Beyond histograms: efficiently estimating radial distribution functions via spectral Monte Carlo

Paul N. Patrone[1, a)] and Thomas W. Rosch[1, b)]
National Institute of Standards and Technology
100 Bureau Drive, Gaithersburg MD 20899

(Dated: 3 January 2017)

Despite more than 40 years of research in condensed-matter physics, *state-of-the-art* approaches for simulating the radial distribution function (RDF) $g(r)$ still rely on binning pair-separations into a histogram. Such methods suffer from undesirable properties, including subjectivity, high uncertainty, and slow rates of convergence. Moreover, such problems go undetected by the metrics often used to assess RDFs. To address these issues, we propose (I) a spectral Monte Carlo (SMC) quadrature method that yields $g(r)$ as an analytical series expansion; and (II) a Sobolev norm that assesses the quality of RDFs by quantifying their fluctuations. Using the latter, we show that, relative to histogram-based approaches, SMC reduces by orders of magnitude both the noise in $g(r)$ and the number of pair separations needed for acceptable convergence. Moreover, SMC reduces subjectivity and yields simple, differentiable formulas for the RDF, which are useful for tasks such as coarse-grained force-field calibration via iterative Boltzmann inversion.

Keywords: Radial distribution function, Monte Carlo methods, simulation, iterative Boltzmann inversion

I. INTRODUCTION

In simulations of condensed matter systems, one can barely overstate the importance of the radial distribution function (RDF) $g(r)$. To name only a few applications, $g(r)$ is used to (i) link thermodynamic properties to microscopic details;[1–3] (ii) compute structure factors for comparison with X-ray diffraction;[4,5] and more recently, (iii) calibrate interparticle forces for coarse-grained (CG) molecular dynamics (MD).[6–11] Indeed, the RDF is such a key property that in the past few years, much work has been devoted to estimating $g(r)$ via parallel processing on GPUs.[12] Given these observations, it is thus surprising that state-of-the-art techniques still construct $g(r)$ by binning simulated pair-separations into histograms, with little thought given to developing more efficient methods.[3,13]

In this article, we address this issue by proposing a spectral Monte Carlo (SMC) method for computing simulated RDFs. The key idea behind our approach is to express $g(r)$ in an appropriate basis set and determine the mode coefficients via Monte Carlo quadrature estimates. Relative to binning, we show that this approach decreases subjectivity of the analysis, thereby reducing both the noise in $g(r)$ and the number of pair separations needed to generate useful RDFs. To support these claims, we also discuss how traditional L^2 (or sum-of-squares) metrics are insufficient for assessing convergence of $g(r)$ and propose a Sobolev norm[14] as an appropriate alternative.

The motivation for this work stems from the fact that $g(r)$ is increasingly being used in settings in which the details of its functional form play a critical role. For example, scientists now routinely simulate untested materials in an effort to tailor their structural properties without the need for expensive experiments;[15,16] in such applications, *objectively* computing RDFs is a key task. Along related lines, structural properties are increasingly being used to calibrate coarse-grained force-fields.[6–11] In iterative Boltzmann inversion (IBI) for MD, for example, this is achieved by updating the ith correction to the CG forces $F(r)$ and energies $U(r)$ via

$$U_{i+1}(r) = U_i(r) + k_BT \ln\left[g_i(r)/g_t(r)\right] \quad (1)$$

$$F_i(r) = -\nabla U_i(r), \qquad g_i(r) = g_i(r, \mathbb{S}[F_i]), \quad (2)$$

where k_BT is the temperature, $U_0(r) = -k_BT \ln\left[g_t(r)\right]$ for a target RDF g_t, and $g_i(r)$ is computed from a CG MD simulation $\mathbb{S}$ that uses $F_i(r)$ as the CG force.[6–11] Ultimately, the success of this strategy relies on being able to differentiate $g(r)$, which requires that simulated RDFs be accurate and relatively noise-free.

In this light, we therefore emphasize that histogram-based RDFs suffer from an inability to objectively control uncertainties. This arises for several reasons. For one, histogram bin-sizes are *subjective* parameters that limit the resolution of small-scale features, and often one must trade this resolution for reduced noise. Smoothing is sometimes used as an alternative to increasing bin-sizes, but this introduces difficult-to-quantify uncertainties that depend on the choice of method. Moreover, finite differences and/or derivatives are known to amplify noise, which renders tasks such as CG force-field calibration more difficult. Given that (i) a corresponding experimental RDF may be unavailable for comparison, and (ii) simulation resources are often at a premium, histogram-based approaches therefore place undue burden on modelers to obtain accurate results.

Conceptually, SMC overcomes these limitations by generalizing the notion of a histogram bin to include or-

[a)] Electronic mail: paul.patrone@nist.gov
[b)] Currently staff at the Johns Hopkins Applied Physics Laboratory. Email: thomas.rosch@jhuapl.edu

thogonal functions, such as cosines and Legendre polynomials. As a result, the corresponding RDFs are virtually guaranteed to be well-behaved because the underlying functions are smooth. Moreover, this approach allows us to invoke known results that (i) establish conditions for the convergence of an orthogonal expansion to $g(r)$, and (ii) estimate decay rates of the associated spectral coefficients. Armed with this, we can ultimately assess the accuracy of our reconstructions by determining when noise (associated with the finite system size) inhibits our ability to determine these coefficients, thereby providing an objective measure of the quality of an RDF.

In order to better place our method in the context of the computational physics literature, we note that "Monte Carlo quadrature" (our main analytical tool) refers to a method of evaluating integrals, not a simulation method *per se*. In more detail, SMC defines the relevant spectral coefficients in terms of integrals that cannot be evaluated analytically. However, these integrals can be well-approximated via random evaluations of their integrands, which is essentially the definition of Monte Carlo quadrature. *In our case, the points at which we perform these evaluations are given to us by the simulation, irrespective of its underlying algorithm.* Therefore, the method we propose can be suitably adapted to Monte Carlo *simulations*, etc., although our main focus here is MD.

While the concept of "Monte Carlo quadrature" may be unfamiliar to some readers, we point out, however, that the method is common in computational science. One has only to note that expectation values (which are often integrals) of continuous physical quantities are typically estimated by averaging a fixed number of random realizations thereof. What is more, this can be done without recourse to histograms; e.g. volumes need not be binned before computing an average density in an NPT simulation. In this light, SMC is therefore a natural extension of methods already employed within the community. To further strengthen this point, we bring the discussion full circle by showing that histogram-based approaches are actually a specific form of SMC that use indicator functions as the spectral basis; see Sec. IV.

