Maximal Use of Minimal Libraries through the Adaptive Substituent Reordering Algorithm

Fan Liang, Xiao-jiang Feng, Michael Lowry, and Herschel Rabitz*

Department of Chemistry, Princeton University, Princeton, New Jersey 08544

Received: September 9, 2004; In Final Form: January 4, 2005

This paper describes an adaptive algorithm for interpolation over a library of molecules subjected to synthesis and property assaying. Starting with a coarse sampling of the library compounds, the algorithm finds the optimal substituent orderings on all of the functionalized scaffold sites to allow for accurate property interpolation over all remaining compounds in the full library space. A previous paper introduced the concept of substituent reordering and a smoothness-based criterion to search for optimal orderings (Shenvi, N.; Geremia, J. M.; Rabitz, H. *J. Phys. Chem. A* **2003**, *107*, 2066). Here, we propose a data-driven root-mean-squared (RMS) criteria and a combined RMS/smoothness criteria as alternative methods for the discovery of optimal substituent orderings. Error propagation from the property measurements of the sampled compounds is determined to provide confidence intervals on the interpolated molecular property values, and a substituent rescaling technique is introduced to manage poorly designed/sampled libraries. Finally, various factors are explored that can influence the applicability and interpolation quality of the algorithm. An adaptive methodology is proposed to iteratively and efficiently use laboratory experiments to optimize these algorithmic factors, so that the accuracy of property predictions is maximized. The enhanced algorithm is tested on copolymer and transition metal complex libraries, and the results demonstrate the capability of the algorithm to accurately interpolate various properties of both molecular libraries.

I. Introduction

High throughput screening (HTS) and combinatorial library synthesis are seeing broad utility, with applications ranging from medicinal chemistry to the fields of immunobiology and materials science.[2−4] While the design of combinatorial libraries varies considerably, more recent trends show movement toward creating a family of structurally and functionally related compounds by placing different functional groups/substituents on one or more molecular scaffolds.[5] In the case of a single scaffold with N substituent sites and S distinct substituents that may be attached at each site, the complete compound library would contain a total of S^N compounds. The synthesized members of the library would be screened for one or more properties, such as binding affinity, toxicity, or absorption in the case of drug discovery.

Deciding which substituents and scaffolds to use for creating a potentially effective library is a major challenge. Even when the scaffolds and substituents are carefully selected, the actual synthesis and screening of the entire library can be laborious, if not impossible, due to limitations in synthesis, purification, and property measurement. Wet lab time and associated costs could be significantly lowered if it were possible to produce and screen only a subset, M, of compounds within the library space and then accurately interpolate the screening properties of the remaining $S^N - M$ compounds, where $M \ll S^N$.

Various methods, based on correlating physicochemical descriptors (e.g., hydrophobicity, electronegativity, etc.), exist for interpolation and even extrapolation for compound libraries and functional group structure with compound behavior.[6] A common example of such chemometric approaches is the generation of quantitative structure−activity relationships (QSAR).[7] There are however, serious limitations on both the interpolative and extrapolative ability of these techniques due to difficulties in selecting the correct and complete descriptor set and quantifying its values.

A previous study[1] proposed a substituent reordering and interpolation algorithm for estimating property values of an entire library from the synthesis and assaying of a small number of randomly sampled library compounds. The critical feature of this new algorithm is the reordering operation that places the substituents of each functionalized site in an optimal sequence, enabling the identification of rational property behavior over the full space for effective interpolation. The reordering algorithm can function without *a priori* knowledge of any descriptor set. Fundamentally, both the reordering and QSAR-type methods are pattern recognition algorithms. However, whereas QSAR seeks structure−activity correlations, the reordering algorithm approaches the problem from a top-down perspective that aims to find the optimal substituent ordering: an ordering derived from an assumption of physical regularity in the overall library, and the M sampled compounds.

A single scaffold compound library that allows S_i substituents at the ith substituent site, where $i = 1, 2, ..., N$, can be visualized as residing in an N-dimensional space, for which the set of $\mathbf{R}_i = \{R_{i1}, R_{i2}, ..., R_{iS_i}\}$ substituents available at the ith scaffold site are assigned to uniformly spaced integer positions along an axis representing that site. Each compound in the library can then be indexed by a N-dimensional vector $\mathbf{X} = \{X_1, X_2, ..., X_i, ..., X_N\}$, where X_i corresponds to a substituent in $\mathbf{R}_i$ and has an integer value between 1 and S_i. The screened property value, y, associated with each compound $\mathbf{X}$ in the library space is

$$y = g(\mathbf{X}). \tag{1}$$

Whether $g(\mathbf{X})$ is amenable to interpolation across the full library

* To whom correspondence should be addressed. E-mail: hrabitz@princeton.edu. Telephone: (609) 258-3917. Fax: (609) 258-0967.

10.1021/jp045926y CCC: $30.25 © 2005 American Chemical Society
Published on Web 03/04/2005

space, when only a small subset of M compounds is known, depends on the regularity of $g(\mathbf{X})$ as a function of $\mathbf{X}$. It is possible that some systems do not possess regular behavior, but such problems are not expected to be amenable to interpolation by any method. Practical experience also suggests that physical and chemical properties do display regular behavior, as characterized by continuity and differentiability at a reasonable level of resolution over the compound library. Thus, we presume that the unknown property relationship, $g(\mathbf{X})$, of the molecular library under study exhibits regular behavior and has the capacity of being reliably interpolated. From an analysis perspective, it is of interest to exploit the reordering algorithm to assess the degree to which typical properties of well-designed molecular libraries have regular behavior.

A crucial point is the dependence of the regularity of $g(\mathbf{X})$ on how the substituent space is defined, which in turn is determined by the integer assignments given to each substituent in the set $\mathbf{R}_i$. The ordering of substituents in $\mathbf{R}_i$ defines the layout of the library space, so that changes in substituent values will likewise change the coordinates X_i, $i = 1, 2, ..., N$, for each compound $\mathbf{X}$. A key aspect of the algorithm is the identification of an optimal ordering for all $\mathbf{R}_1, \mathbf{R}_2, ..., \mathbf{R}_N$ substituent sets of the N scaffold sites so that $g(\mathbf{X})$ in this particularly structured library space can reveal its physical regularity. Once the optimal substituent assignment is acquired, it becomes possible to use sparse data interpolation techniques across the entire library space.

The original algorithm[1] defined the optimal ordering as one which maximizes the smoothness of $g(\mathbf{X})$, considering that smooth $g(\mathbf{X})$ behavior permits better interpolation. This assumption was confirmed in model studies on mathematical functions and in tests with a copolymer library data set, where a definitive correlation was observed between smooth $g(\mathbf{X})$ surfaces and good interpolative ability outside of the M sampled compounds. Here, we present conceptual and methodological developments that increase the effectiveness of the reordering algorithm. First, the smoothness-based criterion is recognized as only one particular measure of the interpolative capacity of $g(\mathbf{X})$. This work proposes a new cost function that directly appraises the quality of interpolation using the root-mean-squared (RMS) difference between interpolated and actual compound measurements on a training set of M compounds. The smoothness and RMS-based costs may be combined to capitalize on the strengths of both alternatives, thereby obtaining better interpolation than offered from using either of the costs alone.

The treatment of input data error also arises as a central practical matter, especially when dealing with HTS that often focuses on fast and possibly less precise screening techniques. Data errors may cause ill-behaved interpolation, and we explore a variety of regularization methods to address this problem. The assessment of data error propagation from the M samples to the interpolated compounds is included as well. Last, a rescaling technique is introduced in conjunction with reordering to enhance interpolation quality by allowing the substituent assignments in $\mathbf{R}_i$ to occur at nonuniform intervals. It is also argued that rescaling may identify and compensate for poorly designed molecular libraries, where functionally disparate compounds are combined in a single library.

In practice, the reordering algorithm is designed for incorporation into an iterative process of laboratory synthesis/assaying and data analysis/interpolation for efficient discovery of desirable compounds, where the number of screened compounds M is slowly increased from a small initial sampling. The goal is to keep the synthesis and screening efforts to a minimum, and the iteration is stopped when the library space is adequately resolved and consistently capable of accurately estimating the properties of newly synthesized compounds. This iterative process would also serve to optimize the parameters influencing the algorithm's efficacy.

Section II explains the algorithmic methodology advances. Section III applies the new methodologies to several test cases, including contact angle (CA) data from a copolymer library[8] and energy emission data from a chromophore transition metal complex library.[9] Concluding remarks are presented in section IV.

II. Algorithm

This section will briefly summarize the basic reordering concept and then introduce new algorithmic tools. Section III will selectively illustrate key aspects of the overall algorithm presented in Figure 1.

II.1. Substituent Reordering. An ordering $Z_i = [Z_{i1,}Z_{i2}, ..., Z_{iSi}]$, for the substituent set $\mathbf{R}_i = \{R_{i1}, R_{i2}, ..., R_{iSi}\}$ available at the ith scaffold site, where $i = 1, 2, ..., N$, is defined as a permuted set of integers 1, 2, ..., S_i. Hence, an ordering for the whole library space is $\mathbf{Z} = [Z_1, ..., Z_i, ..., Z_N]$. As each of the N scaffold sites may be ordered independently, there are a total of $\tilde{\mathbf{Z}} = S_1! \times S_2! \times ... \times S_N!$ possible orderings. The library space is represented in N dimensions, where substituents of the set $\mathbf{R}_i$ are positioned at uniformly spaced integer intervals along the axis representative of the ith scaffold site. This construct places each compound in the library on a lattice point $\mathbf{X} = \{X_1, X_2, ..., X_i, ..., X_N\}$, where $X_i \in \mathbf{R}_i$, within the N-dimensional space. The compound property $g(\mathbf{X})$ topology over the substituent space depends on the substituent ordering $\mathbf{Z}$ assigned to the scaffold axes. While it is possible to define the space using a random ordering, interpolation of $g(\mathbf{X})$ across an arbitrarily ordered library normally has no predictive ability because the regularity of $g(\mathbf{X})$ is scrambled. We seek an optimal ordering $\mathbf{Z}^*$ that uncovers the regularity in $g(\mathbf{X})$ arising from the underlying chemical/physical relationship between the library members and the observed property. The determination of $\mathbf{Z}^*$ can operate without using any descriptor information by minimizing a suitable cost function associated with the M sampled compounds and their property observations.

For any given ordering $\mathbf{Z}$, sparse data interpolation techniques can be used to approximate $g(\mathbf{X})$ with an analytic function $f(\mathbf{X})$, where $g(\mathbf{X})$ is known solely at the M sampled points $\{\tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2, ..., \tilde{\mathbf{X}}_M\}$. For the purposes of testing the algorithm, the M samples were chosen at random, with the sole criterion that each substituent be represented at least once. This may not be the case under all circumstances, namely when a compound library has already been partially synthesized. Under those instances, it may be most expedient to use the samples with assay values already at hand, provided that each substituent is represented among the samples.

