SAND2007-0338J

Efficient Calculation of Molecular Properties from Simulation using Kernel Molecular Dynamics

W. Michael Brown* **and Shawn Martin**

Computational Biology, Sandia National Laboratories, PO Box 5800, M/S 1316, Albuquerque, NM 87185-1316

*Corresponding Author: William Michael Brown, Sandia National Laboratories, PO Box 5800, M/S 1316, Albuquerque, NM 87185-1316, Phone: 505-284-8938, Fax: 505-845-7442, E-Mail: wmbrown@sandia.gov

Classification: PHYSICAL SCIENCE (Chemistry) / BIOLOGICAL SCIENCES (Biophysics)

Page Count (*including title*)**:** 22
Figure Count: 4
Table Count: 1 in Supplementary Material
Word Count: 5694
Character Count: 36460

Abbreviations: kMD, kernel molecular dynamics; MD, molecular dynamics; CoMFA, comparative molecular field analysis; SVM, support vector machine; QSPR, quantitative structure-property relationship

Abstract

Atomistic simulation for molecular property calculation and elucidation of structure-property relationships is limited in scope due to constraints on system sizes, time scales, and energy landscapes. For many relevant processes in biology and chemistry, a sampling of the phase space sufficient for accurate property calculations cannot be obtained. Here, we introduce a novel formalism that utilizes supervised learning to reduce the complexity of simulations required for property calculation in complex processes. In contrast to traditional informatics approaches utilized for studies on proteins or small molecules, learning is achieved based on molecular descriptions that are rooted in the physics of dynamic intermolecular forces. We demonstrate the efficacy of the approach with calculations of the binding affinity of small organics to proteins based on molecular dynamics simulations.

Introduction

The problem of molecular property prediction is central to many fields within biology and chemistry including 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 bioinformatics and molecular biophysics, cheminformatics and computational chemistry, 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 and likewise 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 towards 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, forcing 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 by fitting equations for a given property using higher level abstractions for molecular description that are 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 due to their ability to encode information more closely related to molecular interaction; however, such models require the selection of an "active conformation". While methods 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 is required in order to quantify molecule properties in terms of dynamical molecular interaction fields, rather than 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 and does not 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.

Methodology

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 within a reasonable time. A typical approach to handling such difficulties is to seek higher level formulations based on 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 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 Fig. 1 and give a formal description below.

Analytic kMD

We consider the case where we have a single assay for a given molecular property P that we would like to quantify. Denote by $\mathcal{M} = \{ m_1, m_2, \ldots, \}$ the set of all molecules. For a given molecule $m \subseteq \mathcal{M}$, 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 $\mathcal{M}$ 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 molecule in question. Therefore, any changes affecting the property P should be inherent to the molecule m itself. This suggests the existence of a function $f: \mathcal{M} \rightarrow \mathbb{R}$ 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 $Q = \{ q_1, q_2, \ldots, q_k \}$ 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 $\Phi_{m,q_v} : \mathbb{R}^3 \times [0, t_m] \to \mathbb{R}$ 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.

In order to obtain *f*, we consider kernel methods for learning and therefore require a kernel function $k : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ that gives the similarity between two molecules m_i and m_j in terms of $\Phi_{m_i,q_v}(\mathbf{r},t)$ and $\Phi_{m_j,q_v}(\mathbf{r},t)$. 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 $\rho_{m,q_j}(\mathbf{r},\varphi)$ the probability density function for probe interaction potential such that $\int_{\varphi=c}^{d} \rho_{m,q_v}(\mathbf{r},\varphi) = \Pr(c \le \Phi_{m,q_v}(\mathbf{r},t) \le d)$. We can then define a similarity kernel,

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

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 **r** introduces a frame of reference problem which requires alignment

between molecules. We address this issue by enforcing the Eckart conditions in the form of holonomic restraints in order to separate rotations and translations of the molecules from those due to internal vibrations (15). Additionally, we parameterize 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,

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

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

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

Unfortunately, this is not necessarily a true inner product (a necessary condition for a kernel function) because it is not linear. 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 $\{m_1, m_2, \ldots, m_n\}$ there is a corresponding set of transformations $\{T_1, T_2, \ldots, T_n\}$ 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, provided that data is available where a

property has been measured for a set of training molecules. Here we utilize support vector machines (SVMs) (16) 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 non-linear 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.