Because a central theme of this paper is to compare SMC and histogram-based methods, the reader should be aware that we do not seek to "optimize" the latter in an effort to sharpen the discussion. While this may be a minor deficiency of our presentation, it is worth noting that numerous authors in the statistics community have proposed different criteria for what constitutes an "optimal" bin size, with little consensus among them; see, e.g. Refs. 17–20 and references contained therein. Alone, this observation reinforces one of our central themes, namely that histograms are subjective. What is more, many of these approaches make strong assumptions that are not always valid for RDFs. Given the vastness of this literature, we therefore feel that an in-depth review obscures our overall message. As a compromise, we have chosen examples that we feel explore the realities of histograms while being faithful to the benefits of SMC.

The rest of the manuscript is organized as follows. In Sec. II, we formulate SMC and discuss practical issues related to its implementation. Section III compares a variety of SMC calculations with histogram-based counterparts, highlights the relative efficiency of the former, and motivates the usefulness of Sobolev norms for comparing RDFs. Section IV discusses the deeper connections between SMC and histogram-based approaches and considers potential complications that sometimes arise in practice.

Finally, upon request, we are willing to provide sample Matlab scripts that show how to use SMC and perform associated computations on benchmark systems.[21,22]

II. SPECTRAL MONTE CARLO

A. Formulation

Given that typical RDFs are smooth functions, we propose to express $g(r)$ via the expansion

$$g(r) \approx g_M(r) = \sum_{j=0}^{M} a_j \phi_j(r) \tag{3}$$

where $\phi_j(r)$ are orthogonal basis functions on the domain $[0, r_c]$, r_c is a cutoff radius beyond which we do not model $g(r)$, a_j are coefficients to-be-determined, and M is a mode cutoff. Formally, the a_j are determined by invoking the orthogonality relationship of $\phi_j(r)$, viz.

$$\int_0^{r_c} \mathrm{d}r \; \phi_j(r)\phi_k(r) = \delta_{j,k}, \tag{4}$$

where $\delta_{j,k}$ is the Kronecker delta. This yields

$$a_j = \int_0^{r_c} \mathrm{d}r \; \phi_j(r) g(r) = \int_0^{r_c} \mathrm{d}r \; \phi_j(r) \frac{N(r)}{4\pi r^2 \rho}, \tag{5}$$

where ρ is the bulk number density and $N(r)\mathrm{d}r$ is the expected number of particles in a spherical shell with radius r, thickness $\mathrm{d}r$, and a particle at the origin.

In practice, Eq. (5) cannot be evaluated analytically, since $N(r)$ is unknown. However, MD simulations yield random pair-separations[23] distributed according to $N(r)\mathrm{d}r$. Thus, we replace Eq. (5) by its Monte Carlo quadrature estimate[24]

$$a_j \approx \bar{a}_j = \frac{\mathcal{N}(r_c)}{n_{\text{pairs}}} \sum_{k=1}^{n_{\text{pairs}}} \frac{\phi_j(r_k)}{4\pi r_k^2 \rho}, \tag{6}$$

where $\mathcal{N}(r_c)$ is the expected number of particles in a sphere of radius r_c (given a particle at the origin), r_k is the kth pair separation, and n_{pairs} is the total number of such separations. Conceptually, Eq. (6) can be understood by noting that $(1/n)\sum_{k=1}^{n} \Phi(r_k)$ is the standard formula for the sample average of the function

$\Phi(r_k) = \phi_j(r_k)/(4\pi r_k^2\rho)$ given n simulated outputs r_k. Thus, the prefactor $\mathcal{N}(r_c)$ simply adjusts for the fact that the normalization of $g(r)$ is not one, but instead depends on the number of particles in a sphere of radius r_c.

In order to simplify Eq. (6), note that $n_{\text{pairs}} = n_c n_{\text{ppc}}$, where n_c is the number of MD configurations (i.e. timesteps or "snapshots") used to compute $g(r)$, and n_{ppc} is the number of pairs-per-configuration. The latter is well approximated by

$$n_{\text{ppc}} \approx \mathcal{N}(r_c)\mathcal{N}_{\text{tot}}/2, \tag{7}$$

when $\mathcal{N}_{\text{tot}}$ (the number of particles per configuration) and r_c are large. This identity arises as follows. First, the total number of pair separations is $\binom{\mathcal{N}_{\text{tot}}}{2} \approx \mathcal{N}_{\text{tot}}^2/2$ when $\mathcal{N}_{\text{tot}} \to \infty$. Only considering pairs separated by $r \leq r_c$, we reduce the total number of pairs by a factor of $\mathcal{N}(r_c)/\mathcal{N}_{\text{tot}}$. We require r_c to be large enough so that the relative fluctuations in $\mathcal{N}(r_c)$ are small.

Given this, we next substitute Eq. (7) into Eq. (6) to find[25]

$$\bar{a}_j = \frac{2}{\mathcal{N}_{\text{tot}} n_c} \sum_{k=1}^{n_{\text{pairs}}} \frac{\phi_j(r_k)}{4\pi r_k^2 \rho}, \tag{8}$$

which is the desired estimate of the spectral coefficients. As opposed to histogram-based approaches, this expression provides more objective control over uncertainties in simulated RDFs. Specifically, for many choices of $\phi_j(r)$, the mode coefficients decay as $|a_j| < Cj^{-p}$ (or even e^{-pj}), where the constant C and rate p depend on the smoothness of $g(r)$.[26] Furthermore, for such bases, $g_M(r)$ converges to $g(r)$ uniformly in M.[27] This implies that in principle, the maximum error in $g_M(r)$ is controlled through M. However, Monte Carlo sampling also introduces uncertainty in a_j, which can be estimated via

$$\sigma_j^2 = \frac{4}{(\mathcal{N}_{\text{tot}} n_c)^2} \sum_k \left[\bar{a}_j - \phi_j(r_k)/4\pi r_k^2\rho\right]^2. \tag{9}$$

This suggests that the largest meaningful mode cutoff $M^\star$ can be estimated from $|a_M^\star| = \mathcal{O}(\sigma_M^\star)$, which corresponds to the noise-floor of $\bar{a}_j$ (cf. Sec. II B for a practical algorithm to estimate $M^\star$). Given the uniform convergence of Eq. (3), we then conclude that: (i) the error in $g_M(r)$ is the greater of either $\mathcal{O}(\sigma_M)$ or $\mathcal{O}(a_M)$ for any cutoff; and (ii) $g_M(r)$ can model all features whose characteristic size is greater than r_c/M.

