

12.2.4 The Random Phase Volume Model

The effectiveness of these protocols is demonstrated by combinatorial chemistry experiments as simulated by the Random Phase Volume Model. The Random Phase Volume Model is not fundamental to the protocols; it is introduced as a simple way to test, parameterize, and validate the various searching methods. The model relates the figure of merit to the composition and noncomposition variables in a statistical way. The model is fast enough to allow for validation of the proposed searching methods on an enormous number of samples, yet possesses the correct statistics for the figure-of-merit landscape. The d-dimensional vector of composition mole fractions is denoted by $\mathbf{x}$. Composition mole fractions are nonnegative and sum to unity; so the allowed compositions are constrained to lie within a simplex in $d - 1$ dimensions. For the familiar ternary system, this simplex is an equilateral triangle. The composition variables are grouped into phases centered around N_x points $\mathbf{x}_\alpha$ randomly placed within the allowed composition range (the phases form a Voronoi diagram [33]; see Figure 12.3). The model is defined for any number of composition variables, and the number of phase points is defined by requiring the average spacing between phase points to be $\xi = 0.25$. To avoid edge effects, additional points are added in a belt of width 2 ξ around the simplex of allowed compositions.

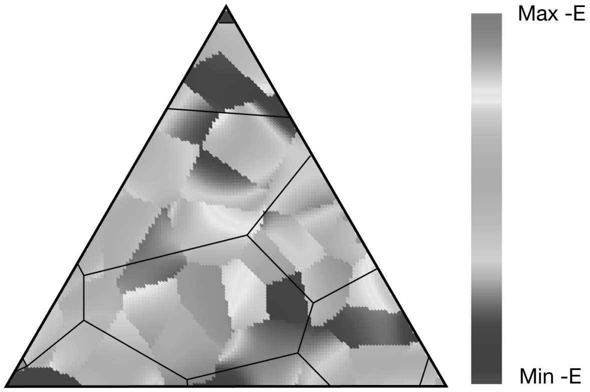

Figure 12.3 The Random Phase Volume Model. The model is shown for the case of three composition variables and one noncomposition variable. The boundaries of the **x** phases are evident by the sharp discontinuities in the figure of merit, *-E*. To generate this figure, the **z** variable was held constant. The boundaries of the **z** phases are shown as thin dark lines. (Reproduced, with permission, from Falcioni, M., and Deem, M. W., *Physics Review E,* 2000, 61: 5948–5952; copyright © 2000. American Physical Society, College Park, MD.)

The figure of merit should change dramatically between composition phases. Moreover, within each phase, α, the figure of merit should also vary with $\mathbf{y} = \mathbf{x} - \mathbf{x}_\alpha$ due to crystallinity effects such as crystallite size, intergrowths, defects, and faulting [24]. In addition, the noncomposition variables should also affect the measured figure of merit. The noncomposition variables are denoted by the b-dimensional vector $\mathbf{z}$, with each component scaled so as to fall within the range $[-1,1]$ without loss of generality. There can be any number of noncomposition variables. The figure of merit depends on the composition and noncomposition variables in a correlated fashion, and so the noncomposition variables also fall within N_z "z-phases" defined in the space of composition variables. There are a factor of 10 fewer noncomposition phases than composition phases. The functional form of the model when $\mathbf{x}$ is in composition phase α and noncomposition phase γ is

$$E(\mathbf{x},\mathbf{z}) = U_\alpha + \sigma_x \sum_{k=1}^{q} \sum_{\substack{i_1+\cdots+i_d=k \\ i_1,\ldots,i_d \geq 0}} f_{i_1\ldots i_d;k} \xi_x^{-k} A^{(\alpha k)}_{i_1,\ldots,i_d} y_1^{i_1},\ldots,y_d^{i_d}$$

$$+ \frac{1}{2}\left(W_\gamma + \sigma_z \sum_{k=1}^{q} \sum_{\substack{i_1+\cdots+i_b=k \\ i_1,\ldots,i_b \geq 0}} f_{i_1,\ldots,i_b;k} \xi_z^{-k} B^{(\gamma k)}_{i_1,\ldots,i_b} z_1^{i_1},\ldots,z_b^{i_b} \right), \quad (12.4)$$

where $f_{i_1,\ldots,i_n;k}$ is a constant symmetry factor, ξ_x and ξ_z are constant scale factors, and U_α, W_γ, $A^{(\alpha k)}_{i_1,\ldots,i_d}$, and $B^{(\gamma k)}_{i_1,\ldots,i_b}$ are random Gaussian variables with zero mean and unit variance. In more detail, the symmetry factor is given by

$$f_{i_1,\ldots,i_n;k} = \frac{k!}{i_1!,\ldots,i_n!}. \quad (12.5)$$

The scale factors are chosen so that each term in the multinomial contributes roughly the same amount: $\xi_x = \xi/2$ and $\xi_z = (\langle z^6\rangle/\langle z^2\rangle)^{1/4} = (3/7)^{1/4}$. The σ_x and σ_z are chosen so that the multinomial, crystallinity terms contribute 40% as much as the constant, phase terms in the root-mean-square sense. For both multinomials $q = 6$. As Figure 12.3 shows, the Random Phase Volume Model describes a rugged figure of merit landscape, with subtle variations, local maxima, and discontinuous boundaries.

12.2.5 Several Monte Carlo Protocols

Six different search protocols are tested with increasing numbers of composition and noncomposition variables. The total number of samples whose figure of merit will be measured is fixed at $M = 100{,}000$, so that all protocols have the same experimental cost. The single-pass protocols grid, random, and low-discrepancy sequence (LDS) are considered. For the grid method, the constants $M_x = M^{(d-1)/(d-1+b)}$ and $M_z = M^{b/(d-1+b)}$ are defined. The grid spacing of the composition variables is $\zeta_x = (V_d/M_x)^{1/(d-1)}$, where

$$V_d = \frac{\sqrt{d}}{(d-1)!} \quad (12.6)$$

is the volume of the allowed composition simplex. Note that the distance from the centroid of the simplex to the closest point on the boundary of the simplex is

$$R_d = \frac{1}{[d(d-1)]^{1/2}}. \qquad (12.7)$$

The spacing for each component of the noncomposition variables is $\zeta_z = 2/M_z^{1/b}$. For the LDS method, different quasi-random sequences are used for the composition and noncomposition variables. The feedback Monte Carlo protocols, Monte Carlo with swap, and parallel tempering are considered. The Monte Carlo parameters were optimized on test cases. It was optimal to perform 100 rounds of 1000 samples with $\beta = 2$ for $d = 3$ and $\beta = 1$ for $d = 4$ or 5, and $\Delta x = 0.1\ R_d$ and $\Delta z = 0.12$ for the maximum random displacement in each component. The swapping move consisted of an attempt to swap all of the noncomposition values between the two chosen samples, and it was optimal to use $P_{\text{swap}} \simeq 0.1$ for the probability of a swap versus a regular random displacement. For parallel tempering it was optimal to perform 100 rounds with 1000 samples, divided into three subsets: 50 samples at $\beta_1 = 50$, 500 samples at $\beta_2 = 10$, and 450 samples at $\beta_3 = 1$. The 50 samples at large β essentially perform a "steepest-ascent" optimization and have smaller $\Delta x = 0.01\ R_d$ and $\Delta z = 0.012$.

12.2.6 Effectiveness of the Monte Carlo Strategies

The figures of merit found by the protocols are shown in Figure 12.4. The random and LDS protocols find better solutions than does grid in one round of experiment. More importantly, the Monte Carlo methods have a tremendous advantage over one pass methods, especially as the number of variables increases, with parallel tempering the best method. The Monte Carlo methods, in essence, gather more information about how best to search the variable space with each succeeding round. This feedback mechanism proves to be effective even for the relatively small total sample size of 100,000 considered here. The advantage of the Monte Carlo methods will become even greater for larger sample sizes. Note that in cases such as catalytic activity, sensor response, or ligand specificity [4] the physical property depends on the exponential of the energy. In these cases, the experimental figure of merit would be exponential in the values shown in Figure 12.4, so that the success of the Monte Carlo methods would be even more dramatic.

The question of how to design and redesign materials discovery experiments has been addressed by analogy with Monte Carlo computer simulation. The Random Phase Volume Model has been used to compare various strategies. The multiple-round Monte Carlo protocols are found to be especially effective on the more difficult systems with larger numbers of composition and noncomposition variables.

An efficient implementation of the search strategy is feasible with existing library-creation technology. Moreover "closing the loop" between library design and redesign is achievable with the same database technology currently used to track and record the data from combinatorial chemistry experiments. These multiple-

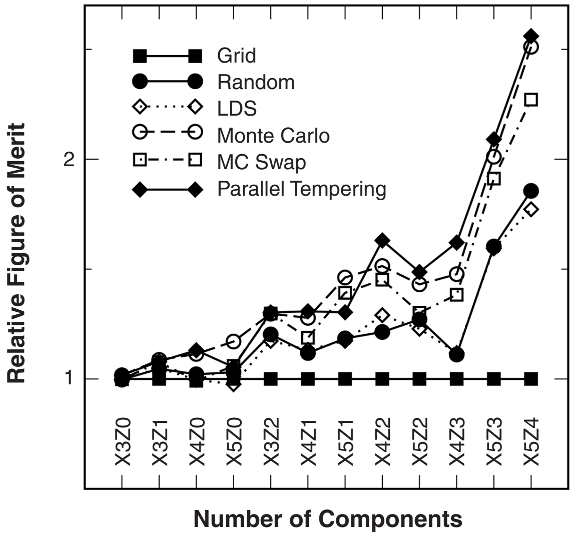

Figure 12.4 The maximum figure of merit found with different protocols on systems with a different number of composition (**x**) and noncomposition (**z**) variables. The results are scaled to the maximum found by the grid searching method. Each value is averaged over scaled results on 10 different instances of the Random Phase Volume Model with different random phases. The Monte Carlo methods are especially effective on the systems with a larger number of variables, where the maximal figures of merit are more difficult to locate. (Reproduced, with permission, from Falcioni, M., and Deem, M. W., *Physics Review E,* 2000, 61: 5948–5952; copyright © 2000. American Physical Society, College Park, MD.)

round protocols, when combined with appropriate robotic controls, should allow the practical application of high throughput experimentation to more complex and interesting systems.

12.3 SMALL-MOLECULE-DISCOVERY[1]

High throughput synthesis is established as one of the methods for the discovery of small molecules, such as drugs or homogeneous catalysts. High throughput or combinatorial methods allow for simultaneous creation of a large number of structurally diverse and complex compounds [34–36], generalizing the traditional techniques of single-compound synthesis. Parallel synthesis and split/pool synthesis [37,38] on solid phase, for example, are two commonly used methods for combinatorial synthesis. Among the high-throughput methodologies, small-molecule combinatorial

[1] Portions of this section are reprinted with permission from Chen, L., and Deem, M. W., *Journal of Chemical Information and Computer Sciences,* 2001, 41: 950–957. Copyright © 2001. American Chemical Society, Washington, DC.

chemistry is the most developed and has been applied successfully in areas such as transition metal complexation [39], chemical genetic screening [40], catalysis [41], and drug discovery [42].

