Published in final edited form as:

J Theor Comput Chem. 2013 December ; 12(8): . doi:10.1142/S0219633613410149.

OPTIMIZATION BIAS IN ENERGY-BASED STRUCTURE PREDICTION

Robert J. Petrella

Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, MA 02138, USA

Department of Medicine, Harvard Medical School, 25 Shattuck Street, Boston, MA 02115, USA

Robert J. Petrella: petrella_fas.harvard.edu

Abstract

Physics-based computational approaches to predicting the structure of macromolecules such as proteins are gaining increased use, but there are remaining challenges. In the current work, it is demonstrated that in energy-based prediction methods, the degree of optimization of the sampled structures can influence the prediction results. In particular, discrepancies in the degree of local sampling can bias the predictions in favor of the oversampled structures by shifting the local probability distributions of the minimum sampled energies. In simple systems, it is shown that the magnitude of the errors can be calculated from the energy surface, and for certain model systems, derived analytically. Further, it is shown that for energy wells whose forms differ only by a randomly assigned energy shift, the optimal accuracy of prediction is achieved when the sampling around each structure is equal. Energy correction terms can be used in cases of unequal sampling to reproduce the total probabilities that would occur under equal sampling, but optimal corrections only partially restore the prediction accuracy lost to unequal sampling. For multiwell systems, the determination of the correction terms is a manybody problem; it is shown that the involved cross-correlation multiple integrals can be reduced to simpler integrals. The possible implications of the current analysis for macromolecular structure prediction are discussed.

Keywords

Energy function; structure prediction; conformational search; optimization bias; sampling; macromolecules; energy correction; probability distribution function; parent distribution

1. Introduction

Energy functions based on physical models, statistical (e.g. structural database) information or both, are now commonly used to help predict molecular structure.^{1–5} Energy-based structure prediction is usually carried out by sampling a potential energy surface, comparing the energies of structures corresponding to different parts of the surface — i.e. different wells — and then choosing the lowest energy structures or clusters of structures (e.g. Ref.

6). For large molecules such as proteins and long nucleic acid polymers, or systems composed of such molecules, it is often difficult to carry out searches that are extensive enough to reliably include the region around the native or minimum-energy conformation. Hence, the development of efficient conformational search methods has been the focus of much research.^{7–12}

However, inclusion of the minimum-energy region is not the only necessary condition for an accurate search-based prediction. The extent of optimization that is carried out for each conformer in the search can also significantly affect the results. In one study, near-native and non-native conformations of a protein were found to be isoenergetic because the non-native conformation had been generated on a finer search grid¹³; it was proposed that, for the purposes of predicting structure, the energies of these conformations could not be directly compared. In effect, the conformer generated on the finer grid was more highly optimized, making a direct comparison of the energies “unfair.” More broadly, if two conformations are generated with different degrees of optimization, the comparison between their energies is said here to be subject to *optimization bias*. Most generally, optimization bias may be defined as the error in an estimate of the differences between or among the local extrema of a quantity in different subpopulations of a parent population, caused by discrepancies in the extent of local sampling. In the current work, it refers more narrowly to the error in an estimate of the differences between or among local energy minima in different regions of a molecular system’s conformational space, caused by discrepancies in the extent of local sampling around those minima. An extreme example of the error introduced by such bias is illustrated in Fig. 1, in which the structure of a Val side chain predicted by the CHARMM energy surface^{14,15} is shown to be biased in favor of a non-native position after that single position has been energy-minimized.

An understanding of this type of bias is important because of an implicit assumption in energy-based structure prediction methods — namely, that the energies of sampled structures correlate with the actual minima of their local energy wells. When this assumption holds, the energies of the sampled structures can be used to compare the energies of their respective local energy wells, and, if enough conformational space is sampled, to identify the global energy minimum and, hence, the native structure.

The current work is essentially an examination of this assumption, motivated in part by the success of systematic, “parallel-generation” searches for structure prediction,^{13,16–19} in which optimization bias is largely avoided. In the main part of the paper, a formalism is developed for describing the bias that is introduced by the unequal sampling of spherically symmetric wells in an energy-based prediction. It is shown that oversampling one of the wells shifts the probability distribution of its minimum sampled energy (MSE) to the left, and, therefore, shifts the probability distribution of the prediction in that well’s favor. In systems for which the distribution of energy states is known, it is shown that the degree of bias can be quantified exactly, and the shifts in the probability distributions can be corrected exactly with the use of an energy correction term. For multiwell systems, the determination of the correction terms is a multibody problem involving multiple integrals; it is shown in an appendix that the multiple integrals can be reduced to an integral over the product of single integrals. In model systems with two identical wells differing only by a random energy shift,

the energy correction term that results in the optimal prediction accuracy can be derived analytically, for a given number of sampling points in each well.

The systems used to illustrate the theory are multi-dimensional, hyperspherical model systems, for which the probability distribution functions (PDFs) of the energy are derivable. These systems are convenient because both the energy surfaces and the PDFs have rather simple mathematical forms and yet allow the study of the effects of dimensionality and steepness on the results. In fact, these effects are shown to be related.

In the second part of the paper, a number of illustrative numerical calculations are carried out. Predictive calculations performed with the two-well system mentioned above show that optimal prediction accuracy is obtained when the sampling is split equally between the two wells. For these systems, when the sampling is not equal, the use of the derived correction term improves the prediction, but never to the accuracy that is obtained under equal sampling. In addition, a serine side chain torsional energy surface in the protein CheY is used to illustrate that even for “real” systems, the magnitude of the prediction errors and the correction terms can be calculated from the probability distributions of the sampled energy minima, which can in turn be calculated from the potential surface. A series of side chain prediction trials for CheY illustrate that the use of “adiabatic maps” (steepest-descent or conjugant-gradient minimized energy surfaces) results in poorer predictions than the use of surfaces optimized with uniform grid searches. Finally, a series of local searches is performed around the native structure of the CheY protein to illustrate the dependence that the MSE can have on the grid spacing. The results of all the calculations are interpreted in light of the previous analysis. Overall, the findings in the study demonstrate that optimization bias can introduce systematic errors into energy-based structure predictions and that avoiding or accounting for such bias may improve prediction accuracy.

2. Theory

In Sec. 2.1, it is shown for the general two-well case that unequal sampling introduces bias into predictions. Section 2.2 describes the model systems that are used in Secs. 2.3–2.7 to illustrate the derived principles.

2.1. Bias caused by oversampling one of two wells, general case

A simple argument can be used to illustrate, in a general way, the effect optimization bias can have on structure prediction.

Let the actual minimum energy of a well i be $E_{i,m}$, and let the MSE after q sampling points in this well be $E_{i,q} = \min\{U_{i,0}, U_{i,1}, U_{i,2}, \dots, U_{i,q}\}$, where $U_{i,j}$ is the energy of sampling point j in well i , and $i, j, q \in \mathbb{Z}$. Assume that two wells, wells 1 and 2, have smooth continuous surfaces and are not flat, so that the PDFs for their MSEs have nonzero widths.

Assume further that for finite, random sampling there is some overlap in the probability distributions of the MSEs for the two wells: i.e. $0 < P(E_{1,q_1} < E_{2,q_2}) < 1$, for any finite q_1, q_2 randomly chosen points.

After p_1 sampling points in each well, assume that well 1 has the lower MSE — i.e. $E_{1,p_1} < E_{2,p_1}$. Now sample well 2 at an additional point. Because of the overlap in the distributions there is a nonzero chance that the single additional point in well 2 will have an energy below E_{1,p_1} :

$$0 < P(E_{1,p_1} < E_{2,p_1+1}) < 1. \quad (1)$$

Thus,

$$P(E_{1,p_1} < E_{2,p_1+1} | E_{1,p_1} < E_{2,p_1}) < 1, \quad (2)$$

where $P(A/B)$ denotes the probability of A conditional on B .

More generally, after $p_2 - p_1$ additional sampling points in well 2 (where $p_2 > p_1$),

$$P(E_{1,p_1} < E_{2,p_2} | E_{1,p_1} < E_{2,p_1}) < 1. \quad (3)$$

From this it follows that:

$$P(E_{1,m} < E_{2,m} | E_{1,p_1} < E_{2,p_2}) > P(E_{1,m} < E_{2,m} | E_{1,p_1} < E_{2,p_2}) \cdot P(E_{1,p_1} < E_{2,p_2} | E_{1,p_1} < E_{2,p_1}). \quad (4)$$

Since $P(A/C) = P(A/B) \cdot P(B/C)$,

$$P(E_{1,m} < E_{2,m} | E_{1,p_1} < E_{2,p_2}) > P(E_{1,m} < E_{2,m} | E_{1,p_1} < E_{2,p_1}), \quad (5)$$

for $p_2 > p_1$.

Hence, given that the sampled (global) minimum resides in well 1, the likelihood that the actual global minimum resides in that well is higher if well 2 has been oversampled. This is a general result; it holds unless the wells are flat, in which case there can be no sampling-related bias, or otherwise nonoverlapping in energy, in which case the bias has no effect on the prediction. As illustrated in the rest of the analysis for hyperspherical systems, oversampling affects the predictions by shifting the probability distribution of the MSEs in favor of the oversampled well.

For structure prediction methods, this means that unequal sampling of two wells can introduce a systematic error. The argument can be extended to demonstrate that this is also true in the 3-, 4-, or multi-well case.

2.2. Energy distribution for a symmetric, multi-dimensional well

Assume an unnormalized distribution of states, $\mathcal{R}(\mathbf{R})d\mathbf{R}$, over a D-dimensional surface $\mathbf{V}^D$, where $\mathbf{R} = \{x_1, x_2, x_3, \dots, x_D\}$. The normalized distribution, $f(\mathbf{R})$, or PDF of the states as a function of spatial position is

$$f(\mathbf{R})d\mathbf{R} = \frac{\mathcal{F}(\mathbf{R})}{\int_{V_D} \mathcal{F}(\mathbf{R})d\mathbf{R}} d\mathbf{R}. \quad (6)$$

Consider now a hyperspherically symmetric distribution — i.e. one in which $\mathcal{R}(\mathbf{R}) = \mathcal{R}(r, \theta_1, \theta_2, \dots, \theta_{D-1}) = \mathcal{R}(r)$, where $r = (\sum_{i=1}^D x_i^2)^{1/2}$, $r \in (0, s)$, $s > 0$, and the θ s are the nonradial coordinates. (All variables will be assumed to be real unless otherwise specified.) Since in this case, from symmetry, $d\mathbf{R} = Cr^{D-1} dr$, where C is a constant that depends on the dimensionality, the PDF is given by

$$f(r)dr = \frac{\mathcal{F}(r)Cr^{D-1}}{\int_0^s \mathcal{F}(r)Cr^{D-1}dr} dr = \frac{\mathcal{F}(r)r^{D-1}}{\int_0^s \mathcal{F}(r)r^{D-1}dr} dr, \quad (7)$$

where C cancels. The function $\mathcal{R}(r)dr$ will be called the *parent spatial distribution*; it can be thought of as the (unnormalized) distribution of states along any single radius in the system. By contrast, the PDF $f(r)dr$ gives the total probability of finding a state between r and $r + dr$ anywhere in the system.

Assume an energy function of the form $E = g(r)$, which describes a hyper-spherically symmetric well. If $E = g(r)$ has an inverse, $r = \bar{g}(E)$, then a corresponding *parent function* can be defined as $\mathcal{F}_E(E) = \mathcal{R}(\bar{g}(E))$. If $g(r)$ is monotonically increasing, the PDF for the energy in the well, $f(E)dE$, is then given by

$$f(E)dE = \frac{\mathcal{F}_E(E)\bar{g}(E)^{D-1}\bar{g}'(E)}{\int_{E_m}^{E_M} \mathcal{F}_E(E)\bar{g}(E)^{D-1}\bar{g}'(E)dE} dE, \quad (8)$$

where $g'(r) = dE/dr$, since $dr/dE = \bar{g}'(E)$. The limits $E_M = g(s)$ and $E_m = g(0)$ are the maximal and minimal energies in the well.^a

The energy function $E = g(r) = Ar^k$, where A , $k > 0$, A is a constant and k is the *order* of the well, will be used throughout the analysis section because it is spherically symmetric, invertible and monotonically increasing with r . Furthermore, energy wells of this type, while simple in structure, have some interesting properties, some of which will be described below.

In the case of Ar^k wells, $\bar{g}(E) = (E/A)^{1/k}$ and $\bar{g}'(E) = (E/A)^{(1-k)/k}/Ak$, which give

$$f(E)dE = \frac{\mathcal{F}_E(E)(E/A)^{(D-1)/k}(1/Ak)(E/A)^{(1-k)/k}}{\int_0^{E_M} \mathcal{F}_E(E)(E/A)^{(D-1)/k}(1/Ak)(E/A)^{(1-k)/k}dE} dE = \frac{\mathcal{F}_E(E)E^{D/k-1}}{\int_0^{E_M} \mathcal{F}_E(E)E^{D/k-1}dE} dE, \quad (9)$$

since $E_m = 0$.

^aNote that the energy distribution along a radius of the system is not identically $\mathcal{F}_E(E)$, but $\mathcal{F}_E(E)\bar{g}'(E)dE$.

The variables D and k on the right-hand side (rhs) of Eq. (9) appear only in the ratio D/k . Therefore, let $\mu = D/k$:

$$f(E)dE = \frac{\mathcal{F}_E(E)E^{\mu-1}}{\int_0^{E_M} \mathcal{F}_E(E)E^{\mu-1}dE}dE. \quad (10)$$

This is the general expression for the PDF of the energy in an Ar^k well, given any parent function of states, $\mathcal{F}_E(E) = \mathcal{F}(\tilde{g}(E))$. Note that the PDF is independent of the prefactor in the energy expression (A), as well as the constant associated with the volume element (C). Further, the fact that D/k can be replaced by μ means that the dimensionality and steepness of these wells are related. For example, the energy distribution for a six-dimensional quadratic (Ar^2) well is identical to that for a nine-dimensional cubic (Ar^3) well. For a fixed-order (k), higher-dimensional Ar^k wells have more high-energy states; by contrast, for a fixed number of dimensions (D) and a fixed maximal energy, higher-order Ar^k wells are flatter (a larger portion of the wells are close to $E = 0$), and therefore have fewer high-energy states, as a fraction of the total (see Fig. 2). These effects exactly cancel. This is a property of the wells, themselves, and is independent of the sampling, and therefore of $\mathcal{F}_E(E)$. Because D , k and, therefore, μ are (positive) reals, fractional orders and dimensionalities can also be represented in the model.

For the next several subsections, the random sampling of states will be carried out using a uniform parent function (corresponding to a uniform spatial distribution throughout the system) — i.e. $\mathcal{F}_E(E)$ is a constant — and a finite boundary for the well. The resulting PDFs will be used as examples in the derivations for the oversampling and corrections. Later in the work, a Boltzmann parent function and an infinite boundary will be used.

