THE JOURNAL OF PHYSICAL CHEMISTRY Letters

pubs.acs.org/JPCL Letter

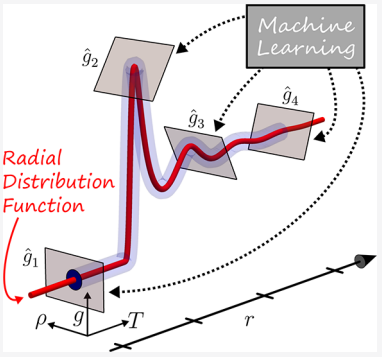
Ex Machina Determination of Structural Correlation Functions

Galen T. Craven,* Nicholas Lubbers, Kipton Barros, and Sergei Tretiak

Cite This: *J. Phys. Chem. Lett.* 2020, 11, 4372−4378 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Determining the structural properties of condensed-phase systems is a fundamental problem in theoretical statistical mechanics. Here we present a machine learning method that is able to predict structural correlation functions with significantly improved accuracy in comparison with traditional approaches. The usefulness of this *ex machina* (from the machine) approach is illustrated by predicting the radial distribution functions of two paradigmatic condensed-phase systems, a Lennard-Jones fluid and a hard-sphere fluid, and then comparing those results to the results obtained using both integral equation methods and empirically motivated analytical functions. We find that application of the developed *ex machina* method typically decreases the predictive error by more than an order of magnitude in comparison with traditional theoretical methods.



Atomistic structural correlations in condensed-phase systems are often determined by complex multibody interactions that give rise to rich collective behaviors at the macroscale. As such, predicting how a system's microscopic structure leads to its macroscopic properties is one of the primary challenges in equilibrium statistical mechanics.[1−5] Solving this problem is paramount in multiple fields and disciplines, as knowledge of the mapping between the microscale and macroscale can lead to enhanced system designs and to the implementation of new system functionalities.[6−13] Because of its importance, theoretical determination of structural correlation functions has been an ongoing research focus for nearly a century.[1,14−20] There are three primary approaches for predicting structural correlations: (a) applying integral equation methods,[19−24] (b) fitting data to empirically motivated functional forms,[25−32] and (c) estimating the correlation functions directly in molecular simulations.[33−38] Integral equation methods often give accurate predictions for the properties of fluids. Currently, however, there is not a significantly robust theoretical framework for these methods that is void of both numerical complexity and the need for *ad hoc* manipulation for applications to complex molecular systems.[19] Fitting data to empirically motivated functional forms can result in simple and accurate predictive models, but the applicability of these functions is limited.[26,27] Moreover, in many systems, applying these two methods is intractable or results in inaccurate predictions, and molecular simulation approaches must be employed. This is often problematic because simulations that include atomistic-level descriptions of the system can incur significant computational costs in order to make simple predictions.

Machine learning (ML) methods have shown significant promise for generating faster and/or improved solutions to a number of problems in physics and chemistry,[39−47] albeit in typically limited application windows.[48] Here we develop an ML method that can be used to predict structural correlation functions in condensed-phase systems with increased accuracy in comparison with traditional analytical approaches. This ML process is trained using a limited set of molecular dynamics (MD) simulation data. After training, the process allows structural correlation functions to be predicted with high accuracy over a broad range of system parameters at a negligible computational cost. We illustrate the power of this method by predicting the pair correlation functions of two historically significant models from statistical mechanics, a Lennard-Jones system and a hard-sphere fluid, and then comparing the results predicted by ML to results generated by various other theoretical methods. Compared with traditional theoretical methods that fit to fixed functional forms,[26−29] our *ex machina* (from the machine) approach can enable significant error reduction, typically an order of magnitude or greater, when trained on comparable amounts of MD data (more specifically, comparable sampling *densities* within the thermodynamic feature space). This magnitude of error reduction is also observed in comparison with integral equation methods.