The parallel synthesis and split/pool synthesis methods search the composition space in a regular, gridlike fashion. As the complexity of the molecular library grows, the number of dimensions in the composition variable space grows, and with a gridlike method, the number of compounds that must be synthesized to search the space grows exponentially. Synthesis and screening of mixtures of compounds can partially alleviate the dimensional curse [43]. However, a mixture approach raises the question of how to deconvolute and interpret the results. The greater the degree of mixing, the stronger the synergistic effects can be in the mixture, and the more difficult it is to identify individual compounds responsible for the activity [44].

The challenge of searching the composition space in an efficient way has led to extensive efforts in the rational design of combinatorial, or high throughput, libraries. A basic assumption in library design is that structurally similar compounds tend to display similar activity profiles. By designing libraries with maximum structural diversity, the potential for finding active compounds in the high throughput screenings can be enhanced. This design approach requires a quantitative account of the structural and functional diversity of the library, and many descriptors have been developed [45]. Optimization of a library to maximum diversity is then driven by a reliable statistical method. Several structurally diverse libraries have been successfully designed along these lines [46,47]. For example, strategies have been presented to optimize the structural diversity of libraries of potential ligands and molecules by using stochastic optimization of diversity functions and a point mutation Monte Carlo technique [48]. Peptide libraries have been designed by using topological descriptors and quantitative structure–activity relationships combined with a genetic algorithm and simulated annealing [49–51]. Diverse libraries of synthetic biodegradable polymers have been designed by using molecular topology descriptors and a genetic algorithm [52]. Similarly, peptoid libraries have been designed by using multivariate quantitative structure–activity relationships and statistical experimental design [53].

The question of how an initial library should be redesigned for subsequent rounds of HTE in light of the results of the first round of screening has remained largely unanswered. In this chapter, it is suggested that Monte Carlo methods provide a natural means for library redesign in HTE. There is a striking analogy between searching configuration space for regions of low free energy in a Monte Carlo simulation and searching composition space for regions of high figure of merit in a high throughput experiment. Importantly, Monte Carlo methods do not suffer the curse of dimensionality. A Monte Carlo approach should, therefore, be exponentially more efficient than a regular, gridlike method for libraries of complex molecules.

GAs are the computational analog of Darwinian evolution. Typically, a GA consists of three basic processes: crossover, mutation, and selection. In the crossover step, new compounds are generated by mixing the compositions of parent compounds. In the mutation step, individual molecules are changed at random. In the selection step, the best molecules are identified for the next round. The application of GAs to combinatorial synthesis and library design has achieved considerable success [54–57]. Nonetheless, unlike Monte Carlo algorithms, GAs do not satisfy detailed balance. Because of this, GAs cannot be guaranteed to sample the variable space in a defined statistical way or to locate optimal molecules. Furthermore, in most experiments, one wants to identify several initially promising molecules in the hope that, among them, a few can survive further stringent screenings, such as patentability or lack of side effects [58]. In the genetic approach, however, all the molecules in the library tend to become similar to each other due to the crossover step. While diversity can be encouraged in a genetic approach [59,60], diversity can never be guaranteed. The Monte Carlo approach, on the other hand, can maintain or even increase the diversity of a molecular library, due to the satisfaction of detailed balance.

Several strategies for small-molecule HTE will be derived by analogy with Monte Carlo methods. These Monte Carlo protocols will be compared to the GA approach. In order to make this comparison and to demonstrate the effectiveness of the Monte Carlo approach, simulated high-throughput experiments are performed. A random-energy model is introduced and used as a surrogate for experimental measurement of the figure of merit. The random-energy model is not fundamental to the protocols; it is introduced as a simple way to test, parameterize, and validate the various searching methods. In an experimental implementation, the random-energy model would be replaced by the value returned by the screen. Various Monte Carlo protocols are introduced, and a means to calculate the diversity of a library is provided. The effectiveness of the protocols is gauged, and some discussion of the results is presented.

12.3.1 The Space of Chemical Variables

The molecules in a high throughput library are uniquely characterized by their composition, such as the identity of the template and ligands. For specificity, the figure of merit of interest will be considered to be a binding constant, but the results are generically valid. A schematic view of the model is presented in Figure 12.5. For simplicity, the small molecule is considered to consist of one template, drawn from a library of templates, and six binding ligands, each drawn from a single library of ligands [61]. Numerous energetic interactions could exist between this molecule and the substrate. It is commonly believed that descriptors can be directly related to compound performance. A large class of descriptors, such as one-dimensional, two-dimensional, three-dimensional, and BCUT descriptors, has been used to measure the diversity between ligands, templates, and molecules in the literature [45–47]. To simplify, consideration is limited to a set of six weakly correlated descriptors for each ligand and template. For example, the descriptors could be hydrogen-bond donors, hydrogen-bond acceptors, flexibility, an electrotopological calculation, clog*P*, and aromatic density [47]. These descriptors are needed only for the definition of the random-energy model. In a real experiment, the figure of merit would be measured directly, and the descriptors would likely not be needed.

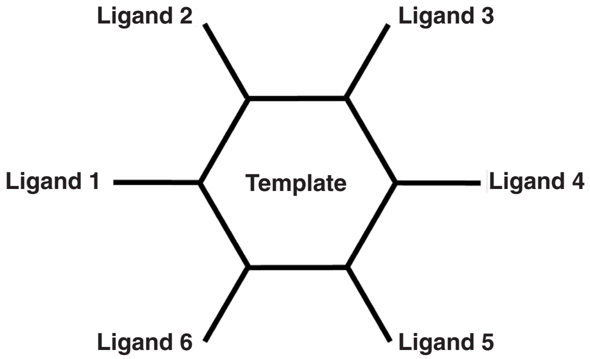

Figure 12.5 Schematic view of the small molecule model. (After [6].)

12.3.2 A Random-Energy Model for the Experimental Screen

To carry out simulated experiments, a figure-of-merit function that mimics the experimental step of measuring the figure of merit is needed. Once constructed or synthesized, the molecules are scored by the model, which takes the composition or molecular descriptors as input.

The basic building block for the random-energy model is a random polynomial of n descriptors, $w_1, \ldots, w_n$:

$$F(w_1, \ldots, w_n, \{G\}) = \sum_{k=0}^{q} \sum_{\substack{i_1+\cdots+i_n=k \\ i_1, \ldots, i_n \geq 0}} f^{1/2}_{i_1, \ldots, i_n; k} \xi^{-k} G_{i_1, \ldots, i_n} w_1^{i_1}, \ldots, w_n^{i_n} \quad (12.8)$$

Here q is the degree of the polynomial, $G_{i_1, \ldots, i_n}$ are the fixed coefficients of the polynomial, and $f_{i_1, \ldots, i_n; k}$ are constant symmetry factors given by equation (12.5). The square root of f is taken here because each term of the unsymmetrized polynomial is considered to be random. The scale factor, ξ, is used to equalize roughly the contributions from each term of the polynomial. Since w_i will be drawn from a Gaussian random distribution of zero mean and unit variance, the value

$$\xi = \left(\frac{\langle x^q \rangle}{\langle x^2 \rangle} \right)^{1/(q-2)} = \left(\frac{q!}{(q/2)!2^{q/2}} \right)^{1/(q-2)} \quad (12.9)$$

is appropriate. The values $q = 6$ and $n = 6$ are used in the random-energy model.

The random-energy model accounts for contributions to substrate binding arising from interactions between the substrate and the template and from interactions between the substrate and each of the ligands. In addition, synergistic effects between the ligands and template are incorporated. Consider, for example, a molecule made from template number m from the template library and ligands number $l_1, \ldots, l_6$ from the ligand library. The template is characterized by six descriptors, $D_1^{(m)}, \ldots, D_6^{(m)}$. Similarly, each ligand is characterized by six descriptors, $d_1^{(li)}, \ldots, d_6^{(li)}$. The template contribution to binding is denoted by E_T and the ligand contributions by E_L. The contribution due to synergistic ligand–ligand interactions is denoted by E_{LL} and the contribution due to synergistic template–ligand interactions by E_{TL}. The total contribution to the figure of merit is, then,

$$E = E_L + E_T + E_{LL} + E_{TL}. \tag{12.10}$$

Each of these factors is given in terms of the random polynomial:

$$E_L = \alpha_1 \sum_{i=1}^{6} F(d_1^{(li)}, \ldots, d_6^{(li)}, \{G_L\}) \tag{12.11}$$

$$E_T = \alpha_2 F(D_1^{(m)}, \ldots, D_6^{(m)}, \{G_T\}) \tag{12.12}$$

$$E_{LL} = \alpha_3 \sum_{i=1}^{6} h_i F(d_{j1}^{(li)}, d_{j2}^{(li)}, d_{j3}^{(li)}, d_{j4}^{(li+1)}, d_{j5}^{(li+1)}, d_{j6}^{(li+1)}, \{G_{LL}^i\}) \tag{12.13}$$

$$E_{TL} = \alpha_4 \sum_{i=1}^{6} h_i\, F(d_{k_1}^{(li)}, d_{k_2}^{(li)}, d_{k_3}^{(li)}, D_{k_4}^{(m)}, D_{k_5}^{(m)}, D_{k_6}^{(m)}, \{G_{TL}^i\}), \tag{12.14}$$

where the $\{G_L\}$, $\{G_T\}$, $\{G_{LL}^i\}$, and $\{G_{TL}^i\}$ are sets of fixed random Gaussian variables with zero mean and unit variance. The α_i are constants to be adjusted so that the synergistic terms will contribute in desired percentages, and h_i is a structural constant indicating the strength of the interaction at binding site, i. The interaction strengths, h_i, are chosen from a Gaussian distribution of zero mean and unit variance for each site on each template. Only synergistic interactions between neighboring ligands are considered in E_{LL}, and it is understood that l_7 refers to l_1 in equation (12.13). In principle, the polynomial in equation (12.13) could be a function of all 12 descriptors of both ligands. It is reasonable to assume, however, that important contributions come from interactions among three randomly chosen distinct descriptors of ligand l_i, $d_{j_1}^{(li)}$, $d_{j_2}^{(li)}$, and $d_{j_3}^{(li)}$, and another three randomly chosen distinct descriptors of ligand l_{i+1}, $d_{j_4}^{(li+1)}$, $d_{j_5}^{(li+1)}$, and $d_{j_6}^{(li+1)}$. Similarly, it is reasonable to assume that template–ligand contributions come from interactions between three randomly chosen distinct descriptors of the ligand, $d_{k_1}^{(li)}$, $d_{k_2}^{(li)}$, and $d_{k_3}^{(li)}$, and another three randomly chosen distinct descriptors of the template, $D_{k_4}^{(m)}$, $D_{k_5}^{(m)}$, and $D_{k_6}^{(m)}$. Both j_i and k_i are descriptor indices ranging from 1 to 6. Assuming that the degrees of freedom of the substrate have been integrated out, these indices depend only on the template.

The parameters in the random-energy model are chosen to mimic the complicated interactions between a small molecule and a substrate. Focus is directed to the case where these interactions are unpredictable, which is typical. That is, in a typical experiment, it would not be possible to predict the value of the screen in terms of molecular descriptors. Indeed, when rational design fails, an intelligent use of HTE is called for. This task of library design and redesign, rather than single-molecule design, is addressed in the next subsection.