For a uniform parent function, the PDF of an Ar^k well is

$$f(E)dE = \frac{E^{\mu-1}}{\int_0^{E_M} E^{\mu-1}dE}dE = \frac{\mu(E/E_M)^{\mu-1}}{E_M}dE,$$

where $E_M = As^k$ is the maximal energy at the boundary s , which is chosen here to be finite. Define $\tilde{E} = E/E_M$, the energy as a fraction of the maximum. Since

$$f(\tilde{E})d\tilde{E} = \mu\tilde{E}^{\mu-1}d\tilde{E} = \frac{d(\tilde{E}^\mu)}{d\tilde{E}}. \quad (11)$$

At the extremes of μ , the PDF $f(\tilde{E})d\tilde{E}$ shifts toward the extreme energy values—toward $\tilde{E} = 1$ for high μ and $\tilde{E} = 0$ for low μ .

The energy surface can be described in a corresponding manner as $\tilde{E} = (r/s)^k = r^{\tilde{k}}$, where $\tilde{E}$ and r are both unitless and vary within $[0, 1]$. It is well-behaved, except at $\tilde{E} = 0$, for $\mu < 1$, where it results in a singularity.

2.3. Distribution of the minimum energy for a multiply sampled well

Under each subsection in the remainder of the theory section, general expressions will be derived first and followed with illustrative applications to Ar^k wells.

Suppose that, for a single trial, the MSE for an energy well having a PDF of $f(E)dE$ is obtained by sampling the well p times. This trial is then repeated many times until the PDF for the MSE converges. Such a PDF will be designated $f_{E;p}(E)dE$, where the first subscript, E , describes the distribution, and the second subscript, p gives the number of sampling points per trial. Hence, $f_{E;3}(E)dE$ denotes the PDF for the MSE resulting from sampling a well at three points. In this notation, the original PDF for the MSE, $f(E)dE$, is equivalent to $f_{E;1}(E)dE$.

For this analysis, an expression for $f_{E;p}(E)dE$, for any arbitrary $p \in \mathbb{Z}$, needs to be derived from the single-point distribution $f(E)dE$. In general, the probability of energy E being the MSE will be the probability that a state with this energy is sampled at the first point, times the product of the probabilities that any of the other $p - 1$ points samples a higher energy, times p . The factor of p arises because any of the points could be chosen first. Hence, the general expression is

$$f_{E;p}(E)dE = pf(E) \left(\int_E^{E_M} f(E')dE' \right)^{p-1} dE = -\frac{\partial}{\partial E} \left(\int_E^{E_M} f(E')dE' \right)^p dE, \quad (12)$$

where the primes indicate a dummy integration variable.

For the particular case of finite Ar^k wells and a uniform parent function, from Eq. (11),

$$f_{\tilde{E};p}(\tilde{E})d\tilde{E} = p\mu\tilde{E}^{\mu-1} \left(\int_{\tilde{E}}^1 \mu\tilde{E}'^{\mu-1} d\tilde{E}' \right)^{p-1} d\tilde{E} = p\mu\tilde{E}^{\mu-1} (1 - \tilde{E}^\mu)^{p-1} d\tilde{E} = \frac{-\partial(1 - \tilde{E}^\mu)^p}{\partial \tilde{E}} d\tilde{E}. \quad (13)$$

This is the general form of the PDF for the minimum sampled energy, as a fraction of the well's maximum, that arises from randomly sampling a finite Ar^k well from a uniform parent function at p points. Figure 3(a) shows $f_{\tilde{E};p}(\tilde{E})$ for $\mu = 1$ and 5, and for three different values of p . For fixed p , a larger μ (higher dimensionality, smaller order) shifts the PDFs to the right; higher-dimensional or steeper spaces have a larger fraction of their surfaces at higher energies. On the other hand, for fixed μ , an increasing number of sampling points, p , shifts the PDFs to the left, as expected. At high μ values and small p , the distributions become narrower and shifted closer to the upper bound.

In general, the average MSE is given by $\langle E \rangle = \int_{E_m}^{E_M} E f_{E;p}(E)dE$.

For finite Ar^k wells and a uniform parent function,

$$\langle \tilde{E} \rangle = \int_0^1 \tilde{E} \frac{-\partial(1 - \tilde{E}^\mu)^p}{\partial \tilde{E}} d\tilde{E} = \frac{p\Gamma(1/\mu)\Gamma(p)}{\mu\Gamma(1/\mu + p + 1)} = \frac{pB(1/\mu, p)}{1 + \mu p}, \quad (14)$$

where Γ is the Gamma function, and B is the Beta function.^b

Two three-dimensional plots of $\langle \tilde{E} \rangle$ as a function of p and μ are shown in Fig. 4. At large values of μ (Panel B, e.g. $\mu = 100$), $\langle \tilde{E} \rangle$ approaches 1 (the maximum) for the entire range of sampling points shown, $p = [1, 100]$. Because of the maximal, constant energy at the boundary, in very steep or high-dimensional wells large fractions of the surfaces have energies close to the maximum, so that a large number of sampling points is required to substantially lower the average MSE. By contrast, at small values of μ (Panel A, $\mu < 1$), $\langle \tilde{E} \rangle$ approaches zero for all p . Hence the absolute effect on the MSE of increased sampling in these systems is smallest at the extremes of μ . As shown in Fig. 1S in Supplementary Material, for a given p , the value of μ for which the effect of increasing p by 1 is maximal varies as $\approx \ln(p)$, as measured by the absolute reduction in the average MSE. For example, raising the number of sampling points from 3 to 4 has the optimal effect in a well having $\mu \approx 1.45$; from 9 to 10, $\mu \approx 2.44$. By contrast, the greatest relative decrease in $\langle \tilde{E} \rangle$ for incrementing p by 1 always occurs as $\mu \rightarrow 0$ — i.e. in flatter, lower-dimensional wells — since the ratio $\langle \tilde{E}_{\mu,p} \rangle / \langle \tilde{E}_{\mu,p+1} \rangle = 1 + 1/(\mu(1 + p_1))$. At any (fixed) value of μ , $\langle \tilde{E} \rangle / p$ is most negative at $p = 1$ (for all positive integral p) and increases with p ; this means that the absolute effect of increasing the number of sampling points on the average MSE is always greatest at smaller numbers of sampling points.

The variance of the distribution,

$$\langle \tilde{E}^2 \rangle - \langle \tilde{E} \rangle^2 = \frac{p\Gamma(2/\mu+1)\Gamma(p)}{\Gamma(2/\mu+p+1)} - \left(\frac{p\Gamma(1/\mu)\Gamma(p)}{\mu\Gamma(1/\mu+p+1)} \right)^2. \quad (15)$$

In these systems, for fixed $\mu > 1$, the variance decreases with $p \rightarrow 1$.^c For fixed p , the smallest variances always occur at very large (because of the finite boundary) or very small μ values.

2.4. Distribution of the difference between the minimum sampled energy of two identical wells

Consider two identical wells sampled at p_1 and p_2 points and having PDFs for the MSE of $f_{E;p_1}(E)$ and $f_{E;p_2}(E)$, respectively. The PDF, $f_{E;p_1,p_2}(E)$, for the difference between the

^bFor integral D and k , this is equivalent to

$$\langle \tilde{E} \rangle = \frac{k!^{(D)}(Dp)!^{(D)}}{(k+Dp)!^{(D)}},$$

where $k!^{(D)}$ is the D th multifactorial of k .

^cWhen $p \gg 1/\mu$, for $\mu \in \mathbb{R}^+$, the mean can be approximated for computational purposes as

$$\langle \tilde{E} \rangle \approx \frac{(p+1)^{-\frac{\mu+1}{\mu}} (2\mu^2(p+1)+\mu-1)\Gamma(1/\mu)}{2\mu^3},$$

from the approximation $\Gamma(x+x) \approx 2x^{-x+1} \Gamma(x)/(2x-x^2+x)$, for $x \gg x$.

^cFor fixed μ greater than ≈ 2.1 , the variance increases for $p \rightarrow 1$ until a maximal variance is reached at $p \approx (5+3\mu)/10$, after which it decreases. For example, when $\mu = 10$, the integral p value giving the maximal variance of 0.00820 occurs at $p = 4$.

MSE of the two wells, $E = E_{2,p_2} - E_{1,p_1}$ is given by the cross-correlation of the individual PDFs. Hence, in general,

$$f_{\Delta E;p_1,p_2}(\Delta E) = \int_{\ell}^{\mathcal{L}} f_{E;p_1}(E) \cdot f_{E;p_2}(E + \Delta E) dE. \quad (16)$$

In the region $0 < E < E_M - E_m$, where the function will be called $f_{\Delta E}^+$, the limits are $(\ell, \mathcal{L}) = (E_m, E_M - E)$. In the region $E_m - E_M < E < 0$, where it will be called $f_{\Delta E}^-$, the limits are $(\ell, \mathcal{L}) = (E_m - E, E_M)$.

Since $E = E_2 - E_1$, the total cumulative probability of well 2 having the lower MSE, $\text{Pr}(2)$, is the total probability for $E < 0$. Namely,

$$\text{Pr}(2) = \int_{E_m - E_M}^0 f_{\Delta E;p_1,p_2}^-(\Delta E) d(\Delta E). \quad (17)$$

In the particular case of Ar^k wells, define $\tilde{E}$ as the energy difference between the two wells as a fraction of the maximum energy. For a uniform parent function, the corresponding PDF of $\tilde{E}$ in the region $-1 < \tilde{E} < 0$ is

$$f_{\Delta \tilde{E};p_1,p_2}^-(\Delta \tilde{E}) = \int_{-\Delta \tilde{E}}^1 \mu^2 p_1 p_2 (\tilde{E}(\tilde{E} + \Delta \tilde{E}))^{\mu-1} \times (1 - \tilde{E}^\mu)^{p_1-1} (1 - (\tilde{E} + \Delta \tilde{E})^\mu)^{p_2-1} d\tilde{E}. \quad (18)$$

In the region $0 < \tilde{E} < 1$, the PDF $f_{\Delta \tilde{E};p_1,p_2}^+(\tilde{E})$ is defined as the same integral, evaluated at the limits $(\ell, \mathcal{L}) = (0, 1 - \tilde{E})$.

For example, for $\mu = 1$ (e.g. a two-dimensional quadratic well) and $p_1 = 1$,

$$f_{\Delta \tilde{E};1,p_2}(\Delta \tilde{E}) = \begin{cases} 1 - (-\Delta \tilde{E})^{p_2}, & \text{for } \Delta \tilde{E} < 0 \\ (1 - \Delta \tilde{E})^{p_2}, & \text{for } \Delta \tilde{E} > 0 \end{cases} \begin{matrix} [f^-] \\ [f^+] \end{matrix}. \quad (19)$$

For $p_2 = 1 (= p_1)$, the plot of $f_{\tilde{E};1,1}$ is symmetric, as shown in Fig. 5(a). The total cumulative probability of well 2 having the lower sampled minimum, $\text{Pr}(2)$, is $1/2$. However, for $p_2 = 2$, the distribution is shifted, as shown in Fig. 5(b), and $\text{Pr}(2) = 2/3$.

More generally, for any finite Ar^k well and a uniform parent function

$$\text{Pr}(2) = \int_{-1}^0 \int_{-\Delta \tilde{E}}^1 \mu^2 p_1 p_2 (\tilde{E}(\tilde{E} + \Delta \tilde{E}))^{\mu-1} (1 - \tilde{E}^\mu)^{p_1-1} \times (1 - (\tilde{E} + \Delta \tilde{E})^\mu)^{p_2-1} d\tilde{E} d(\Delta \tilde{E}).$$

Reversing the order of integration and changing the limits to correspond,

$$\begin{aligned} & \mu^2 p_1 p_2 \int_0^1 (1 - \tilde{E}^\mu)^{p_1-1} \tilde{E}^{\mu-1} \\ & \times \left(\int_{-\tilde{E}}^0 (\tilde{E} + \Delta \tilde{E})^{\mu-1} (1 - (\tilde{E} + \Delta \tilde{E})^\mu)^{p_2-1} d(\Delta \tilde{E}) \right) d\tilde{E} \\ & = \mu p_1 \int_0^1 (1 - \tilde{E}^\mu)^{p_1-1} \tilde{E}^{\mu-1} (1 - (1 - \tilde{E}^\mu)^{p_2}) d\tilde{E} = \frac{p_2}{p_1 + p_2}. \end{aligned}$$

Since the wells here have exactly the same form (and dimension), the probability of either well having the lower MSE depends only upon the fraction of the total sampling that is done in that well, and is independent of the form, itself (e.g. μ).

The general case for two Ar^k wells having different orders and dimensions (μ values), prefactors and minimum and maximum energies is described in Supplementary Material: Sec. “Integration limits for cross-correlation integrals”.

2.5. Correction function

A type I CSD (*correction for sampling discrepancy*), ξ , can be defined as a correction in E , under unequal sampling of two wells, that results in total probabilities of choosing the wells that are equal to those that would occur under equal sampling. Formally, for two identical wells sampled at p_1 and p_2 points, $p_2 > p_1$,^d

$$\int_{E_m - E_M}^{-\xi} f_{\Delta E; p_1, p_2}^{-}(\Delta E) d\Delta E = \int_{E_m - E_M}^{-\xi} \left(\int_{E_m - \Delta E}^{E_M} f_{E; p_1}(E) \cdot f_{E; p_2}(E + \Delta E) dE \right) d\Delta E = 1/2. \quad (20)$$

Since $E_m - E_M < E < 0$ here, this states that the probability of well 2 having the lower MSE, given any $p_1 < p_2$ and an energy correction of ξ , has to be 1/2. The sense of this definition is such that $\xi > 0$, and it is added to the energy of the oversampled well. Note the correction ξ is not in general equal to the difference between the expectation values (i.e. the means over many sampling trials) of the MSE for each well (e.g. Eq. (14)).