The developed ML method can also reduce the computational time needed to generate structural correlation data in comparison with using purely simulation-based approaches. Specifically, this method will be computationally advantageous when (a) knowledge of the atomistic structure is needed over

Received: February 26, 2020
Accepted: May 5, 2020
Published: May 5, 2020

large regions (or at a high density of state points) in the thermodynamic feature space,[46,47,49,50] (b) large batches of structural data and the thermodynamic properties derived from those data are needed in order to fit simple empirical functions for analytical manipulation,[26−29,51−54] and/or (c) thermodynamic properties must be generated quickly in order to efficiently connect with continuum codes.[55−57] As an example, we estimate that the present ML method can generate the atomistic structural data obtained in ref 46 using MD at ∼1/25th of the computational cost. The computational advantage of the given ML approach should be particularly significant when *ab initio* molecular dynamics simulations are needed in order to generate accurate predictions for structural properties.[58,59]

Our goal is to train an ML process to predict a particular structural correlation function G. Common examples of G are the radial distribution function $g(r)$ (see Figure 1) and the triplet correlation function. Formally, we want to use ML to construct the mapping $[\boldsymbol{x}, \boldsymbol{f}] \rightarrow G(\boldsymbol{x}, \boldsymbol{f})$, where $\boldsymbol{x} = \{x_1, x_2, ..., x_S\}$ is a set of S spatial variables and $\boldsymbol{f} = \{f_1, f_2, ..., f_N\}$ is set of N thermodynamic features. Typical spatial variables in $\boldsymbol{x}$ are distances and angles between particles or atoms. Typical features in $\boldsymbol{f}$ are density, pressure, and temperature.

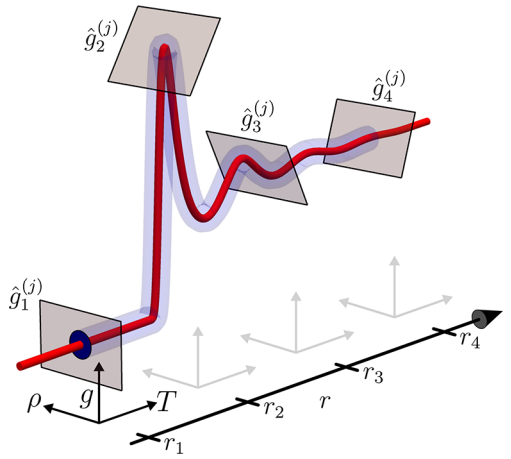


Figure 1. Schematic representation of the application of the developed ML method to predict a radial distribution function g (red curve) that depends on distance r, density ρ, and temperature T. In the body of this Letter, we present a generalized procedure to predict structural correlation functions that could depend on additional variables beyond ρ and T. Each gray plane represents the linear approximation $\hat{g}_k^{(j)}$ to g at a particular $r = r_k$. The index j denotes a small region of $\rho \times T$ space over which the local model $\hat{g}_k^{(j)}$ (see eq 1) interpolates from MD data. The smooth blue tube suggests that interpolation between positions $[r_1, r_2, ...]$ would also be possible.

The ML process that we apply is a segmented linear regression with multivariate function decomposition approach. A full review of ML concepts can be found in ref 60. A schematic representation of this procedure is shown in Figure 1. The data used to train this process consist of inputs of the form $[\boldsymbol{x}_k, \boldsymbol{f}^{(s)}]$, which are mapped to targets $G(\boldsymbol{x}_k, \boldsymbol{f}^{(s)})$, where the subscript k denotes a set of specific values for $\boldsymbol{x}$ and the superscript (s) denotes a specific set of values for $\boldsymbol{f}$. These training data will typically be the results of molecular simulations and/or experimental measurements. The algorithm for the ML approach is as follows: First, the feature space $F = f_1 \times f_2 \times \cdots \times f_N$ is partitioned into n feature subspaces F_j. Second, at each $\boldsymbol{x}_k$, a collection of local least-squares regression processes are trained in these subspaces, where each process in the collection is applied to generate the desired structural correlation function G in a specific subspace. Finally, this collection of processes are combined into a single program that can be used to generate G over the total feature space and over each $\boldsymbol{x}_k$.

To generate an output from this procedure, the ML program takes a feature space input point $p = \boldsymbol{f}^{(p)}$ (this is the point at which G is to be predicted), decides which points from the training data to use in the regression process on the basis of the region of the feature space in which p lies, and then generates the function $G(\boldsymbol{x}, \boldsymbol{f}^{(p)})$. One way of selecting the training data used in each regression process is to employ proximity-based approaches, that is, to train the process using data points that are near the input point p. If the training data are generated on a grid in feature space and $p \in F_j$, one particular choice is to use the points from the training data that form the 2^N vertices of F_j—this is the method used here.