B. Practical considerations about the basis and mode cutoff

Generally speaking, the task of choosing a suitable basis is straightforward. It is well known, for example, that if $g(r)$ is twice differentiable and $g'(0) = g'(r_c) = 0$ (which should approximately hold if r_c is large enough), then $\phi_j(r) = \sqrt{2/r_c}\cos(j\pi r/r_c)$ converges uniformly and yields a series whose derivative converges to $g'(r)$.[28]

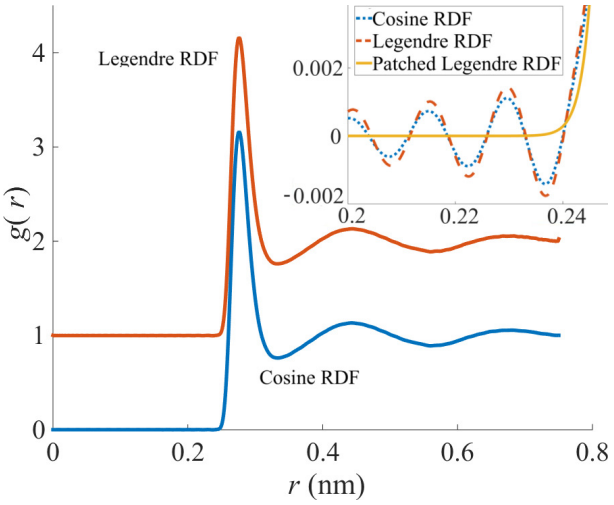

FIG. 1. Two RDFs computed using a cosine basis and Legendre polynomial basis. The latter has been shifted vertically by 1 since it would otherwise obscure the cosine reconstruction. Differences between the reconstructions are on the order of 10^{-3}. The inset shows the behavior of the SMC reconstructions with and without the patching function described in Eq. (10).

Moreover, $a_j \leq \mathcal{O}(j^{-2})$, although exponential convergence is expected when $g(r)$ is infinitely differentiable (cf. Fig. 2).[26] Orthogonal polynomials (e.g. Legendre or Chebyshev) are also reasonable choices, as they provide uniform approximations and similar rates of convergence, irrespective of boundary conditions.[26] For RDFs whose slopes do not vanish as $r \to 0$ or $r \to r_c$ (e.g. due to long-range correlations), such bases may behave better than trigonometric functions; see also the discussion on patchy particles below.

To illustrate these properties, we tested SMC on a 1000-frame simulation of 1000 TIP4P water molecules in LAMMPS.[29–31] To generate the underlying data, we first equilibrated the system for 0.1 ns at 300 K using an NPT simulation. Then, we ran a 1 ns NVT simulation (Nosé-Hoover thermostat) on the final equilibrated system, which had a density of $\rho = 0.996$ g/cm^3. Configurations were output every 1 ps. Figure 1 shows two separate O-O RDFs computed using 101 cosine modes and 145 Legendre polynomials (see below for discussion of choosing M). The latter is shifted up since the curves would be indistinguishable if superimposed. Notably, both RDFs are smooth and generally well behaved.

Although not visible in the main figure, the inset shows that the reconstructions oscillates slightly about zero as $r \to 0$, which is typical of spectral expansions. In all of our applications (e.g. IBI), we have found that this behavior is not problematic for practical computations and is reduced by simultaneously using more data and larger mode cutoffs. Moreover, if necessary one can locally replace the spectral expansion with an exponential

(or similar) patching function of the form

$$\mathfrak{g} = b\exp(-c/r^p), \qquad r < \tilde{r} \tag{10}$$

where $\tilde{r}$ is a user-defined separation to the right of the last negative value of $g_M(r)$. The free parameters b, c, and p can be determined by matching $\mathfrak{g}(\tilde{r})$, $\mathfrak{g}'(\tilde{r})$, and $\mathfrak{g}''(\tilde{r})$ to $g_M(\tilde{r})$, $g'_M(\tilde{r})$, and $g''_M(\tilde{r})$, respectively, which yields

$$\begin{aligned} p &= \tilde{r}\left[\frac{g'_M(\tilde{r})}{g_M(\tilde{r})} - \frac{g''_M(\tilde{r})}{g'_M(\tilde{r})}\right] - 1 \\ c &= \frac{g'_M(\tilde{r})\tilde{r}^{p+1}}{pg_M(\tilde{r})} \\ b &= g_M(\tilde{r})\exp[c/\tilde{r}^p]. \end{aligned}$$

We typically choose $\tilde{r}$ to be the smallest value $\hat{r}$ for which $g_M(\hat{r}) = 0.02$ and $g_M(r) \geq g_M(\hat{r})$ when $r > \hat{r}$.

Figure 2 illustrates an automated method for estimating the largest meaningful mode cutoff $M^\star$. We first define a function of the form

$$\alpha(j) = be^{-pj\Theta(j-c)-pc\Theta(c-j)}, \tag{11}$$

where b, p, and c are positive free parameters and $\Theta(x)$ is the Heaviside step function. Conceptually, $\alpha(j)$ is motivated by the observation that for an infinitely differentiable $g(r)$, the mode weights go as $|a_j| \sim e^{-pj}$ (for some power $p > 0$) until they hit the noise floor, at which point they should be approximately constant. Thus, fitting the logarithm of Eq. (11) (via, e.g. least-squares) to $\log|a_j|$ yields an estimate c of the first mode at which $|a_j|$ no longer decays exponentially. This is illustrated in Fig. 2 at the point where the two lines intersect. Rounding the corresponding c to the nearest integer thereby produces an estimate $M^\star$.