Numerical Approximation to kMD

Here, 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 dataset, MD was performed using the Tinker implementation of the MM3 force-field (17). Specifically, Beeman integration with a Berendsen thermostat was performed for 200,000 femtoseconds (fs) using a time step of 1 fs with a temperature of 300K and a pressure of 1 atmosphere. For the second dataset, the Sybyl implementation of the MMFF94 force-field (18) 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 pair-wise nature of the intermolecular interaction energies in MM3 and MMFF94 and 2) utilize cubic B-splines (19) 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 with a 1Å resolution and with dimensions equal to the bounding box of $\mathbf{r}_m(t)$ plus 15Å.

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 windows (20) with a Gaussian function such that

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

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 probabilties $\mathbf{w}_{m,q_v} = \{\omega_{m,q_v,1}, \omega_{m,q_v,2}, ..., \omega_{m,q_v,n}\}$ from the probability density between a minimum and maximum energy ($\varphi_{\min}$ and $\varphi_{\max}$) 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 $0 \le i \le s$-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 integration (21) 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 parameterized correctly, non-zero 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 **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 **r** with a summation over uniform grid points with a 1Å 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 .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 pair-wise 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^{light} (16) which was modified to accept a pre-calculated 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 1Å 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.

Results

For the first dataset, we analyzed corticosteroid binding globulin (CBG) binding affinity for a set of steroids as described in Ref. (22). This dataset was first compiled for evaluation of CoMFA (13) and has since become a benchmark for 3D-QSPR approaches. The dataset consists of 31 compounds with p*K* values ranging from -5 to -7.881. We chose the MM3 force-field for MD and kMD calculations for this dataset. However, parameters for 11β, 17, 21-trihydroxy-2α-methyl-9α-fluoro-4-pregnene-3,20-dione were not available and therefore this compound was not included in initial tests. Because this compound has also been identified as an outlier (22), 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. (22) as the starting point for MD calculations. Conformations at 50 random time points were taken for continuous PDF calculation (Eq. 8). 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 squared correlation coefficient (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 squared correlation coefficient (r^2) when trained on all molecules was 0.9. For the non-linear model generated with the RBF kernel, a q^2 of 0.86 and an r^2 of 0.93 were obtained (c=3.8, γ=1.3). We next tested the discrete probability

approximation (Eq. 9), 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, γ=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, γ=0.3) for the RBF. When the full dataset is used (31 instead of 30), a q^2 of 0.76 and an r^2 of 0.83 (c=2.7) is obtained for the linear kernel and a q^2 of 0.86 and an r^2 of 0.93 (c=5.42, γ=0.3) is obtained for the RBF. We believe the decrease in accuracy is not due to parameterization, but rather the unique substitution of the steroid ring within this compound. A correlation plot for the RBF model on the full dataset is shown in Fig. 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 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 Fig. 3. In this example, we have interpolated an isosurface showing the locations where negative interaction potentials of the uncharged probe enhance binding affinity according to the model. 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.

The steroid molecules are relatively rigid and share a common scaffold (the cyclophenanthrene nucleus). For this reason, this dataset 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 dataset with a variety of methods is given in Ref. (23). The q^2 of 0.86 on the full dataset 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 dataset). For this same dataset, 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 datasets or methods; however, a discussion of these results has been given (23).

The fact that kMD offers improved accuracy in the calculation of steroid binding affinity is not the sole point of this work. The steroid dataset 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. Therefore, for the second 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 MLSCN 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. Details on the assay are available through the National Library of Medicine PubChem site (assay ID 440). The assay identified 17 active compounds and 9965 inactive compounds (from which 17 were chosen at random). The resulting set of compounds is illustrated in the Supporting Information. For this dataset, we implemented MMFF94 capabilities into kMD and took conformations from 5000 MD time points for PDF generation.

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 steroid activity for the final model is shown in Fig. 4. 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.

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 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 parameterization. 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 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 issue. In this regard, MD and 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.