12.3.3 Monte Carlo Strategies that Include A Priori Knowledge

To initiate the Monte Carlo protocol, the first template and ligand libraries are built. The size of the template library is denoted by N_T and the size of the ligand library by N_L. In a real experiment, in all likelihood, the descriptors for each ligand and template would not even be calculated, as the figure of merit would be directly measured in the screen. In the simulated experiment, however, the model figure of merit is a function of the descriptors, and the values for the six descriptors for each ligand and template are taken from a random Gaussian distribution with zero mean and unit variance. In the simulated experiment, two sets of random interaction descriptor indices are also associated to each template for the interaction terms in equations (12.13) and (12.14).

To give a baseline for comparison, the library is first designed using a random construction. New molecules are constructed by random selection of one template and six ligands from the libraries. Since the properties of each ligand and template are assigned randomly, this first library should be reasonably diverse and comparable to examples in the literature.

For the Monte Carlo schemes, the initial molecular configurations are still assigned randomly. The library is modified in subsequent rounds of HTE, however, by the Monte Carlo protocol. Two kinds of move are possible for each molecule in the library: template changes and ligand changes. Either the template is changed with probability p_{template}, or one of the six ligands is picked randomly to change with probability $1 - p_{\text{template}}$. The probability of changing from template m to m' is denoted by by $T(m \rightarrow m')$ and from ligand i to i' by $t(i \rightarrow i')$. The new configurations are updated according to the acceptance rule at β, the inverse of the protocol temperature. All the samples are sequentially updated in one Monte Carlo round.

For the simple Metropolis method, the transition matrices are

$$T(m \rightarrow m') = 1/N_T \tag{12.15}$$

$$t(i \rightarrow i') = 1/N_L, \tag{12.17}$$

and the acceptance rule is

$$p_{\text{acc}}(c \rightarrow p) = \min[1, \exp(-\beta \Delta E)]. \tag{12.17}$$

To make use of the idea that smaller moves are accepted more often, one could try to choose a modified ligand or template that is similar to the current one, that is, one could use a transition matrix weighted toward the proposed ligands or templates close to the current one in the six-dimensional descriptor space. Interestingly, this refinement turns out not to work any better than does the simple random move. It seems that even a small move in the descriptor space is already much larger than the typical distance between peaks on the figure-of-merit landscape.

Biased Monte Carlo methods have been shown to improve the sampling of complex molecular systems by many orders of magnitude [62]. In contrast to conventional Metropolis Monte Carlo, trial moves in biased schemes are no longer chosen completely at random. By generating trial configurations with a probability that depends on a priori knowledge, the moves are more likely to be favorable and more likely to be accepted. Since small-molecule discovery deals with a discrete configurational space, the implementation of biased Monte Carlo in this case is relatively simple. First, a biasing term is needed for both ligands and templates. Since the form of this term is not unique, one can proceed in several different ways. One strategy is to bias the choice of template and ligand on the individual contributions of the templates and ligands to the figure of merit. One might know, or be able to estimate, these contributions from theory. For the random-energy model, for example,

$$e^{(i)} = \alpha_1 F(d_1^{(i)}, \ldots, d_6^{(i)}, \{G_L\}) \tag{12.18}$$

$$E^{(m)} = \alpha_2 F(D_1^{(m)}, \ldots, D_6^{(m)}, \{G_T\}), \tag{12.19}$$

where $e^{(i)}$ is the bias energy to the ligand i in the library, and $E^{(m)}$ is the bias energy of template m in the library. Alternatively, one can estimate the contribution of each ligand or template to the figure of merit experimentally [63–65]. An electrospray ionization source coupled to a mass spectrometer, for example, can serve this purpose [66]. To measure the contributions experimentally, a preexperiment is performed on 10,000 randomly constructed molecules. This number of compounds will give on average each ligand 60 hits and each template 667 hits. By averaging the figure of merit of the molecules containing a particular ligand or template over the total number of hits, experimental estimates of $e^{(i)}$ and $E^{(m)}$ can be obtained. Using these two methods of bias, three different types of biased Monte Carlo schemes are considered: theoretical biased move, experimental biased move, and mixed biased move. In theoretical bias, both $e^{(i)}$ and $E^{(m)}$ are from the random-energy model. In experimental bias, both $e^{(i)}$ and $E^{(m)}$ are calculated from the preexperiment. In mixed bias, $e^{(i)}$ comes from the random-energy model, while $E^{(m)}$ comes from the preexperiment.

These biases tend to exhibit a large gap between a few dominant templates and ligands and the rest. To ensure the participation of more ligands and templates in the strategy, cutoff energies are introduced for the ligand and template, e_c and E_c. These cutoffs are chosen so that e_c is the 21st lowest ligand energy and E_c is the fourth lowest template energy. The biased energy, $e_b^{(i)}$, for the ith ligand is

$$e_b^{(i)} = \begin{cases} e^{(i)} & \text{if } e^{(i)} > e_c \\ e_c & \text{otherwise,} \end{cases} \tag{12.20}$$

and the biased energy, $E_b^{(m)}$, for the mth template is

$$E_b^{(m)} = \begin{cases} E^{(m)} & \text{if } E^{(m)} > E_c \\ E_c & \text{otherwise.} \end{cases} \tag{12.21}$$

To correct for this bias, Rosenbluth factors are introduced [62]. Since the transition probabilities are the same at each Monte Carlo step, the Rosenbluth factor for the ligand is constant:

$$w(p) = w(c) = \sum_{i=1}^{N_L} \exp(-\beta e_b^{(i)}). \tag{12.22}$$

The probability of transition from ligand i to i' is

$$t(i \to i') = \frac{\exp(-\beta e_b^{(i')})}{w(p)}. \tag{12.23}$$

In the same way, the Rosenbluth factor for the template is

$$W(p) = W(c) = \sum_{m=1}^{N_T} \exp(-\beta E_b^{(m)}). \tag{12.24}$$

The probability of transition from template m to m' is

$$T(m \to m') = \frac{\exp(-\beta E_b^{(m')})}{W(p)}. \tag{12.25}$$

Finally, the remaining, nonbiased part of the figure of merit is defined to be

$$E_b = E - E_b^{(m)} - \sum_{i=1}^{6} e_b^{(li)}. \tag{12.26}$$

To satisfy the detail balance, the acceptance rule becomes

$$p_{\text{acc}}(c \to p) = \min[1, \exp(-\beta \Delta E_b)]. \tag{12.27}$$

The idea of a swapping move that attempts to exchange fragments between two molecules is also used. The probability of attempting a swap instead of a single-molecule move is denoted by p_{swap}. In a swap move, the templates or a pair of ligands can be swapped between two randomly selected molecules. The probability of switching the template or ligand at the same position is given by p_{swap_T} and p_{swap_L}, respectively. The crossover event from genetic algorithms could also be introduced in the swap moves, but this additional move did not improve the results. The acceptance rule for swapping is $p_{\mathrm{acc}}(c \rightarrow p) = \min[1, \exp(-\beta\Delta E)]$.

Parallel tempering is also used in this redesign approach. The acceptance rule for this move is equation (12.3). The number of groups, the number of samples in each group, the value of β_i, and the exchange probability, p_{ex}, are experimental parameters to be tuned.

For comparison, these Monte Carlo protocols are compared to a standard GA approach [54–57]. In the GA, as in the Monte Carlo strategy, multiple rounds of experimentation are performed on a large set of compounds. The difference between the Monte Carlo and the GA lies in how the library is redesigned, that is, how the compounds are modified in each round. In the GA, first two parents are randomly selected. Then the explicit composition of each molecule is listed, i.e., template, ligand 1, . . . , ligand 6. After aligning the sequences from the two parents, a random cut is made and the part of the sequences before the cut is exchanged. Random changes, or mutations, are also performed on the templates and ligands of the offsprings. Finally, since the population is doubled by crossover, the better half of the molecules is selected to survive this procedure and continue on to the next round.

12.3.4 A Measurement of Library Diversity

The diversity of the library as it passes through the rounds of HTE is an important quantity. The diversity, $\mathcal{D}$, is calculated as the standard deviation of the library in the 42-dimensional descriptor space:

$$\mathcal{D}^2 = \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{j=1}^{6}(D_j^{(m(i))} - \langle D_j\rangle)^2 + \sum_{j=1}^{6}\sum_{k=1}^{6}(d_j^{(sk(i))} - \langle d_j^{(sk)}\rangle)^2\right], \qquad (12.28)$$

where $m(i)$ is the index of the template of molecule i, $s_k(i)$ is the index of ligand k of molecule i, and j is the index for the descriptor. The average value in each descriptor dimension is given by $\langle D_j\rangle = N^{-1}\Sigma_{i=1}^{N} D_j^{(m(i))}$ and $\langle d_j^{(sk)}\rangle = N^{-1}\Sigma_{i=1}^{N} d_j^{(sk(i))}$. The diversity of the library will change as the library changes. A larger library will generally possess a higher absolute diversity simply due to the increased number of compounds. This important, but trivial, contribution to the diversity is scaled out by the factor of $1/N$ in equation (12.28).

12.3.5 Effectiveness of the Monte Carlo Strategies

To gauge how the synergistic terms in the figure of merit affect the efficiency of the Monte Carlo protocols, three models with increasingly important synergistic effects are considered. This is done by adjusting the α_i in equations (12.11)–(12.14), so that the root-mean-square values of the terms are in the ratio $E_L : E_T : E_{LL} : E_{TL}$ = 1 : 1 : 0.5 : 0.3 in model I, 1 : 1 : 1 : 0.6 in model II, and 1 : 1 : 2 : 1.2 in model III. Finally, $\alpha_1 = 0.01$. To maintain roughly the same average magnitude of the total energy in each model, $\alpha_1 = 0.01 \times (2.8/3.6) = 0.00778$ in model II, and $\alpha_1 = 0.01 \times (2.8/5.2) = 0.00538$ in model III. It turns out that the results are very similar for all three models, and so the results for the most challenging model III only are presented.

The size of the library is fixed at $N_T = 15$ and $N_L = 1000$. The compositional space of this model has 15×1000^6 distinct molecules. Clearly, it is impossible to search exhaustively even this modestly complex space. The total number of molecules to be synthesized is fixed at 100,000, that is, all protocols will have roughly the same experimental cost. Specifically, 100,000 molecules will be made in the random library design protocol, while in the case of the Monte Carlo or genetic protocols, the number of molecules times the number of simulation rounds is kept fixed at 100,000.

To locate optimal parameters for the protocols, a few short preexperiments were performed. First the energy coefficients in the energy function and the descriptors of the ligand and template libraries were fixed. For simple Metropolis, it is optimal to use 10 samples with 10,000 rounds, suggesting that the system is still far from equilibrium at the random initial configuration. With the biased Monte Carlo method, 100 samples and 1000 rounds is optimal. Focus is restricted on systems with 1000 or 100 rounds, since a minimal number of rounds is typically preferred in experiments. It is more difficult to achieve effective sampling in the system with 100 rounds, and so this system is used when setting optimal parameter values. For parallel tempering, it was optimal to have the samples divided into three subsets, with 30% of the population at β_1, 40% at β_2, and 30% at β_3. The optimal parameters are listed in Table 12.1 for each model. Determination of these parameter values corresponds experimentally to gaining familiarity with the protocol on a new system.

The various Monte Carlo schemes are compared with the random-selection method and the GA. Once the optimal parameters are chosen, the coefficients of the energy function and the descriptor values of the ligand and template libraries are generated differently in each simulated experiment. The simulation results are shown in Figure 12.6. Each data point in the figure is an average over 20 independent runs. This averaging is intended to give a representative performance of the protocols on various figures of merit of experimental interest. Since there is much randomness in the results, the standard deviation of the average is shown as well.