Reversing the order of integration with the appropriate change in limits (see Supplementary Material: Sec. “Change in limits in cross-correlation integrals”),

$$\int_{\xi}^{E_M - E_m} \left(\int_{-E}^{-\xi} f_{E; p_1}(E) \cdot f_{E; p_2}(E + \Delta E) d\Delta E \right) dE = \int_{\xi}^{E_M - E_m} f_{E; p_1}(E) \left(\int_{-E}^{-\xi} f_{E; p_2}(E + \Delta E) d\Delta E \right) dE = 1/2. \quad (21)$$

This form is often convenient because it reduces the inner integrand to a single PDF. In the particular case of (identical) finite Ar^k wells and sampling from a uniform parent function,

^dAn equivalent definition shifts the distributions to the right by ξ by adding $-\xi$ to the argument of the second PDF, with limits of $(E_m - E_M, 0)$ for the outer integral (see also Supplementary Material: Sec. “Energy shift in one Ar^k well”).

$$\begin{aligned}
 & \int_{\tilde{\xi}}^1 \frac{d(1 - \tilde{E}^\mu)^{p_1}}{d\tilde{E}} \times \\
 & \left(\int_{-\tilde{E}}^{-\tilde{\xi}} \frac{d(1 - (\tilde{E} + \Delta\tilde{E})^\mu)^{p_2}}{d\Delta\tilde{E}} d\Delta\tilde{E} \right) d\tilde{E} \quad (22) \\
 & = \int_{\tilde{\xi}}^1 \frac{d(1 - \tilde{E}^\mu)^{p_1}}{d\tilde{E}} ((1 - (\tilde{E} - \tilde{\xi})^\mu)^{p_2} - 1) d\tilde{E} = 1/2.
 \end{aligned}$$

Note that since the energies are normalized here relative to the maximum energy of the well, $\tilde{\xi}$ is similarly normalized. For finite and arbitrary p and μ , the definite integral in Eq. (22) is not generally soluble in closed form; in cases where it is (e.g. $\mu = 1$ for particular p_1) it results in transcendental equations in $\tilde{\xi}$, which are not soluble in closed form. This means $\tilde{\xi}$ can be determined analytically only for specific p_1 and p_2 ; otherwise, it must be approximated.

For $\mu = 1$ and $p_1 = 1$, Eq. (22) gives

$$\frac{\tilde{\xi}^{p_2+1} + p_2}{p_2+1} - \tilde{\xi} = 1/2, \quad (23)$$

for which the following is a good approximate solution^e:

$$\tilde{\xi} \approx \frac{p_2 - 1}{2(p_2 + 1)}, \quad (25)$$

as shown in Fig. 6(a).

The approximate and exact solutions for $\mu = 1$ and $p_1 = 2$ are plotted in Fig. 6(b). For fixed p_1 , the CSD increases with increasing p_2 ; and for fixed $p_2 > p_1$, the correction for $p_1 = 1$ (Fig. 6(a)) is larger than for $p_1 = 2$ (Fig. 6(b)), and continues to diminish with larger p_1 (not shown). Hence, the largest corrections are required when one of the wells is sampled only once.

Determining the CSDs exactly in a multi-well case (i.e. for three wells or more simultaneously) is a many-body problem. For n wells, the determination of the n CSDs involves solving n n -tuple integral constraint equations. The three-well case is considered in the appendix (App. A.1). By integrating over the E_i variables first and E last (as in Eq. (22), for example), the cross-correlation n -tuple integrals involved in these cases can always be reduced to a single integral over a product of $n - 1$ single integrals. See App. A.2.

^eThe general approximate solution for $\mu = 1$ is obtained by solving

$$\sum_{\kappa=0}^{p_1} \binom{p_1}{\kappa} \frac{(-\tilde{\xi})^\kappa p_2}{p_1 + p_2 - \kappa} \approx \frac{1}{2} \quad (24)$$

for $\tilde{\xi}$ given p_1 .

If the form of the single-point energy distribution is known for two identical wells, a limiting (i.e. maximal) correction can be determined for fixed p_1 , in the limit of $p_2 \rightarrow \infty$.

As explained further in Supplementary Material: Sec. “Limiting sampling correction” it may be defined as

$$\int_{E_m - E_M}^{-\xi} f_{\Delta E; p_1, \infty}^{-}(\Delta E) d\Delta E = \int_0^{\xi} f_{E; p_1}(-\Delta E) d\Delta E = 1/2.$$

For finite Ar^k wells and a uniform parent function, the result is:

$$\tilde{\xi} = (1 - 2^{-1/p_1})^{1/\mu}.$$

More generally, for a system of two identical wells that differ in their energies by a constant, U , such that $E_2(r) = E_1(r) + U$, the type I CSD, ξ , can be defined by:

$$\begin{aligned} \int_{\Delta U + E_m - E_M}^{-\xi} \int_{\Delta U + E_m - \Delta E}^{E_M} f_{E; p_1}(E) \cdot f_{E; p_2}(E + \Delta E - \Delta U) dE d\Delta E = \\ \int_{\Delta U + E_m - E_M}^0 \int_{\Delta U + E_m - \Delta E}^{E_M} f_{E; p_1}(E) \cdot f_{E; p_1}(E + \Delta E - \Delta U) dE d\Delta E, \end{aligned} \quad (26)$$

for $U > 0, p_2 > p_1$.

The left-hand side (lhs) gives the PDF for choosing well 2, given $p_1 < p_2$ and a CSD $\tilde{\xi}$, while the rhs gives the PDF for choosing well 2 under equal sampling. This relation is used to calculate the CSDs for uneven sampling of the torsional angle energy surface of Ser 103 (see Sec. 3.2). A more complete treatment of the shifting of the energy in one well is given in Supplementary Material: Sec. “Energy shift in one Ar^k well”.

2.6. Canonical ensemble

In canonical sampling, the parent function is $\mathcal{P}E(E) = e^{-E/T}$, where T is the temperature and the energy is in units of k_B . Hence, for a hyperspherically symmetric system, and an energy function that is invertible and monotonically increasing, from Eq. (8) the PDF for the energy is:

$$f(E) dE = \frac{e^{-E/T} \bar{g}(E)^{D-1} \bar{g}'(E)}{\int_{E_m}^{E_M} e^{-E/T} \bar{g}(E)^{D-1} \bar{g}'(E) dE} dE. \quad (27)$$

These PDFs correspond to canonical ensembles.

For Ar^k wells, this gives

$$f(E) dE = \frac{e^{-E/T} E^{\mu-1}}{\int_0^{E_M} e^{-E/T} E^{\mu-1} dE} dE = \frac{e^{-E/T} E^{\mu-1}}{T^\mu (\Gamma(\mu) - \Gamma(\mu, E_M/T))} dE = -\frac{\partial}{\partial E} \left(\frac{\Gamma(\mu, E/T)}{T^\mu (\Gamma(\mu) - \Gamma(\mu, E_M/T))} \right), \quad (28)$$

where $\Gamma(t_1, t_2); t \in \mathbb{R}$, is the incomplete (upper) Gamma function. An infinite boundary will be used henceforth, since the expressions then simplify to

$$f(E)dE = \frac{e^{-E/T} E^{\mu-1}}{T^\mu \Gamma(\mu)} dE = -\frac{\partial}{\partial E} \left(\frac{\Gamma(\mu, E/T)}{\Gamma(\mu)} \right). \quad (29)$$

From Eq. (12), the PDF for the minimum sampled energy upon sampling an infinite Ar^k well from a canonical parent function p times is

$$f_{E;p}(E)dE = -\frac{\partial}{\partial E} \left(\int_E^\infty f(E')dE' \right)^p dE = -\frac{\partial}{\partial E} \left(\frac{\Gamma(\mu, E/T)}{\Gamma(\mu)} \right)^p dE = \frac{p e^{-E/T} E^{\mu-1} \Gamma(\mu, E/T)^{p-1}}{T^\mu \Gamma(\mu)^p} dE. \quad (30)$$

Since for integral a , $\Gamma(a, b)/\Gamma(a) = e^{-b} \sum_{n=0}^{a-1} b^n/n!$,

$$f_{E;p}(E)dE = \frac{p e^{-pE/T} E^{\mu-1} \left(\sum_{n=0}^{\mu-1} (E/T)^n/n! \right)^{p-1}}{T^\mu \Gamma(\mu)} dE, \quad (31)$$

for integral $\mu \geq 1$. This expression is convenient in expanded form for cases of particular μ .

In the case of $\mu = 1$

$$f_{E;p}(E)dE = \frac{p e^{-pE/T}}{T} dE. \quad (32)$$

The average MSE can be shown to be $\langle E \rangle = T/p$ and the variance $\text{Var}(E) = T^2/p^2$. Hence, as expected, the mean and variance for the MSE distribution both decrease with an increasing number of sampling points and increase with temperature. This can be shown numerically to hold for other values of μ , including fractional ones (from Eq. (30)).

For arbitrary μ and $p = 1$, the mean and variance are μT and μT^2 respectively; hence, they both increase with μ . This can be shown to hold for other values of p , and is in part related to the infinite boundary. For a finite boundary, the variance at first increases with μ , then decreases (results not shown), similar to the case for sampling from a uniform parent function.

Figures 3(b) and 3(c), show plots of the PDF for the MSE, $f_{\hat{E},p}(\hat{E})d\hat{E}$, where $\hat{E} = E/T$, for Ar^k wells sampled from a canonical parent function, with either $\mu = 1$ or $\mu = 5$ and $p = 1, 3$ or 10 . Figure 3(b) shows the result for finite wells; note that the PDFs for E/T here are very similar to those for $\tilde{E}$ in Fig. 3(a) (finite Ar^k wells, uniform parent function), for the same μ and p parameters. Figure 3(c) shows the distribution for wells with an infinite boundary; they have broader distributions than in the other two cases. This suggests that the boundary of the wells has a greater effect on the distributions of the MSEs than whether the parent distribution is canonical or uniform. In all cases, the distributions are shifted to the right with increasing μ , for fixed p , and to the left for increasing p , for fixed μ , as expected. Hence, for very high- μ systems, the distribution of the MSE obtained with a finite number of sampling

points, from both canonical and uniform parent functions, can be substantially above the actual energy minimum.

For two such (identical) Ar^k wells, having $\mu = 1$ and being sampled at p_1 and p_2 points, the following expression, derived from Eq. (21), can be used to determine the type 1 CSD, ξ :

$$\int_{\xi}^{\infty} \frac{p_1 e^{-p_1 E/T}}{T} \times \left(\int_{-E}^{-\xi} \frac{p_2 e^{-p_2 (E+\Delta E)/T}}{T} d\Delta E \right) dE = \frac{p_2}{p_1 + p_2} e^{-\xi p_1/T} = 1/2. \quad (33)$$

The CSD is found to be

$$\xi = \frac{T}{p_1} \ln \left(\frac{2p_2}{p_1 + p_2} \right). \quad (34)$$

The limiting CSD, for $p_2 \rightarrow \infty$ and $\mu = 1$, is therefore

$$\xi = T \ln(2)/p_1. \quad (35)$$

For $\mu > 1$, the limiting CSD can be shown numerically to increase with μ . In particular, for $p_1 = 1$, $\xi \approx T(\ln(2) + \mu - 1)$ and approaches $\xi = \mu T$ for μ very large.^f Hence, the limiting energy penalties under canonical sampling and an infinite boundary become quite large for steep or high-dimensional surfaces.

2.7. Finding the optimal CSD for prediction

In Secs. 2.5 and 2.6, corrections called type I CSDs were used to equalize the total probabilities of selecting two wells under unequal sampling. The current subsection explores whether CSDs can be used to improve prediction. These will be called type II CSDs.

In general, for two wells having minima $E_{1,m}$ and $E_{2,m}$, and with MSEs E_1 and E_2 , respectively, there are four possible outcomes for prediction:

$$\begin{aligned} A: E_1 > E_2 \quad \text{and} \quad E_{1,m} > E_{2,m} \\ B: E_1 > E_2 \quad \text{and} \quad E_{1,m} < E_{2,m} \\ C: E_1 < E_2 \quad \text{and} \quad E_{1,m} > E_{2,m} \\ D: E_1 < E_2 \quad \text{and} \quad E_{1,m} < E_{2,m}. \end{aligned}$$

These cases are depicted in Fig. 7. Cases *A* and *D* describe correct predictions, cases *B* and *C* incorrect predictions. Suppose that in each case well 2 is over-sampled — i.e. $p_2 > p_1$, as depicted in the figure — and a corresponding CSD, ξ , is thus added to E_2 . Let $E = E_2 - E_1$. In cases *A* and *B*, use of the CSD could change the outcome of the prediction, since for $\xi > -E$, it converts the prediction from well 2 to well 1. In case *A*, this worsens prediction

^fThe general solution for any p_1 involves $(\Gamma[\mu, \xi T]/\Gamma[u])^{p_1} = 1/2$. For $p_1 = 1$, since $\lim_{\mu \rightarrow \infty} \Gamma[\mu, \mu]/\Gamma[u] = 1/2$ and $\Gamma[\mu, \xi T]/\Gamma[u]$ decreases monotonically with increasing μT for all fixed μ , $\xi/T \rightarrow \mu$ as $\mu \rightarrow \infty$.

accuracy, since well 2 is the lower-energy well, while in case *B*, it improves prediction accuracy. For $\xi < -E$, there is no effect in either case *A* or *B*. In cases *C* and *D*, use of the CSD never changes the prediction, since the sampled energy in the oversampled well is already higher ($E_2 > E_1$). Hence this type of correction may affect prediction accuracy (for better or worse) only in cases where the MSE of the oversampled well is the lower.

It is of interest whether a single CSD can be applied to improve the overall accuracy of a prediction; and, if so, whether the optimal CSD can be determined *a priori* from the parameters of the problem.

Suppose there are two wells with energies $E_1(r)$ and $E_2(r) = E_1(r) + U$, where $E(r)$ varies in $[0, E_M]$, and the number of sampling points is $p_2 > p_1$, as in previous sections. Suppose further that U is unknown, but varies in the interval $[U_m, U_M]$ with a uniform PDF. Introduce a CSD, ξ , define a corrected energy, $E_2'(r) = E_2(r) + \xi = E_1(r) + U + \xi$, and find the value of the CSD that maximizes the probability of correct prediction.

For wells with a finite boundary, the solution involves five component integrals, depicted in Fig. 8, which arise from all possible positions of $E = 0$ relative to the four quantities: $U + \xi - E_M$ (the minimum value for E), U , $U + \xi$, and $U + \xi + E_M$ (the maximum value for E). The integrals can be expressed as follows:

$$I_1 = \begin{cases} \int_{E_M - \xi}^{\Delta U_M} d\Delta U = \Delta U_M - E_M + \xi, & \text{for } \Delta U_M > E_M - \xi \text{ and } \xi < E_M, \\ 0, & \text{otherwise,} \end{cases} \quad (36a)$$

$$I_2 = \int_0^{\mathcal{L}} \left(1 - \int_{\Delta U + \xi - E_M}^0 f_{\Delta E}^-(\Delta E - \Delta U - \xi) d\Delta E \right) d\Delta U, \quad (36b)$$

$$\text{where } \mathcal{L} = \begin{cases} E_M - \xi, & \text{if } \Delta U_M > E_M - \xi, \\ \Delta U_M, & \text{otherwise,} \end{cases}$$

$$I_3 = \int_{\ell}^{-\xi} \left(1 - \int_0^{\Delta U + \xi + E_M} f_{\Delta E}^+(\Delta E - \Delta U - \xi) d\Delta E \right) d\Delta U, \quad (36c)$$

$$\text{where } \ell = \begin{cases} -\xi - E_M, & \text{if } \Delta U_m < -\xi - E_M, \\ \Delta U_m, & \text{otherwise,} \end{cases}$$

$$I_4 = \begin{cases} \int_{\Delta U_m}^{-\xi - E_M} d\Delta U = -\xi - E_M - \Delta U_m, & \text{for } \Delta U_m < -\xi - E_M, \\ 0, & \text{otherwise,} \end{cases} \quad (36d)$$

$$I_5 = \int_{-\xi}^0 \int_{\Delta U + \xi - E_M}^0 f_{\Delta E}^-(\Delta E - \Delta U - \xi) d\Delta E d\Delta U. \quad (36e)$$

The fraction correct (or probability of correct prediction), P_C , is given by

$$P_C = (I_1 + I_2 + I_3 + I_4 + I_5) / (\Delta U_M - \Delta U_m). \quad (37)$$

Integrals I_1 and I_2 give the (unnormalized) average number of correctly predicted cases for $U > 0$; integrals I_3 , I_4 and I_5 for $U < 0$.

