

Efficient Calculation of Molecular Properties from Simulation Using Kernel Molecular Dynamics

W. Michael Brown,^{*,†} Ariella Sasson,[‡] Donald R. Bellew,[§] Lucy A. Hunsaker,^{||} Shawn Martin,[†] Andrei Leitao,^{||} Lorraine M. Deck,[§] David L. Vander Jagt,^{||} and Tudor I. Oprea^{||}

Computational Biology, Sandia National Laboratories, P.O. Box 5800, M/S 1316, Albuquerque, New Mexico 87185-1316, Department of Computational Biology and Molecular Biophysics, Rutgers State University of New Jersey, Piscataway, New Jersey 08854, and Department of Chemistry and Chemical Biology and Department of Biochemistry and Molecular Biology, University of New Mexico, Albuquerque, New Mexico 87131

Received April 9, 2008

Understanding the relationship between chemical structure and function is a ubiquitous problem within the fields of chemistry and biology. Simulation approaches attack the problem utilizing physics to understand a given process at the particle level. Unfortunately, these approaches are often too expensive for many problems of interest. Informatics approaches attack the problem with empirical analysis of descriptions of chemical structure. The issue in these methods is how to describe molecules in a manner that facilitates accurate and general calculation of molecular properties. Here, we present a novel approach that utilizes aspects of simulation and informatics in order to formulate structure–property relationships. We show how supervised learning can be utilized to overcome the sampling problem in simulation approaches. Likewise, we show how learning can be achieved based on molecular descriptions that are rooted in the physics of dynamic intermolecular forces. We apply the approach to three problems including the analysis of corticosteroid binding globulin ligand binding affinity, identification of formylpeptide receptor ligands, and identification of resveratrol analogues capable of inhibiting activation of transcription factor nuclear factor kappaB.

INTRODUCTION

The problem of molecular property prediction is central to many fields within chemistry and biology. These include protein engineering and function prediction, prediction of environmental fate and toxicity, and the design of novel drugs and materials. Despite the differences in the ultimate goals in fields such as cheminformatics/computational chemistry, bioinformatics/molecular biophysics, environmental science, and materials design, all share a fundamental objective: identifying the relationship between molecular structure and a given property. Finding this relationship facilitates quantification without the cost of synthesis and/or assay. Likewise, this relationship facilitates the design of novel molecules with desired properties. For toxic or pathogenic molecules, the need for accurate computational methods is paramount for safe, low-cost investigations.

Traditionally, there has been a dichotomous approach toward the problem of molecular property prediction. Simulation methods, on the one hand, obtain results from a quantum or classical formulation of molecular mechanics applied to an atomistic model. By employing equations fit at the particle level, these approaches provide a general

method for property prediction with atomic detail. Often, however, the system size and/or time-scale of relevant processes preclude an ergodic sampling from simulation. The result is a limited sampling of the phase-space and predictions based on insufficient statistics. Novel methods for improving the sampling of phase space are therefore an area of active research for both Monte Carlo (MC) and Molecular Dynamics (MD) simulations.^{1–11}

Informatics approaches, on the other hand, can circumvent the time-scale problem. This is achieved by finding higher level abstractions for molecular description that can be correlated directly to a given property. The tradeoffs, when compared to simulation, include 1) the requirement for training data on every property for which a prediction is to be made and 2) a more limited domain of applicability for a given model as determined by the training molecules. An important issue in the informatics approach is the selection of appropriate molecular descriptors composing the feature space. Descriptors based on the molecular graph (whether atom connectivity or protein primary sequence) are commonly employed in informatics models. However, studies investigating model accuracy suggest that such models may only be accurate for calculations on molecules similar in structure to those used for training.¹² Descriptors based on 3-dimensional structure might offer the potential for more general models. This is due to their ability to encode information more closely related to molecular interaction. One issue with 3D methods, however, is the requirement for the selection of an “active conformation”. While methods

* Corresponding author phone: (505)284-8938; fax: (505)845-7442; e-mail: wmbrown@sandia.gov.

[†] Sandia National Laboratories.

[‡] Rutgers State University of New Jersey.

[§] Department of Chemistry and Chemical Biology, University of New Mexico.

^{||} Department of Biochemistry and Molecular Biology, University of New Mexico.

for automating this approach have been developed, the concept of a static molecular conformation responsible for activity is somewhat nebulous.

Here, we present a new formalism, Kernel Molecular Dynamics (kMD), that utilizes both simulation and informatics approaches for molecular property prediction. We address the sampling problem in MD by shrinking the system size down to the molecule in question. In trade, training data are required in order to quantify molecule properties in terms of dynamical molecular interaction fields as opposed to specific intermolecular interactions with the system. The approach has roots in comparative molecular field analysis (CoMFA)¹³ due to its use of interaction fields and in 4D-QSAR,¹⁴ the first method to explicitly utilize MD simulation for regression on molecular properties. It holds advantages in that it does not assume or require a static active conformation as in CoMFA. Nor does it require a similar scaffold for alignment as is typical in 4D-QSAR. While the method is intended to be general in scope, we have chosen to validate the approach in a context relevant to the current National Institute of Health initiative for molecular library screening by using prediction of small organic ligand activity. Specifically, we have analyzed corticosteroid binding globulin ligand binding affinity using a benchmark data set allowing for comparison to other methods. We have used the approach to generate models capable of predicting those analogues of resveratrol capable of inhibiting human tumor necrosis factor alpha induced activation of the transcription factor nuclear factor kappaB (NF- κ B). We have used this model to identify novel inhibitors with more potent activities than resveratrol. Finally, we looked at a much more difficult problem—analysis of results from high-throughput screening. This allowed for calculations on flexible and diverse chemical structures that were not limited by a common scaffold or synthetic scheme.

The Problem of Property Prediction. Perhaps the most intuitive approach for understanding how molecular structure relates to function or activity would be based on a derivation from first principles using particle simulations intended to represent an accurate reflection of the physical processes involved. Unfortunately, the complexity of such calculations based on our current understanding of physics precludes accurate analysis for many processes of interest. The time required for such calculations is just too far out of reach. A typical approach to handling such difficulties is to seek higher level formulations with empirical analysis at a higher level than the physics of individual particles within a system. With this regard, we face the problem of describing the dynamic interactions of a molecule in question with other molecules in the system in a manner that allows for the calculation of desired properties. This description must be canonical in the sense that it allows for a unique and general quantification for any molecule of interest (regardless of the size or structure of the molecule). Also, the description should involve as little information loss as possible.

In kMD, we approach this problem by reducing the complexity of the particle simulation such that it involves only the conformation of the molecule in question; therefore, the approach is built on the idea of 4D-QSAR.¹⁴ We address the problem of intermolecular interaction by considering a probe atom, fragment, or molecule; therefore, the approach is also built on the idea of CoMFA.¹³ For a given probe, we

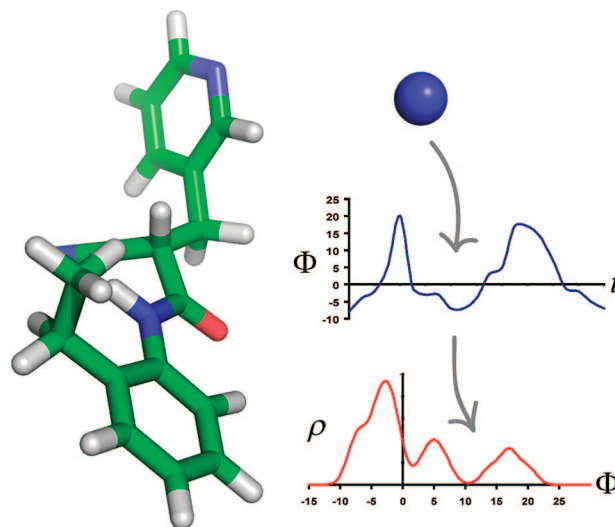


Figure 1. Calculation of dynamic molecular interaction fields in kMD. We measure the potential energy (ρ) of interaction between the probe (blue sphere) at a position relative to a molecule (sticks). This is performed for every conformation of the molecule as identified by simulation to give the upper plot. This plot is not canonical in the sense that the ordering of conformational change is not unique. We therefore transform this function. Here, we have illustrated transformation into a probability density (ρ). Analysis of the probability density for different types of probes at every position surrounding the molecule can be utilized to identify differences in how two molecules will interact with a system. This, in turn, can be utilized to quantify molecular properties of interest.

measure the energy of interaction of the probe with the molecule. By calculating this energy for different probes at all positions surrounding the molecule and for different conformations of the molecule, we obtain a basis for comparison of the differences in how molecules will interact with other molecules in the system. We then seek equations that relate these “dynamic molecular interaction fields” to a property of interest based on existing measurements for a set of molecules. We have illustrated this approach in Figure 1 and give a formal description below.