Although various basis functions may be used to express $f(\mathbf{X})$, radial basis functions (RBF) are convenient because of their ability to interpolate high-dimensional sparse data.[10,11] Many choices for RBFs are possible, and we have found multiquadrics to be quite effective

$$\phi_k(\mathbf{X}) = (|\tilde{\mathbf{X}}_k - \mathbf{X}|^2 + \beta)^{1/2}. \tag{2}$$

Here $\tilde{\mathbf{X}}_k$ locates where the kth RBF is centered, and β is a basis parameter. A total of B basis functions $\phi_1,\phi_2,...,\phi_B$, where $B \leq M$, are used to construct $f(\mathbf{X})$. In this work, each RBF center

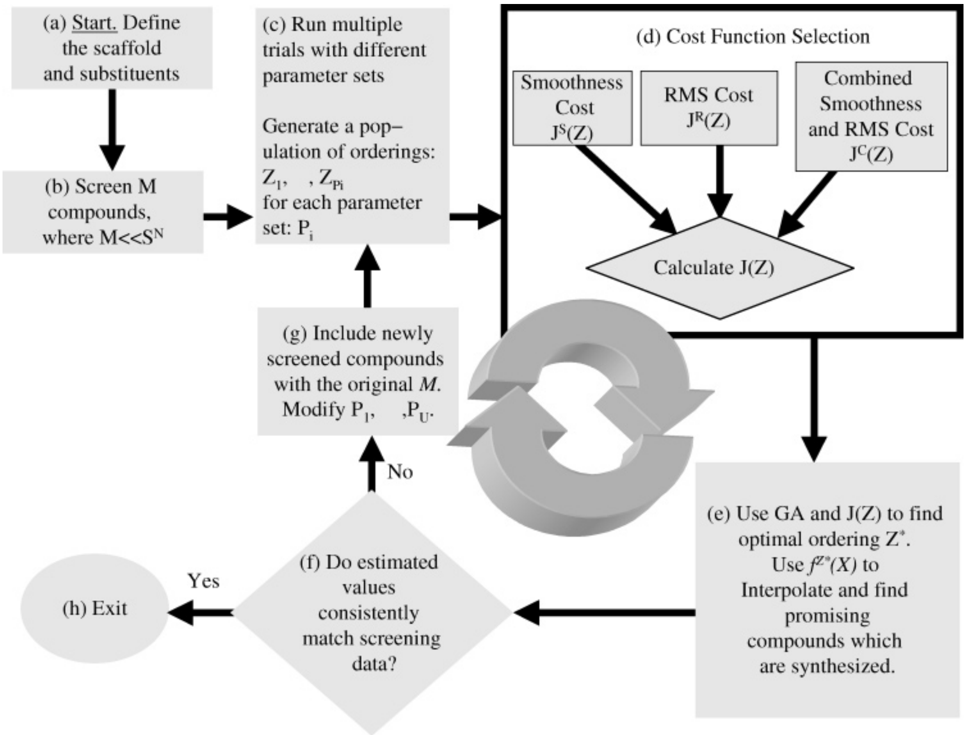
(a) Start. Define the scaffold and substituents
(b) Screen M compounds, where M<<S^N
(c) Run multiple trials with different parameter sets
Generate a population of orderings: Z_1, , Z_Pi for each parameter set: P_i
(d) Cost Function Selection
Smoothness Cost J^S(Z)
RMS Cost J^R(Z)
Combined Smoothness and RMS Cost J^C(Z)
Calculate J(Z)
(e) Use GA and J(Z) to find optimal ordering Z*. Use f^Z*(X) to Interpolate and find promising compounds which are synthesized.
(f) Do estimated values consistently match screening data?
Yes
No
(h) Exit
(g) Include newly screened compounds with the original M. Modify P_1, ,P_U.

Figure 1. After defining the scaffold and substituents (a) for an initial sampling set of M compounds (b), the reordering algorithm is implemented (c) using various trial parameter sets P_1, P_2, ..., P_U. Each of the U trials are associated with a random ordering $\mathbf{Z}$ and a cost $J(\mathbf{Z})$ that may take the form (d) of $J^S(\mathbf{Z})$, $J^R(\mathbf{Z})$, or $J^C(\mathbf{Z})$, representing the smoothness cost (eq 12), the RMS cost (eq 13), and the combined cost (eq 15), respectively. For each parameter set P_i, a genetic algorithm (GA) is employed (e) to find an optimal ordering $\mathbf{Z}^*$ that minimizes the chosen cost, and the properties $f^{\mathbf{Z}^*}(\mathbf{X})$ of all the unsampled compounds in the library are interpolated. On the basis of these predictions, a small number of additional targeted compounds are synthesized and their properties measured (f); the differences between the observed and the interpolated property values indicate the relative effectiveness of different parameter sets. The parameter sets with better interpolation quality are kept and/or properly modified to provide another round of property prediction (g) through $f(\mathbf{X})$, using all the available compounds (including the newly synthesized ones). This iterative process continues until in the final round, all newly synthesized compounds can be reliably estimated, and only marginal gains in the interpolation can be obtained from further iterations. At the latter point the algorithm is finished (h).

$\tilde{\mathbf{X}}_k$ is chosen to lie on one of the M sampled points $\{\tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2, ..., \tilde{\mathbf{X}}_M\}$, although this criterion can be lifted. Thus,

$$f(\mathbf{X}) = \sum_{k=1}^{B} c_k(|\tilde{\mathbf{X}}_k - \mathbf{X}|^2 + \beta)^{1/2} \tag{3}$$

where the coefficient vector $\mathbf{c} = [c_1, c_2, ..., c_k, ..., c_B]$ is found by linear regression at the M sampled points.

$$\min_{\mathbf{c}} \chi^2 \tag{4a}$$

with

$$\chi^2 = \sum_{j=1}^{M} [g(\tilde{\mathbf{X}}_j) - \sum_{k=1}^{B} c_k(|\tilde{\mathbf{X}}_k - \tilde{\mathbf{X}}_j|^2 + \beta)^{1/2}]^2. \tag{4b}$$

The regression analysis reduces to

$$\mathbf{c} = (A^T \cdot A)^{-1} A^T \cdot y, \tag{5}$$

where A is the design matrix of dimension $M \times B$, whose rows and columns are composed of basis functions evaluated at the M observed points and along the B basis centers,

$$A = \begin{Bmatrix} \phi_1(\tilde{\mathbf{X}}_1) & \phi_2(\tilde{\mathbf{X}}_1) & ... & \phi_B(\tilde{\mathbf{X}}_1) \\ \phi_1(\tilde{\mathbf{X}}_2) & \phi_2(\tilde{\mathbf{X}}_2) & ... & \phi_B(\tilde{\mathbf{X}}_2) \\ ... & & & \\ \phi_1(\tilde{\mathbf{X}}_M) & \phi_2(\tilde{\mathbf{X}}_M) & ... & \phi_B(\tilde{\mathbf{X}}_M) \end{Bmatrix} \tag{6}$$

and where y is the vector of M observed compound property values,

$$y = \begin{Bmatrix} g(\tilde{\mathbf{X}}_1) \\ g(\tilde{\mathbf{X}}_2) \\ ... \\ g(\tilde{\mathbf{X}}_M) \end{Bmatrix} \tag{7}$$

The form of the RBFs, the number of sampled points M, the number of basis functions B, and the associated RBF parameter β (or the form of the RBF) are all subject to optimization, which is discussed in section II.6.

To arrive at an optimal ordering $\mathbf{Z}^*$ that allows $f(\mathbf{X})$ to reliably interpolate $g(\mathbf{X})$, we search among the set $\tilde{\mathbf{Z}}$ of all possible orderings for one that minimizes a suitably defined cost function $J(\mathbf{Z})$ (see section II.2):

$$\min_{\mathbf{Z} \in \tilde{\mathbf{Z}}} J(\mathbf{Z}) \tag{8}$$

In this work, genetic algorithms (GAs) are used to minimize the cost $J(\mathbf{Z})$ in the search for $\mathbf{Z}^*$. GAs are a family of commonly used global search algorithms that are especially adept at optimizing mixed combinatorial problems in high dimensions.[12,13] Goldberg's GA template[17] was modified to find the $\mathbf{Z}^*$ that minimizes the cost $J(\mathbf{Z})$. The GA begins with a random population of genes which represent an assortment of unique arbitrary orderings. $J(\mathbf{Z})$ is then evaluated for each of the individual genes and used to direct the search for the optimal gene by the GA[1]. Individuals with the best fitness (i.e., the lowest $J(\mathbf{Z})$) have the highest probability of being propagated to the following generation of genes. Crossovers and mutations are carried out between the parent genes, resulting in a new generation of genes biased in favor of minimizing $J(\mathbf{Z})$. Crossovers involve selecting a Z_i for each of the N substituent sets at random; from either of the two parent genes at the ith site, parts of each ordering from the parent genes may be crossed over with a specified probability of occurrence. Mutations involve a swap of substituent assignments within a Z_i ordering also specified by a separate probability of occurrence.[1]

Through the process of sexual selection guided by the GA, the orderings with lower $J(\mathbf{Z})$ values stay within the gene pool, while those with higher $J(\mathbf{Z})$ values are slowly replaced by more competitive orderings. All genes are unique within each generation, but the diversity between individual genes tends to fall as only fit genes survive in the next generation. Typically, after a few thousand generations, the gene population becomes static in the upper echelon of fit genes; the algorithm converges, and the optimal ordering $\mathbf{Z}^*$ can be identified. Because of the stochastic nature of the GA, different runs may result in slightly different genes/optimized orderings. While finding the global minimum of $J(\mathbf{Z})$ would be ideal, it is not necessary; any ordering capable of giving reliable interpolation over the molecular library is generally an acceptable solution. As a point of speculation, an exception to this situation might arise in the case of molecular libraries among biological molecules (e.g., amino acid mutants of native proteins) associated with their biofunctional property where a unique underlying natural ordering might exist, leading to fundamental chemical/physical insights.

II.2. The Cost Function *J*(Z). The original algorithm[1] defined good orderings as ones which resulted in a smooth topology for $g(\mathbf{X})$. By using smoothness as an indicator of interpolation quality, Shenvi et al. assumed that a maximally smooth function $g(\mathbf{X})$, as represented by $f(\mathbf{X})$, best reflected the physical/chemical property estimation capabilities over the library. The smoothness cost, $J^S(\mathbf{X})$ used to direct the GA search, was based on the norm of the gradient of the function $f(\mathbf{X})$, although higher gradient measures could be used as well. This norm was expressed as the mean square partial derivatives along each substituent axis, summed over all R compounds $\mathbf{X}_1$, ..., $\mathbf{X}_l$, ..., $\mathbf{X}_R$ in the library:

$$J^S(\mathbf{Z}) = \frac{1}{R}\sum_{l=1}^{R}[\nabla f^{\mathbf{Z}}(\mathbf{X}_l)]^2 \tag{9}$$

$$= \sum_{l=1}^{R}\sum_{i=1}^{N}[\nabla_i f(\mathbf{X}_l)]^2 \tag{10}$$

Here, the superscript $\mathbf{Z}$ appears on $f^{\mathbf{Z}}(\mathbf{X}_l)$ to indicate that the interpolation will depend on the specific ordering; unless it is ambiguous, this label will generally be implicitly understood below. For $f(\mathbf{X})$ in eq 3, the gradient along the ith axis evaluated at the lth compound becomes

$$\nabla_i f(\mathbf{X}_l) \equiv \frac{\partial}{\partial X_i} f(\mathbf{X}_l) = -\sum_{k=1}^{B} c_k(|\tilde{\mathbf{X}}_k - \mathbf{X}_l|^2 + \beta)^{-1/2}(\tilde{X}_k^i - X_l^i), \tag{11}$$

where X_l^i is the integer value on the ith axis for the lth compound. Combining eqs 10 and 11 produces,

$$J^S(\mathbf{Z}) = \frac{1}{R}\sum_{l=1}^{R}\sum_{i=1}^{N}[-\sum_{k=1}^{B} c_k(|\tilde{\mathbf{X}}_k - \mathbf{X}_l|^2 + \beta)^{-1/2}(\tilde{X}_k^i - X_l^i)]^2, \tag{12}$$

which may be minimized over $\mathbf{Z}$ to find the smoothest function $f(\mathbf{X})$ for effective interpolation over the library.