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 dataset and the variation in chemical structure. For simulation approaches, the dataset 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.

Acknowledgements

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.

References

1. **Berg BA, Neuhaus T (1992) *Phys Rev Lett* 68:9-12.**
2. **Grubmuller H (1995) *Phys Rev E* 52:2893-2906.**
3. **Hansmann UHE (1997) *Chem Phys Lett* 281:140-150.**
4. **Laio A, Parrinello M (2002) *Proc Natl Acad Sci USA* 99:12562-12566.**
5. **Lyubartsev AP, Martsinovski AA, Shevkunov SV, Vorontsovvelyaminov PN (1992) *J Chem Phys* 96:1776-1783.**
6. **Schlitter J, Engels M, Kruger P, Jacoby E, Wollmer A (1993) *Mol Simul* 10:291-308.**
7. **Stolovitzky G, Berne BJ (2000) *Proc Natl Acad Sci USA* 97:11164-11169.**
8. **Sugita Y, Okamoto Y (1999) *Chem Phys Lett* 314:141-151.**
9. **Voter AF (1997) *J Chem Phys* 106:4665-4677.**
10. **Voter AF, Montalenti F, Germann TC (2002) *Annu Rev Mater Res* 32:321-346.**
11. **Wales DJ, Doye JPK (1997) *J Phys Chem A* 101:5111-5116.**
12. **Sheridan RP, Feuston BP, Maiorov VN, Kearsley SK (2004) *J Chem Inf Comput Sci* 44:1912-1928.**
13. **Cramer RD, Patterson DE, Bunce JD (1988) *J Am Chem Soc* 110:5959-5967.**
14. **Hopfinger AJ, Wang S, Tokarski JS, Jin BQ, Albuquerque M, Madhav PJ, Duraiswami C (1997) *J Am Chem Soc* 119:10509-10524.**
15. **Eckart C (1935) *Phys Rev* 47:552-558.**
16. **Joachims T (1999) in *Advances in Kernel Methods-Support Vector Learning*, eds Scholkopf B, Burges CJC, Smola AJ (MIT Press, Cambridge, MA), pp 169-184.**
17. **Allinger NL, Yuh YH, Lii JH (1989) *J Am Chem Soc* 111:8551-8566.**
18. **Halgren TA (1996) *J Comput Chem* 17:490-519.**
19. **Oberlin D, Scheraga HA (1998) *J Comput Chem* 19:71-85.**
20. **Parzen E (1962) *Ann Math Statist* 33:1065-1076.**
21. **Lepage GP (1978) *J Comput Phys* 27:192-203.**
22. **Wagener M, Sadowski J, Gasteiger J (1995) *J Am Chem Soc* 117:7769-7775.**
23. **Coats EA (1998) *Perspect Drug Discovery Des* 12:199-213.**

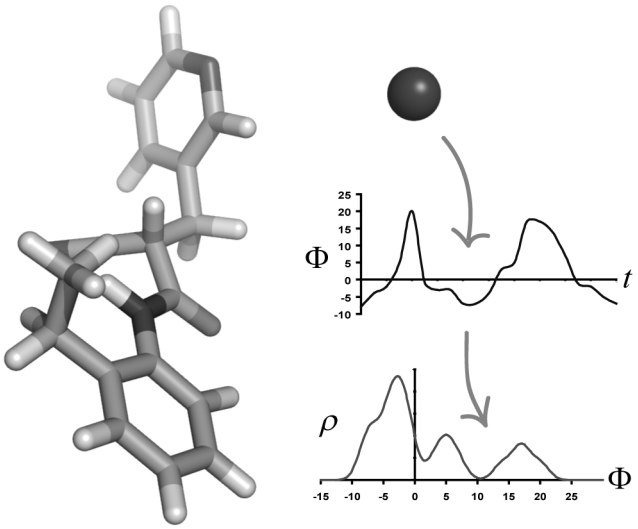


Figure 1. Calculation of dynamic molecular interaction fields in kMD. We measure the potential energy (Φ) of interaction between the probe (blue sphere) at a position relative to a molecule (sticks). This is performed for every conformation of the molecule as identified by simulation to give the upper 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.