Kernel Molecular Dynamics. We consider the case where we have a single assay for a given molecular property P that we would like to quantify. Denote by $M = \{m_1, m_2, \dots\}$, the set of all molecules. For a given molecule $m \in M$, we assume that any molecular property can be quantified based on its dynamic interactions with other molecules in the system. While a traditional simulation approach assumes a function utilizing a subset of M , intended to represent a system of interest, we take advantage of an observation central to the study of quantitative structure–property relationships (QSPR)—for a given assay, the interacting molecules within a system are identical aside from the test molecule. Therefore, any changes affecting the property P should be inherent to the molecule m itself. This suggests the existence of a function $f: M \rightarrow \mathbb{R}$ for property prediction such that $f(m)$ gives P without the requirement for analysis of other molecules in the system. Because it is unlikely that such a function can be derived directly from thermodynamics equations, we trade a reduction in the size of the system for training data such that f can be learned empirically.

In order to obtain computational efficiency, we do not look at explicit interactions between m and molecules in a system but rather the potential for interaction with other molecules

as probed by molecular interaction fields. We therefore consider a smaller set $Q = \{q_1, q_2, \dots, q_k\}$ of probe molecules, atoms, or fragments that are intended to provide, in some sense, a canonical basis for elucidating differences in how molecules interact with any system. The molecular interaction field is given by a function $\Phi_{m,q_v}: \mathbb{R}^3 \times [0, t_m] \rightarrow \mathbb{R}$ that represents the energy of interaction between m and a probe q_v as a function of Cartesian space. Because Φ_{m,q_v} is dependent on the conformation of m , it is a function of the molecule's dynamic conformation, denoted here by $\mathbf{r}_m(t)$ with t in $[0, t_m]$ for a range of conformations between 0 and t_m for each molecule m . We solve for $\mathbf{r}_m(t)$ with simulation. Given a fixed center of mass and net rotation for a molecule m , the equation $\int_{\mathbf{r}_1}^{\mathbf{r}_2} d\mathbf{r} \langle \partial \Phi_{m,q_v}[\mathbf{r}_m(t), \mathbf{r}] / \partial \mathbf{r} \rangle_t$ is related to the approximate potential of mean force between the probe at positions \mathbf{r}_1 and \mathbf{r}_2 with the notable difference that $\mathbf{r}_m(t)$ is never perturbed by the probe at position \mathbf{r} . This, of course, assumes constant temperature and adequate representation of configurations with higher energies.

In order to obtain f , we consider kernel methods for learning and therefore require a kernel function $k: \mathcal{M} \times \mathcal{M} \rightarrow \mathbb{R}$ that gives the similarity between two molecules m_i and m_j in terms of $\Phi_{m_i,q_v}(\mathbf{r}, t)$ and $\Phi_{m_j,q_v}(\mathbf{r}, t)$. This idea of molecular similarity is similar, in some regard, to the concept of CoMSIA.¹⁵ Because we do not impose any limitations on the initial conformation for m or on t_m , the comparison of $\Phi_{m_i,q_v}(\mathbf{r}, t)$ with $\Phi_{m_j,q_v}(\mathbf{r}, t)$ over t is not trivial. We therefore use a canonical transformation of $\Phi_{m,q_v}(\mathbf{r}, t)$ to provide a function that is independent of t and facilitates comparison with an inner product. Two obvious choices include transformation into a frequency domain and transformation into a probabilistic domain. Here, we use the latter and denote by $\rho_{m,q_v}(\mathbf{r}, \varphi)$ the probability density function for probe interaction potential such that $\int_{\varphi=c}^d \rho_{m,q_v}(\mathbf{r}, \varphi) = \text{Pr}(c \leq \Phi_{m,q_v}(\mathbf{r}, t) \leq d)$. We can then define a similarity kernel

$$k_{q_v, T_i, T_j}(m_i, m_j) = \int_{\varphi=-\infty}^{e_v} \int_{\mathbf{r}} \rho_{m_i,q_v}(T_i(\mathbf{r}), \varphi) \rho_{m_j,q_v}(T_j(\mathbf{r}), \varphi) d\mathbf{r} d\varphi \quad (1)$$

that compares at each point in space surrounding the two molecules the probability that a probe molecular interaction potential takes on each value for negative interaction energies. In this function, $e_v < 0$ is a parameter that restricts interaction potential to a finite volume surrounding the molecules. The integral over \mathbf{r} introduces a frame of reference problem that requires alignment between molecules. We address this issue by enforcing the Eckart conditions in the form of holonomic restraints. This allows for the separation of rotations and translations of the molecules from those motions due to internal vibrations.¹⁶ Additionally, we parametrize the kernel with transformations T_i and T_j that represent translation and rotation of the probe atom (or, equivalently, the molecule).

In order to consider all probes, we introduce a summation over v and normalize the similarity to lie between 0 and 1

$$k_{T_i, T_j}(m_i, m_j) = \sum_v \frac{\chi_v k_{q_v, T_i, T_j}(m_i, m_j)}{k_{q_v, T_i, T_i}(m_i, m_i) k_{q_v, T_j, T_j}(m_j, m_j)} \quad (2)$$

where χ_v gives a constant weight specifying the relative importance of probe q_v . The problem of choosing appropriate transformations is a difficult one. Perhaps the most intuitive approach, in terms of the idea of a pairwise molecular

similarity, is to choose transformations independently for each pair such that the similarity is maximized:

$$k'(m_i, m_j) = \max_T k_{T, T}(m_i, m_j) \quad (3)$$

Unfortunately, this is not necessarily a true inner product (a necessary condition for a kernel function) because it is not linear. One solution to this problem is to take the projection of the kernel matrix using only positive eigenvalues. An alternative approach, that facilitates a true inner product, is to use a fixed frame of reference such that the transformation for each molecule is fixed. Thus, for a set of molecules $\{m_1, m_2, \dots, m_n\}$ there is a corresponding set of transformations $\{T_1, T_2, \dots, T_n\}$ that define

$$k(m_i, m_j) = k_{T_i, T_j}(m_i, m_j) \quad (4)$$

We describe one approach for calculating transformations for each molecule below.

The similarity metrics presented allow us to obtain equations for a given property in terms of a molecule's dynamic probe interaction fields. This requires that data for a given property are available for a set of training molecules. Here we utilize support vector machines (SVMs)¹⁷ for learning to provide an equation for f of the form

$$f(m) = \sum_i \alpha_i k(m_i, m) + b \quad (5)$$

where i indexes the molecules in the training set chosen as support vectors, and α_i and b are determined during training. SVMs can be utilized for either regression (where $f(m)$ gives the property) or for classification (where the sign of f represents an assigned class for a given property). Here, we apply both approaches. An additional advantage of SVMs is their ability to obtain nonlinear functions for a property using derived kernels. Here, we consider, in addition to the kernel in eq 4, an RBF kernel defined as

$$k_G(m_i, m_j) = \exp(-(k(m_i, m_i) - 2k(m_i, m_j) + k(m_j, m_j))/2\gamma^2) \quad (6)$$

The ability to calculate an unknown property is useful for screening; however, further intuition into how the structure of a molecule relates to a given property is beneficial for design problems. For the linear SVM, the model can be projected into Cartesian space to allow for visualization in a manner analogous to that used for CoMFA. This can be seen more clearly by rearranging eq 5; for a single probe, neglecting normalization, we obtain

$$f(m) = \int_{\varphi=-\infty}^{e_v} \int_{\mathbf{r}} \left(\sum_i \alpha_i \rho_{m_i,q_v} \right) \rho_{m,q_v} d\mathbf{r} d\varphi + b \quad (7)$$

In this form, it becomes clear that the contribution to f over a range of space $d\mathbf{r}$ and a range of probe interaction potentials $d\varphi$ can be isolated. If we choose q_v such that it represents solely a van der Waals interaction potential, we can extract information in the form of key steric interactions in a given region of space. If we add a separate probe that is charged, we can extract information on Coulombic interaction potential. By plotting isosurfaces of the Shannon entropy of ρ_{m,q_v} , we can obtain insight into how thermal motion influences a given model.