Integral I_1 is independent of the CSD, since $f_E > 0$ over its entire interval, even before an upward shift by ξ (see Fig. 7(d)). For Integral I_2 , $f_{\Delta E}^- < 0$ over part of its interval, so that a CSD can increase the value of the integral (Fig. 7(b)). A CSD can decrease the values of I_3 and I_5 (Fig. 7(a)). Integral I_4 is independent of the CSD since $f_E < 0$ throughout its interval. Overall, the optimal CSD achieves the best balance between increasing I_2 and reducing I_3 and I_5 .

Note the integrals are all independent of U_m and U_M in the regions $U_m < -E_M - \xi$ and $U_M > E_M - \xi$. In these regions, U_m and U_M appear only in the normalization, and hence the optimal CSD depends only on the shape of the wells and the number of sampling points.

The specific forms of integrals I_2 , I_3 and I_5 for the case of Ar^k wells and $\mu = 1$ are given in Supplementary Material: Sec. “Integrals for Ar^k well prediction accuracy model”.

For a uniform parent distribution and finite Ar^k wells of the form $\tilde{E} = r^k$, with $\tilde{U}_M > 1 - \tilde{\xi}$, $\tilde{U}_m < -1 - \tilde{\xi}$, and $\mu = 1$,

$$P_c = 1 + \frac{1}{\Delta \tilde{U}_M - \Delta \tilde{U}_m} \times \left(\frac{2p_2 \tilde{\xi}}{p_1 + 1} \sum_{k=1}^{p_1} \binom{p_1+1}{k} \frac{(-\tilde{\xi})^{k-1}}{p_1+p_2+1-k} + \frac{2(-1)^{p_1+1} \tilde{\xi}^{p_1+p_2+1} p_1!}{\prod_{k=1}^{p_1+1} (k+p_2)} + \frac{2(-1)^{p_1} \tilde{\xi}^{p_1+1} - 1}{p_1+1} + \frac{2}{p_1+p_2+1} - \frac{1}{p_2+1} - \tilde{\xi} \right). \quad (38)$$

For $p_1 = 1$

$$P_c = \frac{[1 - p_2^2(2\tilde{\xi}^2 - 2\tilde{\xi} + 1) + p_2(6\tilde{\xi}^2 - 2\tilde{\xi} + 1) - 4\tilde{\xi}^{2+p_2} + 4\tilde{\xi}^2 + 4\tilde{\xi} + 2]}{2(\Delta \tilde{U}_M - \Delta \tilde{U}_m)(p_2+1)(p_2+2)}. \quad (39)$$

To find the optimal value of $\tilde{\xi}$, let $\frac{\partial P_c}{\partial \tilde{\xi}} = 0$. This gives

$$(p_2+2)(2\tilde{\xi}^{p_2+1} - 2\tilde{\xi}(p_2+1) + p_2 - 1) = 0. \quad (40)$$

This equation yields the same expression for $\tilde{\xi}$ as Eq. (23). Thus, for the case where $\mu = 1$ and $p = 1$, the CSD giving the optimal prediction accuracy is the same as that which equalizes the likelihood of choosing either well — i.e. the types I and II CSDs are the same. However, this is not the general rule for arbitrary μ .

For wells with infinite boundaries, the solution involves three integrals, which can be written as:

$$I_1 = \int_0^{\Delta U_M} \int_{-\infty}^0 f_{\Delta E}^-(\Delta E - \Delta U - \xi) d\Delta E d\Delta U, \quad (41a)$$

$$I_2 = \int_{\Delta U_M}^{-\xi} \left(1 - \int_0^{\infty} f_{\Delta E}^+(\Delta E - \Delta U - \xi) d\Delta E \right) d\Delta U, \quad (41b)$$

and

$$I_3 = \int_{-\xi}^0 \int_{-\infty}^0 f_{\Delta E}^-(\Delta E - \Delta U - \xi) d\Delta E d\Delta U. \quad (41c)$$

The total fraction of cases correctly predicted is

$$P_c = (I_1 + I_2 + I_3) / (\Delta U_M - \Delta U_m).$$

For infinite $A r^k$ wells sampled from a canonical parent distribution $\mu = 1$,

$$P_c = 1 + \frac{1}{\Delta U_M - \Delta U_m} \times \left[-\xi + \frac{T}{p_1 p_2 (p_1 + p_2)} (p_1^2 (e^{p_2(\Delta U_m + \xi)/T} - 1) + p_2^2 (e^{-p_1(\Delta U_M + \xi)/T} - 2e^{-p_1 \xi/T} + 1)) \right] \quad (42)$$

To find the optimal penalty,

$$\frac{\partial P_c}{\partial \xi} = \frac{1}{\Delta U_M - \Delta U_m} \left[\frac{1}{p_1 + p_2} (p_1 e^{p_2(\Delta U_m + \xi)/T} + p_2 (2e^{-p_1 \xi/T} - e^{-p_1(\Delta U_M + \xi)/T})) - 1 \right] = 0,$$

so that

$$p_1 e^{p_2(\Delta U_m + \xi)/T} + p_2 (2e^{-p_1 \xi/T} - e^{-p_1(\Delta U_M + \xi)/T}) = p_1 + p_2.$$

This is the expression for finding the CSD, ξ , that gives the optimal prediction accuracy, in the case of two infinite $A r^k$ wells sampled from canonical parent distributions, $\mu = 1$, and a difference in the minimum energies of the wells that varies randomly with a uniform distribution in the interval $[U_m, U_M]$.

Assume now that for a given trial, either well is equally likely to be lower in energy, so that

$U_m = -U_M$. If there is even sampling — i.e. $p_1 = p_2 = p$ and

$$e^{p(-\Delta U_M + \xi)/T} + 2e^{-p\xi/T} - e^{-p(\Delta U_M + \xi)/T} = 2.$$

Solving for ξ gives

$$\xi = 0 \quad \text{and} \quad \xi = T \ln(2e^{p\Delta U_M/T} - 1)/p.$$

The first solution is the correction that gives the maximal prediction accuracy and the second the minimal one. Hence, similar to the case for finite Ar^k wells sampled from a uniform parent, here the optimal accuracy under equal sampling involves no correction.

3. Calculations

3.1. Use of CSDs to improve prediction in Ark wells

Figure 9 shows the prediction accuracies for a system with two identical Ar^k wells having $\mu = 1$ and differing in energy by a constant, U , that varies randomly in a range of $[-U_M, U_M]$. The wells are sampled at a total of 100 points ($p_1 + p_2 = 100$), either from a uniform parent distribution (Fig. 9(a), finite boundary) or a canonical one (Fig. 9(b), infinite boundary). The solid red and black curves indicate the accuracies for calculations with and without energy corrections, respectively, as a function of p_1 , the number of points sampled in well 1. The numerical and analytic results are indistinguishable on the plots. The prediction accuracies (P_c) are always lowest at $p_1 = 1$: e.g. without corrections, for the uniform parent distribution and $U_M = 0.1$, $P_c = 52.95\%$; for the canonical parent distribution and $T = 25$, $P_c = 53.35\%$. (The lowest possible accuracy is 50%, corresponding to a random choice of well.) In all cases, the accuracies always increase with p_1 , until the maximal prediction accuracy is reached at $p_1 = p_2 = 50$, as expected. For uneven sampling ($p_1 < 50$), the predictions are always improved with the use of the optimal CSD, by up to 12.01% points (at $U_M = 0.5, 1, 2$ and $p_1 = 1$) for the uniform parent distribution and up to 6.7% points (at $T = 1, p = 1$) for the canonical parent distribution, in the range of conditions tested. With either prior sampling distribution, the improvement in prediction due to the optimal CSD, as well as the magnitude of the optimal CSD, itself (dashed lines), diminish as the sampling approaches parity (here, $p_1 = 50$). Note that use of the CSDs here only partially restores the accuracy lost by the sampling discrepancies — i.e. the accuracy with the CSD is always less than the accuracy under sampling parity. For the uniform parent distribution, the prediction accuracies both with and without correction decrease as the range of U narrows (lower U_M). “Flatter” energy landscapes — i.e. smaller differences between local minima — generally yield less accurate structure predictions. For example, the maximal prediction accuracy ($p_1 = 50$) for $U_M = 1$ is 99.03%; for $U_M = 0.1$, it is 90.34%. For the canonical parent distribution, the accuracies decrease (and the magnitudes of the required CSDs increase) as the temperature increases. The maximal accuracy at $T = 1$ is 99.0%; at $T = 25$, it is 78.38%. As shown in Sec. 2.6, canonical sampling at higher temperatures occurs at higher energies, and in infinite wells it results in broader MSE distributions; these, in turn, tend to result in less accurate predictions.

Figure 10 shows the prediction accuracies in the limit of infinite sampling of one of two finite Ar^k wells (here well 2) from a uniform parent distribution, with $U_M = 1$. As expected, the accuracies in the absence of corrections (Fig. 10(a)) are lowest at $p_1 = 1$ and increase monotonically to a value approaching 1 as $p_1 \rightarrow \infty$. For a fixed (finite) number of sampling points, p_1 , in well 1, higher-dimensional/steeper wells (higher μ) are less accurately predicted, since in these cases the distribution of the MSE is shifted further to the right and away from the minimum. For example, at $p_1 = 30$, the accuracies for $\mu = 1/2, 3$ and 25 are

99.90%, 85.74% and 57.33%, respectively. Results for canonical parent distributions demonstrate similar patterns (not shown).

As shown in Fig. 10(b) of the same figure, the prediction accuracies with use of the optimal CSDs are much higher: 87.5%, for all p and μ values. The minimum accuracy occurs at $p_1 = 1$, $\mu = 1$. At smaller μ values, the accuracies are higher because they are higher even without the CSD, as mentioned above. At larger μ values, they are higher because of a similar boundary effect: the spread in the MSE distribution is narrower (see Sec. 2.3, Eq. (15)) and centered near the maximum energies. Narrower spreads in energy distributions generally result in more accurate predictions, provided the proper correction term can be applied. For the flattest wells shown here, $\mu = 1/6$, the accuracies are consistently high (92.8% for all p) and they are about the same for both the corrected and uncorrected predictions (within 3.9×10^{-5} for $p_1 = 2$). The optimal CSDs (Fig. 10(a), dotted lines), themselves, decrease with the number of sampling points and increase with μ .

3.2. Calculations with Ser 103 side chain torsional angle

Calculations were performed with the side chain of Ser 103 in the native state of the protein CheY. The χ_1 torsional energy surface is shown in Fig. 11. Well 3 corresponds to the native minimum (309.3°). The wells are separated by barriers of ≈ 46 – 86 kcal/mol. Figure 12 shows the PDF for the serine side chain torsional energy $f_E(E)$, calculated using Eq. (60) (Supplementary Material: General formula for a pdf, numerical calculation) on a 0.01 kcal/mol grid. The energies around the region of 7 kcal/mol have contributions from the minima of all three wells and therefore have the highest probabilities. The “spikes” in the distribution are near-singularities that correspond to the extrema in the energy surface. Figure 2S A in Supplementary Material, shows the distribution $f_{E;1;1}$ for wells 1 and 2 — i.e. single-point sampling in either well. There are sharp peaks that here correspond to the energy differences between the minima and maxima of the two wells. The highest peak in the distribution occurs at 1.68 kcal/mol, which corresponds to the difference between the minima of the two wells. Even though well 1 has the lower minimum, the probability distribution overall favors the region of $E < 0$, or $E_2 < E_1$. This is due to the fact that the lower-energy region of well 2 is broader, and this effect can dominate at smaller sampling sizes. For the case of $p_1 = 1$ and $p_2 = 10$, the distribution is skewed even further toward more negative E values (Fig. 2S B), although the position of the peak is unchanged. By contrast, in the case of 10 sampling points for each well ($p_1 = p_2 = 10$; Fig. 2S C), the distribution is centered much more closely around the position of the peak.

Using the PDFs for the energies of wells 2 and 3 in the Ser 103 torsional surface and the discrete equivalent of Eq. (26), the (type I) CSDs corresponding to different p_2, p_3 sampling size combinations for the two wells were calculated, relative to a reference of $p_2 = p_3 = 10$. For each combination, the prediction accuracies with and without use of the CSDs were then calculated, over 100,000 random sampling trials. The results are given in Table 1. In the absence of CSDs, for small sampling sizes the prediction accuracy is low, as expected — e.g. $\approx 39\%$ for single-point sampling. It is much higher for larger sampling sizes and $\approx 86\%$ for $p_2 = p_3 = 10$. With the use of the calculated CSDs, however, the prediction accuracy is very nearly 86% for all sampling size combinations. The small deviations are due to

discretization errors around the near-singularities in the PDFs and also some small statistical error in the sampling trials.