12.3.6 Discussion of the Small Molecule Approach

It is clear that for all systems, the Metropolis methods perform better than does random selection. The system with 1000 molecules and 100 rounds is not well equili-

Table 12.1 Optimal Parameters Used in Simulations for the Three Random-Energy Models

Model	β	p_{template}	p_{swap}	p_{swap_T}	p_{swap_L}	β_1	β_2	β_3	p_{ex}
I	30	0.02	0.1	0.05	0.2	5	30	200	0.1
II	30	0.02	0.2	0.2	0.3	5	30	500	0.1
III	50	0.02	0.4	0.2	0.2	5	50	500	0.1

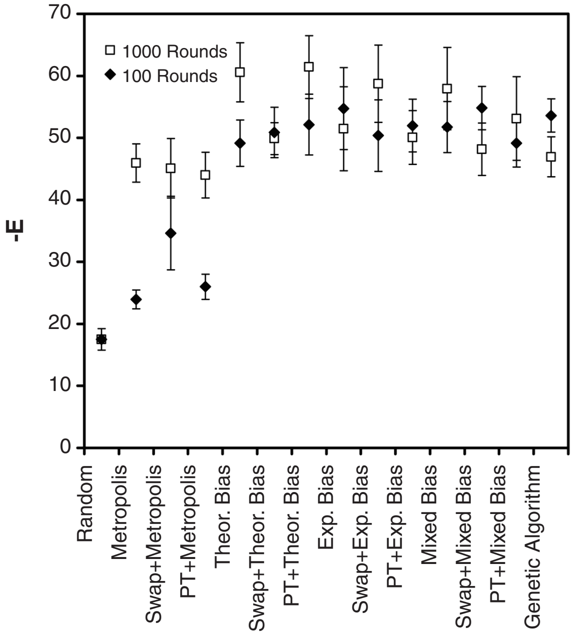

Figure 12.6 Comparison of different Monte Carlo schemes with random and genetic schemes for model III ($E_L : E_T : E_{LL} : E_{TL}$ = 1 : 1 : 1 : 1.2). Data from two cases are shown, one with 1000 molecules and 100 rounds (filled diamonds) and one with 100 molecules and 1000 rounds (unfilled squares). (Reproduced, with permission, from Chen, L. and Deem, M. W. *Journal of Chemical Information and Computer Sciences,* 2001, 41: 950–957; copyright © 2001. American Chemical Society, Washington, DC.)

brated by the Metropolis schemes, and an experiment with 100 molecules and 1000 rounds significantly improves the optimal compounds identified. However, by incorporating a priori knowledge, the biased Monte Carlo schemes are able to equilibrate the experiment with either 1000 or 100 rounds. Interestingly, the theoretical bias and experimental bias methods yield similar results. This result strongly suggests that a minimal number of preexperiments can be very useful, both for the understanding of the structure of the figure-of-merit landscape and for improving the performance in future rounds.

The results produced with the composite moves including swap and parallel tempering are slightly improved relative to those from the plain Monte Carlo schemes. Typically, however, these composite moves significantly improve the sampling of a rough landscape. Indeed, swapping and crossover moves are very effective in protein molecular evolution, where the variable space is extremely large [4]. Perhaps the variable space is not so large in small molecule HTE that these composite moves are required. Alternatively, the random energy model may underestimate the ruggedness of the landscape. The landscape for RNA ligands, for example, is estimated to be extremely rough [66], and composite moves may prove more important in this case.

The GA is relatively easy to use. It does not satisfy detailed balance, however, so there is no theoretical guarantee of the outcome. The optimal figures of merit identified are, nonetheless, comparable to those from the better Monte Carlo methods for all three models. However, due to the crossover and selection steps in the GA, the molecules in the library tend to become similar to each other, which prevents this scheme from sampling the whole variable space. To help elucidate this point, diversity measurements for model I are shown in Figure 12.7. It is clear that the GA has reduced the diversity of the library by 60% relative to the biased schemes. Interestingly, the Monte Carlo simulations actually increase the diversity from the initial random configurations. The biased schemes tend to bring the system to equilibrium relatively quickly, and the diversity measurements are similar for the 100- and

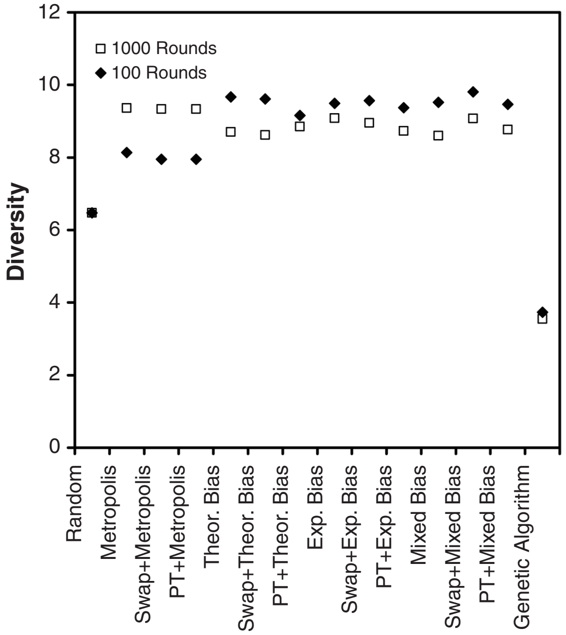

Figure 12.7 Diversity measurement of the final configurations for model I. Data from two cases are shown, one with 1000 molecules and 100 rounds (filled diamonds) and one with 100 molecules and 1000 rounds (unfilled squares). The error bars are negligible. As in equation (12.28), the contribution to the absolute diversity that scales as the square root of the number of molecules per round has been scaled out in this figure. (Reproduced, with permission, from Chen, L. and Deem, M. W. *Journal of Chemical Information and Computer Sciences,* 2001, 41: 950–957; copyright © 2001. American Chemical Society, Washington, DC.)

1000-round experiments. For the Metropolis method, on the other hand, an experiment with 100 rounds is less diverse than an experiment with 1000 rounds. The GA approach finds less favorable figure-of-merit values in the 100-compound 1000-round experiment, presumably due to a greater sensitivity to the $\sqrt{10}$ reduction in the absolute diversity relative to the already small absolute diversity in the 1000-compound 100-round experiment.

The greater the number of potentially favorable molecules in the library space, the greater the diversity of the experimental library will be for the Monte Carlo methods. The GA, on the other hand, will tend to produce a library that contains many copies of a single favorable molecule. A key distinction, then, is that a Monte Carlo strategy will sample many compounds from the figure-of-merit landscape, whereas a GA will tend to produce a single molecule with a favorable figure-of-merit value. How strongly the compounds with high figures of merit are favored in the Monte Carlo strategy is determined by the protocol temperature, since the probability of observing a compound with figure of merit $-E$ is proportional to $\exp(-\beta E)$. The sampling achieved by the Monte Carlo methods is important not only because it ensures that the composition space is thoroughly sampled, but also because it ensures that the library of final hits will be as diverse as possible.

The Monte Carlo methods perform equally well on all three models. The three models were introduced to gauge the impact of unpredictable, synergistic effects in the experimental figure of merit. It might be expected that the a priori bias methods would perform less well as the synergistic effects become more pronounced. That the biased methods perform well even in model III suggests that the bias Monte Carlo approach can be rather robust. In other words, even a limited amount of a priori information is useful in the Monte Carlo approach to library redesign.

Monte Carlo appears to be a fruitful paradigm for experimental design of multiround HTEs. A criticism of HTE has been its mechanical structure and lack of incorporation of a priori knowledge. As shown here, a biased Monte Carlo approach handily allows the incorporation of a priori knowledge. Indeed, the results reveal that biased Monte Carlo schemes greatly improve the chances of locating optimal compounds. For the moderately complex libraries considered here, the bias can be determined equally well by experimental or theoretical means. Although the compounds identified from a traditional GA are comparable to those from the better Monte Carlo schemes, the diversity of identified molecules is dramatically decreased in the genetic approach. GAs, therefore, are less suitable when the list of good molecules is further winnowed by a secondary screen, a tertiary screen, patentability considerations, lack of side effects, or other concerns. Interestingly, composite Monte Carlo moves, such as swap or parallel tempering, bring only a slight improvement to the plain biased Monte Carlo protocols, possibly due to the relatively small size of the composition space in small-molecule HTE. Presumably, as the complexity of the library is increased, these composite moves will prove more useful for the more challenging figures of merit. Although for simplicity the initial library configurations are chosen at random, the sophisticated initial library design strategies available in the literature can be used, and they would complement the multiround library redesign strategies presented here.

12.4 TEMPLATED MATERIALS SYNTHESIS

As a last example of the Monte Carlo approach to library design and redesign, ideas from Sections 12.2 and 12.3 are combined to address the topic of high throughput, templated materials synthesis.

A canonical example of templated materials synthesis is that of zeolite synthesis. Zeolites are widely used in industrial applications such as catalysis, molecular sieving, gas separation, and ion exchange. The catalytic activity and selectivity of these materials are attributed to the highly distributed active sites and the large internal surface area accessible through uniformly sized pores. The present set of zeolite frameworks is incomplete, and additional geometries and chemistry are desired. In addition, even the present zeolite geometries are often imperfect, and there is strong motivation to reduce the defects and faults in these materials. Syntheses of novel zeolites with new frameworks and chemical compositions has therefore received heavy research attention. To date, roughly 150 framework structures are known. Roughly six new structures are added per year. The use of organocation template molecules to induce structure direction is a major tool in the quest for new structure [26,67]. The addition of organic templates such as alkylammonium ions or amines to zeolite synthesis gels affects the rate at which a particular structure is nucleated and grown, and this control can be used to make new structures or framework chemical compositions accessible. Indeed, the nature and extent of interaction between the organic templates and the inorganic components of the zeolite synthesis gel are important factors that influence the final zeolite pore architecture [68,69].

As shown in this chapter, Monte Carlo protocols are efficient methods for library design and redesign in both material discovery and small-molecule design. Material discovery deals with the continuous variables of composition and noncomposition. Small-molecule design deals with the discrete variables of the template and ligand identities. For templated zeolite synthesis, both the continuous variables and the discrete variables are present. All these variables affect the function of the final zeolite material in a correlated way.

Several strategies are described here for templated, high throughput zeolite synthesis. A random energy model is built from those previously introduced. This model will serve as a surrogate for experimental measurement of the figure of merit in simulated experiments. Again, the random energy model is not fundamental to the protocols; it is introduced as a simple way to test, parameterize, and validate the various searching methods. In an experimental implementation, the random-energy model would be replaced by the value returned by the experimental primary screen. Details of the implementation of the Monte Carlo protocols are provided. The effectiveness of the protocols is gauged, and some implications are discussed.

12.4.1 The Random-Energy Model

For demonstration of the proposed protocols in an efficient fashion, a model is necessary in lieu of the real experimental screening process, that is, a model that relates the chemical composition of the zeolite and structure-directing agent to the figure of merit of the material. Again, this model is not essential to the protocols; it is simply a cheap and fast means to evaluate proposed protocols. This model captures the essence of the physical system and provides validation for the protocols.