The formalism for kMD describes a method for understanding molecular properties in terms of the same physics that are utilized for simulation approaches but with calcula-

tions that are accessible to modern computers. A full description of our numerical approximation to kMD is given in the Experimental Section.

RESULTS

Steroid Binding to Corticosteroid Binding Globulin. For the first data set, we analyzed corticosteroid binding globulin (CBG) binding affinity for a set of steroids as described in ref 18. This data set was first compiled for evaluation of CoMFA¹³ and has since become a benchmark for 3D-QSPR approaches. The data set consists of 31 compounds with pK values ranging from -5 to -7.881 . We chose the MM3 force field for MD and kMD calculations for this data set. However, parameters for $11\beta,17,21$ -trihydroxy- 2α -methyl- 9α -fluoro- 4 -pregnene- $3,20$ -dione were not available. Therefore this compound was not included in initial tests. Because this compound has also been identified as an outlier,¹⁸ we felt it was important to include the steroid in final tests as described below, taking torsion parameters from an atom type with the same hybridization.

Initially, we used molecular conformations taken from the work in ref 18 as the starting point for MD calculations. Conformations at 50 random time points were taken for continuous PDF calculation (eq 8 in the Experimental Section). The dynamics trajectories were aligned by hand based on the initial conformation. Alignment and similarity calculations were performed using only local BFGS minimization. The resulting SVM model had a leave-one-out cross-validation coefficient of determination (q^2) of 0.86 using a regularization parameter (c) of 2.8. A SVM regression tube width of $1 \cdot 10^{-7}$ was used for all calculations. The coefficient of determination (r^2) when trained on all molecules was 0.9. For the nonlinear model generated with the RBF kernel, a q^2 of 0.86 and an r^2 of 0.93 were obtained ($c=3.8$, $\gamma=1.3$). We next tested a discrete probability approximation for PDF calculation (eq 9 in the Experimental Section), repeating the above procedure. For this case, we saw only a small loss in accuracy with a q^2 of 0.84 and an r^2 of 0.89 ($c=1.45$) for the linear kernel and a q^2 of 0.84 and an r^2 of 0.9 ($c=2.2$, $\gamma=1.5$) for the RBF.

Finally, we tested the approach as intended, with no user intervention in selection of the starting conformation or alignment. In this case, each steroid was subjected to conjugate gradient minimization in the MM3 force field to generate initial conformations for MD. Each trajectory was centered at the origin and transformed onto its principal axes. Full global alignment was performed using the hybrid GA followed by full local minimization. The resulting model had a q^2 of 0.88 and an r^2 of 0.9 ($c=5.8$) for the linear kernel and a q^2 of 0.94 and an r^2 of 0.96 ($c=5.2$, $\gamma=0.3$) for the RBF. When the full data set is used (31 instead of 30), a q^2 of 0.76 and an r^2 of 0.83 ($c=2.7$) are obtained for the linear kernel, and a q^2 of 0.86 and an r^2 of 0.93 ($c=5.42$, $\gamma=0.3$) are obtained for the RBF. We believe the decrease in accuracy is not due to parametrization but rather the unique substitution of the steroid ring within this compound. A correlation plot for the RBF model on the full data set is shown in Figure 2.

Visualization of the resulting model is important for interpretation, and there are various approaches that might be used to map the model into a space suitable for

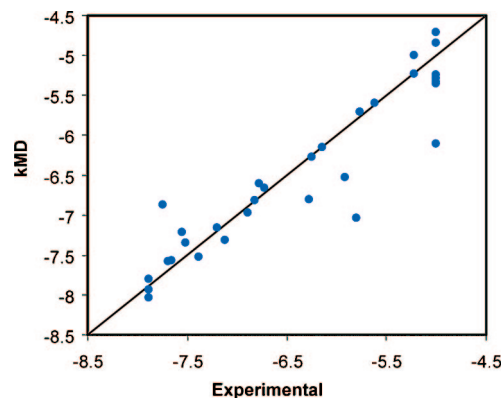


Figure 2. Correlation plot of experimental versus calculated binding affinities resulting from cross-validation using kMD with the RBF kernel on 31 steroids. The resulting q^2 is 0.86. Units are $-\log K$.

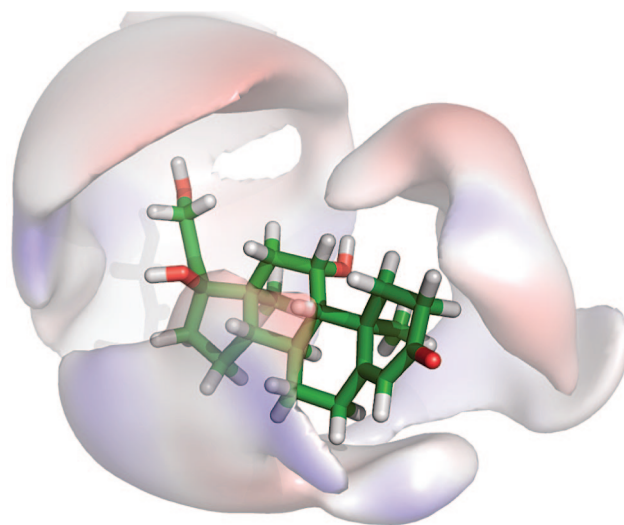


Figure 3. Visualization of the contribution of probe interaction potentials to the binding affinity of steroids with CBG. The surface is interpolated using the uncharged probe and represents locations of probe atom centers where negative interaction potentials increase binding affinity. The surface is shaded blue in regions where positive probe interactions increase binding and is shaded red where negative probe interactions increase binding. The initial conformation of cortisone is shown as a stick model.

visualization. Here, we have used an approach similar to that applied in CoMFA by utilizing eq 7 and averaging the contribution of probe interaction potential over the integration range used for model development. The result for the steroid model is shown in Figure 3. In this example, we have interpolated an isosurface showing the locations where negative interaction potentials of the uncharged probe enhance binding affinity. This surface is then colored red or blue based on the degree to which a positively or negatively charged probe increases binding affinity. It is important to note that figures such as this do not simply illustrate how a probe interacts with a molecule but rather how probes are able to differentiate accuracy between molecules. In an ideal case, this surface should be representative of binding pocket structure and electrostatics simply because it is this structure that is truly determining binding affinity. An alternative approach for isosurface visualization is given in the example below.

We have referred to kMD as an “efficient” method for calculation of molecular properties from simulation, referring

to the idea that much simpler simulations can be utilized for property prediction. When compared to informatics approaches (e.g., similarity calculations on topological descriptors), the method is *much* slower. For the steroid data set, we have broken the run-time down into the precalculations of the probability densities (eq 9), alignment (eq 3), and final integration (eq 1). On a 2.4 GHz Pentium 4 Xeon, the average time for precalculation per molecule was 23.2 CPU s. Alignment required an average of 741 CPU s per pair, and integration required an average of 27 CPU s per pair. Although we believe that a 1–2 order of magnitude decrease in run time can be achieved by optimizing the proof-of-concept code, screening can be achieved on modern high performance computing systems using the code without optimization. The problem, however, is with training. As presented, SVM training will scale $O(n^2)$ with the number of data points (the kernel must be calculated for each pair of molecules). Therefore, “boosting”^{19,20} or alternative methods for improving the scaling will be required in order to facilitate kMD training with large data sets.