The ML process is trained by assuming that in each subspace the dependence of G on each feature in $\boldsymbol{f}$ is locally linear at each $\boldsymbol{x}_k$ value. That is, in the neighborhood about the input point p, with the spatial variables held constant, we assume that the feature-space dependence of the true correlation function can be well-approximated by a hyperplane. In this approximation, the correlation function can be constructed at a specific $\boldsymbol{x}_k$ as

$$G_j(\boldsymbol{x}_k, \boldsymbol{f}) \approx \hat{G}_k^{(j)}(\boldsymbol{x}_k, \boldsymbol{f}) = a_0^{(j)}(\boldsymbol{x}_k) + \sum_{i=1}^{N} a_i^{(j)}(\boldsymbol{x}_k) f_i \tag{1}$$

where $\hat{G}_k^{(j)}$ is a function that is linear in the elements of $\boldsymbol{f}$ (see Figure 1) and $a_i^{(j)}(\boldsymbol{x}_k)$ is the spatially dependent coefficient of feature $f_i \in \boldsymbol{f}$ evaluated at $\boldsymbol{x}_k$, whose value is determined from least-squares regression. This procedure produces G_j at discrete $\boldsymbol{x}_k$ points. A function that is continuous in $\boldsymbol{x}$ can be constructed in each subspace F_j by interpolating between $\boldsymbol{x}_k$ points, leading to

$$G_j(\boldsymbol{x}, \boldsymbol{f}) \approx \hat{G}_k^{(j)}(\boldsymbol{x}, \boldsymbol{f}) \tag{2}$$

The total correlation function G over the entire feature space F can then be constructed as

$$G(\boldsymbol{x}, \boldsymbol{f}) \approx \mathcal{S}[G_1(\boldsymbol{x}, \boldsymbol{f}), G_2(\boldsymbol{x}, \boldsymbol{f}), ..., G_n(\boldsymbol{x}, \boldsymbol{f})] \tag{3}$$

where $\mathcal{S}$ is a functional (a selector) that takes the set $\boldsymbol{G} = \{G_1, G_2, ..., G_n\}$ as an argument and returns the function G_j such that the input point p is an element of F_j. For example, if $p \in F_1$ then $\mathcal{S} = G_1$, if $p \in F_2$ then $\mathcal{S} = G_2$, and so forth. The most expensive computational element of this method will typically be generating the training data. After the ML process is trained—a negligible computational expense—it approximates structural correlation functions at significantly reduced computational cost with respect to molecular simulations.

One of the most important and well-studied structural correlation functions is the pair correlation function $g(r)$, i.e., the radial distribution function (RDF). The RDF plays a significant role in the development of theories of the liquid state because it can be used to directly calculate thermodynamic observables in systems that are dominated by pairwise interactions. To express the RDF in the notation developed above, $G = g$ and $\boldsymbol{x} = \{r\}$ where r is the distance between particles. Because of its importance, the remainder of this

https://dx.doi.org/10.1021/acs.jpclett.0c00627
J. Phys. Chem. Lett. 2020, 11, 4372−4378

Letter focuses on applying the present ML approach to determine the RDFs of Lennard-Jones (LJ) and hard-sphere (HS) fluids, two paradigmatic and historically significant condensed-phase models. Specifically, we generate RDF training data for both of these systems using MD simulations and then use that data to train the ML process described above. Full details of the MD simulation methods used to generate the training data are provided in the Supporting Information.

The pertinent thermodynamic features of a Lennard-Jones system in the canonical ensemble are density ρ and temperature T (both given dimensionless LJ units[36]), and therefore, in this case $\boldsymbol{f} = (\rho, T)$. To generate training data for the LJ system, we partitioned the $\rho \times T$ plane (the feature space) using a grid and then generated RDF data by performing MD at each grid point. The accuracy of the model and the cost to collect the MD data set are determined by the spacing between grid points; we use $\Delta\rho = 0.05$ and $\Delta T = 0.2$ in the respective dimensions. A detailed analysis of how varying the grid spacing affects the accuracy of the ML procedure is given in the Supporting Information. The feature subspaces F_j correspond to quadrilateral grid cells, and the data used to train the local regression process in each subspace are taken from the four grid points at the vertices of F_j in the $\rho \times T$ plane.