The cost $J^S(\mathbf{Z})$ may be viewed as an indirect measure of subsequent interpolation quality, and as an alternative we introduce the direct measure as the root-mean-squared (RMS) difference between the actual function $g(\mathbf{X})$ and the interpolated function $f(\mathbf{X})$ at the M sampled points, $\{\tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2, ..., \tilde{\mathbf{X}}_M\}$.

$$J^R(\mathbf{Z}) = \left[\frac{1}{M}\sum_{j=1}^{M}\frac{(g(\tilde{\mathbf{X}}_j) - f^{\mathbf{Z}}(\tilde{\mathbf{X}}_j))^2}{(g(\tilde{\mathbf{X}}_j))^2}\right]^{1/2} \tag{13}$$

$$= \left[\frac{1}{M}\sum_{j=1}^{M}\frac{(g(\tilde{\mathbf{X}}_j) - \sum_{k=1}^{B} c_k(|\tilde{\mathbf{X}}_k - \tilde{\mathbf{X}}_j|^2 + \beta)^{1/2})^2}{(g(\tilde{\mathbf{X}}_j))^2}\right]^{1/2} \tag{14}$$

Using RMS as the cost function requires that $f(\mathbf{X})$ be determined from $g(\mathbf{X})$ (eqs 4−6) using T sampled compounds, where $T \leq M$, and employing B basis functions, where $B \leq T$. All M data values, including the $M - T$ not used in the regression to find the coefficient vector $\mathbf{c} = [c_1, c_2, ..., c_k, ..., c_B]$, are used in eq 14 to evaluate the capacity of $f(\mathbf{X})$ to interpolate $g(\mathbf{X})$. Note that the value of $J^R(\mathbf{Z})$ in eq 12 will generally be nonzero, even if $T = M$, when the system is overdetermined (i.e., $B < T$). The GA then finds the optimal ordering $\mathbf{Z}^*$ that minimizes the residual $J^R(\mathbf{Z})$ between interpolated and measured values for $g(\mathbf{X})$ at $\{\tilde{\mathbf{X}}_{1,}\tilde{\mathbf{X}}_{2,}..., \tilde{\mathbf{X}}_M\}$.

Last, a third cost function $J^C(\mathbf{Z})$ which combines $J^S(\mathbf{Z})$ and $J^R(\mathbf{Z})$, may be used to exploit the selective advantages from both smoothness- and RMS-based cost functions. The form of $J^C(\mathbf{Z})$ is flexible and can be subject to optimization. In this work, we used

$$\mathrm{J}^C(\mathbf{Z}) = [d_1 + J^S(\mathbf{Z})][d_2 + J^R(\mathbf{Z})], \tag{15}$$

where d_1 and d_2 are positive constants. Optimizing d_1 and d_2 guarantees equal, if not better, property interpolation with a smaller set of synthesized compounds M than that for $J^S(\mathbf{Z})$ or $J^R(\mathbf{Z})$ alone (i.e., if either d_1 is zero and d_2 large, or if d_2 is zero and d_1 large, then the combined cost collapses to $J^S(\mathbf{Z})$ or $J^R(\mathbf{Z})$, respectively). The practical goal is to find optimized values for d_1 and d_2, which is addressed in section II.6. Here, T for $J^R(\mathbf{Z})$, is set equal to M, and $B < T$, so that a common coefficient vector $\mathbf{c}$, and hence $f(\mathbf{X})$, can be used for evaluating both the RMS and smoothness component costs in $J^C(\mathbf{Z})$.

One advantage of employing $J^R(\mathbf{Z})$ or $J^C(\mathbf{Z})$ is the ability to automatically optimize algorithmic parameters such as the basis set size, B, and the basis parameter, β, during the GA search. Optimizing B or β is not directly feasible when using $J^S(\mathbf{Z})$ because treating these parameters as variables for optimization by the GA results in effectively flat $f(\mathbf{X})$ surfaces that are uncontestedly smooth, but useless for interpolation (e.g., when the number of RBFs is driven to $B^* = 1$ or when the RBF parameter β^* is extremely large). Automated searching for B^* and β^* using any form of RMS-based cost (e.g., $J^R(\mathbf{Z})$ and $J^C(\mathbf{Z})$) is possible because RMS considerations protect the GA from selecting extreme unphysical values for B^* and β^*.

II.3. Regularization Techniques. The two parameters B and β are only a portion of the entire set of variables that can affect the interpolative quality of $f(\mathbf{X})$. Ultimately, the choice of parameters needs to ensure that $f(\mathbf{X})$ can adequately represent the general structure of the substituent space without overfitting the scattered sampling data, especially in cases with significant data error. Parameter optimization for function representation is a regularization operation; parameters are sought that allow for the greatest regularity in $f(\mathbf{X})$ without sacrificing interpolation quality. Three alternative regularization methods are given below for determining the coefficient vector **c**.

1. RBF Modifications. Reducing the number, B, of RBFs and/or increasing β will generally lead to a smoother function $f(\mathbf{X})$. In addition, changing the type of RBF from multiquadrics to Gaussians or to other suitable functions may also influence interpolation quality.

2. Singular Value Decomposition (SVD). SVD is a standard regularization technique that increases the robustness of the interpolation with $f(\mathbf{X})$ by removing offending bases that are nearly linearly dependent. SVD factors the design matrix A in the linear least squares inverse problem so that

$$\mathbf{c} = V\Sigma^{-1}U^{T}y = \sum_{k}^{B} \frac{1}{\sigma_k}(U_k \cdot y)V_k \tag{16}$$

where U and V are orthonormal matrixes, with columns $k = 1, 2, ..., B$, Σ is a diagonal matrix containing the singular values $\sigma_1, ..., \sigma_k$, and y is the vector in eq 7 containing measurement values for the M sampled compounds.[14] Small singular values which cause unstable behavior in determining **c** are excised from eq 2. Consequently, overfitting is avoided, and $f(\mathbf{X})$ is less sensitive to noise in the input data.

3. Introduction of a Regularization Term. We may require that $f(\mathbf{X})$ be reasonably smooth by adding a gradient cost to the regression in eqs 4−6 to calculate **c** in $J^{R}(\mathbf{Z})$,

$$\min_{\mathrm{c}}\left\{\left[\sum_{m=1}^{T}\left(\frac{g(\tilde{\mathbf{X}}_m) - \sum_{k}^{B} c_k\phi_k(\tilde{\mathbf{X}}_m)}{g(\tilde{\mathbf{X}}_m)}\right)^2\right] + \Omega\left[\sum_{l=1}^{R}\sum_{i=1}^{\mathrm{N}}\left(\sum_{k=1}^{B} c_k \frac{\partial\phi_k(\tilde{\mathbf{X}}_l)}{\partial\tilde{X}_i}\right)^2\right]\right\} \tag{17}$$

where Ω is the weight associated with the smoothness regularization term. Note that this introduction of smoothness is distinct from the notion introduced for $J^{S}(\mathbf{Z})$, which used smoothness as a measure for optimizing the substituent ordering at each scaffold site. Equation 17 has also introduced the normalization $g(\tilde{\mathbf{X}}_m)$ in the denominator of the first term such that relative error is used as the metric.

The best choice among these regularization techniques and their associated parameters is application dependent, and this feature can be exploited by using the adaptive methodology presented in section II.6.

II.4. Considerations of Data Errors. The treatment of data error becomes a central issue in practical applications. Early efforts at combinatorial chemistry focused largely on rapid synthesis and screening. Although these methods were expedient, the error associated with the screened data values was often large. Current emphasis has increasingly moved toward obtaining better quality synthesis and screening. Nevertheless, the influence of data error generally always needs to be considered when the compound properties are used for interpolation.

The input data errors are factored into the regression of eq 4 by introducing a weight ω_j,

$$\omega_j = \frac{1}{\delta_j} \tag{18}$$

where δ_j is the observed or estimated standard deviation of $g(\tilde{\mathbf{X}}_j)$. Thus, the cost function determining c is now

$$\min_{\mathbf{c}}\left[\sum_{j=1}^{M,T}\omega_j\left(\frac{g(\tilde{\mathbf{X}}_j) - \sum_{k}^{B} c_k\phi_k(\tilde{\mathbf{X}}_j)}{g(\tilde{\mathbf{X}}_j)}\right)^2\right], \tag{19}$$

where the sum goes to M when utilizing $J^{S}(\mathbf{Z})$ or $J^{C}(\mathbf{Z})$ as the cost and to T when $J^{R}(\mathbf{Z})$ is used as the cost.

The variance of the interpolated property for the lth compound in the library is $\tilde{\delta}^2(f(\mathbf{X}^l))$. The variance may be expressed as

$$\tilde{\delta}_l^{\,2} \equiv \tilde{\delta}^2(f(\tilde{\mathbf{X}}^l)) = \sum_{k}^{B}\sum_{k'}^{B} \delta^2(c_k,c_{k'})\phi_k(\tilde{\mathbf{X}}_l)\phi_{k'}(\tilde{\mathbf{X}}_l), \tag{20}$$

where $\delta^2(c_k,c_{k'})$ is the covariance of coefficients c_k and $c_{k'}$[14]

$$\delta^2(c_k,c_{k'}) = \sum_{j}^{M}(\delta_j)^2\left(\frac{\partial c_k}{\partial y_j}\right)\left(\frac{\partial c_{k'}}{\partial y_j}\right), \tag{21}$$

and δ_j^2 is the variance of the jth property observation. A derivation of the error propagation formulation is included in the Appendix for the more general case incorporating the regularization term in eq 3.

The treatment above introduced data error at the level of determining the vector **c** and later, at the end of the algorithm in estimating the interpolated property standard deviation $\tilde{\delta}_l$. The latter standard deviation could also be used to provide a degree of resolution for $f^{\mathbf{Z}}(\mathbf{X})$ when deciding on whether one ordering **Z** is better than another **Z′** The present formulation of the algorithm did not employ this feature.