To consider template-assisted zeolite synthesis, a model that combines ideas from Sections 12.2 and 12.3 is used. The figure of merit is naturally given by a sum of a zeolite energy and an organic molecular energy:

$$E = E_{\text{zeolite}} + E_{\text{molecule}}. \tag{12.29}$$

The d zeolite framework composition variables x_{i} are certainly key variables in E_{zeolite}. Typical elemental compositions in zeolite frameworks include silicon, aluminum, oxygen, phosphorus, germanium, and boron, among others. The Random Phase Volume Model from Section 12.2 is used to capture the dependence of the figure of merit on these composition variables. Noncomposition variables also affect the measured figure of merit, and they are accounted for by the Random Phase Volume Model. When the zeolite is in composition phase α and noncomposition phase γ, the contribution to the figure of merit from the zeolite is given by

$$E_{\text{zeolite}}(\mathbf{x}, \mathbf{z}) = H(\mathbf{x} - \mathbf{x}^{\alpha}, \{G^{\alpha}\}) + \tfrac{1}{2} H(\mathbf{z}, \{G^{\gamma}\}). \tag{12.30}$$

Here $\{G^{\alpha}\}$ and $\{G^{\gamma}\}$ are phase-dependent, Gaussian random variables with zero mean and unit variance. The parameters $\{G^{\alpha}\}$ are different in each composition phase, and the parameters $\{G^{\gamma}\}$ are different in each noncomposition phase. The H function is the familiar random polynomial:

$$H(w_1, \ldots, w_n, \{G\}) = G_0 + \sigma_H \sum_{k=1}^{q} \sum_{\substack{i_1 + \cdots + i_n = k \\ i_1, \ldots, i_n \geq 0}} f_{i_1 \ldots i_n; k}\, \xi_H^{-k} G_{i_1, \ldots, i_n} w_1^{i1} \ldots w_n^{i_n} \tag{12.31}$$

The degree $q = 6$ is chosen for the polynomial. The coefficients G_0 and $G_{i_1, \ldots, i_n}$ are denoted in compact notation by $\{G\}$. The symmetry factors $f_{i_1 \ldots i_n; k}$ are given by equation (12.5). The scale factor ξ_H is chosen so that each term in the multinomial contributes roughly the same amount in the root-mean-square sense. For the composition variables $\xi_H = \xi/2$, and for the noncomposition variables $\xi_H = (\langle z^6 \rangle / \langle z^2 \rangle)^{1/4} = (3/7)^{1/4}$. The σ_H are chosen so that the multinomial terms contribute 40% as much as the corresponding constant, phase terms, G_0 in Equation (12.31), in the root-mean-square sense.

The organic structure-directing molecules are characterized by composition, such as the identity of the ligands and template, as shown in Figure 12.8. As in Section 12.3, the small molecule is considered to consist of one template and six binding ligands. There is a template library and a ligand library. Six weakly correlated descriptors are used to describe the characteristics of the molecule, as in Section 12.3.

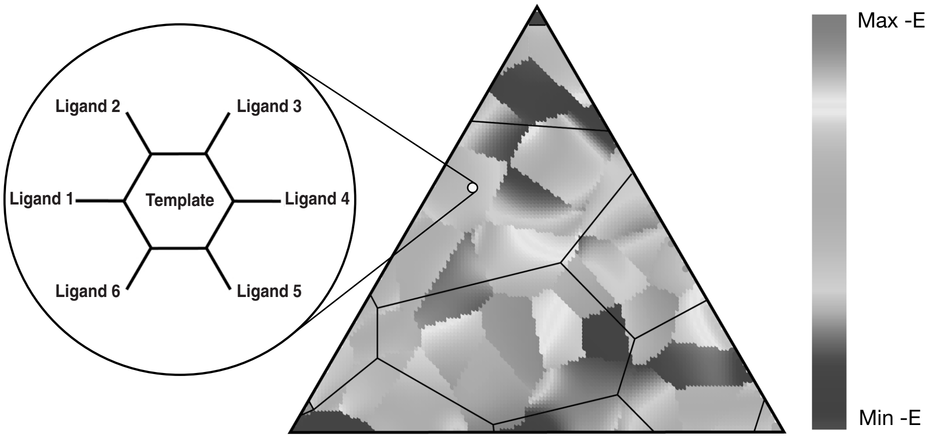

Figure 12.8 Schematic of the random-energy model for templated materials synthesis. (Adapted, with permission, from Chen, L., and Deem, W. *Mol. Phys.,* 2002, 100: 2175–2181, copyright © 2002, Taylor & Francis, Ltd., London.)

As in Section 12.3, a random-energy model is used to account for the contributions to structure direction arising from interactions between the zeolite, template, and ligands. The total molecular contribution to the figure of merit is, then,

$$E_{\text{molecule}} = E_L + E_T + E_{LL} + E_{TL}. \tag{12.32}$$

Each of these terms is given in the form of a random polynomial, where the coefficients in the random polynomial now depend on which zeolite phase is being made:

$$E_L(\mathbf{x}, \mathbf{z}) = \sum_{i=1}^{6} F[d_1^{(li)}, \ldots, d_6^{(li)}, \{G_L(\mathbf{x}, \mathbf{z})\}], \quad G_L(\mathbf{x}, \mathbf{z}) = \lambda_{x,1} G_L^{\alpha} + \tfrac{1}{2} \lambda_{z,1} G_L^{\gamma} \tag{12.33}$$

$$E_T(\mathbf{x}, \mathbf{z}) = F[D_1^{(m)},\ldots, D_6^{(m)}, \{G_T(\mathbf{x}, \mathbf{z})\}], \quad G_T(\mathbf{x}, \mathbf{z}) = \lambda_{x,2} G_T^{\alpha} + \tfrac{1}{2} \lambda_{z,2} G_T^{\gamma} \tag{12.34}$$

$$E_{LL}(\mathbf{x}, \mathbf{z}) = \sum_{i=1}^{6} h_i(\mathbf{x}, \mathbf{z})\, F[d_{j1}^{(li)}, d_{j2}^{(li)}, d_{j3}^{(li)}, d_{j4}^{(li+1)}, d_{j5}^{(li+1)}, d_{j6}^{(li+1)}, \{G_{LL}(\mathbf{x}, \mathbf{z})\}],$$
$$G_{LL}(\mathbf{x}, \mathbf{z}) = \lambda_{x,3}\, G_{LL}^{i,\alpha} + \tfrac{1}{2} \lambda_{z,3}\, G_{LL}^{i,\gamma} \tag{12.35}$$

$$E_{TL}(\mathbf{x}, \mathbf{z}) = \sum_{i=1}^{6} h_i(\mathbf{x}, \mathbf{z})\, F[d_{k1}^{(li)}, d_{k2}^{(li)}, d_{k3}^{(li)}, D_{k4}^{(m)}, D_{k5}^{(m)}, D_{k6}^{(m)}, \{G_{TL}(\mathbf{x}, \mathbf{z})\}],$$
$$G_{TL}(\mathbf{x}, \mathbf{z}) = \lambda_{x,4}\, G_{TL}^{i,\alpha} + \tfrac{1}{2} \lambda_{z,4}\, G_{TL}^{i,\gamma}. \tag{12.36}$$

The building block for our random-energy model as a function of those descriptors is equation (12.8). The $\{G\}$, again, will be composed of a composition and a noncomposition piece, as indicated in equations (12.33)–(12.36). The orders of the polynomials are $q = 6$ and $n = 6$. The scaling factors are as in equation (12.9). In general, the polynomial coefficients, $\{G\}$ in equations (12.33)–(12.36), are functions of the zeolite composition and noncomposition variables $\mathbf{x}$ and $\mathbf{z}$. However, for simplicity, it is assumed that the coefficients are phase-dependent only. The coefficients $\{G_L^\alpha\}$, $\{G_L^\gamma\}$, $\{G_T^\alpha\}$, $\{G_T^\gamma\}$, $\{G_{LL}^{i,\alpha}\}$, $\{G_{LL}^{i,\gamma}$, $\{G_{TL}^{i,\alpha}\}$, and $\{G_{TL}^{i,\gamma}\}$ are chosen from random Gaussian distributions with zero mean and unit variance. These coefficients are different in each phase. The strength of the interaction at structure directing site i is given by the structural constant h_i, which depends on the zeolite variables $\mathbf{x}$ and $\mathbf{z}$ as

$$h_i(\mathbf{x}, \mathbf{z}) = H(\mathbf{x} - \mathbf{x}^\alpha, \{G_i^\alpha\}) + \tfrac{1}{2} H(\mathbf{z}, \{G_i^\gamma\}). \tag{12.37}$$

The coefficients $\{G_i^\alpha\}$ and $\{G_i^\gamma\}$ are sets of phase-dependent fixed Gaussian random variables with zero mean and unit variance. The parameter σ_H is adjusted so that the multinomial terms contribute in the mean-square sense roughly the same as the constant, phase terms. It is assumed that the interaction indices, j_m and k_n, in equations (12.35) and (12.36) depend only on the template, since the molecular details of the zeolite have been suppressed. The constants $\lambda_{x,i}$ are adjusted so that the terms depending on $\mathbf{x}$ contribute as $E_L : E_T : E_{LL} : E_{TL} = 1 : 1 : 2 : 1.2$ in the root-mean-square sense. Furthermore, the $\lambda_{x,i}$ are adjusted so that the total molecular contribution from the terms depending on $\mathbf{x}$ account for roughly 15% of the constant, compositional phase term in E_{zeolite} in the root-mean-square sense. The values of $\lambda_{z,i}$ are similarly adjusted so that the contribution from the terms depending on $\mathbf{z}$ is 15% of the constant, noncomposition phase term in E_{zeolite} in the root-mean-square sense.

As always, the energy returned by the model is minimized. That is, the figure of merit, $-E$, is sampled by $\exp(-\beta E)$.

12.4.2 Multiround Monte Carlo Strategies

As in the small-molecule case, the template and ligand libraries are built first. The size of the template library is denoted by $N_T = 15$, and the size of the ligand library by $N_L = 1000$. As in Section 12.3, the values of the six descriptors of each ligand and template are extracted from a Gaussian random distribution with zero mean and unit variance. Two sets of random interaction descriptor indices are also associated to each template for the interaction terms in equations (12.35) and (12.36).

As with previous examples, the total number of samples to be synthesized is kept fixed at 100,000 for all protocols. This condition ensures that the total experimental cost is kept roughly constant.

Both grid and random single-pass protocols are considered. Most high throughput experiments tend to perform a single-pass grid search on all continuous variables. To mimic this procedure, a grid search is used on the composition and noncomposition variables. For the discrete molecular variables, however, a random approach is used to pick the template and ligands from the organocation libraries. In the random single-pass protocol, all variables are searched at random, i.e., the composition variables and noncomposition variables are chosen at random as well.

Several multipass Monte Carlo protocols are considered. Unlike single-pass protocols, multipass protocols allow one to learn about the characteristics of the system as the experiment proceeds. The multiple samples considered per round allow for a rather diverse population and increase the opportunity for several zeolite samples to survive more elaborate tests for application performance, tests that are only roughly correlated with the primary screen. A Monte Carlo protocol is used so as to maintain diversity in the samples. To be consistent with current experimental technology, 1000 samples are synthesized for a total of 100 rounds. The initial sample configurations are assigned by the random protocol.