A comparison between the LJ RDFs predicted using ML and measured using MD simulations is shown in Figure 2 for various state points in the $\rho \times T$ plane. The curves, from top to bottom, are the results for state points $\rho = 1.06$, $T = 2.47$; $\rho = 0.625$, $T = 1.71$; and $\rho = 0.255$, $T = 2.87$. In all cases, there is excellent agreement between the predicted and measured RDFs. In fact, at the given level of visual fidelity, the two functions are indistinguishable. This illustrates the effectiveness of using the presented ML method to generate structural correlation functions. The temperature of each of these state points is above the LJ critical temperature,[61−63] and each point lies in the supercritical fluid region of the phase diagram. We have also confirmed that a similar level of agreement is observed between the ML and MD RDFs in other regions of the phase diagram.

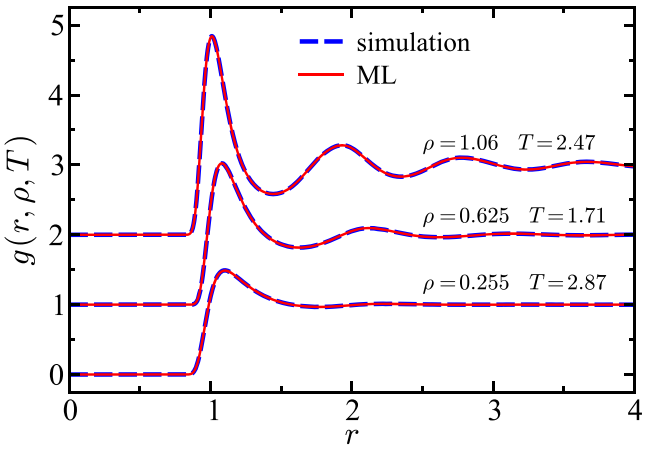


Figure 2. Radial distribution function of an LJ system measured using MD (dashed blue) and predicted using the ML method developed in this Letter (red). Results are shown for the state points labeled in the plot. Some of the plots have been vertically shifted for visual clarity. All quantities are expressed in LJ units.

To quantify the error between between the RDF predicted theoretically, g_{theory}, and the RDF measured in simulation, g_{MD}, and to make comparisons to other theoretical methods, we computed the error

$$\Delta g(r) = g_{\text{theory}}(r) - g_{\text{MD}}(r) \tag{4}$$

for the ML method developed here and two other theoretical approaches discussed below.[26,27] Figure 3 shows the results for $\Delta g(r)$ in the LJ system at the same state points used in Figure 2. Each subplot corresponds to the results from a different state point. The three colored curves in each subplot correspond to the results generated by ML and by applying the expressions of Goldman[26] and Morsali et al.[27]—two of the most prominent analytical functions used to generate the LJ RDF. These expressions are constructed by fitting LJ simulation data to empirically motivated functions. At each state point, the ML method generates significantly lower error. This is especially pronounced at $r \approx 1$ where the empirical functions generate significant errors. It is noteworthy that the errors generated through application of the two analytical functions are manifested as quasi-oscillations (aperiodic systematic divergences) whereas the error from ML takes a noiselike form.

For a particular grid spacing, the accuracy of the ML procedure is not significantly affected by the distance to the nearest training point in a feature space grid cell. This implies that the data from the training points is weighted in a way that