II.5. Substituent Rescaling. There is no requirement that the substituent orderings need to derive from permutations on a set of consecutive integer numbers as introduced in section II.1. The ordering $Z_i = [Z_{i1}, Z_{i2}, ..., Z_{iS_i}]$, for the ith substituent set $\mathbf{R}_i$ may just as reasonably be constructed from permutations of nonconsecutive integer values, in which case substituents $\{R_{i1}, R_{i2}, ..., R_{iS_i}\}$ could have breaks in the uniform interval placement along the ith scaffold axis. This added flexibility of rescaling the distance between substituents in the N-dimensional library space can provide for better interpolation of $g(\mathbf{X})$. In addition, rescaling may allow greater insight into the physicochemical relationship between substituents available at one scaffold site and between substituents allowed on other scaffold sites. To illustrate this point for the ith scaffold site, consider the case where the molecular property associated with R_{i1} is more similar to that of R_{i4} than R_{i4} is to R_{i3} and the optimal ordering placed them in the sequence R_{i1}, R_{i4}, R_{i3} adjacent one another. Under these conditions, it is reasonable that the interval between R_{i1} and R_{i4} along the ith axis be smaller than that between R_{i4} and R_{i3} to allow for better interpolation.

The rescaling and reordering operations may be carried out either sequentially or simultaneously during execution of the algorithm. Under sequential operations, the reordering step would first find an optimal substituent assignment **Z***. Next, the integer space along the *i*th axis for placement of a substituent is increased by an integer magnifying factor of $G_i > 1$ units. Thus, the overall library space is expanded by $g = \Pi(\langle i \rangle, \langle N \rangle, \langle G_i \rangle)$. Within this augmented space, the substituents are allowed to slide along their respective axis to any integer position bounded between their neighbors. A second GA operating with one of the cost functions given in section II.2 is used to find the best scaling of the substituents to produce the rescaled ordering $\mathbf{Z}^{*\ddagger}$. Note that while this sequential method does not change the relative substituent positioning found by the first reordering step, the overall process would be repeated on each cycle of the algorithm seeking the minimum value of $J(\mathbf{Z})$.

The second method would perform the reordering and rescaling operations simultaneously. Starting from the population of randomly initialized genes/orderings, a single GA is used to find an ordering that minimizes a cost chosen from section II.2. However, the search space here is *g*-fold larger than the original $\Pi_i^N S_i$ compound space as each substituent axis is expanded by a factor of G_i. Substituents along one scaffold site can assume any position among the $S_i \times G_i$ possibilities. Thus, rescaling and reordering are accomplished simultaneously as the intervals between substituents are not limited while substituent reordering is also allowed to take place.

Besides aiding library interpolation, the rescaling operation may also identify the existence of a poorly designed library. Large intervals between optimally reordered and rescaled substituents may indicate that that region of the compound library had not been adequately sampled by synthesis. In the extreme case, the library may be grouping highly disparate functional groups together that should actually belong in separate libraries.

II.6. Adaptive Algorithm for Molecular Library Interpolation. Sections II.1–II.5 presented the basic concepts of the molecular discovery algorithm. In practical applications, it may be difficult a priori to determine the algorithmic parameters that optimally enhance interpolation quality in any particular case. These parameters include the number of sampled points *M*, the number, type, and associated parameters of the RBFs, cost function choices, choice of regularization technique and associated parameters, and rescaling distances. An adaptive procedure is prudent for optimizing these parameters for high-accuracy interpolation by $f(\mathbf{X})$. The adaptive procedure would efficiently utilize information from laboratory data to guide the choice of algorithmic parameters while enhancing interpolation quality. Similar algorithms have been successfully employed in other areas, including the optimal identification of quantum mechanical systems and biochemical reaction networks.[15,16]

Figure 1 shows the general components of the adaptive algorithm. A minimal set of *M* compounds $\{\tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2, ..., \tilde{\mathbf{X}}_M\}$ is first synthesized and assayed. Next, the algorithm is run with multiple parameter sets P_1, P_2, ..., P_U where each set defines the bases, cost function, etc. Once the optimal orderings $\mathbf{Z}_1^*$, $\mathbf{Z}_2^*$, ..., $\mathbf{Z}_U^*$ have been found from each trial, a suitable number of promising compounds can be identified from their respective interpolated property values and synthesized. Thus, the aptitude of $f^{\mathbf{Z}_1^*}(\mathbf{X})$, $f^{\mathbf{Z}_2^*}(\mathbf{X})$,..., $f^{\mathbf{Z}_U^*}(\mathbf{X})$ to predict property values of the newly synthesized compounds can be evaluated. The parameter sets that give better interpolation can be kept or suitably modified and applied to the next iterative cycle, where the additionally synthesized compounds are added to the original *M* sample set. This iterative approach would continue until an accurate interpolation is achieved consistently for the newly synthesized compounds, such that further optimization of *P* has marginal benefits.

Implementing this adaptive algorithm calls for interfacing with laboratory synthesis and assaying. The present work aims to lay the foundations for such interfacing efforts. The illustrations below in section III will demonstrate the influence of algorithmic parameter changes on interpolation quality, and the observed behaviors strongly suggest that employing the adaptive algorithm will be beneficial in practice.

For an initial sampling set of M compounds, the reordering algorithm is implemented using various trial parameter sets P_1, P_2, ..., P_U. Each of the *U* trials is associated with a random ordering $\mathbf{Z}$ and a cost $J(\mathbf{Z})$ that may take the form of $J^S(\mathbf{Z})$, $J^R(\mathbf{Z})$, or $J^C(\mathbf{Z})$, representing the smoothness cost (eq 12), the RMS cost (eq 13), and the combined cost (eq 15), respectively. For each parameter set P_i, a genetic algorithm (GA) is employed to find an optimal ordering $\mathbf{Z}^*$ that minimizes one of the costs, and the properties $f^{\mathbf{Z}^*}(\mathbf{X})$ of all the unsampled compounds in the library are interpolated. On the basis of these predictions, a small number of additional targeted compounds are synthesized and their properties measured; the differences between the observed and the interpolated property values indicate the relative effectiveness of different parameter sets. The parameter sets with better interpolation quality are kept and/or properly modified to provide another round of property prediction, using all the available compounds (including the newly synthesized ones). This iterative process continues until, in the final round, all newly synthesized compounds can be reliably estimated and only marginal gains in the interpolation can be obtained from further iterations.

III. Illustrations

A series of numerical experiments using available laboratory data were run to test the algorithmic advances in section II and the effects of parameter changes on the resultant library property interpolation quality. The efficiency and applicability of the cost function alternatives ($J^S(\mathbf{Z})$, $J^R(\mathbf{Z})$, and $J^C(\mathbf{Z})$) were studied during trials with a transition metal complex library.[9] This latter case provided a partial test of the fully iterative algorithm in Figure 1, as new compounds were synthesized based on initial runs, and the outcome of the algorithm was also shown to be capable of revealing chemical insights about the library members. The influence of algorithmic parameters on interpolation, rescaling, and data error was explored with laboratory data from a copolymer library.[8] In some cases the latter data were modified to help illustrate a specific point or capability of the algorithm, when no suitable data for that purpose were readily available.

The transition metal complex library is based on Ir(III) complexed with cyclometalating and bis-diimine ligands. The library of compounds, derived from a single molecular scaffold, were synthesized to discover practical electroluminophores for organic light-emitting diode fabrication.[9] The two scaffold sites ($N = 2$) and 10 available substituents ($S = \{10, 10\}$) at either site yield a whole library of 100 compounds. The entire library was synthesized and screened for high photoluminescent quantum yields (ϕ) and high emission energies (λ).

The copolymer library consists of a single molecular backbone with two scaffold sites ($N = 2$). The first site accommodates one of 14 diol substituents, while the second allows one of 8 diacid substituents ($S = \{14, 8\}$). The entire library of 8 × 14 = 112 compounds was synthesized and screened for the glass transition temperature (T_g) and hydrophobicity, as measured by the air–water contact angle (CA). Reynolds used quantitative structure–property relationships (QSPR) to isolate key descriptors in predicting CA and T_g values across the library.[8] The smoothness-based reordering algorithm was applied

TABLE 1: Algorithmic Parameters and RMS Values across the Entire Library Space for the Transition Metal Compounds with the Emission Energy Data[a]

	smoothness cost	RMS cost	combined cost
RMS ($\mathbf{Z}^{\mathrm{Rand}}$)	0.078	0.144	0.080
RMS ($\mathbf{Z}^*$)	0.027	0.029	0.029
$\mathbf{Z}^* = \{ Z_1$	2, 1, 4, 8, 3, 10, 9, 5, 6, 7	6, 9, 3, 2, 4, 8, 1, 10, 5, 7	7, 6, 5, 9, 10, 3, 8, 2, 4, 1
Z_2	5, 7, 1, 6, 3, 2, 4, 8, 10, 9	4, 2, 3, 7, 5, 1, 6, 8, 10, 9	4, 2, 3, 6, 7, 1, 5, 8, 10, 9
B	30^{F}	9^{GA}	29^{GA}
M	30	30	30
T	N/A	10	30
β	1^{F}	50^{F}	100^{F}
Ω	0^{F}	0^{GA}	0^{GA}
$J(\mathbf{Z}^*)$	0.005	0.010	0.101
smoothness	0.005	0.007	0.008
generations	146	564	746

[a] Superscripts of F or GA for B, β (eq 2), and Ω (eq 17) values indicate whether the value was fixed or whether the value was optimized by a GA. $J^C(\mathbf{Z})$ was run using $d_1 = 1$ and $d_2 = 0.1$ and "generations" indicates the number of generations needed for the optimal ordering $\mathbf{Z}^*$ to converge. $\mathbf{Z}^{\mathrm{Rand}}$ refers to a random ordering, and RMS is with respect to the entire library of compounds. The smoothness $J^S(\mathbf{Z}^*)$ of each optimal surface was calculated, regardless of whether smoothness was used to guide the algorithm.

to the T_g data set, producing similar interpolation quality.[1] In this paper, we use the CA data as another illustration of the efficacy of the reordering algorithm.

The GA,[17] employed to find $\mathbf{Z}^*$ in all the following tests, used a population of 100 genes/orderings. Thirty-five of the most fit (i.e., those having the lowest cost $J(\mathbf{Z})$) genes were maintained after each generation, and the rest were replaced by new genes synthesized under 80% crossover and 50 mutation probabilities. The GA ran for a maximum of 2000 generations, which sufficed to allow the genes with the best $J(\mathbf{Z})$ values to converge (convergence often occurred at far fewer generations as illustrated later in Table 1).

III.1. Cost Function Comparisons. The cost function alternatives $J^S(\mathbf{Z})$, $J^R(\mathbf{Z})$, and $J^C(\mathbf{Z})$ were tested on both the emission energy λ of estimated error ~10% and quantum yield ϕ data of estimated error ~15−20% from the transition metal library. Much of the analysis was placed on emission energy data. The optimal parameter set was cost function dependent, and repeat trials with different parameter values were carried out to find the optimal orderings $\mathbf{Z}^{*S}$, $\mathbf{Z}^{*R}$, and $\mathbf{Z}^{*C}$ that gave the best interpolation. The quality of the interpolation was measured by calculating the RMS across the entire library, which was only used for testing purposes and would not be needed in real applications when only partially screened libraries are available (i.e., in application of the full adaptive algorithm in Figure 1). The interpolation quality of $f^{\mathbf{Z}^*}(\mathbf{X})$ was determined by comparing it to $g(\mathbf{X})$ in a so-called truth plot of one outcome vs the other (cf., Figure 3).