The steroid molecules are relatively rigid and share a common scaffold (the cyclophenanthrene nucleus). For this reason, this data set is very amenable to 3D-QSPR approaches; there is little ambiguity in the selection of molecular conformations and alignments. In fact, increasing the number of MD samples from 50 to 500 offers little improvement in the model accuracy as obtained by kMD. Nonetheless, most 3D-QSPR approaches have accuracies that are sensitive to steroid conformation, and kMD results are comparable or superior to previous methods. A thorough review of QSPR results on the steroid data set with a variety of methods is given in ref 21. The q^2 of 0.86 on the full data set is directly comparable to a value of 0.63 for Spatial Autocorrelation and 0.63 for Molecular Similarity. CoMFA was originally benchmarked on a subset of the first 21 steroids to produce a q^2 of 0.734 (after correction of errors in the original data set). For this same data set, we are able to achieve a q^2 of 0.90 ($r^2=0.99$). Direct comparisons to other methods are not possible due to differences in the data sets or methods; however, a discussion of these results has been given.²¹

The fact that kMD offers comparable or improved accuracy in the calculation of steroid binding affinity is not the sole point of this work. The steroid data set has been carefully analyzed with conformations and alignments selected to produce accurate results for 3D-QSPRs. Our approach does not require user-bias in the selection of active conformation or alignment but rather considers the dynamic nature of molecular structure. For many realistic applications, this should be advantageous in that it is not straightforward to limit flexible molecules with different structures to static conformations.

Inhibition of NF- κ B Activation by Substituted *trans*-Stilbenes. For the second test case, we looked at the ability of resveratrol and substituted *trans*-stilbene and *cis*-stilbene analogues to inhibit the activation of NF- κ B. Resveratrol (3,4',5-trihydroxystilbene - Figure 4) is a polyphenolic phytochemical that is abundant in red wine. The chemical received substantial attention following the proposal that it was responsible for the French Paradox,^{22–24} the low incidence of cardiovascular disease in a French population with high intake of saturated fat. In addition to its potential

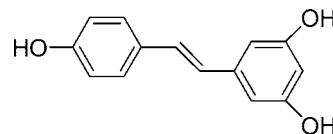


Figure 4. Structure of resveratrol.

cardioprotective effects, resveratrol has been identified as a chemopreventive agent active at all three stages of carcinogenesis²⁵ and has been shown to aid in ischemic injuries, stress resistance, and expansion of lifespan (reviewed in ref 26). Resveratrol is known to inhibit human tumor necrosis factor α -mediated activation of NF- κ B, a transcription factor well-known for its regulation of anti-inflammatory genes and promotion of antiapoptotic states in cancer cells.²⁷ Although resveratrol activity is often attributed to its antioxidant properties, we have shown in a recent study²⁷ that substituted *trans*-stilbene analogues lacking antioxidant activity can be more potent inhibitors of NF- κ B activation. We were therefore interested to see if we could generate a model capable of differentiating the inhibitory ability of NF- κ B activation.

Initially, we attempted to generate a model capable of classifying chemicals as active or inactive and therefore used SVM classification rather than regression. Our original report contained the activities for 73 resveratrol analogues²⁷ (compounds 1–73 in Supporting Information Table 1). Since that time, 89 additional analogues have been synthesized and tested which are reported here (compounds 74–162 in Supporting Information Table 1). Fifteen compounds were chosen from the best inhibitors along with the bottom 15 inhibitors for model generation (as noted in Supporting Information Table 1). The activities for this decision were measured as percent inhibition of TNF- α induced activation of NF- κ B at a 15 μ M concentration using the Panomics NF- κ B stable reporter cell line 293/ NF- κ B-luc (see the Experimental Section). For this data set, we implemented MMFF94 capabilities into kMD and took conformations from 5000 MD time points for PDF generation. Using the linear SVM (regularization parameter $c = 14$), we were able to achieve a cross-validation accuracy of 90% (final accuracy of 100%). For the RBF kernel, the same results were achieved ($\gamma=0.1$). In each case, the specificity was 100%, and the sensitivity was 83.3%. We were able to achieve similar results using a larger data set of 60 compounds (30 active/30 inactive). In this case, the cross-validation accuracy was 85% (95% final accuracy), and for the RBF kernel the cross-validation accuracy was 85% with a final accuracy of 98.33% ($c=6$, $\gamma=0.6$).

IC₅₀s have been measured for 15 of the compounds, and, therefore, in addition to the ability to identify inhibitors, we also tested the ability to calculate the actual IC₅₀s. Unfortunately, we were unable to obtain an accurate model. This was likely due to the fact that the IC₅₀ range was small (0.15–1.5 μ M). Nonetheless, the classification model offers high potential for screening candidate molecules and identification of novel inhibitors until the data set can be expanded in order to provide SAR information and quantitative prediction of activity. To test this, we performed predictions on a set of candidate molecules (Supporting Information Table 2) using a consensus model consisting of the regular and RBM kernels for both models (consisting of 30 or 60 training molecules). Of the candidate molecules, 6

Table 1. New Resveratrol Analogues Sorted by Percent Inhibition of NF- κ B Activation at 15 μ M^a

ID	% of Control
168	3.4 \pm 0.07
167	4.5 \pm 0.7
184	12 \pm 0.9
164	13 \pm 0.7
183	14 \pm 0.3
186	15 \pm 4.8
178	16 \pm 4.0
166	17 \pm 1.8
179	21 \pm 1.5
180	28 \pm 2.5
165	39 \pm 14
182	40 \pm 0.2
170	43 \pm 5.6
181	55 \pm 4.2
163	56
171	57 \pm 8.7
185	57 \pm 8.9
173	67 \pm 8.0
177	67 \pm 0.8
172	71 \pm 10
174	71 \pm 12
169	74 \pm 6.1
175	97 \pm 13
176	102 \pm 2.8

^a kMD predictions from the consensus model are colored as follows: red - *active* according to all models, green - *maybe active* (debated by models), and blue - *inactive* according to all models.

were predicted to be active by all models and were synthesized for testing. All 6 were found to be more potent inhibitors than resveratrol (Supporting Information Table 2). In order to test the false negative rate, we synthesized and assayed the remaining compounds. Although the most potent inhibitor was misclassified, the model otherwise did very well in classifying the relative activities of the compounds (Table) and likewise in the identification of novel potent inhibitors.

Ligand Binding to Formylpeptide Receptor. Both the steroid test case and the resveratrol test case involved chemicals that share a common scaffold. Therefore, for the final application we chose a much more difficult problem: classification of high-throughput screening results. These molecules are unlikely to be limited in flexibility and unlikely to share a common substructure or scaffold because they are not obtained from a limiting synthetic scheme. Here, we have used screening results from the formylpeptide receptor (FPR)

ligand binding assay and the NIH Molecular Libraries Screening Center 10K ST1 compound set. The FPR family of G-protein coupled receptors contributes to the localization and activation of tissue-damaging leukocytes at sites of chronic inflammation and has been proposed as a prospective target for therapeutic intervention against malignant gliomas.^{28,29} Details on the assay are available through the National Library of Medicine PubChem site (assay ID 440). At the time of our investigation, the assay had identified 17 active compounds and 9965 inactive compounds (from which 17 were chosen at random). The resulting set of compounds is illustrated in Tables 2 and 3.

Despite the high flexibility and variation in structure, we were able to predict ligand binding activity with a leave-one-out accuracy of 82%. In this case, both the specificity and the sensitivity were 0.85 corresponding to 3 false positives and 3 false negatives. We obtained an accuracy of 100% when all molecules were used for training. A visualization of the impact of probe interaction potential on RBP activity for the final model is shown in Figure 5. In this case, we have calculated isosurfaces for each probe independently in order to clearly illustrate regions where probe electrostatic and steric interactions influence activity.

Following the development of this model, subsequent experimental screens led to the identification of new hits as reported in PubChem assays 723 and 724. We performed a blind test of the kMD model with the classification of 10 molecules from these assays (5 of which were hits). Classification of PubChem compound IDs 1185399, 3092570, 5291419, 1036526, 1561677, 5766371, 5290341, 5291826, 5291891, and 5299897 gave an accuracy of 80% despite the fact that the activities of these compounds were unknown at the time of model development. It is important to note that the compounds were selected for these assays by application of computational screening techniques (Free-Wilson PLS analysis, 2D substructure search, and 3D ROCS/EON similarity search) based on chemotypes from the previously identified FPR ligands. Although this introduces bias into the newly screened compounds, their 3-dimensional structures are still very diverse, and the compounds do not share a scaffold with the ligands used for training. Additionally, these methods were only able to identify 15 confirmed hits from a selection of 1276 screened compounds.

Due to the small number of hits from the initial screen available for the development of unbiased models, we tested the cross-validation performance with 2 additional groups of 17 randomly selected negatives. Group 1 resulted in a cross-validation accuracy of 79%, and group 2 resulted in a cross-validation accuracy of 74%. The structures and CIDs of these molecules are given in Supporting Information Table 3.