3.3. Adiabatic versus rigid rotation mapping in side chain prediction

In a large, complex molecule like a protein it is of interest to compare predictions using adiabatic maps — i.e. surfaces produced by energy-minimizing the structure at each grid position with gradient-based methods — to the native (unminimized) energy surface or a surface locally optimized with other methods. A series of predictive calculations was carried out here for the rotatable side chains of CheY under various conditions, using the CHARMM potential energy function with the all-atom model, and including the 95 water molecules present in the X-ray crystal structure. The results are given in Table 2. The prediction accuracy (last column) is given as the number of angles that are correctly predicted to within $\pm 40^\circ$ of the native position. In general, the worst results are obtained with the adiabatic maps, particularly with deeper minimization (102 out of 181 angles were correct using the expanded Lovell rotamer library and 100 steps of SD). The best results are obtained in the searches without optimization (136–148 angles correct), where only the χ angles are varied, and the rest of the internal coordinates are fixed at the native values. Trials in which the bond angles and an improper dihedral angle in the side chains were varied on a grid for each point on the (χ_1, χ_2) surface give results which are slightly less accurate than the unoptimized (χ_1, χ_2) surfaces (131–132 versus 136 angles correct for the unmodified Lovell libraries). This result is expected, because the discretization of the bond and improper dihedral angular spaces introduces small errors into the predictions (relative to the use of the native angular values). Separately, for the rigid rotations without optimization, the prediction accuracy according to the $\pm 40^\circ$ criterion is essentially unchanged as the grid spacing in the searches is reduced from 20° to 2° , although of course the accuracy as measured by the average angular deviation from the native value drops (not shown). Stricter criteria (e.g. $\pm 30^\circ$ or 20°) result in an approximately uniform reduction in the prediction accuracies. This means that a 10° or 20° grid size is generally sufficient to locate the minimum on the side chain dihedral angle rigid-rotation surface; the finding is consistent with previous observations.²⁰ Some of the calculations were repeated after removal of the structural waters and use of an implicit (EEF1²¹) solvation function and a polar-hydrogen model (results not shown). The predictions were generally less accurate than for the calculations using the explicit structural waters and an all-atom model, but the same patterns were observed. In summary, the findings here suggest that the accuracy of protein side chain predictions using optimized dihedral angle surfaces is related to the manner in which the optimization occurs. Optimization by local grid-based sampling results in more accurate predictions than gradient-based minimization, particularly for deeper minimizations.

3.4. Dependence of the MSE on grid spacing in a model tertiary structure prediction problem

The dependence of the MSE on the grid spacing in a systematic search can be illustrated in a fairly pronounced way by performing simple conformational searches around the native structure of a globular protein in a reduced tertiary structure prediction problem. Using the Z module facility in the CHARMM program, a series of searches with the EEF1 solvation function were carried out for the protein CheY, by varying 36 dihedral angles in the loops

between the 10 secondary structural elements (five α helices and five β strands), whose internal structure was fixed, as in previous work.¹⁹ Here, the searches were conducted only locally about the native structure of the protein. For each series of calculations, the search used a single grid spacing, S . A starting structure for the protein was generated by randomly varying the 36 dihedral angles in a 10° range about the native position. This yielded a conformation that was close to the native in RMSD but much higher in energy. In a preliminary step, several thousand low-energy conformations for each individual loop (and its pairs of flanking secondary structural elements) were first generated on a grid with spacing S , centered at the native positions. Then, in one set of calculations (Table 3A), 100 structures from a band of 5 kcal/mol above the minimum were randomly selected for each loop and used in a series of single-loop (two secondary structural element) searches on the starting structure; the energy evaluations included the entire molecule. (Calculations using larger numbers of structures, e.g. 1000 per loop, yielded similar results and are not shown.) In a second set of calculations (Table 3B), five structures were randomly selected in a band within 5 kcal/mol of the minimum for each loop and a search of all of the loops (10 secondary structural elements, 5⁹ structures) was carried out. In both cases A and B, the minimum-energy conformer for the entire search was selected as the result. This was repeated for $S = 0.125, 0.25, 0.5, 1.0, 2.0$ and $3.0 \text{ \AA}$. In both series, the energy and the RMSD from native decrease, for the most part monotonically, from the largest to the smallest grid size, although the effect is more pronounced in the series with the sparser local sampling (B). In the first series (A), the energy decreases from 772.4 kcal/mol (above that of the native structure) to 2.4 kcal/mol, as the grid spacing decreases from $3.0 \text{ \AA}$ to $0.125 \text{ \AA}$; and in the second series, the energy decreases from 195629.4 kcal/mol to -4.3 kcal/mol. These results illustrate an extreme example of the dependence of the MSE on grid spacing that can occur in a local conformational search of a compact structure like that of a globular protein.

4. Summary and Discussion

The use of potential energy functions to predict structure in large molecular systems is becoming more widespread. Energy-based methods using atomistic models have been used for fitting structures to low-resolution data,^{22–26} refining moderate to high-resolution X-ray data,^{27–29} providing low-to-moderate resolution models for use in crystal structure refinements,^{30,31} folding some small proteins in MD simulations,³² and docking macromolecules.⁴ However, the success rate and accuracy of these and other applications of physics-based methods to macromolecular prediction problems remain substantially less than 100% (e.g. see Refs. 33 and 34). For whole-protein, *ab initio* structure prediction,³⁵ the limitations are thought to be due mainly to the size of the conformational spaces of the target macromolecular systems,^{6,36} for which locating the global minimum-energy structures requires extensive search. By contrast, the current study has examined the possibility of errors in energy-based predictions due to biases introduced by differences in the extent of local sampling around conformers being compared in a search. The work was motivated in part by observations made in prior studies using systematic searches,^{13,17,19,20} in which the best results were obtained when structures were generated “in parallel” — i.e. from a common starting structure, rather than in a serial, chain-like manner,^{37,38} as is commonly done in Monte Carlo-based structure prediction methods.^{39,40} Another observation made in

the course of those studies, and further detailed in the current work, was that the MSE in a search depends on the grid spacing — i.e. smaller grid spacings tend to result in lower sampled minima — raising the question of whether a direct comparison of the energies of structures generated with different grid spacings is valid. An additional observation, unpublished previously but illustrated here, was that predictions of protein structure based on energy-minimized surfaces, or so-called adiabatic maps, are usually no better and often worse than those based on unminimized surfaces. The current analysis is an attempt to shed light on all of these observations.

It has been shown here that for energy wells that are not flat, unequal sampling results in optimization bias — an error, favoring the oversampled well(s), in the estimate of the difference(s) between the actual minimum energies. Provided there is an overlap in the energy ranges of the wells, this bias introduces a systematic error in energy-based predictions. This is a rather general result, independent of the nature of the energy surface^{15,41–45} or the model used to represent the molecular system. Broadly, optimization bias is related to selection bias,⁴⁶ but while the latter arises from the selection of a subpopulation that is not representative of the larger parent population, often through the application of prior knowledge, and results in skewed statistics (e.g. averages), optimization bias arises from discrepancies in the extent of local sampling between different subpopulations (here, regions of conformational space), does not require prior knowledge of the subpopulations, and results in errors in estimates of the differences between local extrema (here local energy minima). The occurrence of optimization bias is not limited to energy surfaces: it may arise in any problem involving a search for extrema over a surface having two or more local minima (or maxima) and hence may also have relevance to areas of inquiry outside that under study here.

In the current work, for simple double-well or multiwell systems, it has been shown that if the form of the energy surface is known, the effects of optimization bias on the probability of selecting one well over another (as that likely to contain the minimum) can be corrected by essentially imposing energy penalties on the wells that are oversampled; these energy correction terms can be calculated or derived from the energy surface, itself. In systems of three or more wells, the determination of these terms is a multibody problem — i.e. not pairwise decomposable — and can be carried out by solving a system of n equations involving n -tuple cross-correlation integrals. Furthermore, it has been shown here that, at least in model systems, the application of such energy correction terms can improve prediction accuracy. This has been illustrated with multi-dimensional systems in which the

energy surfaces are of the form Ar^k , where A is a constant, $r = (\sum_{i=1}^D x_i^2)^{1/2}$ and D is the dimensionality. For systems of two wells of this form, differing only by a randomly assigned shift in energy, analytic and numerical evidence indicates that the prediction accuracy is optimal when the number of sampling points is equal. In the case of unequal sampling, the prediction accuracy can be improved by imposing on the oversampled well a correction term that can be derived from the potential energy function and the number of sampling points; however, the accuracy thus achieved appears always to be less than that under equal sampling. At least in cases of highly skewed sampling, the loss of prediction accuracy and

the magnitude of the required correction terms are greater in steeper or higher-dimensional wells, which have larger fractions of their states at high energies.

In the Ar^k systems, the current work has demonstrated that the worst predictions, and the largest optimization bias, occurs when the sampling is most skewed — in particular, when one of the wells is sampled only once and the other(s) multiple times. The bias diminishes as the number of sampling points increases either in the undersampled well or in both wells. Hence, for chain-based search methods such as conventional Monte Carlo search,^{47,48} simulated annealing^{49–52} or “model-based search” methods^{11,38} in which an energy basin is typically sampled more than once before a transition is made to another basin, the largest optimization bias, at least in these model systems, occurs at the transition. In effect, the structure in the old well is more highly optimized than the first structure sampled in the new well. In this view, optimization bias contributes to the well-described problem of trapping in MC simulations.^{49,53–56} More generally, the current results suggest that for a system with a large conformational space, such as an atomistic representation of a protein, in which thorough sampling of all local minima is unlikely, the resulting unequal or uneven sampling in chain-based methods may render the predictions subject to systematic errors. Furthermore, for such large systems, which have very large and complicated energy surfaces, it is unclear at present how to determine the magnitudes of oversampling penalties *a priori*, and further investigation will be required. In the absence of such correction terms, the current analysis suggests that for systems in which sampling of the lowest-energy regions will be incomplete or sparse, avoiding the introduction of systematic errors requires unbiased or equal sampling.

The “parallel-generation” approach to conformational search mentioned above, described as the Z method in previous work,¹⁹ avoids optimization bias by applying similar types and numbers of changes to a single common starting structure, so that the degree of optimization of the generated structures is essentially the same. However, the current analysis also demonstrates that, even in parallel-generation searches, optimization bias can be introduced if the procedure is not carried out properly — for example, if the energies of structures generated with different grid spacings in separate stages of a search protocol or regions of conformational space are directly compared, without the appropriate corrections. The dependence that the minimum sampled energy can have on the grid spacing has been illustrated here with a series of calculations around the native structure of the protein CheY.

Predictions of protein side chain conformations based on “adiabatic maps” — surfaces minimized with energy-gradient methods (e.g. steepest descent) — were found in the current study to be worse than predictions based on unminimized surfaces, or ones optimized with local grid-type searches. This is consistent with prior experience and suggests that gradient-based energy minimization of globular proteins introduces “random” optimization bias into energy comparisons, owing to the complexity and ruggedness of the surface. Since the local topography varies significantly from one point on the surface to another, a fixed number of minimization steps results in different degrees of optimization for different conformers, in a manner that appears difficult to predict. Prediction methods using gradient-based energy minimization on atomistic representations of globular proteins are less common than would otherwise be expected. Most successful side-chain prediction methods, for example, involve

some type of rigid rotation or rotamer placement,^{16,57–59} without gradient-based minimization. Gradient-based minimization has been used to predict the structure of protein mutants^{60,61} and as part of X-ray refinement protocols (e.g. Ref. 62) but these studies involve local minimization of individual structures, not large-scale conformational searches, and hence are not subject to optimization bias, which arises in comparisons. In accurate template-free modeling of proteins, gradient-based minimization has been used with the Rosetta program during the all-atom refinement stage of the predictions as one part of a Monte Carlo minimization procedure,⁶³ but only in torsional angle space.^{2,64,65} An additional source of bias can be introduced in adiabatic maps when the gradient-based minimization involves degrees of freedom that are not being explicitly searched on the grid (e.g. energy-minimizing all degrees of freedom in a side chain while explicitly searching only the dihedral angles on the grid). Failure to restore the original, unminimized structure at each grid point in these cases will result in predictions that tend to favor the structures generated later in the search, which are, in effect, more highly optimized.¹⁹ (The studies here restored the original structure at each grid point.)

To avoid optimization bias, searches need not be carried out “systematically” — for example, changes can be made randomly to selected degrees of freedom to generate candidate structures — provided the regions of conformational space being compared are sampled/optimized equally. Likewise, in the design and implementation of structural libraries (e.g. side chain rotamer or backbone dihedral angle libraries) based on statistics^{18,66} that are to be used for energy-based predictions, all the conformers for a given subsystem should be optimized to the same extent. For example, if a set of Lys side chain rotamers contains a member that is more highly optimized than the others, energy-based predictions using the library will favor, to an unphysical extent, protein conformations stabilizing that particular Lys rotamer. A similar introduction of bias can occur in the optimization of X-ray crystal structures that have disordered regions. If there are errors in the structure, modeling the disordered regions before optimizing the rest of the structure will tend to stabilize the initial, incorrect X-ray conformation, thereby making the errors more difficult to identify.¹⁷

The Ar^k well studies carried out here have also illustrated that predictions tend to be more accurate for flatter, lower-dimensional (here, lower- μ) wells on more rugged surfaces. The correct well is more likely to be selected if a greater fraction of its surface is close in energy to the minimum and if it is separated in energy from the other wells by a larger gap. For an individual well, an increase in sampling has the least effect on the MSE if the well is flatter and lower in dimension; for a given well shape, it has the greatest effect at small numbers of sampling points. The latter suggests evenly distributed sampling for optimal efficiency.

For the Ar^k systems, themselves, the order or degree of the well, and its dimensionality, D , are linked in that all of the distributions related to the MSE can be expressed in terms of $D/k = \mu$. Hence, the study of high-dimensional wells in this framework can be greatly simplified by studying their projections onto 1D surfaces.

In systems of n wells, each cross-correlation n -tuple multiple integral involved in determining the energy correction terms was found to be reducible to a single integral over a product of $(n - 1)$ single integrals, thereby greatly diminishing the complexity of the

integration or the time for numerical computation. This reducibility arises from the independence of the sampling in each well; the $\{E_1, E_2, \dots\}$ variables depend only on the primary variable (E) and are independent of one another. Hence, the integrals are reducible provided the integration over the primary variable is the last.

The approach taken in the current analysis does not take entropic effects into account; an approach that does is significantly more complicated and will be the subject of further study. Nonetheless, the inclusion of entropic terms is expected to preserve the sign, if not the magnitude, of the effects of oversampling.

In conclusion, the results of this and previous, related work^{13,17,19,20} indicate that in a system having potential energy wells with overlapping energy ranges, for the purposes of structure prediction the energy of a particular conformation cannot be interpreted on its own; the manner in which the conformation was generated needs also to be considered. The same energy in two conformations of the same system — or, indeed, the same conformation generated with different sampling densities — can have different meanings with regard to prediction of the native structure, because the conformer energies are being used as proxies for the energies of the respective local minimum-energy conformers or set of conformers, and their relationship to those minima is statistical. This contextual effect becomes less important the more complete the sampling becomes — it is small in cases of systems with small conformational spaces undergoing thorough sampling, but it can be significant in systems like macromolecules in atomistic representations, which tend to have extremely large conformational spaces, undergoing less complete sampling.