Figure 3 shows an example of how SMC can reconstruct RDFs with discontinuities due to, e.g. excluded volume effects. The underlying system is described in Ref. 32. Interestingly, each particle is a hard sphere (radius $r = 1/2$, dimensionless) with 5 attractive or "sticky" patches; thus, in addition to the discontinuity associated with the excluded volume, we also expect correlations to be large for values of r immediately to the right of the discontinuity. To illustrate this, we analyze a trajectory with 800 particles and 500 frames and a (cubic) box-length of 10 (dimensionless) units.[33] In order to correctly account for the discontinuity, we only decompose $g(r)$ on the interval $1 \leq r \leq 5$ and use a Legendre basis (which we map from $[-1, 1]$ to $[1, 5]$). For $r < 1$ we set $g(r) = 0$. Notably, the SMC reconstruction is able to successfully predict a sharp transition in the correlations. For comparison purposes, we also show histogram reconstructions having a bin widths of 0.1, 0.025, and 0.001 in Fig. 4.

When compared to Figs. 1 and 2, the inset to Fig. 3 reveals an interesting feature of the shifted Legendre representation. Notably, the RDFs for both water and the patchy system have roughly the same number of oscillations, whereas far fewer modes are required for the latter. Heuristically, we can understand this observation by appealing to the rule of thumb that a mode $\phi_j(r)$ has j oscillations. Thus, if the smallest characteristic length scale of interest (associated with a feature of the RDF) has a length $\ell < 1$ (taking $r_c = 1$ non-dimensional), then at a minimum we need $1/\ell$ modes to adequately reconstruct

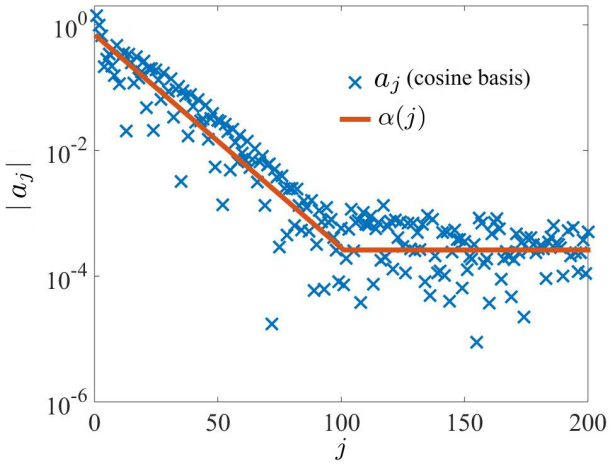

FIG. 2. A method for estimating the largest meaningful mode cutoff $M^\star$ by finding the noise floor. Here we fit $\log[\alpha(j)]$ to $\log|a_j|$ to estimate the mode c at which the weights no longer obey the power law discussed above Eq. (9). For the cosine modes applied to 1000 frames of a 1000-molecule TIP4P water simulation, we find that only 101 modes are required to hit the noise floor.

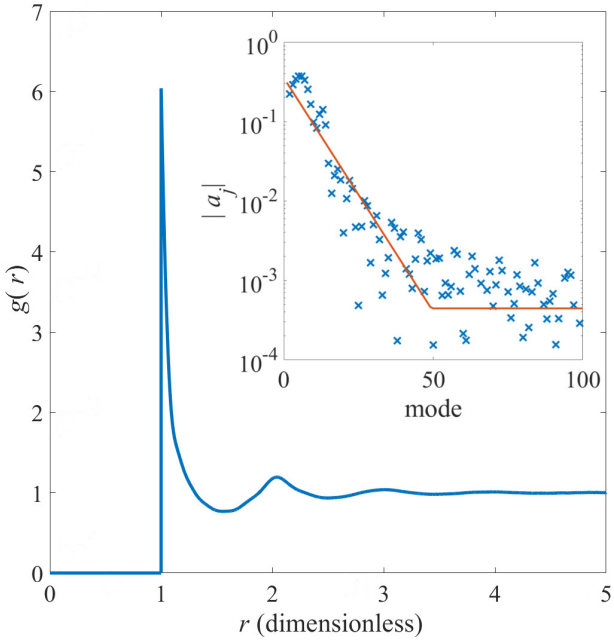

FIG. 3. SMC radial distribution function for patchy, hard-sphere particles discussed in Ref. 32. See main text for details of the spectral reconstruction.

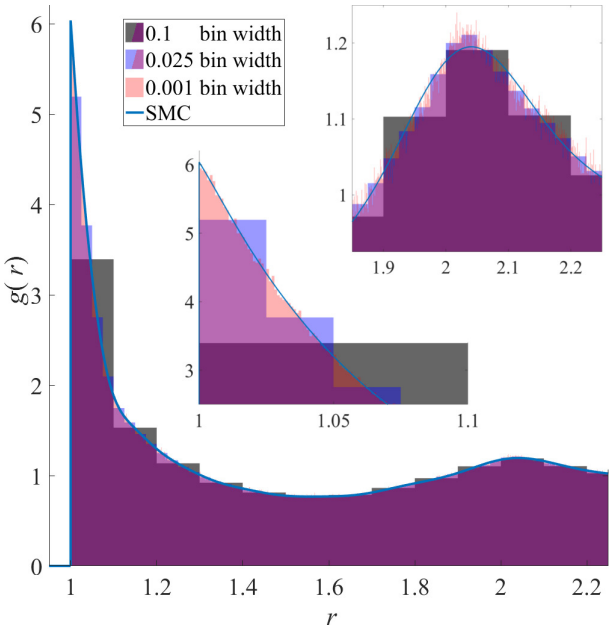

FIG. 4. Comparison of SMC versus three different histogram reconstructions for the RDF of the patchy system. The insets further resolve the first and second peaks. The 0.1 and 0.025 bin-width histograms under-resolve the RDF relative to SMC, especially at the peak, whereas the 0.001 bin-width RDF is noisy.