Once the algorithmic parameters were optimized for each cost function variant, we found that comparable interpolation quality was attained by $f^{\mathbf{Z}^{*S}}(\mathbf{X})$, $f^{\mathbf{Z}^{*R}}(\mathbf{X})$, and $f^{\mathbf{Z}^{*C}}(\mathbf{X})$, using the same subset of M sampled compounds. A randomly selected set of $M = 30$ compounds was approximately the minimum sample size needed to attain good interpolation for the emission energy data. The RMS values improved significantly from initial random orderings (RMS^S = 0.078, RMS^R = 0.144, RMS^C = 0.080, where the RMS values are different because different optimized parameters were used for each cost function evaluation) to the optimized orderings $\mathbf{Z}^*$ (RMS^S = 0.027, RMS^R = 0.029, RMS^C = 0.029), and similar improvements were observed for trials involving quantum yield data as well. As an example, for trials executed using the newly proposed cost J^R($\mathbf{Z}$), the topographic surface in Figure 2 interpolated from $f(\mathbf{X})$ using a random ordering produced the truth plot in Figure 3. The quality of the interpolation in Figure 3 from the random ordering is clearly unsatisfactory. The optimized ordering $\mathbf{Z}^*$ produced the topographic surface $f^{\mathbf{Z}^*}(\mathbf{X})$ in Figure 4, which is distinctly different from that of the random ordering in Figure 2. The corresponding truth plot for the optimal ordering is shown in Figure 5. The truth plots in Figures 3 and 5 show that

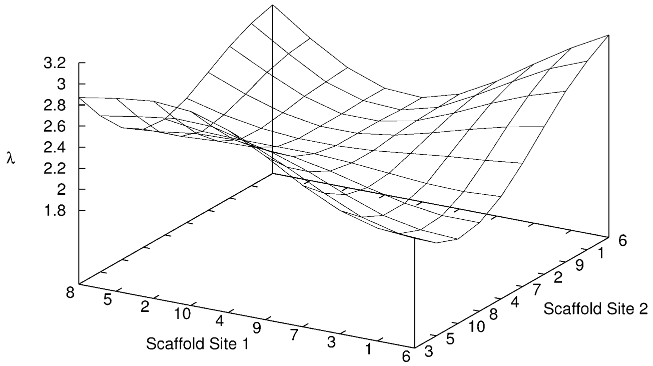

Figure 2. Interpolated surface $f(\mathbf{X})$ for the emission energy λ (eV) data of the transition metal complex library, generated from a random ordering $\mathbf{Z}$ shown by the substituent labels on each site.

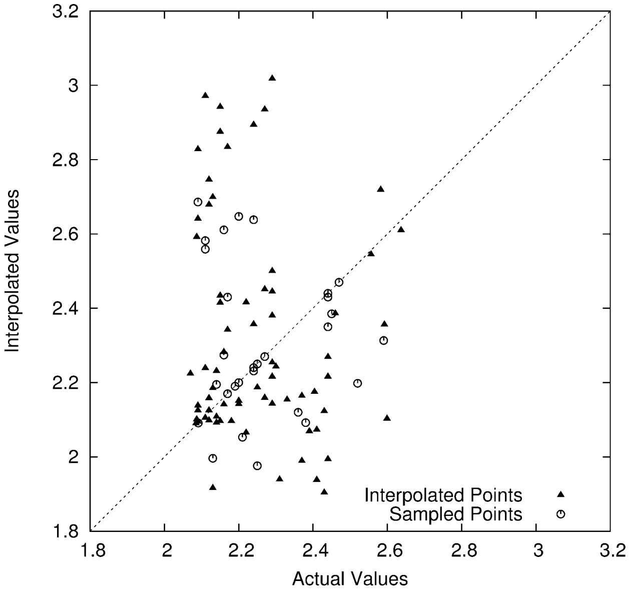

Figure 3. Truth plot for the actual emission energy values $g(\mathbf{X})$ against the interpolated values $f(\mathbf{X})$ generated with the random ordering $\mathbf{Z}$ in Figure 2. The $M = 30$ sampled compounds are represented by open circles, which lay close to the line $f(\mathbf{X}) = g(\mathbf{X})$ denoting perfect interpolation.

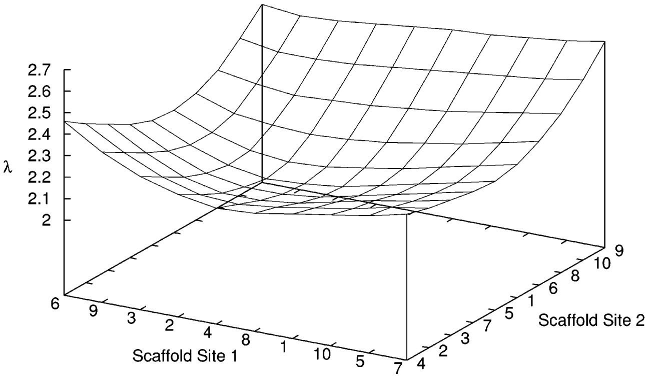

Figure 4. Interpolated surface $f(\mathbf{X})$ generated from the optimal ordering for the emission energy data λ (eV) of the transition metal complex library using $J^R(Z)$ as the cost function. The optimal ordering is indicated on each site axis, as compared to the random ordering in Figure 2. The algorithmic parameters are listed in Table 1.

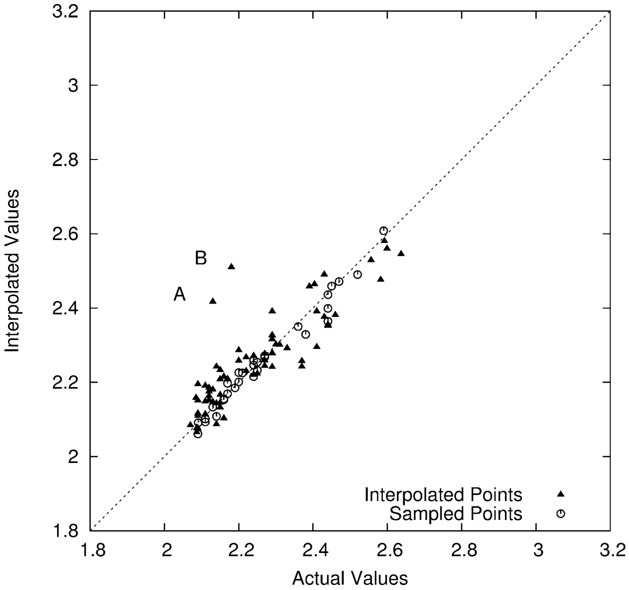

Figure 5. Truth plot of $g(\mathbf{X})$ against $f^{\mathbf{Z}^*}(\mathbf{X})$ values from the surface in Figure 4 using the optimal ordering $\mathbf{Z}^*$. The two outliers labeled as A and B are correctly identified by the algorithm as being physically distinct from the other library members, as explained in the text. The scatter in the estimated property is within the experimental error. The results should be compared to the truth plot of Figure 3, shown on the same scale, that uses a random ordering, and it is evident that the reordering algorithm greatly improved the estimated property values.

interpolation quality improves markedly after the optimal ordering $\mathbf{Z}^*$ is found. Table 1 features the orderings $\mathbf{Z}^{*S}$,$\mathbf{Z}^{*R}$, and $\mathbf{Z}^{*C}$ as well as the algorithmic parameters under which the orderings were found. Note that all three costs converge to similar optimal orderings, including that using smoothness and combined costs create nearly identically structured spaces although mirror images of one another in $\mathbf{Z}_1^{*S}$ and $\mathbf{Z}_1^{*C}$. The similar $\mathbf{Z}^*$'s suggest that these optimal orderings reflect some intrinsic physicochemical pattern relating the substituents themselves. We also observe (a) that a smoothness driven cost $J^S(\mathbf{Z}^*)$ finds a similar ordering pattern as an RMS driven cost $J^R(\mathbf{Z})$, and (b) the $f^{\mathbf{Z}^{*R}}(\mathbf{X})$ surface in Figure 4 is quantitatively smooth (i.e., the smoothness index, defined as the postfacto determined cost in eq 10, has the value of 0.007 vs that of 0.005 for $f^{\mathbf{Z}^{*S}}(\mathbf{X})$), although determining the cost $J^R(\mathbf{Z}^*)$ does not directly involve smoothness. Both points further support the assumption that a smooth property function $f(\mathbf{X})$ correlates with improved interpolability, while also showing that an RMS-based cost is effective as well. In practical applications involving the iteration in Figure 1, it is anticipated that when the smoothness cost is effective it can be more efficient as the procedure does not require using a subset of data to guide the determination of $\mathbf{Z}^*$.

Two evident interpolation outliers, labeled A and B, can be seen in Figure 5, where compound A contains ligands [ppy,dppZ] ([bis-(2-phenylpyridine)dipyridophenazine-iridium-(III)]chloride) and compound B contains [thpy,dppE] ([(bis-(diphenylphosphino)ethene)-bis-(2-thienylpyridine)-iridium(III)]-chloride). The outliers were originally attributed to possible laboratory measurement error, and this was the case for dppE compounds, whose weak emissions made accurate measurements difficult. However, an ensuing analysis also revealed that compounds containing dppE ligands experience a one-step emission decay process rather than the two-step emission pathway characteristic of all other compounds in the library.[9] Both compounds A and B, and other compounds containing either dppE or dppZ, consistently showed up as outliers in many other library data interpolations, including those using quantum yield ϕ data. Removing all the dppE compounds ameliorated interpolation quality, lowering the RMS associated with the best ordering from ~0.030 to ~0.025, but removing dppZ data did not contribute to a substantial improvement in interpolation. By removing the dppE-containing compounds, the library evidently becomes more specific to a class of compounds with closely related properties and thereby allows interpolation that better resolves the space. This case illustrates the capability of the reordering algorithm operating without any knowledge of descriptor sets to uncover incongruous substituents in the library design that derive from distinct physicochemical discrepancies.

III.2. Error Analysis. Equation 20 was used to calculate error propagated from the M laboratory data points valued at $g(\tilde{\mathbf{X}}_i)$ $\pm$ δ_i, where $i = 1, 2, ..., M$ and δ_i is the standard deviation of the ith data point. For illustration here, the CA values of the copolymer library will be used as data. The interpolated CA values are given as $f(\mathbf{X}_j) \pm \tilde{\delta}_j$, where $j = 1, 2, ..., R$, and $\tilde{\delta}_j$ is the standard deviation of the estimated property of the jth compound. The error weighting factor ω_i in eq 14 was defined as $1/\delta_i$ in all the analyses, although other weights can also be used to scale the data. The actual CA data[8] is estimated to have an error of $\delta_i = 1\%$, and in this case the analysis from eq 20, with $M = 35$ and $B = 30$ gives an interpolation error $\tilde{\delta}_j \simeq 1\%$ rather uniformly over the full library showing that there is no magnification of the input data error.