Both the zeolite variables and the components of the structure-directing agent are changed at each round of the Monte Carlo protocol. In the simplest approach, the composition and the noncomposition variables are perturbed about their original values using the traditional Metropolis-type method. For the noncomposition variables, periodic boundary conditions are used. For the composition variables, reflecting boundary conditions are used, as discussed in Appendix 12.B. For the organocation variables, either the template is changed with probability p_{template}, or one of the six ligands is randomly changed with probability $1 - p_{\text{template}}$ in each Monte Carlo move. The proposed samples are accepted or rejected according to the acceptance rule at β, the inverse of the protocol temperature. In one Monte Carlo round, all the samples are sequentially updated. Either a simple Metropolis method or a biased Monte Carlo method can be incorporated for the organocation variables. In the simple Metropolis method, the proposed new template or ligand is selected randomly with uniform probability from the library. After the zeolite and organocation variables have been modified in an unbiased protocol, the acceptance probability is

$$p_{\text{acc}}(c \rightarrow p) = \min[1, \exp(-\beta \Delta E)]. \tag{12.38}$$

Biased Monte Carlo schemes allow the generation of new samples with a probability that depends on the figure of merit of the new sample. In the small-molecule high throughput experiment design of Section 12.3, biased energy forms were constructed from either theory or preexperiments. Since the total figure of merit is a function of both the organocation structure and the zeolite composition and noncomposition variables, it is not feasible to construct the bias from preexperiments. The theoretical bias is therefore used. The bias energy for template m for a zeolite at compositional phase α and noncompositional phase γ is

$$E^{(m)}(\mathbf{x}, \mathbf{z}) = F[D_1^{(m)}, \ldots, D_6^{(m)}, \{G_T(\mathbf{x}, \mathbf{z})\}], \qquad G_T(\mathbf{x}, \mathbf{z}) = \lambda_{x,2} G_T^{\alpha} + \tfrac{1}{2}\lambda_{z,2} G_T^{\gamma}. \tag{12.39}$$

The bias for ligand i is

$$e^{(i)}(\mathbf{x}, \mathbf{z}) = F[d_1^{(i)}, \ldots, d_6^{(i)}, \{G_L(\mathbf{x}, \mathbf{z})\}], \qquad G_L(\mathbf{x}, \mathbf{z}) = \lambda_{x,1} G_L^{\alpha} + \tfrac{1}{2}\lambda_{z,1} G_L^{\gamma}. \tag{12.40}$$

Cutoff energies are chosen as in Section 12.3.3. The biased energy, $e_b{}^{(i)}$, for the ith ligand then becomes

$$e_b^{(i)}(\mathbf{x}, \mathbf{z}) = \begin{cases} e^{(i)}(\mathbf{x}, \mathbf{z}) & \text{if } e^{(i)} > e_c \\ e_c(\mathbf{x}, \mathbf{z}) & \text{otherwise.} \end{cases} \tag{12.41}$$

Similarly, the biased energy, $E_b^{(m)}$, for the mth template becomes

$$E_b^{(m)}(\mathbf{x}, \mathbf{z}) = \begin{cases} E^{(m)}(\mathbf{x}, \mathbf{z}) & \text{if } E^{(m)} > E_c \\ E_c(\mathbf{x}, \mathbf{z}) & \text{otherwise.} \end{cases} \tag{12.42}$$

If the ith ligand in the organocation is to be changed at the proposed new zeolite values $\mathbf{x}'$ and $\mathbf{z}'$, the biased probability for selecting ligand l_i' from the library is

$$f[E(p)] = \frac{\exp[-\beta e_b^{(l_i')}(\mathbf{x}', \mathbf{z}')]}{\sum_{j=1}^{N_L} \exp[-\beta e_b^{(j)}(\mathbf{x}', \mathbf{z}')]}. \tag{12.43}$$

To satisfy detailed balance, the bias for the reverse move is needed:

$$f[E(c)] = \frac{\exp[-\beta e_b^{(l_i)}(\mathbf{x}, \mathbf{z})]}{\sum_{j=1}^{N_L} \exp[-\beta e_b^{(j)}(\mathbf{x}, \mathbf{z})]}. \tag{12.44}$$

Similarly, the bias probability for selecting template m' at proposed new zeolite values $\mathbf{x}'$ and $\mathbf{z}'$ is

$$f[E(p)] = \frac{\exp[-\beta\, E_b^{(m')}(\mathbf{x}', \mathbf{z}')]}{\sum_{j=1}^{N_T} \exp[-\beta\, E_b^{(j)}(\mathbf{x}', \mathbf{z}')]}. \tag{12.45}$$

The bias for the reverse move is needed to satisfy detailed balance:

$$f[E(c)] = \frac{\exp[-\beta E_b^{(m)}(\mathbf{x}, \mathbf{z})]}{\sum_{j=1}^{N_T} \exp[-\beta\, E_b^{(j)}(\mathbf{x}, \mathbf{z})]}. \tag{12.46}$$

The acceptance rule in the biased scheme is

$$p_{\text{acc}}(c \to p) = \min\left\{1, \frac{f[E(c)]}{f[E(p)]} \exp(-\beta\, \Delta E)\right\}. \tag{12.47}$$

The parallel tempering move is also considered. The acceptance rule for a parallel tempering exchange move is equation (12.3). As always, this exchange step is experimentally cost-free, and it can be effective at allowing the protocol to escape from local optima.

12.4.3 Effectiveness of the Monte Carlo Strategies

The size of the organocation library is fixed at $N_T = 15$ and $N_L = 1000$. The parameters $\lambda_{x,i}$ and $\lambda_{z,i}$ are adjusted so that the synergistic terms will contribute in the ratio $E_L : E_T : E_{LL} : E_{TL} = 1 : 1 : 2 : 1.2$ in the root-mean-square sense in the composition and noncomposition phases. The relative contributions from the zeolite and molecule are fixed by requiring that they be roughly of the same order for the optimal configurations obtained by the grid or random protocols. So that this occurs, the parameters are found by trial and error, due to the complicated molecular terms. The σ_H values are first adjusted, then the $\lambda_{x,i}$ and $\lambda_{z,i}$ in equations (12.33)–(12.36) are adjusted. As mentioned, the $\lambda_{x,i}$ and $\lambda_{z,i}$ are adjusted so that the total molecular contribution from either type of phase accounts for roughly 15% of the corresponding constant phase term in E_{zeolite} in the root-mean-square sense.

A few short preexperiments are performed to locate optimal parameters for the protocols. The optimal value for the probability of changing a template is $p_{\text{template}} = 0.02$, since the size of the template library is relatively small. The maximum random displacements are $|\Delta\, \mathbf{x}| = 0.1/\sqrt{d-1}$ and $|\Delta \mathbf{z}| = 0.2$ in the composition space and noncomposition space. The optimal inverse protocol temperature for simple Metropolis Monte Carlo is $\beta = 50$ for $d = 3$ and $d = 4$ and $\beta = 20$ for $d = 5$. The optimal inverse protocol temperature for the biased Monte Carlo schemes is $\beta_b = 500$ for $d = 3$ and $d = 4$ and $\beta_b = 200$ for $d = 5$. It is optimal to have the samples divided into three subsets when biased Monte Carlo is combined with parallel tempering, with 25% of the population at $\beta_1 = \frac{1}{2}\,\beta_b$, 50% at $\beta_2 = \beta_b$, and 25% at $\beta_3 = 2\,\beta_b$. The switching probability, p_{ex}, is 0.1. Determination of these parameters corresponds to gaining familiarity with a new protocol experimentally.

The protocols are tested with a constant organocation library size and increasing numbers of composition and noncomposition variables. Results are shown in Figure 12.9. From this figure, it can be seen that the multiround Monte Carlo protocols are better than single-pass protocols such as grid and random. The simple Metropolis method finds optimum samples that are twice as good as those from the grid or random protocols. The biased Monte Carlo is even more efficient, and the optimum figures of merit from biased Monte Carlo far exceed those from simple Metropolis with even 1000 rounds. The main contribution to the optimal figures of merit is coming from the ability of biased Monte Carlo to find favorable designs for the structure-directing agent. Only for the more complicated $d = 4$ and $d = 5$ systems is parallel tempering noticeably effective. It is more important to keep all 1000 of the samples at the optimum temperature for the relatively simple systems, such as $d = 3$. In either case, since most of the complexity arises from the organocation library variables, the impact of using parallel tempering is small.

Using a model for high throughput zeolite synthesis, multipass Monte Carlo methods are shown to work better than single-pass protocols. Sophisticated biased Monte Carlo schemes are highly efficient and significantly better than simple Metropolis Monte Carlo. For complicated systems with five or more framework chemical compositional variables, parallel tempering is the best method.

Interestingly, the complexity of the chemical space is largely determined by the

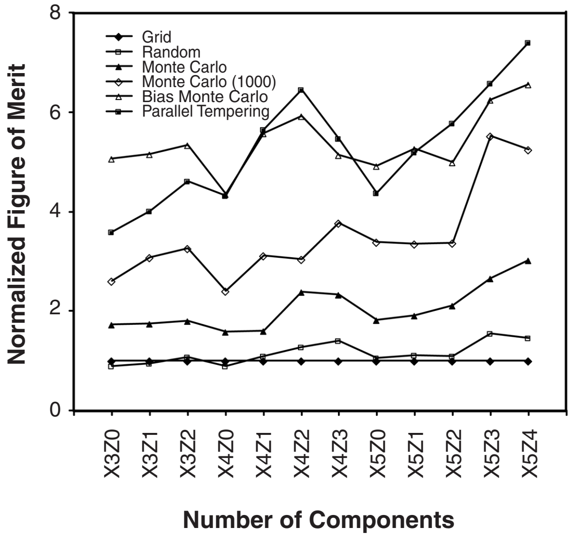

Figure 12.9 The optimal figure of merit found with different protocols with different number of composition (**x**) and noncomposition (**z**) variables. The complexity of the structure-directing agent libraries is identical in all cases. The values are scaled to the optimum found by the grid searching method. Each scaled value is averaged over results from 10 instances. (Reproduced, with permission, from Chen, L. and Deem, M. W., *Molecular Physics,* 2002, 100: 2175–2181, copyright © 2002, Taylor & Francis Ltd., London.)

complexity of the template. The composition and noncomposition variables play a minor role compared to the template. Physically, this means that templates are almost essential to successful zeolite synthesis, which is in accord with experiment. In addition, choice of template is crucial to effective materials performance, because without an appropriate template, no material is made, which is also in accord with experiment. The robotic technology for both template synthesis and zeolite synthesis [21] is available, and it should only be a matter of time before the Monte Carlo protocols are put to use.

SUMMARY

To summarize, there are essentially four main challenges to a successful HTE program. The first challenge is the design of sufficiently robust and transferable chemistries so that parallel synthesis is possible. The second challenge is the design of rapid and accurate robotic screens. The third challenge is the design and maintenance of the databases necessary to track the experimental data produced. The fourth and final challenge is deciding what materials to make, given the ability to synthesize and screen in high throughput. This last challenge of library design and redesign has been addressed by this chapter, and by this book as a whole. It is this last topic to which theory seems most likely to make a contribution.

This chapter has described the Monte Carlo approach to library design and redesign. An alternative approach, with which some comparison was made, is use of GAs. The difference between Monte Carlo and GAs is not in the proposal of new samples, often called the mutation event in GAs. Any mutation event possible in GAs is also a possible move in a Monte Carlo strategy. The difference between the two approaches lies in the selection step. Both GAs and Monte Carlo can perform well. Only Monte Carlo, however, is guaranteed to sample from the figure of merit, the guarantee being provided by Markov chain theory. The benefit of sampling becomes apparent when secondary and tertiary performance screens are performed on hits identified by the primary screen.