DISCUSSION

kMD is intended to provide an approach for the calculation of molecular properties that is more efficient than traditional MD simulation and more accurate than traditional informatics approaches due to the explicit consideration of dynamic molecular conformation. Aside from efficiency, there are several advantages of kMD over traditional simulation. First, specifics of the interacting system are not directly relevant, and, therefore, the

Table 2. Active Compounds Used for kMD Studies As Taken from the FPR High-Throughput Screening Results (PubChem Assay ID 440)

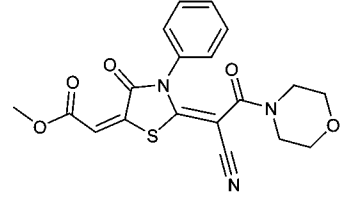
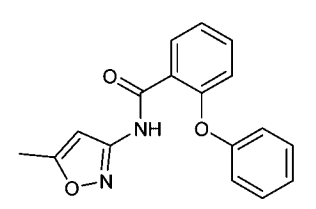
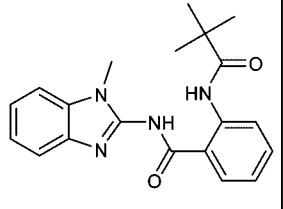
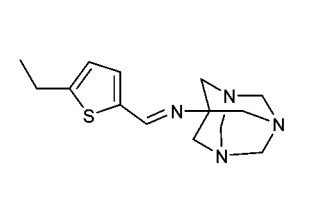
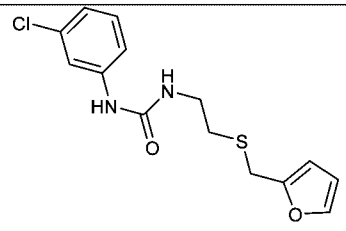
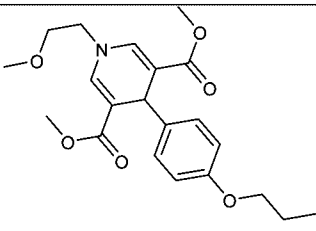
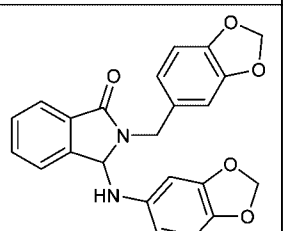
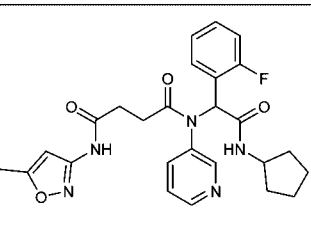
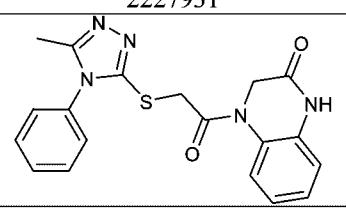
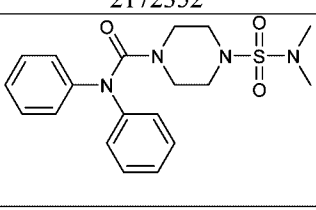
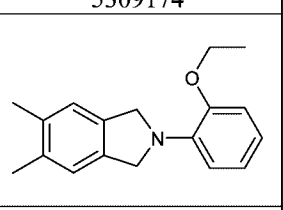
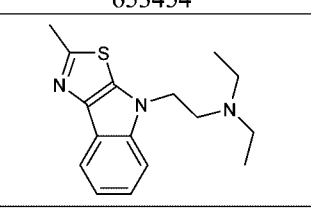
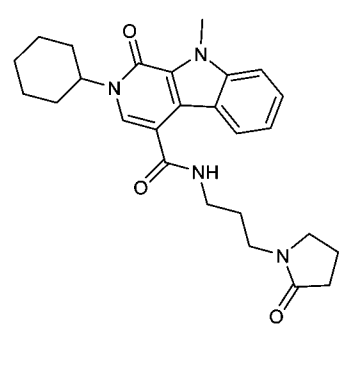
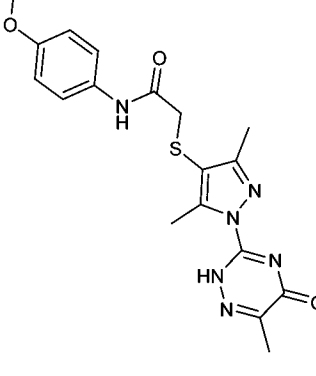
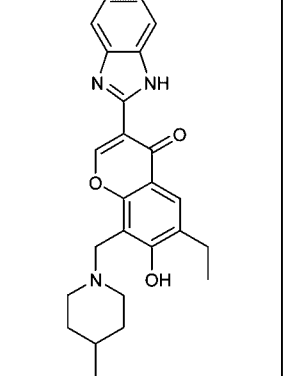
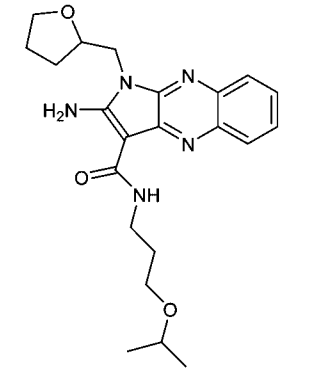
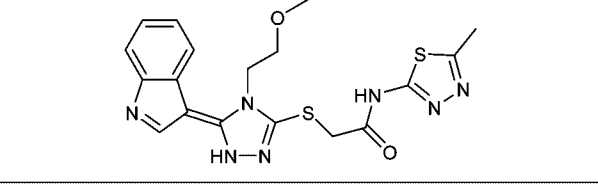
3243503	5739108	658816	666948
5389834	2911378	443084	661728
3242292	2949891	658811	5766180
2947766	2236750	655756	646700
4781			

approach is not sensitive to initial configurations of the system; in fact, the structures of interacting molecules do not need to be known. Second, although molecular mechanics force fields are utilized for simulation and to quantify interaction potentials, molecular properties are not directly derived from energies that result from atom type parametrization. Therefore, this “learning” aspect of property prediction in kMD might allow for calculations that are robust in the face of parameter uncertainty. At the least, problems due to atom type extrapolation should

reveal themselves during training in the form of poor accuracies. Generating such statistics using traditional simulation is often too expensive.

Of course, these advantages do not come without trade-offs. First is the requirement for training data. We do not know the amount of training data required for model development; however, we certainly expect an increase in accuracy with an increase in the amount of data. For certain problems, enough data will not be available. The increase in publicly available databases and high-throughput methods

Table 3. Inactive Compounds Used for kMD Studies As Taken from the FPR High-Throughput Screening Results (PubChem Assay ID 440)

			
661715	652253	1076051	779986
			
2227931	2172352	5309174	653454
			
2109473	1302022	954417	664312
			
3245762	1244388	1391485	663856
			
5771345			

should help with many cases. Second, kMD achieves efficient calculations at the cost of atomic-detail in time-resolved intermolecular interactions within a system. While we have implemented methods for identifying key interactions in the model that contribute to activity, it is important to note that kMD utilizes a simulation where a given molecule is unperturbed by interacting molecules within the system. While kMD can reveal interactions that differentiate accuracy, these interactions are not necessarily a reflection of actual interactions within a given system. Nonetheless, these models reveal information pertaining to how the structure of a given molecule relates to activity and, likewise,

information useful for the design of novel molecules with desired traits. Consider, for example, the problem of engineering a protein with improved ligand selectivity. A kMD model can be trained by assigning desired ligands into one class and problem ligands into another. The resulting model can identify important intermolecular interactions (in the form of probe interaction energies) that differentiate the two classes and can be used to guide the design of improved binding pockets.

kMD addresses the sampling problem in simulation approaches by reducing the complexity of the simulation—the problem of rough energy landscapes might still be an