5. Computational Details

5.1. Side chain predictions using minimized and unminimized surfaces

In order to test the effect of minimization on energy-based structural predictions, a series of predictive calculations was performed on the 100 rotatable side chains of the CheY protein (2chf, 1.8Å resolution structure taken from the RCSB,⁶⁷ 128 total residues), one side chain at a time, while keeping the protein backbone fixed. In the main set of calculations, the all-atom protein model (c27⁴²) in the CHARMM program^{14,15} was used. The 95 structural waters in the X-ray structure were included and represented with the TIP3P model.⁶⁸ In each case, the side chains were rotated by varying the χ_1 and χ_2 dihedral angles, either systematically on a $360^\circ \times 360^\circ$ grid (or in the case of Val, Ser and Thr, by varying only χ_1 on a 360° grid) at a fixed step size from 2° to 20° , or according to one of 2 rotamer libraries. One library is that of Lovell *et al.*⁶⁶; the second one was generated from this library by including conformers at $\chi_1 \pm 20^\circ$ and $\chi_2 \pm 20^\circ$ for each conformer in the original library, so that for side chains with at least two χ angles, there is a nine-fold expansion. Trials were conducted with different methods of optimization for the (χ_1, χ_2) surfaces. In some trials, the structures were minimized, either with steepest descent or conjugant gradient, for a fixed number of steps at each (χ_1, χ_2) gridpoint, to generate “adiabatic” energy surfaces. In other trials, the optimization was carried out with systematic grid-based local sampling of the bond angles and improper angles of the side chain for each (χ_1, χ_2) gridpoint. The angles that were varied were the C–C α –C β and C α –C β –C γ bond angles and the N–C–C α –C β

improper dihedral angle. In a different subset of these trials, the predictions were made solely on the basis of the rigid-rotation χ_1 , χ_2 maps, without local minimization.

5.2. Software

The CHARMM program,^{14,15} and the Z module¹⁹ in particular, were used for the amino acid and protein conformational searches and energy evaluations, as mentioned. The software used for random sampling of the multi-dimensional model systems, numerical calculation of the PDFs, and calculation of the cross-correlation multiple integrals and CSDs in multi-well systems was developed by the author. Mathematica 8.04⁶⁹ was used for creating figures as well as performing some numerical and symbolic calculations. Adobe CS5 and Grace were also used to create figures.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The author thanks Professor Martin Karplus of the Department of Chemistry and Chemical Biology at Harvard for his reading of the manuscript and for his support. He thanks Guishan Zheng for technical assistance. Victor Ovchinnikov had useful comments on the manuscript. NIH grants R01 RR023920 and R01 GM103695 were administered through the group of Charles L. Brooks, III, Department of Chemistry, University of Michigan.

6. Glossary and Abbreviations

CSD —	correction for sampling discrepancy. A correction term added to a statistical result (in the current work, a minimum sampled energy) to account for a discrepancy in the degree or extent of sampling between one subpopulation (here, a set of points in a region of an energy surface) and another.
MSE —	minimum sampled energy. The minimum energy over all sampled conformers in a search — i.e. the MSE over q sampling points is $E_q = \min \{U_0, U_1, U_2, \dots; U_q\}$, where U_j is the energy of sampling point j and $j, q \in \mathbb{Z}$. In the current work, an MSE is often taken to be local to a single energy well.
Optimization bias —	generally, the error in an estimate of the difference(s) between or among local extrema of a quantity in different subpopulations of a parent population (or in different regions, states, or component subsystems of a system), caused by discrepancies in the extent of local sampling. In the current work it refers more narrowly to the error in an estimate of the differences between or among local energy minima in different regions of a molecular system's conformational space, caused by discrepancies in the extent of local sampling around those minima.
PDF —	probability distribution function.

Parent spatial
distribution —

for a spherically symmetric system, the distribution of states along a single radius.

Appendix A. Multiple-Well Cases

A.1. General three-well case

For three wells, namely, wells 1, 2 and 3, with MSEs E_1 , E_2 and E_3 , obtained by sampling p_1 , p_2 and p_3 points, respectively, three individual PDFs, $f_{E;p_1}$, $g_{E;p_2}$ and $h_{E;p_3}$, for the MSEs can be defined as per Eq. (12). Analogously to Eq. (16), then, a PDF can be defined over a two-dimensional surface spanned by the differences between the MSEs of wells 1 and 2 ($E_1 = E_2 - E_1$) and wells 1 and 3 ($E_2 = E_3 - E_1$). This PDF is

$$f_{\Delta E_1, \Delta E_2; p_1, p_2, p_3}(\Delta E_1, \Delta E_2) = \int_{\ell}^{\mathcal{L}} f_{E;p_1}(E) \cdot g_{E;p_2}(E + \Delta E_1) \cdot h_{E;p_3}(E + \Delta E_2) dE, \quad (\text{A.1})$$

where the lower and upper limits, ℓ and $\mathcal{L}$, take on different values in each of 6 regions. They are given in Table 1S in Supplementary Material.

In the current analysis, systems of three or more wells are not pairwise decomposable. As an illustration, consider that if the wells in a three-well system have the same form, the probability of well 1 having the lowest minimum is

$$\text{Pr}(1) = \frac{p_2 p_3 (2p_1 + p_2 + p_3)}{(p_1 + p_2)(p_1 + p_3)(p_1 + p_2 + p_3)},$$

which is not equal to the product of the pairwise probabilities of well 1 having a lower MSE than either of the other two wells, $p_1^2 / ((p_1 + p_2)(p_1 + p_3))$.

Analogously to the two-well case, if the sampling of the three wells is unequal — i.e. $p_1 \neq p_2 \neq p_3$, three CSD terms can be defined, $\xi_{1,2}$, $\xi_{2,3}$ and $\xi_{1,3}$, which correct the total probabilities of each of the wells having the lowest MSE. For example, $\xi_{1,2}$ is the energy penalty applied to well 2 relative to well 1. Because of the multibody nature of the problem, each CSD (e.g. $\xi_{1,3}$) cannot be determined from a consideration of only two wells (1 and 3). Instead, the three CSDs must be determined by solving three simultaneous triple integral (or summation) constraint equations. The general equations for wells of arbitrary type are:

$$\begin{aligned} \int \int \int_{R_1} f_{E;p_1}(E) \cdot g_{E;p_2}(E + \Delta E_1) \cdot h_{E;p_3}(E + \Delta E_2) dE d\Delta E_2 d\Delta E_1 = \\ \int \int \int_{Q_{1'}} f_{E;p_1}(E) \cdot g_{E;p_1}(E + \Delta E_1) \cdot h_{E;p_1}(E + \Delta E_2) dE d\Delta E_2 d\Delta E_1, \end{aligned} \quad (\text{A.2})$$

where $E_1 = E_2 - E_1$ and $E_2 = E_3 - E_1$,

$$\begin{aligned} \int \int \int_{R_2} f_{E;p_1}(E + \Delta E_1) \cdot g_{E;p_2}(E) \cdot h_{E;p_3}(E + \Delta E_2) dE d\Delta E_2 d\Delta E_1 = \\ \int \int \int_{Q_{1'}} f_{E;p_1}(E + \Delta E_1) \cdot g_{E;p_1}(E) \cdot h_{E;p_1}(E + \Delta E_2) dE d\Delta E_2 d\Delta E_1, \end{aligned} \quad (\text{A.3})$$

where $E_1 = E_1 - E_2$ and $E_2 = E_3 - E_2$, and

$$\begin{aligned} & \int \int \int_{R_3} f_{E;p_1}(E+\Delta E_1) \cdot g_{E;p_2}(E+\Delta E_2) \cdot h_{E;p_3}(E) dE d\Delta E_2 d\Delta E_1 = \\ & \int \int \int_{Q_{I'}} f_{E;p_1}(E+\Delta E_1) \cdot g_{E;p_1}(E+\Delta E_2) \cdot h_{E;p_1}(E) dE d\Delta E_2 d\Delta E_1, \end{aligned} \quad (\text{A.4})$$

where $E_1 = E_1 - E_3$ and $E_2 = E_2 - E_3$.

The R_1, R_2, R_3 and $Q_{I'}$ terms represent the three-dimensional regions of integration. The rhs of the three equations gives the probability that the MSE will be located in well 1, 2 or 3, respectively, under equal sampling and without energy correction. For wells with $E \in [0, 1]$, this set of definitions conveniently places the two-dimensional portion of $Q_{I'}$ that is in the $\{E_1, E_2\}$ plane in quadrant I for all three equations; specifically, $\{0 < E_1 < 1, 0 < E_2 < 1\}$ (Fig. 13(d)). The limits of the two integrals in this region are listed in Table 1S in Supplementary Material.

The lhs of the three equations give the probability that the MSE will be found in well 1, 2 or 3, respectively, under unequal sampling and with energy corrections. For the case of $p_1 < p_2 < p_3$, the 2D portions of R_1, R_2 and R_3 in the $\{E_1, E_2\}$ plane correspond to the shaded areas in Figs. 13(a)–13(c), and sections S_1, S_2 and S_3 , respectively, in Table 2S in Supplementary Material. The table lists the limits, which involve the CSDs. (In this example, for S_1 , ξ_1 in the table corresponds to $\xi_{1,2}$, and ξ_2 to $\xi_{1,3}$; for S_2 , ξ_1 corresponds to $-\xi_{1,2}$ and ξ_2 to $\xi_{2,3}$, and for S_3 , ξ_1 corresponds to $-\xi_{1,3}$ and ξ_2 to $-\xi_{2,3}$.) The S_1 integral has seven parts, the S_2 integral has three parts and the S_3 integral two parts.

More specifically, in this example, the integrals for S_3 (lhs of Eq. (A.4)), corresponding to the region $\{E_1 > \xi_{1,3}, E_2 > \xi_{2,3}\}$, for $0 > \xi_{2,3} > \xi_{1,3}$, are as follows:

For $E_2 > E_1 > \xi_{1,3}$ (region marked “1” in Fig. 13(c)),

$$\mathcal{J}_1 = \int_{\xi_{1,3}}^1 \int_{\Delta E_1}^1 \int_0^{1-\Delta E_2} h_{E;p_3}(E) \cdot f_{E;p_1}(E+\Delta E_1) \cdot g_{E;p_2}(E+\Delta E_2) dE d\Delta E_2 d\Delta E_1, \quad (\text{A.5})$$

and for $\{E_1 > E_2 > \xi_{1,3}\} \cup \{\xi_{1,3} > E_1 > \xi_{2,3}, 1 > E_1 > \xi_{1,3}\}$ (region marked “2” in Fig. 13(c)),

$$\mathcal{J}_2 = \int_{\xi_{1,3}}^1 \int_{\xi_{2,3}}^{\Delta E_1} \int_0^{1-\Delta E_1} h_{E;p_3}(E) \cdot f_{E;p_1}(E+\Delta E_1) \cdot g_{E;p_2}(E+\Delta E_2) dE d\Delta E_2 d\Delta E_1, \quad (\text{A.6})$$

One of the integral constraint equations is thus

$$\text{Pr}(3) = \mathcal{J}_1 + \mathcal{J}_2 = \int \int \int_{Q_{I'}} f_{E;p_1}(E+\Delta E_1) \cdot g_{E;p_1}(E+\Delta E_2) \cdot h_{E;p_1}(E) dE d\Delta E_2 d\Delta E_1.$$

The remaining two constraint equations may be derived in a similar fashion from Eqs. (A.2) and (A.3), with the integrals of the lhs having regions of integration S_1 and S_2 , respectively.

For finite, identical Ar^k wells, sampling from a uniform parent distribution, with $\mu = 1$, and $(p_1, p_2, p_3) = (1, 2, 3)$, the solution is $(\xi_{1,2}, \xi_{2,3}, \xi_{1,3}) = (0.19245; 0.05826; 0.13419)$. Since

each equation is of degree six in the CSD in this case, the analytic solution has a degree in the hundreds, and thus solving the equations numerically is far more practical. For $\mu < 1$ (e.g. quadratic 1D wells), the integrals are not in general soluble in closed form, and the entire procedure must be carried out numerically.

A.2. Reduction of cross-correlation multiple integrals

The “natural” order of integration for cross-correlation (or convolution) integrals may be taken as the integration over the primary variable first, since this gives the PDF over all i . Here, for $n + 1$ wells, integration of the cross-correlation multiple integrals over the MSE (E) gives the PDF over $\{E_1, E_2, \dots, E_n\}$. However, for the purposes of determining specific probabilities when $n = 2$, the integrals can be simplified by changing the order of integration. Because the sampling of each energy well is independent, the “ E ” variables are not interdependent; they depend only on the primary variable, E . Hence, integrating over the E variables first allows the integral to be split. For example, in the case given above, $\mathcal{J}_1$ and $\mathcal{J}_2$ can be combined and the order of integration reversed with appropriate changes in limits (see Table 2S B, S_3), so that the total probability of well 3 having the lowest MSE (given by the integral over $E_1 > \xi_{1,3}$, $E_2 > \xi_{2,3}$, when $0 < \xi_{2,3} < \xi_{1,3}$) is

$$\text{Pr}(3) = \int_0^{1-\xi_{1,3}} \int_{\xi_{2,3}}^{1-E} \int_{\xi_{1,3}}^{1-E} h_{E;p_3}(E) \cdot f_{E;p_1}(E+\Delta E_1) \cdot g_{E;p_2}(E+\Delta E_2) d\Delta E_1 d\Delta E_2 dE.$$

The inner two integrals are now independent of each other, so that the triple integral can be split:

$$\int_0^{1-\xi_{1,3}} h_{E;p_3}(E) \left(\int_{\xi_{1,3}}^{1-E} f_{E;p_1}(E+\Delta E_1) d\Delta E_1 \right) \times \left(\int_{\xi_{2,3}}^{1-E} g_{E;p_2}(E+\Delta E_2) d\Delta E_2 \right) dE.$$

The two integrals in parentheses can be evaluated individually, and then the product of the results and $f_{E;p_1}(E)$ can be integrated with respect to dE . Thus, provided the inner integrals are soluble in closed form, the procedure reduces a triple integral to a single integral.

More specifically, in the case of $A r^k$ wells with a finite boundary ($r \in [0, 1]$),

$$\begin{aligned} \text{Pr}(3) &= \int_0^{1-\xi_{1,3}} \frac{\partial(-(1-\tilde{E}^\mu)^{p_3})}{\partial \tilde{E}} \times \\ &\quad \left(\int_{\xi_{1,3}}^{1-\tilde{E}} \frac{\partial(-(1-(\tilde{E}+\Delta \tilde{E}_1)^\mu)^{p_1})}{\partial \Delta \tilde{E}_1} d\Delta \tilde{E}_1 \right) \\ &\quad \times \left(\int_{\xi_{2,3}}^{1-\tilde{E}} \frac{\partial(-(1-(\tilde{E}+\Delta \tilde{E}_2)^\mu)^{p_2})}{\partial \Delta \tilde{E}_2} d\Delta \tilde{E}_2 \right) d\tilde{E} \\ &= \int_0^{1-\xi_{1,3}} \frac{\partial(-(1-\tilde{E}^\mu)^{p_3})}{\partial \tilde{E}} ((1-(\xi_{1,3}+\tilde{E})^\mu)^{p_1}) \times ((1-(\xi_{2,3}+\tilde{E})^\mu)^{p_2}) d\tilde{E}, \end{aligned}$$

where $\tilde{E} = E/E_M$. In numerical calculations, the equivalent reduction of the triple sums to double sums can reduce the time complexity of the calculation from $O(N^3)$ to $O(N^2)$, if the individual integrals cannot be solved in closed form, or $O(N)$ if they can.