Data of such high quality with $\tilde{\delta}_j \simeq 1\%$ is rare, and the behavior of the calculated confidence intervals of $f(\mathbf{X})$ under other circumstances was explored by giving artificial error estimates to the M measurements of the CA property. A single optimal ordering $\mathbf{Z}^*$, compound sampling set ($M = 40$), and basis set ($B = 30$) was used for all the trials. The sensitivity to data error was tested using various estimated error distributions, including: (a) uniform 10 measurement error, and (b) uniform 10% error with the exception of one randomly selected compound whose error was inflated to 200%. Error propagation for 10% uniform data error resulted in confidence intervals for $f(\mathbf{X})$ on the order of ~10−15% across the unsynthesized compounds. The lack of error amplification is significant for practical property estimation. The second trial showed that the data point with grossly inflated error was effectively eliminated in the regression process in eq 17 due to the weight factor ω_i. In this trial, the standard deviation $\tilde{\delta}_j$ for $f(\mathbf{X}_j)$ was nearly identical to the situation where the high error data was actually removed from the M compounds. The region closely neighboring the high error datum did have slightly inflated values of $\tilde{\delta}_j \approx 15\%$, in contrast to $\tilde{\delta}_j \approx 10\%$ for the case where the datum was simply eliminated. In the opposite extreme of setting ω_i to be a constant value, the interpolation quality will be significantly influenced by the single datum of 200% error. These results suggest that employing the weight $\omega_i \approx 1/\delta_i$ can properly manage the problem of errant data values.

While the confidence intervals are reasonable in these trials, it is possible to resolve the space further by employing the regularization term in eq 17. This added smoothness criteria, with the weight coefficient Ω, can aid in the management of data with significant errors. When the smoothness regularization term was applied with $\Omega = 0.001$ to the case of uniform 10% error under identical conditions to those given above, interpolation quality was marginally improved (RMS = 0.031 for $\Omega = 0$, RMS = 0.029 for $\Omega = 0.001$), and the error estimate for $f(\mathbf{X})$ decreased slightly (i.e., the mean standard deviation was $\tilde{\delta} = 10\%$ for $\Omega = 0$ and $\tilde{\delta} = 9.5\%$ for $\Omega = 0.001$). The gains were somewhat larger for trials using increased input data error. For uniform 20% error, the RMS decreased from 0.031 for $\Omega = 0$ to 0.028 for $\Omega = 0.001$; the error estimates of $f(\mathbf{X})$ likewise diminished, this time from a mean value of $\tilde{\delta} = 20\%$ for $\Omega = 0$ to $\tilde{\delta} = 18\%$ for $\Omega = 0.001$.

The regularizing smoothness term functions to screen the data noise, and thereby prevents $f(\mathbf{X})$ from overfitting the data. While the gains from using regularization are apparent, the efficacy of the technique is sensitive to the weight parameter Ω: values of Ω that are too large will disregard the structure of $g(\mathbf{X})$ and generate an overly smooth surface $f(\mathbf{X})$ that has poor interpolative capacity; however, values of Ω that are too small will overfit $g(\mathbf{X})$. A systematic methodology of determining Ω is proposed in section III.4.

III.3. The Rescaling Technique. The effectiveness of rescaling was tested with both the sequential reordering/rescaling and simultaneous reordering/rescaling forms of the algorithm, introduced in section II.5, on the CA data of the copolymer library. From the total library of 112 compounds a sample of 35 was randomly selected as data, and $J^S(\mathbf{Z})$ was employed as the cost. The maximum possible number of RBFs ($B = M = 35$) was used, and the GA crossover and mutation probabilities were increased to 85% to compensate for the enlarged search space with rescaling.

To clearly demonstrate the rescaling technique, we artificially altered the data to model a situation with an irregular surface, because the original library is sufficiently interpolated with reordering alone, such that adding in rescaling had only a marginally improved effect. In practice such surface irregularities may arise in instances of poor library design, where grossly dissimilar compounds are combined in a single library without sufficient transitional compounds to relate the functional groups. Three separate trials were run in which an artificial step was introduced to the CA property of half of the compounds. This value inflation was performed to an optimal sequence $\mathbf{Z}^*$, and all the compound property values associated with the first four diacids in $\mathbf{Z}^*$ were additively increased by units of 10, 30, and 90, respectively, for the three examples. Randomly inflating compound values without taking $\mathbf{Z}^*$ into account would have destroyed the expected situation in a real library, where the underlying regularity for $g(\mathbf{X})$ exists and is disrupted by a substituent data gap. For this modified library, rescaling significantly enhanced interpolation quality. The RMS values improved for all three cases under both the sequential and simultaneous reordering/rescaling methods.

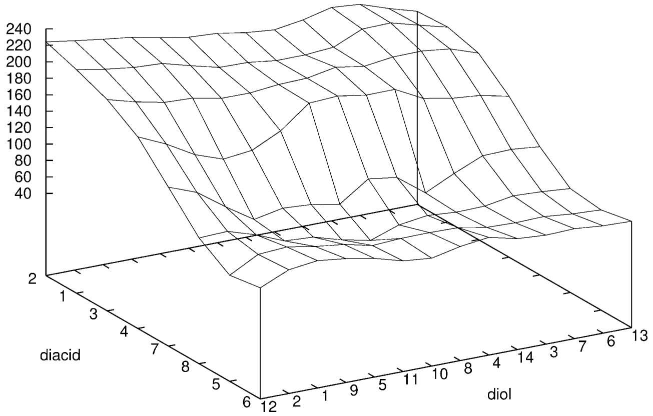

Figure 6. CA property $f(\mathbf{X})$ interpolated from a random ordering $\mathbf{Z}^* = \left\{\begin{matrix} Z_1\text{: } 2, 1, 3, 4, 7, 8, 5, 6 \\ Z_2\text{: } 12, 2, 1, 9, 5, 11, 10, 8, 4, 14, 3, 7, 6, 13 \end{matrix}\right\}$ and using the artificially gapped data of 90 units for the diacid substituents 1−4. The smoothness cost is $J^S(\mathbf{Z}^*) = 900$.

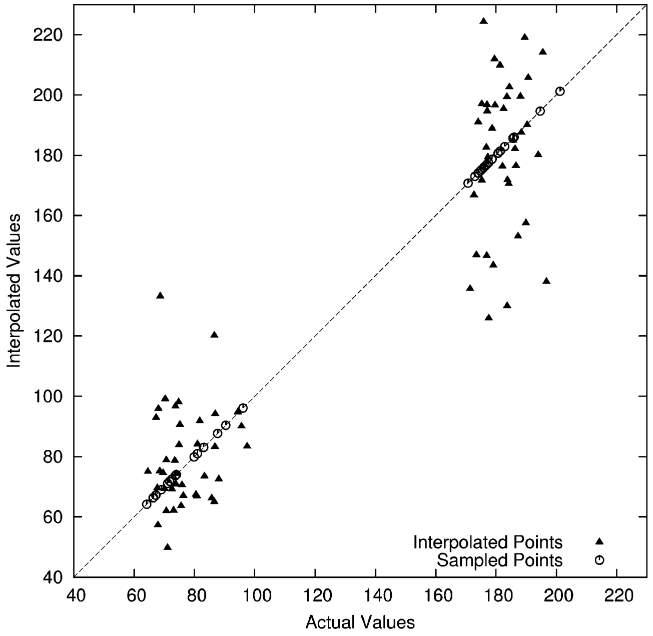

Figure 7. Truth plot for $f(\mathbf{X})$ using the random ordering in Figure 6. The RMS value is 19.

Here, we detail the sequential reordering/rescaling technique in the case of the 90 unit gapped library (Figures 6−8). In Figures 6 and 7, interpolation from an initial random gene (ordering) population gave an RMS of 19.4, and a $J^S(\mathbf{Z})$ value of 900. The data gap is very evident in Figure 6, and the truth plot in Figure 7 shows poor interpolation quality. After optimization with ordering alone, the value of, $J^S(\mathbf{Z}^*)$ reduced to 479, although interpolation quality actually became worse (RMS = 29.4). This behavior arises because smoothness alone as a criterion is not consistent with having a substantial data gap or jump as in this case. However, employing the second GA for rescaling minimized $J^S(\mathbf{Z}^{*\ddagger})$ to 1.1, and dramatically improved the interpolation quality (RMS = 7.9) as seen in Figure 8. The scaling factor of $G = 10$ was used in this case.

Table 2 shows results from all three gapped libraries indicating that rescaling may be necessary for libraries with large dynamic range and poorly sampled regions. Using reordering alone with $J^S(\mathbf{Z})$ in this case, may result in worse interpolation than for a random ordering. However, subsequent application of rescaling, which did not change the relative positioning of the substituents, markedly improved interpolation capacity upon

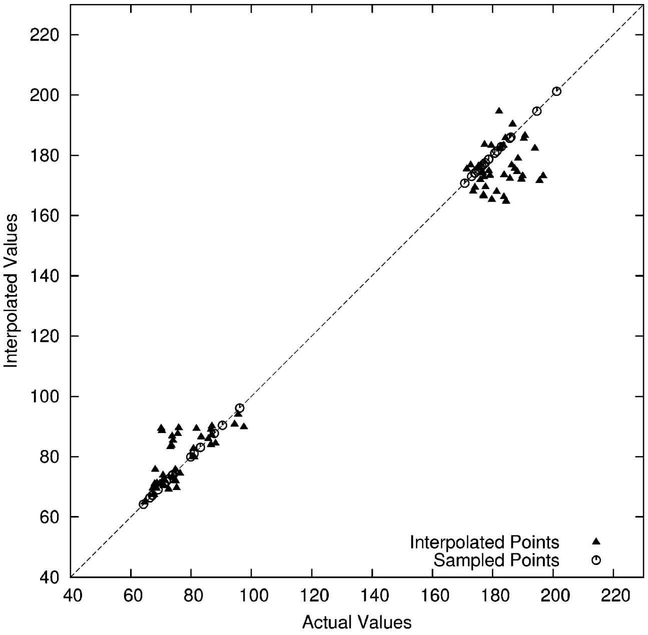

Figure 8. Truth plot for the CA property $f(\mathbf{X})$ interpolated from the optimized ordering and rescaling of the diacid substituents with

$$\mathbf{Z}^* = \left\{\begin{array}{l} Z_1\text{: } 1, 18, 19, 12, 78, 73, 75, 79 \\ Z_2\text{: } 42, 121, 87, 41, 49, 19, 1, 118, 101, 132, 91, 113, 62, 48 \end{array}\right\}$$

using the artificially gapped CA data of 90 units. $J^S(\mathbf{Z}^*) = 1.1$, RMS $= 7.9$. The results are much improved over that using random ordering and no rescaling in Figure 7.