Despite the many successes of HTE, the method has been criticized as a simple machinery, lacking incorporation of a priori knowledge when compared with the traditional synthetic approach. A priori knowledge, such as chemical intuition, previous database or experimental information, well-known theory, patentability, or other specific constraints, is indispensable to an efficient library design and is the traditional province of the synthetic chemist. It is fascinating that the Monte Carlo approach to HTE can naturally incorporate such knowledge in the experimental design through the technique of biased Monte Carlo.

The Monte Carlo strategies discussed in this chapter are complementary to other approaches to library design and redesign. For example, while the random construction is superior to the grid approach, even more sophisticated methods that take into account concerns such as diversity or a priori chemical knowledge can be used for the initial library design. Similarly, if a purely spatially addressable synthesis is too expensive in an application, the Monte Carlo protocol can be used to design a few freely addressable samples, and a less expensive combinatorial method can be used to produce a cloud of samples about these few points. If the figure of merit is locally smooth, but globally rough, interpolation procedures can be used to provide a good experimental estimate for the bias that will assist a Monte Carlo protocol to sample the chemical and experimental space. Finally, when promising hits are identified by a Monte Carlo experiment, gradient-type optimization methods can be used to refine the final process and material variables.

ACKNOWLEDGMENT

It is a pleasure to acknowledge the contributions of my collaborators Marco Falcioni and Ligang Chen. This work was supported by the National Science Foundation.

APPENDIX 12.A: THE SIMPLEX OF ALLOWED COMPOSITIONS

In any high throughput experiment involving elemental compositions as variables, some means of searching the composition space must be developed. Typically, the composition space is defined by the mole fractions of each element. If there are d mole fractions, there are only $d - 1$ variables to be searched, as the mole fractions sum to unity. Since there are only $d - 1$ variables among the d mole fractions, x_{i}, the changes to the mole fractions that occur during the library redesign must be made in a correlated fashion. In particular, the changes to the mole fractions must be made in the $d - 1$ dimensional hyperplane that is consistent with the constraint $\Sigma_{i=1}^{d} x_{\mathrm{i}} = 1$. Mathematically, this constraint can be written as as $\mathbf{x} \cdot \mathbf{u}_d = 1/\sqrt{d}$, where the vector $\mathbf{u}_d = (1/\sqrt{d}, 1/\sqrt{d}, \ldots, 1/\sqrt{d})$.

It is convenient to perform a change of variables from the mole fractions to variables that lie in the allowed hyperplane. This change of variables then allows each of the new variables to be changed independently. What this amounts to mathematically is a rotation of the coordinate system describing the mole fractions to a coordinate system in which the first $d - 1$ components are orthogonal to $\mathbf{u}_d$. Defining the original coordinate system by the unit vectors

$$\begin{aligned}\mathbf{e}_1 &= (1, 0, \ldots, 0, 0)\\ \mathbf{e}_2 &= (0, 1, \ldots, 0, 0\ )\\ &\ \ \vdots\\ \mathbf{e}_{d-1} &= (0, 0, \ldots, 1, 0),\end{aligned} \tag{12.A.1}$$

the new coordinate system is identified by the Gram–Schmidt procedure:

$$\begin{aligned}\mathbf{u}_1 &= \frac{\mathbf{e}_1 - (\mathbf{e}_1 \cdot \mathbf{u}_d)\ \mathbf{u}_d}{|\mathbf{e}_1 - (\mathbf{e}_1 \cdot \mathbf{u}_d)\ \mathbf{u}_d|}\\ \mathbf{u}_2 &= \frac{\mathbf{e}_2 - (\mathbf{e}_2 \cdot \mathbf{u}_d)\mathbf{u}_d - (\mathbf{e}_2 \cdot \mathbf{u}_1)\mathbf{u}_1}{|\mathbf{e}_2 - (\mathbf{e}_2 \cdot \mathbf{u}_d)\mathbf{u}_d - (\mathbf{e}_2 \cdot \mathbf{u}_1)\mathbf{u}_1|}\\ &\ \ \vdots\\ \mathbf{u}_i &= \frac{\mathbf{e}_i - (\mathbf{e}_i \cdot \mathbf{u}_d)\mathbf{u}_d - \sum_{j=1}^{i-1} (\mathbf{e}_i \cdot \mathbf{u}_j)\ \mathbf{u}_j}{|\mathbf{e}_i - (\mathbf{e}_i \cdot \mathbf{u}_d)\mathbf{u}_d - \sum_{j=1}^{i-1} (\mathbf{e}_i \cdot \mathbf{u}_j)\ \mathbf{u}_j|}, \qquad i < d.\end{aligned} \tag{12.A.2}$$

The relationship between the original mole fractions and the new variables, w_i, is given by a simple rotation:

$$\mathbf{x} = R\mathbf{w}, \tag{12.A.2}$$

where R is the rotation matrix, and the value of R_{ij} is given by the ith component of $\mathbf{u}_j$. How this coordinate system rotation works for a three-component system is shown in Figure 12.10. To ensure that the mole fractions sum to unity, $w_d = 1/\sqrt{d}$. These new variables, w_i, $i < d$, can be changed independently. Any changes made to

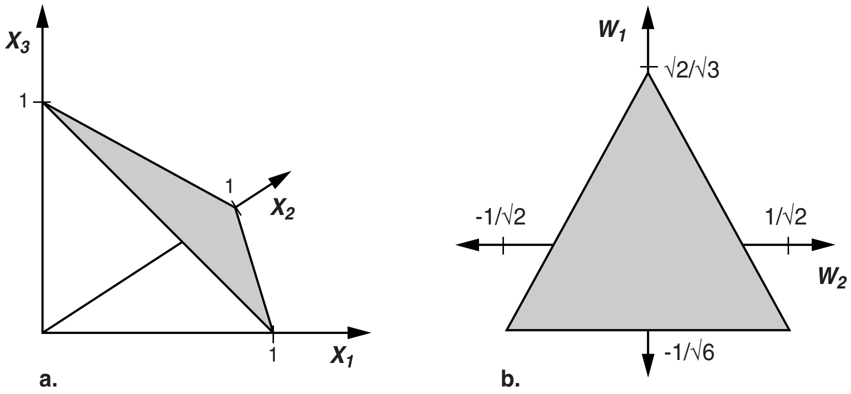

Figure 12.10 The allowed composition range of a three-component system is shown in the (a) original composition variables, x_i, and (b) Gram-Schmidt variables, w_i. (Reproduced, with permission, from Deem, M. W., *Advances in Chemical Engineering,* Vol. 28, A. Chakraborty, Ed., 81–121; copyright © 2001. Academic Press, San Diego.)

w_i, $i < d$, when the mole fractions are calculated from equation (12.A.3), will satisfy the constraint that the mole fractions sum to one. Some changes, however, may violate the constraint that the mole fractions must all be positive. In such cases, the procedure described in Appendix 12.B is used to map the invalid mole fractions to valid mole fractions.

APPENDIX 12.B: REFLECTING BOUNDARY CONDITIONS

As shown in Figure 12.10, the valid mole fractions form a simplex. Not only do the mole fractions sum to one, but they also are all nonnegative. If a particular sample has a composition that is near the edge of the allowed simplex, very often the random change to the composition that occurs in the Monte Carlo protocol will lead to a new composition with invalid values. To remedy this situation, it is best to have a mapping of such invalid values back to the allowed composition simplex. With such a mapping, all of the attempted Monte Carlo moves will at least be valid, and experimental effort will not be wasted.

There is a procedure that ensures all Monte Carlo moves lead to valid new composition points. The procedure involves reflecting an invalid composition point about the $d - 2$-dimensional hyperplanes that define the edges of the allowed simplex. Figure 12.11 shows the procedure for a move in a three-component system that requires only a single reflection. The procedure requires the unit normals, $\mathbf{n}_i$, to each of the faces of the simplex as well as the constants, c_i, that define the faces by the equation $\mathbf{x} \cdot \mathbf{n}_i = c_i$. As shown in Figure 12.11, in the case of a three-component system there are three such unit normals and three such constants. We initially define $\mathbf{y} = \mathbf{x}(c)$. The procedure executes the following steps:

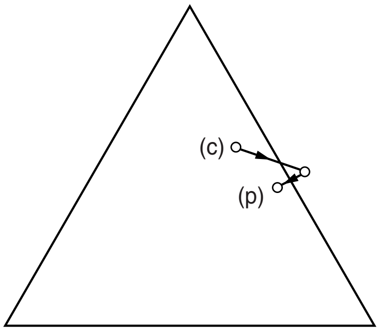

Figure 12.11 Schematic of the reflecting boundary conditions used to obtain valid composition variables. (Reproduced, with permission, from Chen, L. and Deem M. W., *Molecular Physics,* 2002, 100: 2175–2181, copyright © 2002, Taylor & Francis Ltd., London.)

- Determine whether the proposed new composition, $\mathbf{x}$, is in the allowed region. If so, define the new composition $\mathbf{x}(p)$ to be $\mathbf{x}$ and stop.
- If not, find the face i for which the quantity $t_i = (c_i - \mathbf{y}\cdot\mathbf{n}_\mathrm{i})/[(\mathbf{x} - \mathbf{y})\cdot\mathbf{n}_i]$ is minimal, taking into account only those faces for which both t_i and $(\mathbf{x} - \mathbf{y})\cdot\mathbf{n}_i$ are positive.
- Define $\mathbf{y} \leftarrow \mathbf{y} + t_i(\mathbf{x} - \mathbf{y})$.
- Reflect the composition point through face i by the equation $\mathbf{x} \leftarrow \mathbf{x} - 2(\mathbf{x} \cdot \mathbf{n}_i - c_i)\mathbf{n}_i$.
- Go back to the first step.

At each step of this algorithm, the magnitude of the composition vector decreases by a finite amount, and so this procedure converges. The procedure is also reversible. In other words, for each forward move $\mathbf{x}(c) \to \mathbf{x}$ that leads to the new composition $\mathbf{x}(p)$, there is always a backward move $\mathbf{x}(p) \to \mathbf{x}'$ that leads to the original composition $\mathbf{x}(c)$. Furthermore, when the original Monte Carlo move is in the allowed move sphere $|\mathbf{x}(c) - \mathbf{x}| < \Delta x$, the reverse Monte Carlo move is as well: $|\mathbf{x}(p) - \mathbf{x}'| < \Delta x$. This reversibility is important, because it allows the procedure to satisfy detailed balance. These reflecting boundary conditions provide a simple modification of the spherical move so that new compositions are always within the allowed simplex. These boundary conditions are analogous to the periodic boundary conditions that are more typically used in atomistic simulations.

These reflecting boundary conditions can be given a geometrical interpretation. Essentially, they are a form of billiards in the d – 1-dimensional simplex. In each move, we chose a displacement $\mathbf{x} - \mathbf{x}(c)$. We can imagine the composition variable as a small ball that moves at a constant speed along this trajectory. The motion is continued until either a boundary is reached or the entire length of the move has been traveled. If a boundary of the simplex is reached, the ball reflects off the hyperplane by Newtonian mechanics, and the trajectory is continued in the new direction. The reflections continue until the ball has traveled a distance equal to the chosen length of the displacement, $|\mathbf{x} - \mathbf{x}(c)|$. The new point $\mathbf{x}(p)$ is then given by the location of the ball at the end of the trajectory.