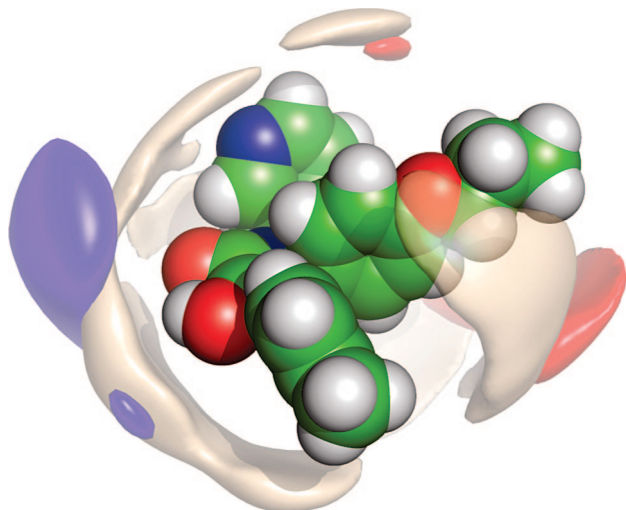


Figure 5. Visualization of the contribution of probe interaction potentials to the binding affinity of ligands with FPR. In this case, surfaces are interpolated for each probe separately. Wheat represents key uncharged probe interactions, blue represents key interactions for the positive probe, and red represents key interactions for the negative probe.

issue. Here, we have performed “vanilla” molecular dynamics simulations using single geometries as a starting point, and we have not performed any weighting based on conformational energy. This approach assumes that the sampling is sufficient to accurately represent probabilistic molecular conformation. Although we were able to obtain accurate models with this approach, this may not be the case in general. In this regard, more sophisticated MD or MC approaches that offer improved phase-space sampling should also benefit conformational studies in kMD. Sampling problems resulting from local minima traps can be identified in kMD studies by assessing self-similarity. That is, the kernel computed with simulations of the same molecule at different starting conformations should be approximately 1. In cases where this is not true, the simulation must be adjusted to improve sampling, or multiple simulations need to be utilized in order to account for local minima. Additional approaches might be utilized to improve the sampling efficiency in kMD. Approaches that take advantage of molecular vibration theory to allow for an analytic treatment of high-frequency vibrations can improve sampling efficiency for molecules with a single equilibrium configuration and provide a natural separation of internal motion from overall rotation and translation.^{30,31} These approaches might also be utilized to allow for efficient kMD simulations in explicit solvent.

We have performed initial studies of kMD on the binding affinity of small ligands to proteins because it facilitates comparison to alternative informatics approaches and because it is an important component of a current NIH Roadmap. Analysis of the FPR screening results represents a difficult problem for informatics approaches due to the small size of the data set and the variation in chemical structure. For simulation approaches, the data set is very large; simulation of ligand binding for a single molecule remains a challenge with MC and MD approaches on modern high-performance computers. Despite the small size of the simulations used in this study, the approach is also potentially applicable to more ambitious problems such as protein engineering. Assuming

that the interaction fields can be limited to a region of interest (e.g., a binding pocket), kMD calculations can be performed for large proteins. Given conformational data, the time complexity of the alignment and similarity calculations are independent of the number of atoms. kMD can also potentially be applied to the molecular recognition problem. By including positive interaction potentials and inverting the sign of the potentials for one of the molecules involved, the kernel presented represents an objective function for the docking problem allowing for ligand and receptor flexibility.

Despite its importance, the problem of extrapolation is often swept under the rug. The use of predictive models is limited if one cannot determine the domain over which the predictions are valid. In molecular physics applications, these problems arise due to the fitting of empirical potentials along with the approximations necessary for an accessible numerical implementation of the mathematic formulations. In informatics approaches, these issues are on the forefront due to the limited chemical diversity in molecules available for training. Although cross-validation provides some measure of predictive ability, the nature from which the experimental data is derived is likely to limit this accuracy to a relatively small portion of the chemical space.

Here, we have used an encoding based on dynamic potential surfaces in order to side-step the problem of extrapolation due to novel chemical structures. Indeed, we were able to achieve good results with the diverse HTS chemicals. The problem of extrapolation remains, however, as highlighted by the change in accuracy coupled to the change in the random set of negatives. The issue of extrapolation in kMD is related to the intermolecular interaction potentials seen during training. Extreme examples of extrapolation include calculation on a molecule that is much larger than those used for training or calculation on a molecule with a novel mechanism of action (e.g., allosteric regulation at a different binding site). Information theoretic methods for identifying problems due to extrapolation are currently being investigated and hopefully will also lead to quantitative methods for selecting molecules to be used for training. For example, in this work, we have selected training molecules by using all that were available for training (steroid data set) or by selecting random molecules with low and high activities. In general, supervised learning can provide improved results by selecting a training set with a wide range of activities. While this provides a straightforward approach for reducing extrapolation, selecting a training set with chemical diversity can be an important criterion for the development of general models.³² For kMD, this might be accomplished using existing methods for assessing chemical diversity or new methods that identify chemical diversity in terms of the diversity in the molecule’s dynamic force fields. Likewise, these approaches have the potential to identify molecules for which predictions will likely result in an inaccurate extrapolation. Certainly, this is an important area for future investigation.

In addition to training set selection and domain of applicability, future research into the use of alternative interaction fields could lead to more accurate kMD models. Here, we have used charged and uncharged probes with a fixed van der Waals radius (see the Experimental Section). These are the traditional probes used in CoMFA as they account for van der Waals and electrostatic interactions. It

has been shown however that the use of additional interaction fields can improve the accuracy of CoMFA, and this strategy might also benefit kMD. For example, although differences in hydrogen-bonding potential might be detectable using only Coulombic interactions, many force fields include an explicit anisotropic hydrogen-bonding term in addition to van der Waals and Coulombic interactions. A hydrogen-bonding probe for 3D-QSAR models has been described and used to generate improved models.³³ Additional work to incorporate lipophilicity fields,³⁴ desolvation fields,³⁵ and molecular orbital fields³⁶ has also been performed and allows for the potential to include effects such as the influence of desolvation more directly into a kMD model. Other quantities such as the thermodynamic entropy for some processes are only included implicitly in a model through changes in the dynamic interaction fields. Of course, any measurable property can also be included explicitly as weighted descriptors in the supervised learning to allow for expansion of the model beyond interaction fields.

EXPERIMENTAL SECTION

Kernel Molecular Dynamics. We obtain the dynamic representation of molecular conformation $\mathbf{r}_m(t)$ for a molecule in isolation using molecular dynamics, such that t represents time. For the first data set, MD was performed using the Tinker implementation of the MM3 force field.³⁷ Specifically, Beeman integration with a Berendsen thermostat was performed for 200,000 fs (fs) using a time step of 1 fs with a temperature of 300 K and a pressure of 1 atm. For the second data set, the Sybyl implementation of the MMFF94 force field³⁸ was used due to the generality of the force field in terms of small organic atom types. Starting conformations for the MD simulations for each molecule were generated using BFGS minimization to an rms gradient of 0.01.

In order to perform efficient calculations of intermolecular interaction fields Φ_{m,q_v} , we 1) limit the probes to single atoms, taking advantage of the pairwise nature of the intermolecular interaction energies in MM3 and MMFF94 and 2) utilize cubic B-splines³⁹ to represent energy. B-splines were chosen so that the time complexities for integration (eq 1) and alignment (eqs 2 and 3) are independent of the number of atoms in the molecule and because the resulting interpolation is C^2 continuous allowing for derivative calculation during alignment. Three probe atoms were used: a neutral atom, an atom with a +0.5 charge, and an atom with a -0.5 charge all with a constant van der Waals radius equal to that of carbon in the respective force fields. For each probe atom, the interaction field corresponding to a given conformation was calculated utilizing B-splines on a uniform grid. As with CoMFA, the grid dimensions should be large enough to include interaction sites surrounding any conformation of the molecule. Here we have used dimensions equal to the bounding box of $\mathbf{r}_m(t)$ plus 15 Å with a 1 Å resolution for the grids. As described below, the Vegas integration for numerical solution of eq 1 is parametrized so that edge effects resulting from the finite grid sizes are negligible.