TABLE 2: Smoothness J^S(Z) and Interpolation Quality (RMS) for Three Artificially Gapped CA Libraries, Obtained from a Random Ordering (Z), Optimal Ordering Alone (Z*), and the Optimal Ordering (Z*‡) found by Reordering and Rescaling

library	ordering type	$J^S(\mathbf{Z})$	RMS
10 gapped	random ordering	532	35.4
	reordered **Z***	53	14.3
	reordered and rescaled **Z***‡	0.20	10.8
30 gapped	random ordering	1035	40.2
	reordered **Z***	116	23.5
	reordered and rescaled **Z***‡	0.3	14.5
90 gapped	random ordering	900	19.1
	reordered **Z***	479	29.4
	reordered and rescaled **Z***‡	1.1	7.9

finding the optimal rescaled locations of the diacid substituents. In this case, the rescaling algorithm naturally discovered the data gap and separated the compounds with diacids 1−4 from 5−9 to give a smooth surface (not shown here).

While employment of real data is needed to fully verify the rescaling technique, the present simulations indicate the capability of the technique even under rather extreme data gap conditions. In instances of gapped data, incorporation of rescaling may significantly improve interpolation quality and also give an indication of flaws in the library design. If desired property values lie in the identified substituent gap, then careful choices for additional substituents may be warranted.

III.4. Influence of Regularization Techniques and the Adaptive Algorithm. The efficacy of the reordering algorithm is affected by the choice of algorithmic parameters. A parameter set P^* is defined to contain optimal parameter assignments if it allows for the best interpolation of library properties, given a sampling of M compounds. Section II.3 delineated some of the parameters involved in the regularization techniques, including basis set definition, singular value decomposition (SVD), and an additional smoothness term, (eq 17). These procedures exert regularization during the regression to obtain the coefficient vector **c**. Many other parameters also identified in the paper, with the exception of M, also exert regularization influence. Some operate on the cost function level, such as the number (T) of regression data for $J^R(\mathbf{Z})$, and the choice of cost function $J(\mathbf{Z})$ and its associated parameters. Others, such as the weight factors ω_j for least squares in eq 19 affect the coefficient calculations as well. Ultimately, all the aforementioned parameters serve to manage the interpolation function $f(\mathbf{X})$ so that the sampled data is not overfit, especially in the cases of high estimated error, and to ensure that the general library structure is correctly represented.

The parameters which gave an optimal level of regularization for the CA copolymer data were found through numerous trials of the algorithm. In the process, we observed the influence of individual parameters on interpolation quality and the interaction that exists between the parameters themselves. For the trial tests here, the interpolative capacity for a given parameter set P is quantified by the RMS between the predicted and actual values across a fully screened library. The following trials tested how the choice of RBFs, the number of functions constituting the basis set (B), and the associated basis parameter (β, and γ below), influence interpolation quality, using $M = 35$ sampled compounds and the smoothness-based cost function $J^S(\mathbf{Z})$.

For RBF selection, the behavior of two commonly utilized RBFs types, multiquadrics $\phi_l(\mathbf{X}) = (|\tilde{\mathbf{X}}_l - \mathbf{X}|^2 + \beta)^{1/2}$ and Gaussians $\phi_k(\mathbf{X}) = \exp(-\gamma|\tilde{\mathbf{X}}_k - \mathbf{X}|^2)$, was explored with the copolymer library. The basis parameters β and γ, and the number (B) of RBFs were systematically varied for multiquadrics and Gaussians, respectively, to observe the impact of such adjustments on the RMS. Multiquadric basis sets proved to be particularly robust, as a wide range of β and B values allowed accurate property predictions. For B ranging from 10 to 35, comparable RMS values of 10.0−10.6 were obtained. Within this range of B, the RMS was also insensitive to significant variations in β ($\beta = [0.1, 3]$). Basis sets composed of fewer than 10 RBFs, however, were particularly sensitive to β. If β was too small, the RBFs would vary too rapidly and lose interpolation ability over the space between the RBF centers. Alternately, if β was too large, the slowly varying behavior of the bases would preclude fine detail from being expressed. Thus, an optimal B and β combination is needed.

Repeated tests with Gaussian RBFs showed similar trends and comparable interpolative capacity. Gaussian basis sets of 5−10 RBFs and small γ (~0.005) were optimal for good interpolation results, as a small set of slowly varying basis functions is adequate at representing the limited dynamic range and characteristic behavior of $g(\mathbf{X})$ for the copolymer CA values. Overall, there was no substantial difference in choosing Gaussians over multiquadric RBFs in this study. The analysis of RBF type in constructing a basis set producing good-quality property interpolation suggests that multiquadrics and Gaussians are robust for libraries with generally smooth surface properties.

A similar diagnostic was applied to the transition metal complex library emission energy data values (λ), using $J^R(\mathbf{Z})$ and $M = 30$. $J^R(\mathbf{Z})$, in contrast to $J^S(\mathbf{Z})$, allows for automated optimization of regularization parameters using a GA, as $J^R(\mathbf{Z})$ guards against the excessive regularization encouraged by $J^S(\mathbf{Z})$ (see the discussion in section II.2). We allowed the algorithm to incorporate B and Ω as parameters for optimization by the GA. The RBF parameter β (only the multiquadric RBF was tested in this case), as well as the selection of T compounds for $J^R(\mathbf{Z})$ was tested at specific values of $\beta = [1, 10, 50, 100]$ and

$T = [1, 5, 10, 15, 20]$. As for the previous copolymer trials, the robust behavior allowed a range of algorithmic parameters to yield similar interpolation quality. Optimal interpolation (RMS of ~0.03) could be attained for T^* over the range.[10,15] Lowering T^* below 10 worsened interpolation since B^* needed to be at least 10 to adequately cover the space, and $B \leq T$.

An interaction between B and β was noted here, as was the case for the CA data trials. Initializing β to a large value (e.g., $\beta = 100$) would direct the GA to find solutions with a small number of basis functions (e.g., $B^* = 9$). Conversely, a small initialized β (e.g., $\beta = 10$) would result in a large number of basis functions (e.g., $B^* = 30$). Other interactions between algorithmic parameters exist as well, although they are not as clearly manifest as the B and β interaction. For example, reducing the number of basis functions B should exert a similar effect to employing SVD, which can also reduce the effective number of basis functions. Alternately, comparisons can be drawn between the regularization smoothness weight Ω in eq 17 and the choice of cost function: employing $J^R(\mathbf{Z})$ and a moderate Ω is similar to combining $J^R(\mathbf{Z})$ and $J^S(\mathbf{Z})$ into $J^C(\mathbf{Z})$ (eq 15). Finally, the choice of T when using $J^R(\mathbf{Z})$ operates to regularize as well. Choosing a small fraction of M as the T regression points emphasizes global interpolation and precludes overfitting the scattered data.

Despite the attractive robustness to the choice of P, determining an optimal choice P^* has distinct advantages for improving interpolation quality. The interplay among the algorithmic parameters indicates that P^* may be difficult to specify using intuition, experience, or random trials in real applications because interpolation quality cannot be measured outside of the immediately available M sampled compounds. However, as detailed in section II.6, the molecular discovery process would naturally be carried out in an adaptive fashion judiciously drawing in additional data to provide an efficient procedure that performs parameter optimization within the algorithm coincident with the goal of providing high-quality property interpolation. The feasibility of this adaptive approach has been demonstrated with success in other optimal control and identification areas,[18,19] and its successful implementation here is strongly suggested by the explorations in this work.

IV. Conclusions

This paper introduced various techniques that enhance the applicability and robustness of the original substituent reordering algorithm[1] for molecular library interpolation. Options for guiding cost functions along with their associated parameters were studied. The influence of various regularization techniques on interpolation quality was also investigated, and a rescaling technique was presented for handling poorly structured libraries. The treatment of data error propagation was also addressed. Property estimation tests using these techniques were applied to both a copolymer and a transition metal complex library, and the results provided solid evidence for the enhanced capabilities of these techniques.

The observed influence of various parameters on the algorithmic performance indicates the need for efficient algorithm optimization of all operations. The adaptive algorithm in Figure 1 provides the means to optimize the operational parameters and further enhance interpolation accuracy in a systematic fashion, using the least number of synthesized compounds.

The fundamental pattern recognition capacity of the reordering algorithm operates independently of any descriptor knowledge and, hence, provides the capability for a broad field of applicability. Future applications may be particularly promising in the area of drug discovery where the growing focus is on the quality of the compounds and property measurements. This focus plays well into the reordering algorithm, which is able to extract better interpolation results from higher-quality synthesis and screening methods over even a modest number of sampled compounds. The algorithm can also be expanded to include competitive property optimization (e.g., in the case of pharmaceuticals this may include absorption, distribution, metabolism, excretion, and toxicity (ADMET) considerations[20,21]). In this case, each property would have its own reordered representation, and all of the different property representations would jointly be used when seeking to collectively identify the likely compounds for additional synthesis balancing the competitive criteria as well as possible. The treatment of multiproperty objectives may arise in many areas of application, such as in artificial peptide discovery, for which the protein backbone serves as a scaffold and the amino acid residues are the substituents at each scaffold site. The algorithm could effectively make use of the high-dimensional model representation technique in this case, where the number N of functionalized backbone sites could be large.[22]

This work has emphasized the capabilities of the reordering algorithm to extract as much information as possible from a small fraction of a synthesized library and its observed properties. This perspective is a natural step in fully developing the reordering algorithm. However, in ultimate practice, the reordering algorithm could likely be beneficially combined with modern chemometric tools.[20,21] These chemometric tools might most effectively be used to expand the substituent set, if the reordering algorithm along with laboratory synthesis/testing indicates that no compounds with attractive properties exist in the full library. Armed with a solid estimate of the properties over the full library should provide a better means to calibrate and utilize chemometric tools as a complement to the reordering algorithm's capabilities.

While only compounds with two scaffold sites were illustrated in section III, the reordering algorithm is capable of treating compounds with many ($N > 2$) substituent sites on the molecular backbone. It may be argued that the fractional amount of sampling needed to adequately cover the library search space should decrease as the dimensionality of the problem grows, giving rise to greater efficiency as the search space expands. This argument rests on the assumption of regular smooth property behavior existing in the reordered substituent space along with the nearly invariant scaling with dimension N using statistical sampling of the compounds to provide data to represent such functions. The latter attractive scaling has been seen in other applications of high dimensional interpolation.[18,19] An explicit demonstration of this scaling with dimension for the reordering algorithm would be a significant advance. Many libraries also build on multiple molecular scaffolds rather than the single backbone cases we have studied in this work. Similar reordering arguments should be applicable to this extension as well.[23] Finally, we hope that this work stimulates direct laboratory testing of the reordering algorithm in its fully iterative closed loop form depicted in Figure 1.

Acknowledgment. This work is supported by the National Science Foundation. We also thank Stefan Bernhard for his helpful insights.

V. Appendix

This appendix (a) summarizes the weighted least-squares procedure for determination of the coefficient vector **c** employing the regularization term and (b) provides the formulation for treating data error propagation.