REFERENCES

1. Manousiouthakis, V.; Deem, M. W. *J. Chem. Phys.,* 1999, 110: 2735–2756.
2. Kauffman, S.; Levin, S. *J. Theor. Biol.,* 1987, 128: 11–45.
3. Perelson, A. S.; Macken, A. *Proc. Natl. Acad. Sci. USA,* 1995, 92: 9657–9661.
4. Bogarad, L. D.; Deem, M. W. *Proc. Natl. Acad. Sci. USA,* 1999, 96: 2591–2595.
5. Falcioni, M.; Deem, M. W. *Phys. Rev. E,* 2000, 61: 5948–5952.
6. Chen, L.; Deem, M. W. *J. Chem. Inf. Comput. Sci.,* 2001, 41: 950–957.
7. Deem, M. W. In A. Chakraborty, Ed., *Advances in Chemical Engineering*, Vol. 28, Academic Press, San Diego, 2001, 81–121.
8. Pirrung, M. C. *Chem. Rev.*, 1997, 97: 473–488.
9. Weinberg, W. H.; Jandeleit, B.; Self, K.; Turner, H. *Curr. Opin. Chem. Bio.*, 1998, 3: 104–110.
10. McFarland, E. W., Weinberg, W. H. *TIBTECH*, 1999, 17: 107–115.
11. Xiang, X.-D.; Sun, X.; Briceño, G.; Lou, Y.; Wang, K. A.; Chang, H.; Wallace-Freedman, G.; Chang, S.-W.; Schultz, P. G. *Science*, 1995, 268: 1738–1740.
12. Briceño, W. G.; Chang, H.; Sun, X.; Schultz, P. G.; and Xiang, X.-D. *Science*, 1995, 270: 273–275.
13. Danielson, E.; Golden, J. H.; McFarland, E. W.; Reaves, C. M.; Weinberg, W. H.; Wu, X. D. *Nature*, 1997, 389: 944–948.
14. Danielson, E.; Devenney, M.; Giaquinta, D. M.; Golden, J. H.; Haushalter, R. C.; McFarland, E. W.; Poojary, D. M.; Reaves, C. M.; Weinberg, W. H.; Wu, X. D. *Science*, 1998, 279: 837–839.
15. Wang, J.; Yoo, Y.; Gao, C.; Takeuchi, I.; Sun, X.; Chang, H.; Xiang, X.-D.; Schultz, P. G. *Science*, 1998, 279: 1712–1714.
16. Francis, M. B.; Jamison, T. F.; Jacobsen, E. N. *Curr. Opin. Chem. Biol.*, 1998, 2: 422–428.
17. Dickinson, T. A.; Walt, D. R.; White, J.; and Kauer, J. S. *Anal. Chem.*, 1997, 69: 3413–3418.
18. Menger, F. M.; Eliseev, A. V.; Migulin, V. A. *J. Organic Chem.*, 1995, 60: 6666–6667.
19. Burgess, K,; Lim, H.-J.; Porte, A. M.; and Sulikowski, G. A. *Angew. Chem. Int. Ed.*, 1996, 35: 220–222.
20. Cole, B. M.; Shimizu, K. D.; Krueger, C. A.; Harrity, J. P. A.; Snapper, M. L.; Hoveyda, A. H. *Angew. Chem. Int. Ed.*, 1996, 35: 1668–1671.
21. Akporiaye, D. E.; Dahl, I. M.; Karlsson, A.; Wendelbo, R. *Angew. Chem. Int. Ed.*, 1998, 37: 609–611.
22. Reddington, E.; Sapienza, A.; Gurau, B.; Viswanathan, R.; Sarangapani, S.; Smotkin, E. S.; Mallouk, T. E. *Science*,1998, 280: 1735–1737.
23. Cong, P.; Doolen, R. D.; Fan, Q.; Giaquinta, D. M.; Guan, S.; McFarland, E. W.; Poo-

jary, D. M.; Self, K.; Turner, H.; Weinberg, W. H. *Angew. Chem. Int. Ed.*, 1999, 38: 483–488.

24. Van Dover, R. B.; Schneemeyer, L. F.; Fleming, R. M. *Nature*, 1998, 392: 162–164.
25. Novet, T.; Johnson, D. C.; Fister, L. *Adv. Chem. Ser.*, 1995, 245: 425–469.
26. Zones, S. I.; Nakagawa, Y.; Lee, G. S.; Chen, C. Y.; Yuen, L. T. *Microporous Mesoporous Mat.*, 1998, 21: 199–211.
27. Niederreiter, H. *Random Number Generation and Quasi-Monte Carlo Methods,* Society for Industrial and Applied Mathematics, Philadelphia, 1992.
28. Bratley, P.; Fox, B. L.; Niederreiter, H. *ACM Trans. Math. Softw.*, 1994, 20: 494–495.
29. Geyer, C. J. In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, American Statistical Association, New York, 1991, 156–163.
30. Swendsen, R. H.; Wang, J. S. *Phys. Rev. Lett.*, 1986, 57: 2607–2609.
31. Marinari, E.; Parisi, G.; Ruiz-Lorenzo, J. In *Spin Glasses and Random Fields*, Vol. 12, *Directions in Condensed Matter Physics*, A. P. Young, Ed., World Scientific, Singapore, 1998, 59–98.
32. Falcioni, M.; Deem, M. W. *J. Chem. Phys.*, 1999, 110: 1754–1766.
33. Sedgewick, R. *Algorithms,* 2nd ed., Addison-Wesley, New York, 1988.
34. Balkenhohl, F.; von dem Bussche-Hünnefeld, C.; Lansky, A.; Zechel, C. *Angew. Chem. Int. Ed. Engl.*, 1996, 35: 2288–2377.
35. Jandeleit, B.; Schaefer, D. J.; Powers, T. S.; Turner, H. W.; Weinberg, W. H. *Angew. Chem. Int. Ed.*, 1999, 38: 2494–2532.
36. McFarland, E. W.; Weinberg, W. H. *Trends Biotechnol.*, 1999, 17: 107–115.
37. Furka, A.; Sebestyen, F.; Asgedom, M.; Dibo, G. *Int. J. Peptide Protein Res.*, 1991, 37: 487–493.
38. Ohlmeyer, M. H. J.; Swanson, R. N.; Dillard, L. W.; Reader, J. C.; Asouline, G.; Kobayashi, G. A. R.; Wigler, M.; Still, W. C. *Proc. Natl. Acad. Sci. USA*, 1993, 90: 10922–10926.
39. Francis, M. B.; Jamison, T. F.; Jacobsen, E. N. *Curr. Opin. Chem. Biol.*, 1998, 2: 422–428.
40. Tan, D. S.; Foley, M. A.; Stockwell, B. R.; Shair, M. D.; Schreiber, S. L. *J. Am. Chem. Soc.*, 1999¸ 121: 9073–9087.
41. Weinberg, W. H.; Jandeleit, B.; Self, K.; Turner, H. *Curr. Opin. Chem. Bio.*, 1998, 3: 104–110.
42. Schreiber, S. L. *Science*, 2000, 287: 1964–1969.
43. Nazarpack-Kandlousy, N.; Zweigenbaum, J.; Henion, J.; Eliseev, A. V. *J. Comb. Chem.*, 1999, 1: 199–206.
44. Desai, M. C.; Zuckermann, R. N.; Moos, W. H. *Drug Dev. Res.*, 1994, 33: 174–188.
45. Ajay, W.; Walters, P.; Murcko, M. A. *J. Med. Chem.*, 1998, 41: 3314–3324.
46. Bures, M. G.; Martin, Y. C. *Curr. Opin. Chem. Biol.*, 1998, 2: 376–380.
47. Drewry, D. H.; Young, S. S. *Chemometrics Intell. Lab.*, 1999, 48: 1–20.
48. Hassan, M.; Bielawski, J. P.; Hempel, J. C.; Waldman, M. *Mol. Diversity*, 1996, 2: 64–74.
49. Zheng, W.; Cho, S. J.; Tropsha, A. *J. Chem. Inf. Comput. Sci.*, 1998, 38: 251–258.
50. Zheng, W.; Cho, S. J.; Tropsha, A. *J. Chem. Inf. Comput. Sci.*, 1998, 38: 259–268.

51. Zheng, W.; Cho, S. J.; Waller, C. L.; Tropsha, A. *J. Chem. Inf. Comput. Sci.*, 1999, 39: 738–746.

52. Reynolds, C. H. *J. Comb. Chem.*, 1999, 1: 297–306.

53. Linusson, A.; Wold, S.; Nordén, B. *Mol. Diversity*, 1999, 4: 103–114.

54. Gobbi, A.; Poppinger, D. *Biotechnol. Bioeng.*, 1998, 61: 47–54.

55. Weber, L.; Wallbaum, S.; Broger, C.; Gubernator, K. *Angew. Chem. Int. Ed. Engl.*, 1995, 34: 2280–2282.

56. Sheridan, R. P.; Kearsley, S. K. *J. Chem. Inf. Comput. Sci.*, 1995, 35: 310–320.

57. Singh, J.; Ator, M. A.; Jaeger, E. P.; Allen, M. P.; Whipple, D. A.; Soloweij, J. E.; Chowdhary, S.; and Treasurywala, A. M. *J. Am. Chem. Soc.*, 1996, 118: 1669–1676.

58. Brennan, M. B. *Chemical Engineering News*, June 5, 2000, 78(23): 63–73.

59. Brown, R. D.; Clark, D. E. *Expert Opin. Therap. Pat.*, 1997, 8: 1447–1459.

60. Brown, R. D.; Martin, Y. C. *J. Med. Chem.*, 1998, 40: 2304–2313.

61. Katritzky, A. R.; Kiely, J. S.; Hébert, N.; Chassaing, C. *J. Comb. Chem.*, 2000, 2: 2–5.

62. Frenkel, D.; Smit, B. *Understanding Molecular Simulation: From Algorithms to Applications,* 2nd edition, Academic Press, San Diego, 2002.

63. Shuker, S. B.; Hajduk, P. J.; Meadows, R. P.; Fesik, S. W. *Science*, 1996, 274: 1531–1534.

64. Fejzo, J.; Lepre, C. A.; Peng, J. W.; Bemis, G. W.; Ajay; Murcko, M. A.; Moore, J. M. *Chem. Biol.*, 1999, 6: 755–769.

65. Maly, D. J.; Choong, I. C.; Ellman, J. A. *Proc. Natl. Acad. Sci. USA*, 2000, 97: 2419–2424.

66. Griffey, R. H.; Hofstadler, S. A.; Sannes-Lowery, K. A.; Ecker, D. J.; Crooke, S. T. *Proc. Natl. Acad. Sci. USA*, 1999, 96: 10129–10133.

67. Davis, M. E.; Zones, S. I. In *Synthesis of Porous Materials: Zeolites, Clays, and Nanostructures*, Occelli, M. L.; Kessler, H., Eds., Marcel Dekker, New York, 1997, 1–34.

68. Burkett, S. L.; Davis, M. E. *Chem. Mater.*, 1995, 7: 920–928.

69. Burkett, S. L.; Davis, M. E. *Chem. Mater.*, 1995, 7: 1453–1463.

70. Chen, L.; Deem, M. W. *Mol. Phys.*, 2002, 100: 2175–2181.