Because MD calculations are performed with numerical integration using discrete time steps, we must estimate the underlying probability density ρ_{m,q_j} for a given interaction field. This was performed using Parzen windows⁴⁰ with a Gaussian function such that

$$\rho_{m,q_v}(\mathbf{r}, \varphi) = \frac{1}{\tau} \sum_{i=1}^{\tau} \frac{1}{\sigma\sqrt{2\pi}} e^{-(\varphi - \varphi(\mathbf{r}, \tau))^2/2\sigma^2} \quad (8)$$

where τ ranges over a random set of MD time points. Because the integral product in eq 1 has a time complexity of $O(\tau^2)$, we also considered a more efficient approach for alignment whose complexity is independent of τ . We did this by calculating a vector of discrete probabilities $\mathbf{w}_{m,q_v} = \{\omega_{m,q_{v1}}, \omega_{m,q_{v2}}, \dots, \omega_{m,q_{vs}}\}$ from the probability density between a minimum and maximum energy (φ_{\min} and φ_{\max}) such that

$$w_{m,q_v,i}(\mathbf{r}) = \int_{\varphi_{\min} + i(\Delta\varphi)/s}^{\varphi_{\min} + (i+1)(\Delta\varphi)/s} \rho_{m,q_v}(\mathbf{r}, \varphi) d\varphi \quad (9)$$

for $0 \leq i \leq s-1$, where $\Delta\varphi = \varphi_{\max} - \varphi_{\min}$. In this case, the integration of $\rho_{m_i,q_v}\rho_{m_j,q_v}$ over φ in eq 1 is replaced with a dot product $\mathbf{w}_{m_i,q_v} \cdot \mathbf{w}_{m_j,q_v}$, and the interpolating B-splines calculate a given probability rather than a given potential energy. For the investigations here, $s=10$ was used with a range of -1 to -0.1 for the uncharged probe and -20 to -2 for the charged probes.

The integrals in eq 1 were calculated using Vegas integration⁴¹ with 5 stages utilizing a total of 100,000 function evaluations. In this case, the integration intervals must be finite; however, if the approach is parametrized correctly, nonzero probabilities outside the integration ranges will be negligible, and the integrals will be equivalent to a certain precision. In order to achieve this, we integrated over φ using two times the smallest potential found as the lower limit of integration. We integrated over \mathbf{r} using the bounding box of the interpolation grids plus 50% of the largest interpolation grid. Because the Vegas integration is expensive and an analytic form for the derivative of eq 2 has not been obtained, during alignment we evaluated eq 2 by replacing the integral over \mathbf{r} with a summation over uniform grid points with a 1 Å spacing over the same integration range. In this case analytic similarities and derivatives can be calculated to allow for efficient alignment.

We performed the alignments represented by eq 3 using a hybrid genetic algorithm (GA) with a local search operator that enforced a 0.02 probability of performing 3 iterations of BFGS conjugate gradient minimization. A consistent initial positioning for each trajectory was obtained using principal-axes transformations. An initial population size of 50 was used, and the population was seeded with 8 genomes that result in no translation of the trajectory and all permutations of 0° and 180° rotations along each of the principal axes. This asserted that the optimization would evaluate similarities with each trajectory centered at the origin and rotated onto its principal axes. Power law scaling with an exponent of 1.5 was utilized for fitness evaluation, and the probabilities for crossover and mutation were set to 0.9 and 0.02, respectively. The GA was evaluated for 50 generations followed by full BFGS minimization to a gradient of 0.001 with a maximum of 100 iterations.

In order to evaluate eq 4, a single transformation for each trajectory was calculated based on the full pairwise similarity matrix given by eq 3. In this process, each molecule is placed into a unique set. Sets are merged by aligning each of the molecules in one set to another set using the single transformation identified to align the two molecules in the respective sets that have the highest similarity as calculated using eq 3. The process is repeated until only one set remains. Eq 5 was obtained using the SVM implementation in

SVM^{light17} which was modified to accept a precalculated kernel matrix and to compute full cross-validation statistics for both regression and classification. Finally, model visualization was accomplished based on eq 7, by calculating coefficients for each probe on a regular 1 Å grid. Isosurfaces were calculated using the Marching Cubes algorithm with cubic B-spline interpolation for vertex and normal placement. Surface triangulations and molecular conformations were rendered using the software Pymol 0.97.

Synthesis of Resveratrol Analogues. Details on the synthesis of the resveratrol analogues along with compound data and NMR characterization for new compounds are given in the Supporting Information.

NF- κ B Assay. An NF- κ B reporter stable cell line derived from human 293T embryonic kidney cells (293T/NF- κ B-luc) (Panomics, Inc., Redwood City, CA) was grown in a humidified atmosphere at 37 °C in 5% CO₂/95% air. The cells were maintained in Dulbecco's Modified Eagle's Medium (DMEM - high glucose containing 4 mM glutamine) supplemented with 10% fetal bovine serum (FBS), 1 mM sodium pyruvate, 100 units/mL of penicillin, 100 µg/mL of streptomycin, and 100 µg/mL of hygromycin (Gibco/Invitrogen, Carlsbad, CA) to maintain cell selection. One day prior to treatment, the 293T/NF- κ B-luc cells were plated into 24-well cell culture plates (Costar, Cambridge, MA) at approximately 70% confluency in the above media without hygromycin. The following day cells were fed fresh media 1 h prior to treatment. Media with or without recombinant tumor necrosis factor alpha (TNF- α) (R&D Biosciences/Clontech, Palo Alto, CA) was then applied to the cells at 20 ng/mL followed by immediate treatments with resveratrol or substituted stilbene. The cells were placed again in a humidified atmosphere at 37 °C in 5% CO₂/95% air for 7 h. Plate wells were gently washed with phosphate buffered saline, pH 7.4, and lysed with 1x passive lysis buffer (Promega, Madison, WI). The subsequent lysates were analyzed with the Luciferase Assay System (Promega) utilizing a TD-20/20 luminometer (Turner Designs, Sunnyvale, CA). The firefly luciferase relative light units were normalized to protein (mg/mL) with BCA Protein Assay Kit (Pierce, Rockford, IL) and standardized to percent of control (TNF- α control).

For assays of cell viability, cells were treated similarly as above and with 15 µM substituted *trans*-stilbene. After washing, cells were treated with 100 µL of media and 20 µL of CellTiter 96 AQueous One Solution reagent for 1 h and then read at 490 nm with a Spectromax plate reader.

ACKNOWLEDGMENT

We thank Aidan Thompson and Steve Plimpton for their critical review of the work. Funding for this work was provided by Sandia National Laboratories under DOE contract DE-AC04-94AL85000 and through an interagency agreement (IAG) DW89921601 with the Environmental Protection Agency. Sandia is a multiprogram laboratory operated by Sandia Corp., a Lockheed Martin Company, for the U.S. Department of Energy (DOE)'s National Nuclear Security Administration. Ariella Sasson is pleased to announce support from the DOE CSGF fellowship DE-FG02-97ER25308. Funding for the synthesis and assay of resveratrol analogues was supported by grant BC043125 from the U.S. Army/DOD Breast Cancer Program.

Supporting Information Available: Structures and activities for resveratrol analogues and candidate molecules are detailed along with experimental data describing the synthesis and verification of new compounds and PubChem compound IDs and structures for inactive compounds used in additional models generated for the formylpeptide receptor assay. This material is available free of charge via the Internet at <http://pubs.acs.org>.