1. Regularized Least Squares. Let H be the merit function that combines weighted least squares with regularization defined in terms of smoothness.

$$H = \left[\sum_{m=1}^{T} \omega_m \left(\frac{g(\tilde{\mathbf{X}}_m) - \sum_{k=1}^{B} c_k \phi_k(\tilde{\mathbf{X}}_m)}{g(\tilde{\mathbf{X}}_m)}\right)^2\right] + \Omega\left[\sum_{l=1}^{R}\sum_{i=1}^{N}\left(\sum_{k=1}^{B} c_k \frac{\partial \phi_k(\tilde{\mathbf{X}}_l)}{\partial X_i}\right)^2\right] \quad (22)$$

We seek the coefficients $\mathbf{c} = [c_1, c_2, ..., c_B]$ that minimize H, which is expressed as

$$\frac{\partial H}{\partial c_{k'}} = 0 = \sum_{m=1}^{T}(-2\omega_m)\left(\frac{g(\tilde{\mathbf{X}}_m) - \sum_{k=1}^{B} c_k \phi_k(\tilde{\mathbf{X}}_m)}{g(\tilde{\mathbf{X}}_m)}\right)\phi_{k'}(\tilde{\mathbf{X}}_m) + 2\Omega\sum_{l=1}^{R}\sum_{i=1}^{N}\left(\sum_{k=1}^{B} c_k \frac{\partial \phi_k(\tilde{\mathbf{X}}_l)}{\partial X_i}\right)\frac{\partial \phi_{k'}(\tilde{\mathbf{X}}_l)}{\partial X_i} \quad (23)$$

Rearranging eq 23,

$$\sum_{m=1}^{T}(2\omega_m)\phi_{k'}(\tilde{\mathbf{X}}_m) = \sum_{k}^{B} c_k\left[\sum_{m=1}^{T}(2\omega_m)\frac{\phi_k(\tilde{\mathbf{X}}_m)\phi_{k'}(\tilde{\mathbf{X}}_m)}{g(\tilde{\mathbf{X}}_m)} + 2\Omega\sum_{l=1}^{R}\sum_{i=1}^{N}\frac{\partial \phi_k(\tilde{\mathbf{X}}_l)}{\partial X_i}\frac{\partial \phi_{k'}(\tilde{\mathbf{X}}_l)}{\partial X_i}\right] \quad (24)$$

and defining the coefficient matrix in brackets on the RHS of eq 24 as $\alpha_{kk'}$, produces

$$\underbrace{\begin{bmatrix} \sum_{m=1}^{T}(2\omega_m)\phi_1(\tilde{\mathbf{X}}_m) \\ \sum_{m=1}^{T}(2\omega_m)\phi_2(\tilde{\mathbf{X}}_m) \\ \cdots \\ \sum_{m=1}^{T}(2\omega_m)\phi_B(\tilde{\mathbf{X}}_m) \end{bmatrix}}_{\beta} = \underbrace{\begin{bmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1B} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2B} \\ \cdots & \cdots & \cdots & \cdots \\ \alpha_{M1} & \cdots & \cdots & \alpha_{MB} \end{bmatrix}}_{\alpha}\underbrace{\begin{bmatrix} c_1 \\ c_2 \\ \cdots \\ c_B \end{bmatrix}}_{\mathbf{c}} \quad (25)$$

which may be solved

$$\mathbf{c} = \alpha^{-1}\beta \quad (26)$$

2. Coefficient Covariance $\delta(c_k,c_{k'})$. To solve for the variance $\tilde{\delta}^2$ of $f(\mathbf{X})$ in eq 20, we need to first determine the covariance matrix of the coefficients,

$$\delta^2(c_k,c_{k'}) \equiv \sum_{m=1}^{T}(\delta_m)^2\frac{\partial c_k}{\partial[g(\tilde{\mathbf{X}}_m)]}\frac{\partial c_{k'}}{\partial[g(\tilde{\mathbf{X}}_m)]} \quad (27)$$

where $(\delta_m)^2$ is the variance of the property associated with the sample compound m. Let $C_{kk'} \equiv [\alpha]^{-1}_{kk'}$, then

$$c_k = \sum_{k''=1}^{B} C_{kk''}\beta_{k''} \quad (28)$$

$$\frac{\partial c_k}{\partial[g(\tilde{\mathbf{X}}_m)]} = 2\omega_m\sum_{k''=1}^{B} C_{kk''}\phi_{k''}(\tilde{\mathbf{X}}_m) \quad (29)$$

Substituting eq 29 in eq 27 gives

$$\begin{aligned}\delta^2(c_k,c_{k'}) &= \sum_{m=1}^{T}(\sigma_m)^2[2\omega^m\sum_{k''=1}^{B} C_{kk''}\phi^m_{k''}][2\omega^m\sum_{k'''=1}^{B} C_{k'k'''}\phi^m_{k'''}] \\ &= \sum_{k''=1}^{B}\sum_{k'''=1}^{B} C_{kk''}C_{k'k'''}[\sum_{m=1}^{T}2\omega^m(\delta^m)^2\phi^m_{k''}\,\phi^m_{k'''}] \quad (30)\end{aligned}$$

Here $\phi^m_{k''} \equiv \phi_{k''}(\tilde{\mathbf{X}}_m)$.

3. The estimated property variance $\tilde{\delta}^2[f(\mathbf{X}^l)]$ for the lth compound is

$$\tilde{\delta}^2[f(\mathbf{X}^l)] \equiv \frac{1}{K}\sum_{p}^{K}[f(\mathbf{X}^l_p) - \bar{f}(\mathbf{X}^l)]^2 \quad (31)$$

where $l = 1, ..., R$, and R is the number of compounds in the library and $p = 1, ..., K$ is the number of input data observations. It is understood that $f(\mathbf{X}^l_p)$ depends on p through the input data, generating the coefficient vector $\mathbf{c}$. Here, $\bar{f}(\mathbf{X}^l)$ denotes the estimated property with the mean vector $\bar{\mathbf{c}}$. Taylor expanding to first order produces,

$$[f(\mathbf{X}^l) - \bar{f}(\mathbf{X}^l)] = (c_1 - \bar{c}_1)\left(\frac{\partial[f(\mathbf{X}^l)]}{\partial c_1}\right) + ... + (c_B - \bar{c}_B)\left(\frac{\partial[f(\mathbf{X}^l)]}{\partial c_B}\right) \quad (32)$$

where the derivatives are evaluated with the nominal vector $\bar{\mathbf{c}}$.

From eqs 31 and 32,

$$\tilde{\delta}^2[f(\mathbf{X}^l)] = \frac{1}{K}\sum_{p}^{K}\left[(c_1 - \bar{c}_1)\left(\frac{\partial[f(\mathbf{X}^l)]}{\partial c_1}\right) + ... + (c_B - \bar{c}_B)\left(\frac{\partial[f(\mathbf{X}^l)]}{\partial c_B}\right)\right]^2 \quad (33)$$

$$\begin{aligned}\tilde{\delta}^2[f(\mathbf{X}^l)] = \frac{1}{K}\sum_{p}^{K}&(c_1 - \bar{c}_1)^2\left(\frac{\partial[f(\mathbf{X}^l)]}{\partial c_1}\right)^2 + ... \\ &+ 2(c_1 - \bar{c}_1)(c_2 - \bar{c}_2)\left(\frac{\partial[f(\mathbf{X}^l)]}{\partial c_1}\right)\left(\frac{\partial[f(\mathbf{X}^l)]}{\partial c_2}\right) + ... \\ &+ (c_2 - \bar{c}_2)^2\left(\frac{\partial[f(\mathbf{X}^l)]}{\partial c_2}\right)^2 + ... \quad (34)\end{aligned}$$

It is understood that the coefficient components $c_1, ..., c_B$ depend on the p^{th} observation of the data. Recognizing that

$1/K\sum_p^K(c_k - \bar{c}_k)^2$ and $1/K\sum_p^K(c_{k'} - \bar{c}_{k'})(c_k - \bar{c}_k)$ are the variance and covariances of c_k, $c_{k'}$, we finally have

$$\tilde{\delta}^2[f(\mathbf{X}^l)] = \sum_k^B \sum_{k'}^B \delta^2(c_k,c_{k'})\phi_k^l\,\phi_{k'}^l. \tag{35}$$

References and Notes

(1) Shenvi, N.; Geremia, J. M.; Rabitz, H. *J. Phys. Chem. A* **2003**, *107*, 2066.

(2) Schultz, J. S. *Biotechnol. Prog.* **1996**, *12*, 729.

(3) Liu, R.; Enstrom, A. M.; Lam, K. S. *Exp. Hematol.* **2003**, *31*, 11.

(4) Danielson, E.; Golden, J. H.; McFarland, E. W.; Reaves, C. M.; Weinberg, W. H.; Wu, X. D. *Nature* **1997**, *389*, 944.

(5) Dolle, R. E. *J. Comb. Chem.* **2003**, *5*, 693.

(6) Matter, H.; Baringhaus, K. H.; Naumann, T.; Klaubunde, T.; Pirard, B. *Comb. Chem. High Throughput Screening* **2001**, *4*, 453.

(7) Parvu, L. *J. Cell. Mol. Med.* **2003**, *7*, 333.

(8) Reynolds, C. H. *J. Comb. Chem.* **1999**, *1*, 297.

(9) Lowry, M. S.; Hudson, W. R.; Pascal, R. A., Jr.; Bernhard, S. *J. Am. Chem. Soc.* **2004**, *126*, 14129.

(10) Cherrie, J. B.; Beatson, R. K.; Newsam, G. N. *SIAM J. Sci. Comput.* **2002**, *23*, 1549.

(11) Morse, B. S.; Yoo, T. S. et al. *Proceedings of the International Conference on Shape Modeling & Applications*; IEEE Computer Society, Washington, D.C., 2001.

(12) Chu, P. C. *A.* Genetic Algorithm Approach for Combinatorial Optimization Problems. Ph.D. Thesis, The Management School, Imperial College of Science, London, 1997.

(13) Ambati, B. K.; Ambati, J.; Mokhtar, M. M. *Biol. Cybernetics* **1991**, *65*, 31.

(14) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical recipes in C*, 2nd ed; Cambridge University Press: New York, 1992.

(15) Feng, X. J.; Rabitz, H. *Biophys. J.* **2004**, *86*, 1270.

(16) Geremia, J. M.; Rabitz, H. *Phys. Rev. Lett.* **2002**, *89*, 263902.

(17) Goldberg, D. E. *Genetic Algorithms in Search, Optimization, and Machine Learning*; Addison-Wesley: Reading, MA, 1989.

(18) Li, G.; Wang, S. W.; Rabitz, H. *J. Phys. Chem.* **2002**, *106*, 8721.

(19) Li, G.; Wang, S. W.; Rabitz, H.; Wang, S. K.; Jaffe, P. *Chem. Eng. Sci.* **2002**, *57*, 4445.

(20) Clark, D. E.; Grootenhuis, P. D. *Curr. Opin. Drug Discovery Dev.* **2002**, *5*, 382.

(21) Walters, W. P.; Namchuk, M. *Nat. Rev. Drug Discovery* **2003**, *2*, 259.

(22) Rabitz, H.; Alis, O. *J. Math. Chem.* **1999**, *25*, 197.

(23) E. Lieberman; Rabitz, H. Manuscript in preparation.