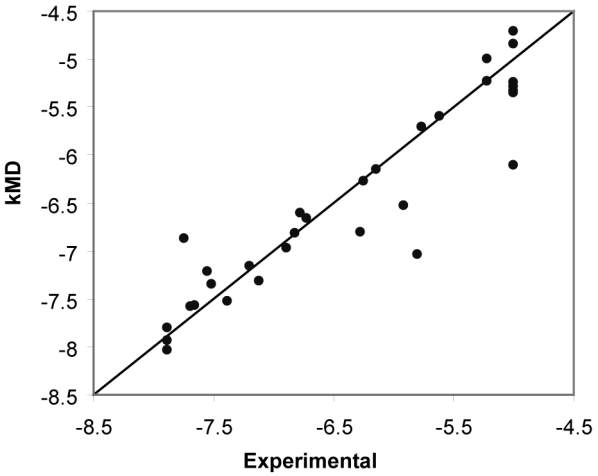


Figure 2. Correlation plot of experimental versus calculated binding affinities resulting from cross-validation using kMD with the RBF kernel on 31 steroids. The resulting q^2 is 0.86. Units are –log K.

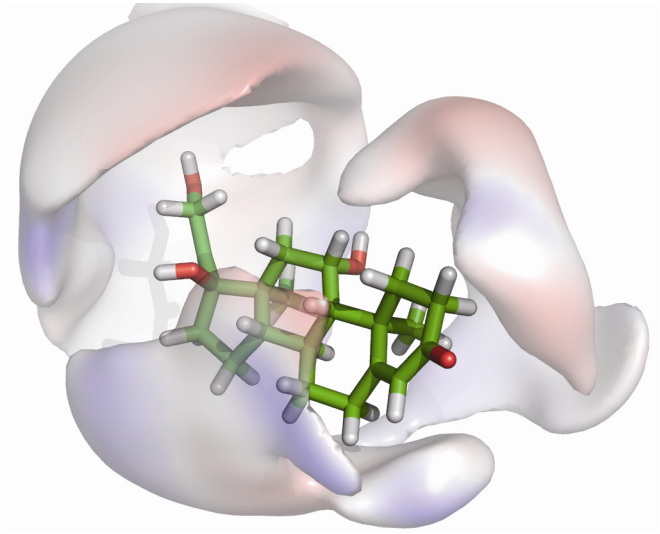


Figure 3. Visualization of the contribution of probe interaction potentials to the binding affinity of steroids with CBG. The surface represents an interpolation of equation coefficients projected onto a grid based on Eq. 7. The surface is interpolated using the uncharged probe and represents locations 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.