Higher-order (e.g. quadruple, quintuple) cross-correlation (and convolution) integrals can also be reduced in this fashion. For n wells, the n -tuple multiple integral for the region S_n , which is analogous to S_3 in the 3-well case, has $E_i > 0$, for all $i = \{1, 2, \dots, n-1\}$, and $\xi_{1,n} > \xi_{2,n} > \dots > \xi_{n-1,n} > 0$. It can be reduced to a single integral over a product of $n-1$ single integrals as:

$$\begin{aligned} \Pr(n) = & \int_0^{1-\xi_{1,n}} f_{E;p_n}(E) \left(\int_{\xi_{1,n}}^{1-E} f_{E;p_1}(E+\Delta E_1) d\Delta E_1 \right) \\ & \times \left(\int_{\xi_{2,n}}^{1-E} f_{E;p_2}(E+\Delta E_2) d\Delta E_2 \right) \\ & \times \dots \left(\int_{\xi_{n-1,n}}^{1-E} f_{E;p_{n-1}}(E+\Delta E_n) d\Delta E_n \right) dE. \end{aligned} \quad (\text{A.7})$$

References

1. Allen B, Mayo S. Dramatic performance enhancements for the faster optimization algorithm. *J Comput Chem.* 2006; 27:1071–1075. [PubMed: 16685715]
2. Das R, Baker D. Macromolecular modeling with Rosetta. *Ann Rev Biochem.* 2008; 77:363–82. [PubMed: 18410248]
3. Zhu J, Fan H, Periole X, Honig B, Mark A, Alan E. Refining homology models by combining replica-exchange molecular dynamics and statistical potentials. *Proteins.* 2008; 72:1171–1188. [PubMed: 18338384]
4. Komazin-Meredith G, Petrella R, Santos W, Filman D, Hogle J, Verdine G, Karplus M, Coen D. The human cytomegalovirus UL44 C clamp wraps around DNA. *Structure.* 2009; 16:1214–1225. [PubMed: 18682223]
5. Ma J. Explicit orientation dependence in empirical potentials and its significance to side-chain modeling. *Acc Chem Res.* 2009; 42:1087–1096. [PubMed: 19445451]
6. Bradley P, Malmstrom L, Qian B, Schonbrun J, Chivian D, Kim D, Meiler J, Misura K, Baker D. Free modeling with Rosetta in CASP6. *Proteins.* 2005; (Supl 7):128–134. [PubMed: 16187354]
7. Xia Y, Huang E, Levitt M, Samudrala R. *Ab initio* construction of protein tertiary structure using a hierarchical approach. *J Mol Biol.* 2000; 300:171–185. [PubMed: 10864507]
8. Pillardy J, Czaplewski C, Wedemeyer W, Scheraga H. Conformation-family Monte Carlo (CFMC): An efficient computational method for identifying the low-energy states of a macromolecule. *Helvetica Chimica Acta.* 2000; 83:2214–2230.
9. Zhang Y, Kihara D, Skolnick J. Local energy landscape flattening: Parallel hyperbolic Monte Carlo sampling of protein folding. *Proteins.* 2002; 48:192–201. [PubMed: 12112688]
10. Paluszewski M, Hamelryck T, Winter P. Reconstructing protein structure from solvent exposure using tabu search. *Alg Mol Biol.* 2006; 1:1–14.
11. Brunette T, Brock O. Guiding conformation space search with an all-atom energy potential. *Proteins.* 2008; 73:958–972. [PubMed: 18536015]
12. McAllister S, Floudas C. An improved hybrid global optimization method for protein tertiary structure prediction. *Comput Optim Appl.* 2010; 45:377–413. [PubMed: 20357906]
13. Petrella R, Karplus M. A limiting-case study of protein structure prediction: Energy-based searches of reduced conformational space. *J Phys Chem B.* 2000; 104:11370–11378.
14. Brooks B, Bruccoleri R, Olafson B, States D, Swaminathan S, Karplus M. CHARMM: A program for macromolecular energy minimization and dynamics calculations. *J Comp Chem.* 1983; 4:187–217.
15. Brooks B, CL Brooks I, Mackerell A, Nilsson L, Petrella R, Roux B, Won Y, et al. CHARMM: The biomolecular simulation program. *J Comput Chem.* 2009; 30:1545–1614. [PubMed: 19444816]
16. Xiang Z, Honig B. Extending the accuracy limits of prediction for side-chain conformations. *J Mol Biol.* 2001; 311:421–430. [PubMed: 11478870]

17. Petrella R, Karplus M. The energetics of off-rotamer protein sidechain conformations. *J Mol Biol.* 2001; 312:1161–1175. [PubMed: 11580256]
18. Jacobson M, Pincus D, Rapp C, Day T, Honig B, Shaw D, Friesner R. A hierarchical approach to all-atom protein loop prediction. *Proteins.* 2004; 55:351–367. [PubMed: 15048827]
19. Petrella R. A versatile method for systematic conformational searches: Application to CheY. *J Comput Chem.* 2011; 32:2369–2385. [PubMed: 21557263]
20. Petrella R, Lazaridis T, Karplus M. Protein sidechain conformer prediction: A test of the energy function. *Folding Des.* 1998; 3:353–377.
21. Lazaridis T, Karplus M. Effective energy function for proteins in solution. *Proteins.* 1999; 35:133–152. [PubMed: 10223287]
22. Brunger A, Clore G, Gronenborn A, Karplus M. Three-dimensional structure of protein determined by molecular dynamics with interproton distance restraints: Application to crambin. *Proc Natl Acad Sci USA.* 1986; 83:3801–3805. [PubMed: 3459158]
23. Tama F, Miyashita O, CL Brooks I. Flexible multi-scale fitting of atomic structures into low-resolution electron density maps with elastic network normal mode analysis. *J Mol Biol.* 2004; 337:985–999. [PubMed: 1503365]
24. Delarue M, Dumas P. On the use of low-frequency normal modes to enforce collective movements in refining macromolecular structural models. *PNAS.* 2004; 101:6957–6962. [PubMed: 15096585]
25. Poon B, Chen X, Lu M, Vyas N, Quijcho F, Wang Q, Ma J. Normal mode refinement of anisotropic thermal parameters for a supramolecular complex at 3.42-crystallographic resolution. *Proc Natl Acad Sci USA.* 2007; 104:7869–7874. [PubMed: 17470791]
26. Schroder G, Levitt M, Brunger A. Super-resolution biomolecular crystallography with low-resolution data. *Nature.* 2010; 464:1218–1222. [PubMed: 20376006]
27. Brunger AT, Kuriyan J, Karplus M. Crystallographic R factor refinement by molecular dynamics. *Science.* 1987; 235:458–460. [PubMed: 17810339]
28. Lin M, Head-Gordon T. Reliable protein structure refinement using a physical energy function. *J Comput Chem.* 2011; 32:709–717. [PubMed: 20925085]
29. Bell J, Ho K, Farid R. Significant reduction in errors associated with nonbonded contacts in protein crystal structures: Automated all-atom refinement with primex. *Acta Crystallogr D.* 2012; 68:935–952. [PubMed: 22868759]
30. Shrestha R, Berenger F, Zhang K. Accelerating *ab initio* phasing with *de novo* models. *Acta Crystallogr D.* 2011; 67:804–812. [PubMed: 21904033]
31. Simoncini D, Berenger F, Shrestha R, Zhang K. A probabilistic fragment-based protein structure prediction algorithm. *PLoS ONE.* 2012; 7:e38799.
32. Lindorff-Larsen K, Piana S, Dror R, Shaw D. How fast-folding proteins fold. *Science.* 2011; 334:517–520. [PubMed: 22034434]
33. Lensink M, Wodak S. Blind predictions of protein interfaces by docking calculations in CAPRI. *Proteins.* 2010; 78:3085–3095. [PubMed: 20839234]
34. Dill K, MacCallum J. The protein-folding problem, 50 years on. *Science.* 2012; 338:1042–1046. [PubMed: 23180855]
35. MacCallum J, Perez A, Schnieders M, Hua L, Jacobson M, Dill K. Assessment of protein structure refinement in CASP9. *Proteins.* 2011; 79(S10):74–90. [PubMed: 22069034]
36. Shenoy S, Jayaram B. Proteins: Sequence to structure and function—current status. *Curr Protein Pept Sc.* 2010; 11:498–514. [PubMed: 20887265]
37. Markov, M. *Dynamic Probabilistic Systems, 1: Markov Chains.* New York: John Wiley and Sons; 1971. reprinted in Appendix B.
38. Brunette T, Brock O. Improving protein structure prediction with model-based search. *Bioinformatics.* 2005; 21(Suppl 1):i66–i74. [PubMed: 15961500]
39. Bradley P, Baker D. Improved beta-protein structure prediction by multilevel optimization of non-local strand pairings and local backbone conformation. *Proteins.* 2006; 65:922–929. [PubMed: 17034045]

40. Mamonov AB, Bhatt D, Cashman D, Ding Y, Zuckerman D. General library-based Monte Carlo technique enables equilibrium sampling of semi-atomistic protein models. *J Phys Chem.* 2009; 113:10891–10904.
41. Cornell W, Cieplak P, Bayly C, Gould I, KM Merz J, Ferguson D, Spellmeyer D, Fox T, Caldwell J, Kollman P. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J Am Chem Soc.* 1995; 117:5179–5197.
42. Neria E, Fischer S, Karplus M. Simulation of activation free energies in molecular systems. *J Chem Phys.* 1996; 105:1902.
43. Lyne P, Hodoscek M, Karplus M. A hybrid qm-mm potential employing hartree-fock or density functional methods in the quantum region. *J Phys Chem A.* 1999; 103:3462–3471.
44. Wei W. Differential geometry based multiscale models. *Bull Math Biol.* 2010; 72:1562–1622. [PubMed: 20169418]
45. Schmid N, Eichenberger A, Choutko A, Riniker S, Winger M, Mark A, van Gunsteren W. Definition and testing of the gromos force-field versions 54a7 and 54b7. *Euro Biophys J.* 2011; 40:843–856.
46. Heckman J. Sample selection bias as a specification error. *Econometrica.* 1979; 47:151–161.
47. Metropolis N, Ulam S. The Monte Carlo method. *J Am Stat Assoc.* 1949; 44:335–341. [PubMed: 18139350]
48. von Freyberg B, Braun W. Efficient search for all low energy conformations of polypeptides by Monte Carlo methods. *J Comput Chem.* 1991; 12:1065–1076.
49. Kirkpatrick S, Gelatt C, Vecchi M. Optimization by simulated annealing. *Science.* 1983; 220:671–680. [PubMed: 17813860]
50. Chou K, Carlacci L. Simulated annealing approach to the study of protein structures. *Protein Eng.* 1991; 4:661–667. [PubMed: 1946323]
51. Huber G, McCammon J. Weighted-ensemble simulated annealing: Faster optimization on hierarchical energy surfaces. *Phys Rev E.* 1997; 55:4822–4825.
52. Liu Y, Beveridge D. Exploratory studies of *ab initio* protein structure prediction: Multiple copy simulated annealing, AMBER energy functions, and a generalized born/solvent accessibility solvation model. *Proteins.* 2002; 46:128–146. [PubMed: 11746709]
53. Berg B, Neuhaus T. Multicanonical algorithms for first order phase transitions. *Phys Lett B.* 1991; 267:249–253.
54. Hansmann U. Parallel tempering algorithm for conformational studies of biological molecules. *Chem Phys Lett.* 1997; 281:140–150.
55. Andricioaei I, Straub J. On Monte Carlo and molecular dynamics methods inspired by Tsallis statistics: Methodology, optimization, and application to atomic clusters. *J Chem Phys.* 1997; 907:9117–9124.
56. Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. *Proc Natl Acad Sci USA.* 2000; 97:10383–10388. [PubMed: 10984534]
57. Xu J, Berger B. Fast and accurate algorithms for protein side-chain packing. *JACM.* 2006; 53:533–557.
58. Krivov G, Shapovalov M, Dunbrack R. Improved prediction of protein side-chain conformations with scwrl4. *Proteins.* 2009; 77:778–795. [PubMed: 19603484]
59. Liang S, Zhou Y, Grishin N, Standley D. Protein side chain modeling with orientation dependent atomic force fields derived by series expansions. *J Comput Chem.* 2011; 32:1680–1686. [PubMed: 21374632]
60. Kini R, Evans H. Molecular modeling of proteins—a strategy for energy minimization by molecular mechanics in the AMBER force-field. *J Biomol Struct Dyn.* 1991; 9:475–488. [PubMed: 1687724]
61. Machicado C, Bueno M, Sancho J. Predicting the structure of protein cavities created by mutation. *Protein Eng.* 2002; 15:669–675. [PubMed: 12364581]
62. Brunger A, Karplus M, Petsko G. Crystallographic refinement by simulated annealing: Application to crambin. *Acta Crystallogr A.* 1989; 45:50–61.
63. Li Z, Scheraga H. Monte Carlo-minimization approach to the multiple minima problem in protein folding. *Proc Natl Acad Sci USA.* 1987; 84:6611–6615. [PubMed: 3477791]

64. Rohl C, Strauss C, Misura K, Baker D. Protein structure prediction using Rosetta. *Method Enzymol.* 2004; 383:66–93.
65. Raman S, Vernon R, Thompson J, Tyka M, Sadreyev R, Pei J, Kim D, Kellogg E, DiMaio F, Lange O, Kinch L, Sheffler W, Kim B, Das R, Grishin NV, Baker D. Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins.* 2009; 77S9:89–99. [PubMed: 19701941]
66. Lovell S, Word J, Richardson J, Richardson D. The penultimate rotamer library. *Proteins.* 2000; 40:389–408. [PubMed: 10861930]
67. Berman H, Henrick K, Nakamura H. Announcing the worldwide protein data bank. *Nature Struc Biol.* 2003; 10:980.
68. Jorgensen W, Chandrasekhar J, Madura J, Impey R, Klein M. Comparison of simple potential functions for simulating liquid water. *J Chem Phys.* 1983:926–935.
69. Wolfram Research, Inc. Mathematica, Version 8.0. Champaign, IL: 2010.

Fig. 1.