REFERENCES AND NOTES

- (1) Berg, B. A.; Neuhaus, T. Multicanonical Ensemble - A New Approach to Simulate 1st-Order Phase-Transitions. *Phys. Rev. Lett.* **1992**, *68*, 9–12.
- (2) Grubmüller, H. Predicting Slow Structural Transitions in Macromolecular Systems - Conformational Flooding. *Phys. Rev. E* **1995**, *52*, 2893–2906.
- (3) Hansmann, U. H. E. Parallel Tempering Algorithm for Conformational Studies of Biological Molecules. *Chem. Phys. Lett.* **1997**, *281*, 140–150.
- (4) Laio, A.; Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12562–12566.
- (5) Lyubartsev, A. P.; Martsinovski, A. A.; Shevkunov, S. V.; Vorontsov-ovlyaminov, P. N. New Approach to Monte-Carlo Calculation of the Free-Energy - Method of Expanded Ensembles. *J. Chem. Phys.* **1992**, *96*, 1776–1783.
- (6) Schlitter, J.; Engels, M.; Kruger, P.; Jacoby, E.; Wollmer, A. Targeted Molecular-Dynamics Simulation of Conformational Change - Application to the T[−]R Transition in Insulin. *Mol. Simul.* **1993**, *10*, 291–308.
- (7) Stoltovitzky, G.; Berne, B. J. Catalytic Tempering: A Method for Sampling Rough Energy Landscapes by Monte Carlo. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 11164–11169.
- (8) Sugita, Y.; Okamoto, Y. Replica-Exchange Molecular Dynamics Method for Protein Folding. *Chem. Phys. Lett.* **1999**, *314*, 141–151.
- (9) Voter, A. F. A Method for Accelerating the Molecular Dynamics Simulation of Infrequent Events. *J. Chem. Phys.* **1997**, *106*, 4665–4677.
- (10) Voter, A. F.; Montalenti, F.; Germann, T. C. Extending the Time Scale in Atomistic Simulation of Materials. *Annu. Rev. Mater. Res.* **2002**, *32*, 321–346.
- (11) Wales, D. J.; Doye, J. P. K. Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms. *J. Phys. Chem. A* **1997**, *101*, 5111–5116.
- (12) Sheridan, R. P.; Feuston, B. P.; Maiorov, V. N.; Kearsley, S. K. Similarity to Molecules in the Training Set is a Good Discriminator for Prediction Accuracy in QSAR. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1912–1928.
- (13) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular-Field Analysis (CoMFA) 0.1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.
- (14) Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B. Q.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. Construction of 3D-QSAR Models using the 4D-QSAR Analysis Formalism. *J. Am. Chem. Soc.* **1997**, *119*, 10509–10524.
- (15) Klebe, G.; Abraham, U.; Mietzner, T. Molecular Similarity Indexes in a Comparative-Analysis (CoMSIA) of Drug Molecules to Correlate and Predict their Biological-Activity. *J. Med. Chem.* **1994**, *37*, 4130–4146.
- (16) Eckart, C. Some Studies Concerning Rotating Axes and Polyatomic Molecules. *Phys. Rev.* **1935**, *47*, 552–558.
- (17) Joachims, T. Making Large-Scale SVM Learning Practical. In *Advances in Kernel Methods-Support Vector Learning*; Scholkopf, B., Burges, C. J. C., Smola, A. J., Eds.; MIT Press: Cambridge, MA, 1999; pp 169–184.
- (18) Wagener, M.; Sadowski, J.; Gasteiger, J. Autocorrelation of Molecular-Surface Properties for Modeling Corticosteroid-Binding Globulin and Cytosolic Ah Receptor Activity by Neural Networks. *J. Am. Chem. Soc.* **1995**, *117*, 7769–7775.
- (19) Pavlov, D.; Mao, J.; Dom, B. In *Scaling-Up Support Vector Machines using Boosting Algorithm*, Proceedings of the 15th International Conference on Pattern Recognition, Barcelona, Spain, 2000; IEEE: Barcelona, Spain, 2000; pp 219–222.
- (20) Zhou, Y. P.; Jiang, J. H.; Lin, W. Q.; Zou, H. Y.; Wu, H. L.; Shen, G. L.; Yu, R. Q. Boosting Support Vector Regression in QSAR Studies of Bioactivities of Chemical Compounds. *Eur. J. Pharm. Sci.* **2006**, *28*, 344–353.
- (21) Coats, E. A. The CoMFA Steroids as a Benchmark Dataset for Development of 3D QSAR Methods. *Perspect. Drug. Discovery Des.* **1998**, *12*, 199–213.

- (22) Dore, S. Unique Properties of Polyphenol Stilbenes in the Brain: More than Direct Antioxidant Actions; Gene/protein Regulatory Activity. *Neurosignals* **2005**, *14*, 61–70.
- (23) Kundu, J. K.; Surh, Y. J. Molecular Basis of Chemoprevention by Resveratrol: NF-kappa B and AP-1 as Potential Targets. *Mutat. Res.-Fund. Mol. M.* **2004**, *555*, 65–80.
- (24) Shimizu, M.; Weinstein, I. B. Modulation of Signal Transduction by Tea Catechins and Related Phytochemicals. *Mutat. Res.-Fund. Mol. M.* **2005**, *591*, 147–160.
- (25) Jang, M. S.; Cai, E. N.; Udeani, G. O.; Slowing, K. V.; Thomas, C. F.; Beecher, C. W. W.; Fong, H. H. S.; Farnsworth, N. R.; Kinghorn, A. D.; Mehta, R. G.; Moon, R. C.; Pezzuto, J. M. Cancer Chemopreventive Activity of Resveratrol, A Natural Product Derived from Grapes. *Science* **1997**, *275*, 218–220.
- (26) Baur, J. A.; Sinclair, D. A. Therapeutic Potential of Resveratrol: The in vivo Evidence. *Nat. Rev. Drug Discovery* **2006**, *5*, 493–506.
- (27) Heynekamp, J. J.; Weber, W. M.; Hunsaker, L. A.; Gonzales, A. M.; Orlando, R. A.; Deck, L. M.; Jagt, D. L. V. Substituted trans-Stilbenes, Including Analogues of the Natural Product Resveratrol, Inhibit the Human Tumor Necrosis Factor Alpha-Induced Activation of Transcription Factor nuclear factor KappaB. *J. Med. Chem.* **2006**, *49*, 7182–7189.
- (28) Gao, J. L.; Lee, E. J.; Murphy, P. M. Impaired Antibacterial Host Defense in Mice Lacking the N-formylpeptide Receptor. *J. Exp. Med.* **1999**, *189*, 657–662.
- (29) Zhou, Y.; Bian, X. W.; Le, Y. Y.; Gong, W. H.; Hu, J. Y.; Zhang, X.; Wang, L. H.; Iribarren, P.; Salcedo, R.; Howard, O. M. Z.; Farrar, W.; Wang, J. M. Formylpeptide Receptor FPR and the Rapid Growth of Malignant Human Gliomas. *J. Natl. Cancer I* **2005**, *97*, 823–835.
- (30) Janezic, D.; Praprotnik, M.; Merzel, F. Molecular Dynamics Integration and Molecular Vibrational Theory. I. New Symplectic Integrators. *J. Chem. Phys.* **2005**, *122*.
- (31) Praprotnik, M.; Janezic, D. Molecular Dynamics Integration Meets Standard Theory of Molecular Vibrations. *J. Chem. Inf. Model.* **2005**, *45*, 1571–1579.
- (32) Golbraikh, A. Molecular Dataset Diversity Indices and their Applications to Comparison of Chemical Databases and QSAR Analysis. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 414–425.
- (33) Kim, K. H. 3D-Quantitative Structure-Activity-Relationships - Describing Hydrophobic Interactions Directly from 3D Structures using a Comparative Molecular-Field Analysis (CoMFA) Approach. *Quant. Struct.-Act. Relat.* **1993**, *12*, 232–238.
- (34) Du, Q. S.; Liu, P. J.; Mezey, P. G. Theoretical Derivation of Heuristic Molecular Lipophilicity Potential: A Quantum Chemical Description for Molecular Solvation. *J. Chem. Inf. Model.* **2005**, *45*, 347–353.
- (35) Norinder, U. Recent Progress in CoMFA Methodology and Related Techniques. *Perspect. Drug. Discovery Des.* **1998**, *12*, 25–39.
- (36) Waller, C. L.; Marshall, G. R. 3-Dimensional Quantitative Structure-Activity Relationship of Angiotensin-Converting Enzyme and Thermolysin Inhibitors 0.2. A Comparison of CoMFA Models Incorporating Molecular-Orbital Fields and Desolvation Free-Energies Based on Active-Analog and Complementary-Receptor-Field Alignment Rules. *J. Med. Chem.* **1993**, *36*, 2390–2403.
- (37) Allinger, N. L.; Yuh, Y. H.; Lii, J. H. Molecular Mechanics - The MM3 Force-Field for Hydrocarbons. 1. *J. Am. Chem. Soc.* **1989**, *111*, 8551–8566.
- (38) Halgren, T. A. Merck Molecular Force Field. 1. Basis, Form, Scope, Parameterization, and Performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490–519.
- (39) Oberlin, D.; Scheraga, H. A. B-Spline Method for Energy Minimization in Grid-Based Molecular Mechanics Calculations. *J. Comput. Chem.* **1998**, *19*, 71–85.
- (40) Parzen, E. Estimation of a Probability Density-Function and Mode. *Ann. Math. Statist.* **1962**, *33*, 1065–1076.
- (41) Lepage, G. P. New Algorithm for Adaptive Multidimensional Integration. *J. Comput. Phys.* **1978**, *27*, 192–203.

CI8001233