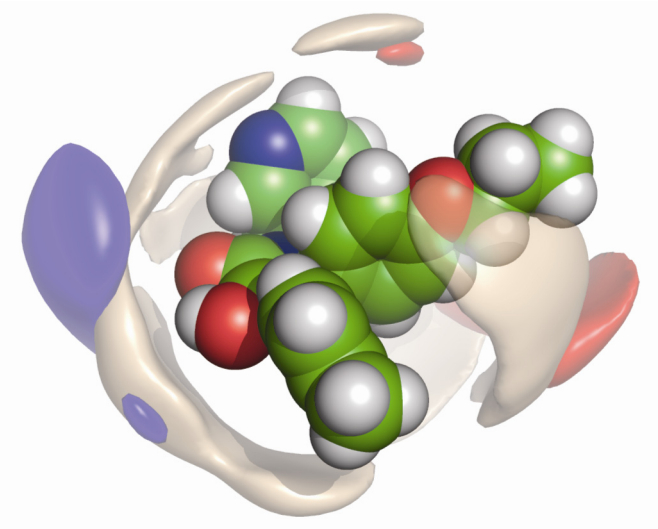
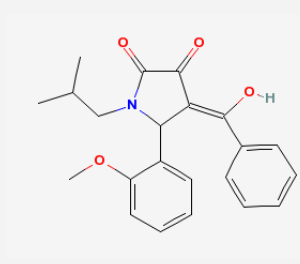
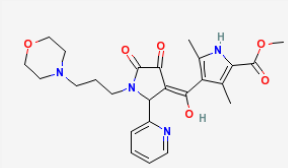
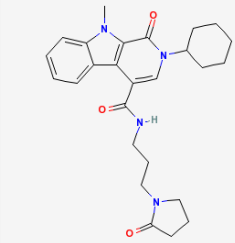
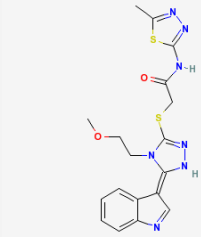
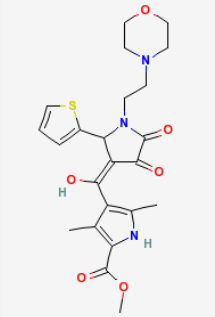
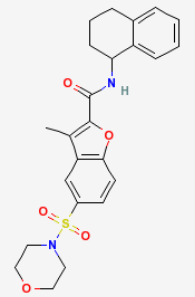
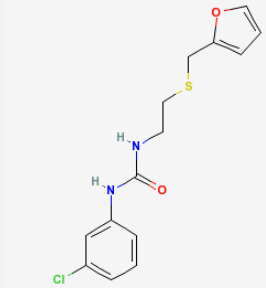
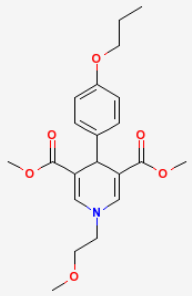
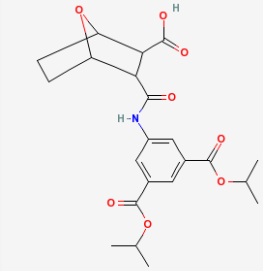
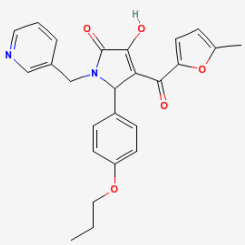
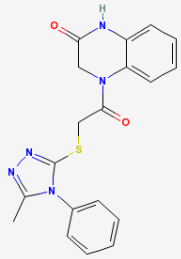
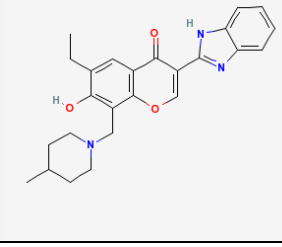
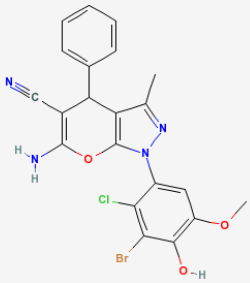
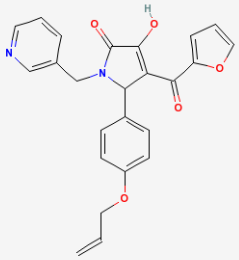
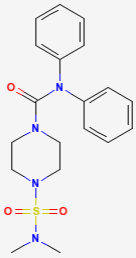
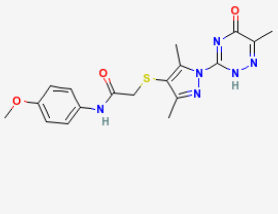
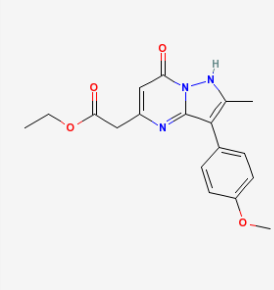
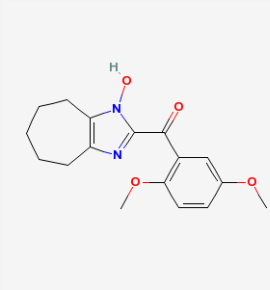
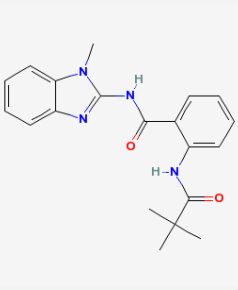
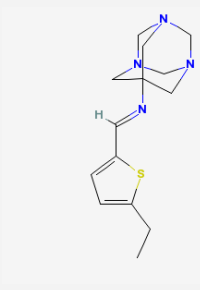
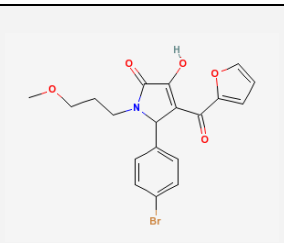
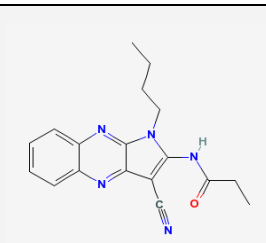
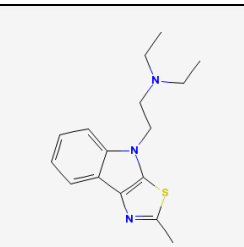
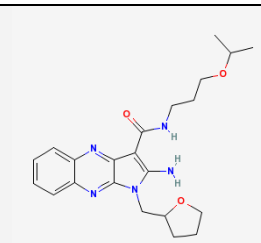
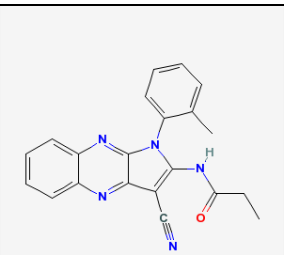
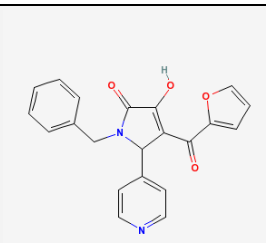
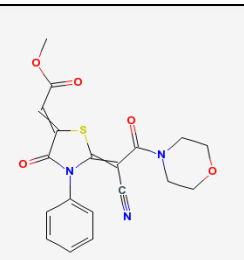