Extreme example of optimization bias in structure prediction. (a) and (c) show the CHARMM energy surface of the Val 20 side chain as a function of the χ_1 dihedral angle, in either the valine dipeptide in vacuum (a) or the native CheY protein environment (c). In both cases, the starting structure is taken from the native CheY coordinates. Note that the global minima are very near the native position of $\chi_1 = -178.40^\circ$. (b) and (d) show the results of the same calculations, except that the starting configurations are obtained by rotating χ_1 to $= 300^\circ$ and then minimizing the structures with 250 steps of Steepest Descent and an additional 250 steps of Adopted-Basis Newton–Raphson. In either environment, the minimization about the non-native conformation biases the landscape in favor of that conformation. In the vacuum case (dipeptide, (a) and (b)), the change in the energy differences between native and non-native conformers due to the minimization is on the order of 10 kcal/mol. In the protein interior (c) and (d), the change is ≈ 150 kcal/mol, and the overall landscape is flatter after minimization. (Note the change in energy scale.)

Fig. 2.

Projection of the energy surfaces in multidimensional Ar^k wells onto one dimension. Each curve is the one-dimensional r^k surface, $r \in [0, 1]$, having the equivalent energy distribution

of the D -dimensional Ar^k well indicated, where $r = (\sum_{i=1}^D x_i^2)^{1/2}$. For example, the surface of a 12-dimensional Ar^k well of shape $E(r) = r^2$ yields the same energy distributions as a one-dimensional surface of shape $r^{1/6}$ (top curve). For fixed D , smaller exponents (orders, k) correspond to steeper curves, with larger fractions of structures at higher energies.

Fig. 3. Probability distribution functions for the MSE in D -dimensional energy wells of the form $E(r) = Ar^k$, with $\mu = D/k = 1$ (red) or $\mu = 5$ (black) and for either 1, 3, or 10 sampling points per trial (as indicated by the number labels by the curves). (a): uniform parent function, finite boundary ($r_{\max} = s = 1$). (b): canonical parent function, finite boundary. (c): canonical parent function, infinite boundary.

(a)

(b)

Fig. 4. Three-dimensional plots of the mean MSE as a function of dimensionality/steepness μ and number of sampling points per trial, $p \in [1, 100]$, for Ar^k wells having a finite boundary and having been sampled from a uniform parent distribution. (a): $\mu \in [0, 1.0]$. (b): $\mu \in [1, 100]$.

Fig. 5. Probability distribution functions for the difference between the minimum sampled energies, $\tilde{E} = \tilde{E}_2 - \tilde{E}_1$, of two wells for which $\mu = 1$ and $p_1 = 1$. (a) The sampling is equal; i.e. $p_2 = p_1 = 1$. (b) The sampling is unequal; i.e. $p_1 = 1, p_2 = 2$.

Fig. 6.

Plots of the energy correction term (type I CSD) for two Ar^k wells having $\mu = 1$ and different numbers of sampling points, p_1 and p_2 , as a function of p_2 . (a) $p_1 = 1$, and (b) $p_1 = 2$. The dots indicate the exact values for the correction, calculated for each (p_1, p_2) pair (Eq. (22), Sec. 2.5). The solid lines are the analytic approximations (see text). Note the differences in scale.

Fig. 7.

Depiction of the four possible outcomes of energy-based prediction in a two-well system and the effect of an energy correction term (type II CSD). In (a) and (d), the predictions are correct; in (b) and (c), they are incorrect. In each of these examples, well 2 has been oversampled relative to well 1; hence, a CSD, or "penalty", ξ , is applied to well 2. In (c) and (d), the MSE for well 2, E_2 , is higher than that for E_1 , even before the application of the CSD, so that the CSD has no effect. In (a), the CSD may reduce prediction accuracy, by converting a correct prediction to an incorrect one. In (b), the CSD may improve prediction accuracy, by doing the opposite.

Fig. 8.

Diagram of the five integrals used for calculating prediction accuracy in finite energy wells. The function being integrated, f_E , is the PDF for the difference between the MSEs of two wells ($E = E_2 - E_1$) whose forms differ only by a randomly varying shift in energy, U — i.e. $E_2(r) = E_1(r) + U$. The shaded regions in each case correspond to the contribution to correct prediction (always on the same side of $E = 0$ as U). Because the energy has finite upper and lower bounds, the f_E curve has two main branches, f^- and f^+ , which may have different analytic forms. E_M is the maximal energy for well 1 (the minimum energy is 0).

The addition of a CSD, or energy correction, ξ , shifts all the curves to the right. For integrals I_1 and I_4 this has no effect on prediction accuracy; because the integral is the entire area under the curve in these cases, the contribution to prediction accuracy is 1, by the definition of a PDF. For integral I_2 the correction improves accuracy. For I_3 and I_5 it diminishes accuracy. See also text.

(a)

(b)

Fig. 9.

Prediction accuracy in a system with two wells and 100 total sampling points, with and without energy corrections (CSDs). The wells are of the form $E_1 = Ar^k$ and $E_2 = Ar^k + U$, U varies randomly in the interval $[-U_M, U_M]$, with a uniform prior distribution, and $u = D/k = 1$, where D is the dimensionality of the wells. The solid black curves indicate the prediction accuracy without CSDs; the red curves indicate the accuracy with the optimal CSD, ξ , which is, itself, indicated with the dotted black lines. The prediction accuracies correspond to the left-hand scale, while the CSDs correspond to the right-hand scale. The

independent variable (horizontal axis) is the number of sampling points, p_1 , for well 1 (and $p_2 = 100 - p_1$). (a) shows the results for finite Ar^k wells, sampled from a uniform parent distribution, and for various values of U_M , which are shown next to the curves. The CSDs for $U_M = 1$ and 2 are identical, as expected, and those for $U_M = 0.5$ are close to those. (b) shows the results for infinite Ar^k wells sampled from a canonical parent distribution, with $U_M = -U_m = 1$, and at various temperatures ("T"), which are indicated. The curves connect points calculated individually from analytically derived expressions; the analytic results were confirmed with computational trials.

(a)

(b)

Fig. 10.

Plots of the prediction accuracy in a system with two finite Ar^k wells sampled from uniform parent distributions, in which the second well is sampled at an infinite number of points. (a) The prediction accuracies (solid curves, left-hand scale) for various μ values (italic labels) as a function of p in the absence of any correction terms. The dotted lines give the analytically derived optimal CSDs (ξ , right-hand scale) for the same μ values. The optimal CSDs for $\mu = 1/6$ are all close to zero. (b) The prediction accuracies using the optimal CSDs shown in (a), for the same μ values (italic labels). Note the change in scale.

Fig. 11.

Plot of the χ_1 torsional side chain energy for Ser 103 in the protein CheY. The minima are $(\chi_1, E(\chi_1)) = (309.3^\circ, 0 \text{ kcal/mol}), (70.0^\circ, 2.67 \text{ kcal/mol}), (213.9^\circ, 4.31 \text{ kcal/mol})$; the maxima are $(\chi_1, E(\chi_1)) = (22.9^\circ, 85.96 \text{ kcal/mol}), (144.6^\circ, 55.47 \text{ kcal/mol}), (269.7^\circ, 46.60 \text{ kcal/mol})$. The well boundaries are defined as the χ values at the peaks of the barriers. Note that well 2 has a broader, though higher, minimum than the other two wells.

Fig. 12.

Plot of the PDF for the Ser 103 torsional energy. The six spikes in the distribution correspond to the three minima and three maxima in the energy surface (see Fig. 11).

Fig. 13.

Plots of the regions of integration for the cross-correlation triple integrals used to compute energy correction terms (type II CSDs) for a three-well system with unequal sampling. The shaded areas indicate the regions in the $\{\Delta E_1, \Delta E_2\}$ plane over which the integrals are valid. In (a)–(c), they correspond to the integrals on the lhs of Eqs. (A.2)–(A.4), when $p_1 < p_2 < p_3$ — i.e. the probabilities of correct prediction under unequal sampling. They also correspond to the integrals whose limits are given in Table 2S of Supplementary Material, sections S_1 , S_2 and S_3 , respectively. In (c), the areas marked “1” and “2” correspond to integrals 1 and 2

under section S_3 of the table. Note the lower boundaries of the regions involve the correction terms, ξ_1 and ξ_2 . By contrast, the shaded area in (d) indicates the 2D portion of the region of integration for the integrals on the rhs of the equations — i.e. the probability of correct prediction under equal sampling, with no correction terms — and corresponding to integrals 1 and 2 in Table 1S of Supplementary Material.

The use of (type I) CSDs to correct the effects of sampling size discrepancies in the prediction of the side chain position of Ser 103 in the protein CheY.

Table 1

p_2	p_3	Expected correct	Observed correct	% Error	CSD (kcal/mol)	Observed cor w/CSD	% Error
1	1	0.3941	0.3860	2.1085	58.5701	0.8642	0.3411
1	2	0.6053	0.5922	2.2221	29.2454	0.8617	0.6264
1	3	0.7288	0.7177	1.5500	13.6015	0.8617	0.6276
1	5	0.8569	0.8494	0.8840	0.7674	0.8585	1.0062
1	10	0.9513	0.9585	-0.7509	-4.2382	0.8598	0.8511
2	1	0.3000	0.2908	3.1736	66.6519	0.8639	0.3678
2	2	0.4953	0.4794	3.3288	35.7780	0.8635	0.4143
2	3	0.6279	0.6136	2.3271	20.0985	0.8612	0.6930
2	5	0.7867	0.7744	1.5938	5.1547	0.8604	0.7761
2	10	0.9281	0.9315	-0.3732	-2.8966	0.8618	0.6159
3	1	0.2584	0.2469	4.6701	69.5775	0.8633	0.4411
3	2	0.4418	0.4240	4.1933	38.1660	0.8609	0.7269
3	3	0.5746	0.5549	3.5476	22.8917	0.8621	0.5785
3	5	0.7450	0.7277	2.3762	7.1797	0.8607	0.7410
3	10	0.9119	0.9114	0.0487	-1.9921	0.8622	0.5762
5	1	0.2221	0.2102	5.6285	71.9876	0.8661	0.1198
5	2	0.3914	0.3702	5.7196	39.9580	0.8631	0.4608
5	3	0.5213	0.5007	4.1296	25.3682	0.8611	0.7023
5	5	0.6991	0.6814	2.6003	9.0536	0.8605	0.7726
5	10	0.8911	0.8890	0.2441	-0.9848	0.8619	0.6054
10	1	0.1954	0.1819	7.4582	73.8329	0.8653	0.2066
10	2	0.3516	0.3321	5.8926	41.1960	0.8644	0.3179
10	3	0.4765	0.4526	5.2829	27.3714	0.8650	0.2437
10	5	0.6560	0.6352	3.2751	10.6292	0.8631	0.4620
10	10	0.8671	0.8624	0.5435	0.0000	0.8642	0.3399

Note: Predictions are made on the basis of varying numbers of sampling points in wells 2 and 3 (columns 1 and 2) of the Ser 103 side chain torsional angle surface. In the absence of the CSD, the expected and observed prediction accuracies (columns 3 and 4) vary widely, depending on the number of sampling points in either well. An energy CSD (column 6) is then calculated on the basis of the known energy surface and an effective sampling size of $p_2 = p_3 = 10$ (10 sampling points in either well). When this term is included, the prediction accuracies are all within of 0.67% points of the observed accuracy in the $p_2 = p_3 = 10$ case. Key: p_2, p_3 — number of sampling points in wells 2 and 3, respectively. Expected correct — fraction of predictions expected to be correct, based on cross-correlation

calculations (without use of CSD). Observed correct — fraction of predictions observed to be correct (without CSD). % Error — the fractional error between the expected and observed results. CSD — the calculated type I CSD for each row's (p_2, p_3) values, relative to $(p_2, p_3) = (10, 10)$ (calculated using Eq. (26) in main text). Observed cor w/CSD — the fraction of predictions using the calculated CSD that are observed to be correct. Results for the last column (% error) are based on the expected result for $p_2 = 10, p_3 = 10$ (0.8671).

Table 2

Results of CheY side chain prediction studies.

# Trial	Rotamers	# Conf	Min	# Correct
1	lovell	538	0	136
2	lovell	538	a1,a2,2deg	132
3	lovell	538	a1,a2,1deg	131
4	lovell	538	a1,a2,i1,2deg	131
5	lovell	538	sd 25	112
6	lovell	538	sd 25, cfix	104
7	lovell2	4500	0	142
8	lovell2	4500	sd 100	102
9	20deg	29602	0	145
10	15deg	51100	0	145
11	10deg	111592	0	148
12	10deg	111592	sd 5	102
13	10deg	111592	sd 10	88
14	10deg	111592	sd 5, cfix	136
15	10deg	111592	sd 25, cfix	130
16	10deg	111592	cg 5, cfix	133
17	10deg	111592	cg 25, cfix	118
18	5deg	433036	0	146
19	2deg	2657080	0	146

Note: Key: Trial — the trial number. Rotamers — the type of dihedral angle library used. lovell = original Lovell library. lovell2 = expanded Lovell library. 10 deg = a systematic 10-degree grid search. 5 deg = a 5-degree search, and so on. # Conf — the number of conformations in the library. Min — the method used for minimizing the surface. sd = steepest descent; cg = conjugate gradient; cfix indicates that everything but the side chain was fixed during the SD or CG minimization. a1,a2 indicates the C–CA–CB angle and the CA–CB–CG angles were searched systematically on a grid. a1,a2,i1 indicates that, in addition, the N–C–CA–CB improper angle was searched on a grid. #Correct = the number of correct dihedral angles out of 181 total χ_1 and χ_2 angles in the 100 rotatable side chains (includes polar and nonpolar, buried and exposed). The criterion for correct prediction was $\pm 40^\circ$ from the native value.

Table 3

Extreme examples of the grid-spacing dependence of the minimum sampled energy in a conformational search.

S (°)	A		B	
	RMSD (Å)	MSE (kcal/mol)	RMSD (Å)	MSE (kcal/mol)
3.000	3.637	772.36	2.162	195629.36
2.000	2.758	322.22	1.586	5654.44
1.000	1.287	123.48	0.639	323.82
0.500	0.587	55.65	0.301	14.12
0.250	0.370	39.39	0.189	-4.71
0.125	0.234	2.43	0.184	-4.31

Note: Two series of local conformational searches were performed in the protein CheY, varying the relative orientation of the 10 secondary structural elements around 36 backbone dihedral angles. The searches were similar to those performed in Ref. 19, but here the searches were very local to the native position and repeated for various grid spacings. A — the searches were performed varying two secondary structural elements at a time. B — varying all 10 secondary structural elements at a time (but with fewer conformations per element). S — the grid spacing. RMSD — all-atom root mean square deviation of the MSE structure from the native structure. MSE — minimum sampled energy; here, relative to the energy of the native structure.