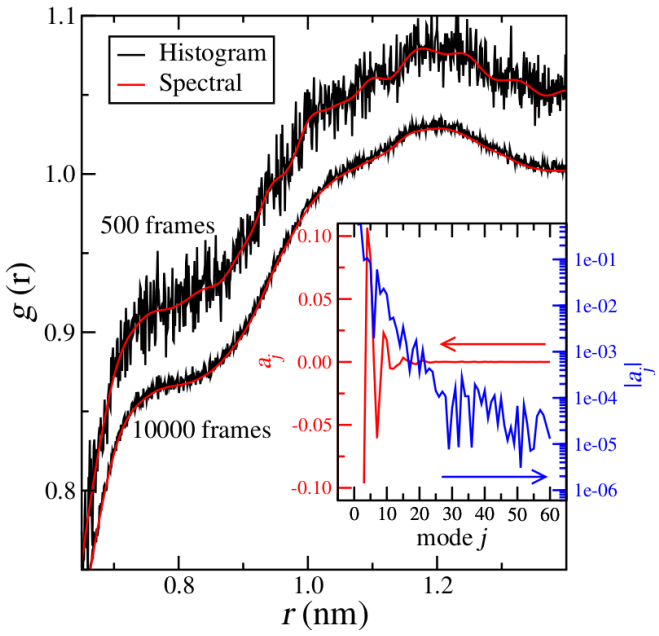

FIG. 5. RDF of atomistic polystyrene (PS) in CG coordinates using the histogram method (black, rough curves) and SMC (red, smooth curves). The upper and lower pairs are calculated with $n_c = 500$ (shifted up by 0.05) and $n_c = 10^4$ snapshots. The inset displays the spectral coefficients a_j (left scale) and $\log|a_j|$ (right scale) for the first 60 modes.

that feature. In this light, large sub-domains where $g(r) \approx 0$ (i.e. for small r) followed by a sharp rise require many modes to accurately reconstruct, since the characteristic length-scale of the transition is small. However, by limiting the spectral reconstruction to the domain $[1, 5]$, the analysis of Fig. 3 eliminates the first and most rapid transition due to the discontinuity, thereby allowing us to decrease M. While we do not pursue this observation further, it nonetheless suggests that shifted spectral representations may be useful for addressing the oscillatory behavior in the inset of Fig. 1.

III. MOTIVATING SOBOLEV NORMS THROUGH SAMPLE COMPUTATIONS

Figure 4 illustrates an issue that first motivated SMC; without *a priori* knowledge of the system at hand, it may be difficult to accurately reproduce an RDF using histograms. Furthermore, the figure makes it obvious that increasing the bin width trades uncertainty along the vertical axis for uncertainty along the horizontal axis, which may be unacceptable in practical settings. To further explore this issue, we consider a few benchmark systems and sample computations that are often performed with RDFs. Using this discussion, we highlight another key theme of our work: traditional sum-of-squares methods for comparing and assessing RDFs may be inadequate in practical settings.

To begin, we compute the RDFs of a CG molecular dynamics polystyrene (PS) model run in LAMMPS. We first run a 10 ns, atomistic NVT simulation of amorphous, atactic PS (10 chains of 50 monomers) interacting through the pcff forcefield at 800 K and $\rho = 0.758$, with configurations output every 1 ps.[34] This trajectory is then mapped into CG coordinates at a resolution of 1 CG bead per monomer (located at the center of mass), so that $\mathcal{N}_{\text{tot}} = 500$. Each bead is roughly 0.5 nm in diameter. Next, we calculate CG RDFs via (i) a histogram with 1400 bins, each having a width of 1 pm on the interval $0 \le r \le 1.4$ nm, and (ii) SMC with a cosine basis. Figure 5 shows the results of these computations for $n_c = 500$ and $n_c = 10^4$. The benefits of the spectral approach are readily apparent, especially when $n_c = 500$. For $n_c = 10^4$, noise in the histogram method decreases by roughly a factor of 4 or 5 (as expected from the central limit theorem), but SMC is still dramatically smoother.

To further illustrate the smoothness of $g_M(r)$, we use iterative Boltzmann inversion [cf. Eq. (2)] to calibrate CG MD forces for PS using first the histogram method and then SMC. For the former, we used a central finite-difference scheme to approximate the F_i on a grid with a 1 pm resolution.[35] For SMC, we took $M = 60$ and computed all forces F_i analytically. More specifically, note that when Eq. (3) is differentiable [and more importantly, when the derivative converges to the derivative of

$g(r)]$, one can rewrite Eqs. (2) as

$$U_{i+1}(r) = U_i(r) + k_B T \ln\left[\frac{\sum_j a_{i,j}\phi_j(r)}{g_t(r)}\right] \quad (12)$$

$$F_{i+1}(r) = F_i(r) + F_0(r) - k_B T \frac{\sum_j a_{i,j}\phi'_j(r)}{\sum_j a_{i,j}\phi_j(r)}, \quad (13)$$

where $a_{i,j}$ is the jth mode for the ith IBI update.[36] Here we took the respective $n_c = 10^4$ RDFs in Fig. 5 as the target g_t; that is, we do not mix binning and SMC computations. Figure 6 shows the results of these computations. Notably, the top subplot shows that after five iterations of IBI, the histogram-based force has extreme, high-frequency noise (despite taking $n_c = 10^4$), whereas the SMC force does not.

To make this comparison more quantitative, we define

$$||g||^2_{L^2} = \frac{1}{r_c}\int_0^{r_c} \mathrm{d}r\; g(r)^2 \approx \sum_{j=1}^{n_{\mathrm{bins}}} \frac{g_j^2 \Delta r_j}{r_c} \quad (14)$$

where the sum is used for the histogram reconstructions, g_j is the RDF evaluated in the jth bin, n_{bins} is the number of bins, and Δr_j is the width of the jth bin. Many works invoke $||g - g_t||_{L^2}$ (or variants thereof) to assess when a given RDF is sufficiently converged to g_t.[6,37,38] However, Fig. 6 shows that both the histogram and SMC RDFs converge in L^2 to their respective g_t at about the same rate, suggesting that this norm is not strongly affected by high-frequency fluctuations. Moreover, $||g - g_t||_{L^2}$ does not assess the extent to which the force $F(r)$ converges when using algorithms such as IBI. To account for such effects, we instead invoke a Sobolev norm[14]

$$||g||^2_{H^1} = ||g||^2_{L^2} + ||g'(r)||^2_{L^2}, \quad (15)$$

where we approximate $g'(r) \approx (g_{j+1} - g_{j-1})/(r_{j+1} - r_{j-1})$ for the histogram reconstructions (r_j are the bin centers). Physically, the second term of Eq. (15) assesses how smoothly $g \to g_t$; equivalently, $||g - g_t||^2_{H^1}$ determines when CG forces (as opposed to just energies) are converging in Eq. (2). This extra information reveals a stark difference between the histogram and SMC reconstructions insofar as the former does not improve in an H^1 sense (i.e. the IBI forces never converge). Moreover, given that the H^1 and L^2 norms of the SMC reconstruction quickly overlap, it is clear that the difference with the H^1 norm of the histogram reconstruction is due to its high-frequency content.[39]