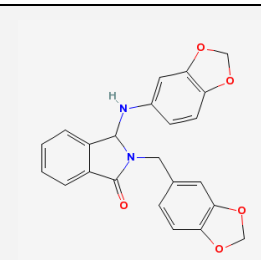
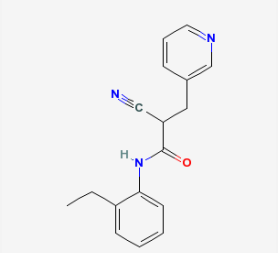
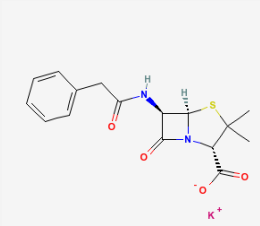
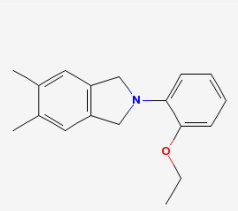
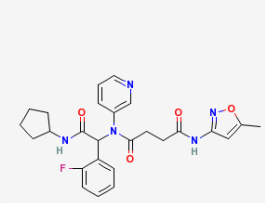
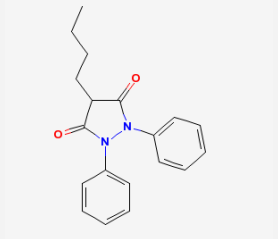
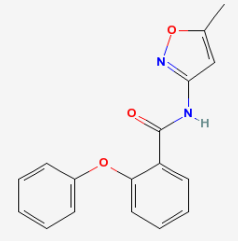


Figure 4. Visualization of the contribution of probe interaction potentials to the binding affinity of ligands with FPR. In this case, surfaces are interpolated for each probe separately. Wheat represents key uncharged probe interactions, blue represents key interactions for the positive probe, and red represents key interactions for the negative probe.

Active and inactive compounds used for kMD studies as taken from the FPR high-throughput screening results (PubChem Assay ID 440).

Active Compounds		Inactive Compounds	
			
5766180	5739108	3245762	5771345
			
5389834	3243503	2227931	2172352
			
3242292	2949891	2109473	1391485
			
2947766	2911378	1302022	1244388

Active Compounds		Inactive Compounds	
			
2236750	666948	1076051	779986
			
661728	658816	664312	663856
			
658811	655756	661715	5309174
			
646700	443084	954417	653454
			
4781		652253	