To test the robustness of SMC and compare with smoothing techniques, we also used 120 cosine modes to construct the O-O $g(r)$ for a 5000-molecule TIP4P water simulation; cf. Fig. 7. After 0.6 ns of equilibration, we ran a 0.2 ns NVT production run and output configurations every 1 ps ($n_c = 200$). We take the corresponding 120-mode SMC reconstruction as a baseline for comparison, given its known convergence properties. For histogram-based approaches, we first partitioned the

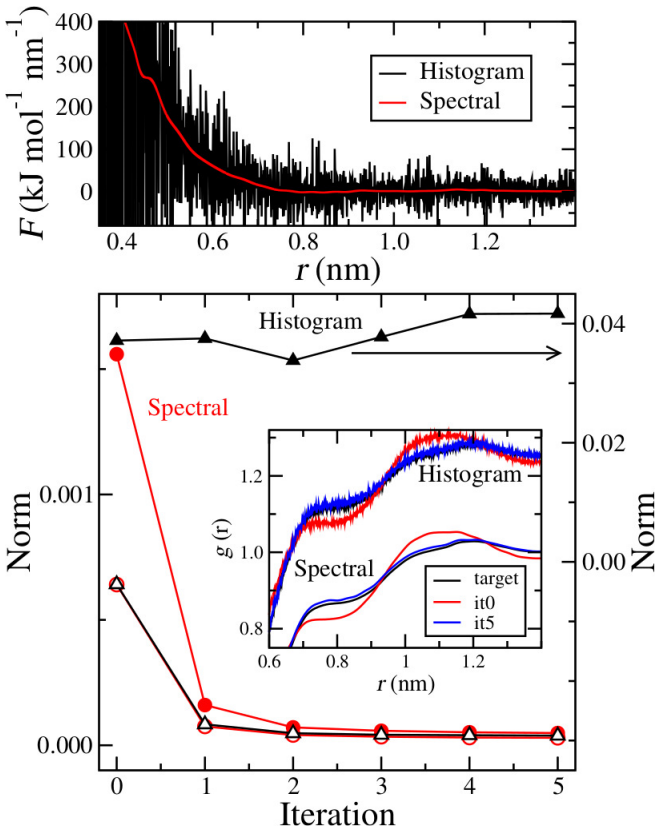

FIG. 6. *Top:* CG force for PS calculated via IBI with $n_c = 10^4$. The black curve (rough) is the histogram method result, whereas the red (smooth) curve is the SMC result. *Bottom:* $||g_i - g_t||^2_{L^2}$ (open symbols) and $||g_i - g_t||^2_{H^1}$ (closed symbols) norms for the histogram (triangle) and SMC (circle) methods as a function of IBI iteration. Note that $||g - g_t||^2_{H^1}$ for the histogram method uses the right axis and is off the scale of the left axis. The inset shows the corresponding RDFs. The black curve is the target RDF; it0 (red) and it5 (blue) denote the initial CG RDF and the RDF after 5 IBI iterations.

domain $0 \le r \le 0.9$ nm into 1800 intervals. After binning pair-separations from the first 20 frames, we used two separate smoothing algorithms to reduce noise: (i) a n-point moving mean with $n = 5$ and $n = 15$; and (ii) a Gaussian-kernel that convolves the histogram with $K(x) = \exp[-0.5(x/h)^2]$ for $h = 1$ pm and $h = 5$ pm. Figure 7 illustrates the key problem tied to the subjectivity of such methods: too little smoothing yields noisy RDFs (bottom inset), whereas too much washes out relevant features (top inset).

This figure also suggest that as a function of n_c, SMC converges to $g(r)$ more quickly than histogram-based approaches. To quantitatively test this, we estimated $g(r)$ for CG PS (cf. Figs. 5 and 6) as a function of n_c and computed the corresponding L^2 and and H^1 norms relative to the $n_c = 10^4$ case (which now acts as g_t). Figure 8 shows the results of this exercise. Most apparent, every norm decays as roughly $1/n_c$. Intuitively we expect this from the central limit theorem, since the variance in

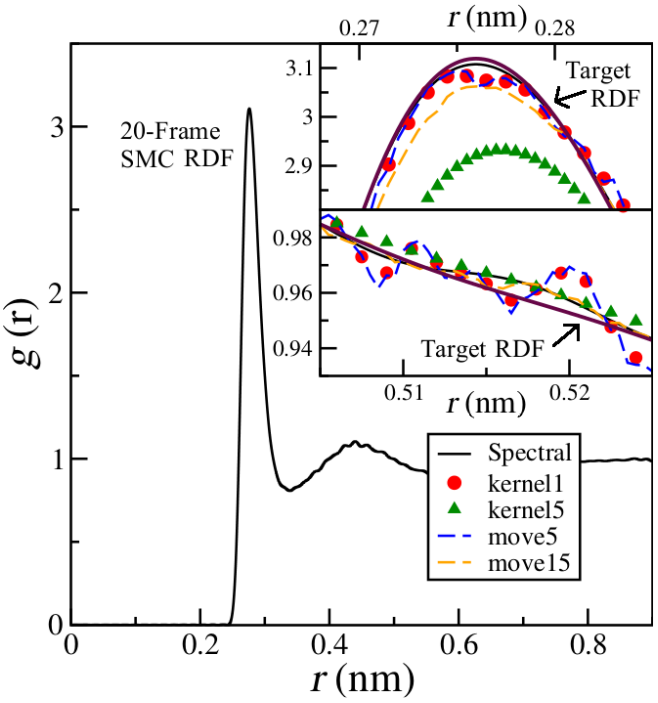

FIG. 7. Comparison of RDFs constructed using 20 frames of a 5000-molecule water simulation. The main figure shows that SMC captures both the sharp peak and the rapid transition around $r = 0.25$ nm. The insets compares the 20-frame SMC, kernel-smoothed, and moving-average RDFs relative to a 200-frame SMC RDF (dark purple). See main text for discussion.

an average of N independent, identically distributed random variables should decay as the inverse of N. However, the SMC norms (circles) are at least an order of magnitude or more smaller than their histogram counterparts (squares and triangles). This suggests that the overhead required to generate pair separations can be reduced by a factor of 10 or more simply by using SMC.

IV. CONCLUDING DISCUSSION

A. Limitations of SMC

As discussed previously, the inset of Fig. 1 illustrates a problem that can arise when spectral expansions attempt to represent functions that behave like typical RDFs at small r, e.g. as $\exp(-1/r^p)$ for $p > 0$. Thus, SMC may not be useful for representing the small-separation behavior of $g(r)$. Such problems, however, are not necessarily unique to SMC. In more detail, the condition $g(r) = \mathcal{O}[\exp(-1/r^p)]$ as $r \to 0$ implies that the probability of finding two particles separated by a small distance is a exceedingly small, and therefore a rare event. Thus, particle-based methods that return finite collections of pair separations do not provide enough data to adequately sample such regions, irrespective of the method for estimating $g(r)$.

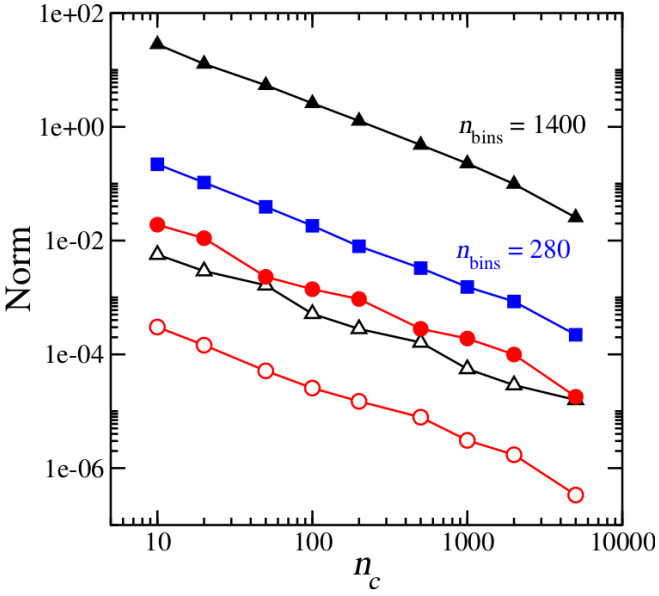

FIG. 8. $||g - g_t||^2_{L^2}$ and $||g - g_t||^2_{H^1}$ as a function of n_c for the 5th IBI update to the PS model in Fig. 6. Here g_t is the $n_c = 10^4$ RDF. Squares denote histogram estimates computed using 280 bins. All other symbols correspond to 1400 bins and have the same meanings as in previous figures.

B. Limitations of the H^1 norm

Figure 8 shows that increasing the histogram bin-width leads to *seemingly* smoother reconstructions of $g(r)$. This arises from the fact that more data points contribute to any given bin, thereby decreasing fluctuations. However, this *does not necessarily improve* the accuracy of such reconstructions, since bin counts are then averages taken over increasingly large domains. Thus, the H^1 norm we propose should be used with caution, since it is likely not a valid assessment of histograms when the number of bins becomes too small. Along similar lines, we do not pursue quantitative comparison with convergence rates of smoothed histograms; such an analysis would require quantification of the uncertainties induced by smoothing, which can be highly non-trivial to estimate.

C. Connection between SMC and histograms

Analytically, the connection between SMC and histograms can be understood by framing the latter in the context of Eq. (8). Specifically, Eq. (8) reduces to a histogram bin count when the $\phi_j(r)$ are indicator functions $I_{[r_j, r_{j+1}]}$, i.e. constants on an interval (i.e. bin) $[r_j, r_{j+1}]$ and zero otherwise. These observations suggest that the ϕ_j act as a generalized histogram "bin." The fact that $\phi_j(r_k)$ may be non-zero for multiple j indicates that each pair separation r_k contributes to multiple "bins," albeit in unequal amounts. Stated differently, SMC bins data according to the characteristic wavelengths with which

the r_k fall on the domain $[0, r_c]$.

D. Typical computation times and associated benefits

The aforementioned connection between SMC and histograms suggests that both methods should be comparable in terms of computation times, which we generally find to be true. The 1000-molecule system discussed in Figs. 1 and 2 provides a good benchmark with $M = 200$. Using a highly vectorized code (which is available upon request) on a 16-core machine, we can serially analyze 1000 frames in less than five minutes with Legendre polynomials and about 75 seconds using cosines. Given that the total number of pair separations is $\binom{1000}{2} \times 1000 \approx 500,000,000$, the script handles between 100 million and 400 million pair separations per minute. If we parallelize over frames, the total computation time drops for both methods drops to roughly 45 seconds, or 750 million pair separations per minute. It is also likely that typical computation times can be further reduced, since SMC can be parallized on GPUs. Given that we never needed more than 145 modes to faithfully reconstruct an RDF, it is reasonable to assume that SMC is competitive with binning in terms of computational cost-per-pair-separation.[40]

It is also worth noting that cost-per-pair-separation may underestimate the true savings of SMC. Figure 8 shows that in order to reach the same level of convergence (in either L^2 or H^1) SMC requires roughly 1/10 to 1/100 as many frames as the corresponding binned RDFs. Considering that atomistic MD simulations may take days to run, the real savings in our approach may come from needing fewer frames (and thus a shorter simulation) to generate an acceptably RDF.[41]

Finally, we note that that this computational speed may nonetheless require large amounts of RAM. For a typical 1000-molecular simulation, we frequently require on the order of several GB of free memory to construct relevant matrices of pair separations and mode-weights. On a 15,625 molecule system, we have seen that 30 GB or more RAM may be required to run a fully vectorized version of our SMC code. In the sample scripts that are available upon request, we attempt to indicate potential memory pitfalls and suggest methods for overcoming them. Moreover, we emphasize that these issues are specific to our programming style and can easily be circumvented through appropriate modifications. Thus, we do not feel that this issue poses a serious problem to the adoption of SMC.

E. Extensions

It is straightforward to generalize SMC to arbitrary distribution functions. For example, assume that x is a continuous random variable with mean zero and unit variance.[42] Denoting its probability density as $P(x)$, we can express this quantity in terms of a spectral expansion analogous to Eq. (3). Choosing the $\phi_j(x)$ to be normalized Hermite functions (not Hermite polynomials!), for example, would be suitable for densities whose shapes somewhat resemble Gaussian distributions. Noting that realizations of x are drawn from $P(x)$ directly, this yields

$$a_j = \int \mathrm{d}x \; \phi_j(x) P(x) \approx \frac{1}{N} \sum_{i=1}^{N} \phi_j(x_i), \qquad (16)$$

where the x_i are the N realizations of the random variable x. Related ideas and extensions are currently being written in another manuscript.

Acknowledgments: the authors thank Timothy Burns, Andrew Dienstfrey, and Vincent Shen for useful feedback during preparation of this manuscript. We also thank Debra Audus for sharing data related to the patchy particle system in Ref. 32. *This work is a contribution of the National Institute of Standards and Technology and is not subject to copyright in the United States.*

[1] J. G. Kirkwood and F. P. Buff, The Journal of Chemical Physics **19**, 774 (1951).
[2] K. E. Newman, Chem. Soc. Rev. **23**, 31 (1994).
[3] M. Allen and D. Tildesley, *Computer simulation of liquids*, Oxford science publications (Clarendon Press, 1987).
[4] J. Yarnell, M. Katz, R. Wenzel, and S. Koenig, Physical Review A **7**, 2130 (1973).
[5] N. Ashcroft and N. Mermin, *Solid State Physics*, HRW international editions (Holt, Rinehart and Winston, 1976).
[6] C.-C. Fu, P. M. Kulkarni, M. Scott Shell, and L. Gary Leal, The Journal of Chemical Physics **137**, 164106 (2012).
[7] C. Peter and K. Kremer, Soft Matter **5**, 4357 (2009).
[8] B. Bayramoglu and R. Faller, Macromolecules **45**, 9205 (2012).
[9] F. Muller-Plathe, ChemPhysChem **3**, 754 (2002).
[10] W. G. Noid, Journal of Chemical Physics **139** (2013).
[11] D. Reith, M. Putz, and F. Muller-Plathe, Journal of Computational Chemistry **24**, 1624 (2003).
[12] B. G. Levine, J. E. Stone, and A. Kohlmeyer, J. Comput. Phys. **230**, 3556 (2011).
[13] D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications*, Computational science series (Elsevier Science, 2001).
[14] L. Evans, *Partial Differential Equations*, Graduate studies in mathematics (American Mathematical Society, 2010).
[15] P. Biswas, D. N. Tafen, F. Inam, B. Cai, and D. A. Drabold, Journal of Physics: Condensed Matter **21**, 084207 (2009).
[16] H. Khandelia, A. A. Langham, and Y. N. Kaznessis, Biochimica et Biophysica Acta (BBA) - Biomembranes **1758**, 1224 (2006).
[17] D. W. Scott, Biometrika **66**, 605 (1979).
[18] D. Freedman and P. Diaconis, Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete **57**, 453 (1981).
[19] H. A. Sturges, Journal of the American Statistical Association **21**, 65 (1926).
[20] D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization* (John Wiley, 1992).
[21] MATLAB, *version 8.1.0 (R2013a)* (The MathWorks Inc., Natick, Massachusetts, 2013).
[22] Certain commercial software is identified in this paper in order to specify the computational procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.
[23] Here we only specify MD because that is the method used in all of our examples. However, the analysis is agnostic to how the

pair-separations are generated. Thus, any method that outputs this type of data can be analyzed with SMC.

[24] C. P. Robert and G. Casella, "Introducing monte carlo methods with r," (Springer New York, New York, NY, 2010) Chap. Monte Carlo Integration, pp. 61–88.

[25] Interestingly, related methods have been developed for density-of-state calculations under the name kernel polynomial method[43].

[26] J. Boyd, *Chebyshev and Fourier Spectral Methods: Second Revised Edition*, Dover Books on Mathematics (Dover Publications, 2001).

[27] Uniform convergence of $g_M(r)$ to $g(r)$ means that for any ϵ, there is an M such that $|g_M(r) - g(r)| < \epsilon$ holds for all r. Moreover, if $g(r)$ has p derivatives, often $|a_j| \leq \mathcal{O}(j^{-p})$; if $g(r)$ has infinitely many derivatives, the $|a_j|$ usually decay exponentially[26].

[28] W. Strauss, *Partial Differential Equations: An Introduction* (Wiley, 1992).

[29] W. Jorgensen, J. Chandrasekhar, J. Madura, R. Impey, and M. Klein, Journal of Chemical Physics **79**, 926 (1983).

[30] H. W. Horn, W. C. Swope, J. W. Pitera, J. D. Madura, T. J. Dick, G. L. Hura, and T. Head-Gordon, The Journal of Chemical Physics **120**, 9665 (2004).

[31] S. Plimpton, Journal of Computational Physics **117**, 1 (1995).

[32] D. J. Audus, F. W. Starr, and J. F. Douglas, The Journal of Chemical Physics **144** (2016).

[33] The trajectory we analyzed corresponds to the parameters $T = 0.22$, $\rho = 0.8$, and $\epsilon_i = 0$ in Ref. 32.

[34] H. Sun, S. Mumby, J. Maple, and A. Hagler, Journal of the American Chemical Society **116**, 2978 (1994).

[35] For completeness, we upload this histogram-based computation as a tabulated forcefield into LAMMPS, which then interpolates the forces using a cubic spline on a finer grid with a spacing of 0.05 pm, or equivalently, 28,000 points.

[36] Again for completeness, we evaluated the SMC Eq. (13) on a fine grid with spacing of 0.05 pm (or equivalently, 28,000 points) and uploaded this directly to LAMMPS as a tabulated forcefield.

[37] T. C. Moore, C. R. Iacovella, and C. McCabe, Journal of Chemical Physics **140** (2014).

[38] R. Faller, Polymer **45**, 3869 (2004).

[39] Admittedly, we have not attempted to "optimize" the histogram approach in any way. However, unless the bin-size is so large as to under-resolve the RDF, it is likely that the associated reconstructions will always exhibit some level of fluctuations; see, for example, Fig. 4. Given that finite-differencing only amplifies noise, our conclusion – non-smooth, histogram-based RDFs yield excessively noisy forces – should generally hold independent of the bin-size.

[40] The difference in times between Legendre and cosines reconstructions is likely due to the way that these functions are internally computed within the script. We have not made significant attempts to understand these differences.

[41] Note that this consideration also applies to IBI when the target RDF is computed from an atomistic simulation.

[42] The mean-zero and unit-variance assumptions are not restrictive. Given N realizations x_i of an arbitrary random variable x, we can always rescale by the sample mean and sample variance to yield a random variable that approximately satisfies these assumptions. This step is often important when applying SMC to probability densities. The rescaling can always be undone after $P(x)$ is computed for the normalized quantity.

[43] L. Lin, Y. Saad, and C. Yang, SIAM Review **58**, 34 (2016).