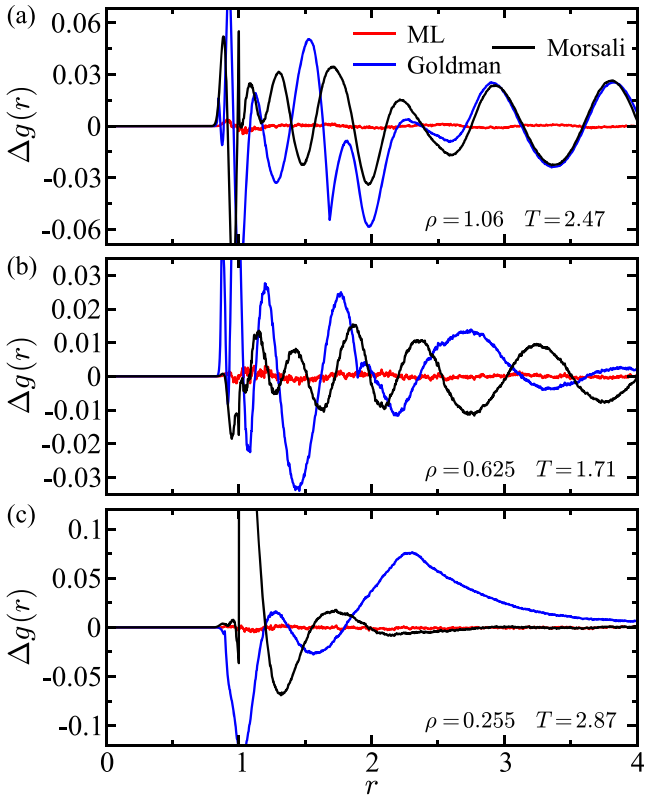


Figure 3. Difference between the predicted and measured RDF, $\Delta g(r)$, of a LJ fluid as a function r for the state points labeled in the individual subplots. Each subplot contains the results for three theoretical methods used to predict the RDF: the ML method developed in this manuscript (red), the expression by Morsali et al. (black), and the Goldman expression (blue). All quantities are expressed in LJ units.

https://dx.doi.org/10.1021/acs.jpclett.0c00627
J. Phys. Chem. Lett. 2020, 11, 4372−4378

generates uniform RDF error over the grid cell area. The Supporting Information contains a detailed analysis of how variation in the grid spacing affects the accuracy of the predicted RDF.

In order to quantify how the error generated by each empirical function compares with that for the ML approach, we computed the mean absolute error,

$$\mathrm{MAE} \equiv \langle |\Delta g(r)| \rangle \tag{5}$$

for each method over the interval $0 \le r \le 4$ at each state point shown in Figure 3 and then calculated the ratios of the MAEs. The ratios extracted from the data shown in Figure 3a for $\rho = 1.06$, $T = 2.47$ are $\mathrm{MAE_{Mor}}/\mathrm{MAE_{ML}} \approx \mathrm{MAE_{Gold}}/\mathrm{MAE_{ML}} \approx 25$. This state point is in a regime with significant structural correlations, as quantified by the magnitudes of the primary and secondary peaks of the RDF shown in the top curve of Figure 2. In this highly structured regime, ML reduces the error by greater than an order of magnitude in comparison with the other methods. For comparison, using the RDF data from the nearest point in the training set ($\rho = 1.05$, $T = 2.4$) to approximate the RDF at this state point produces an error ratio $\mathrm{MAE_{nearest}}/\mathrm{MAE_{ML}} \approx 10$, illustrating that the accuracy of the method arises from the ML scheme and not from the density of the training data. The MAE ratios extracted from the data shown in Figure 3b for state point $\rho = 0.625$, $T = 1.71$ are $\mathrm{MAE_{Mor}}/\mathrm{MAE_{ML}} \approx 15$ and $\mathrm{MAE_{Gold}}/\mathrm{MAE_{ML}} \approx 20$. For the low-density state point $\rho = 0.255$, $T = 2.87$ shown in Figure 3c, the error ratios are $\mathrm{MAE_{Mor}}/\mathrm{MAE_{ML}} \approx 30$ and $\mathrm{MAE_{Gold}}/\mathrm{MAE_{ML}} \approx 40$.

We also examined the performance of the ML method and how the training data grid spacing affects this performance using a test set consisting of MD results from ∼200 randomly sampled state points in the region $\mathcal{R} = [0.05, 1.1] \times [1.2, 5.0]$ in $\rho \times T$ space. Applying the ML method to this test set using few as ∼50 training points to cover $\mathcal{R}$ generates over an order of magnitude decrease in the average MAE compared with the Morsali and Goldman expressions. Also, when only nine training points are used, the ML method produces an MAE that is lower than that produced by the other theoretical methods. Full details of these calculations can be found in the Supporting Information. In general, we find that ML typically decreases the predicted RDF error for an LJ fluid by greater than an order of magnitude in comparison with the empirical functions that are developed from human-guided fitting procedures.

Finite-size effects in the training data limit the range of interpolation in the spatial variables. These effects, and the errors that arise from them, may be significant in systems where the training data are generated using simulations of a small number of particles. A major advantage of applying the present ML approach is that only a limited number of training points are needed in order to generate accurate predictions for the RDF over a large area in the thermodynamic feature space. Therefore, the computational expense to mitigate these size effects is small because large systems can be simulated at a sparse number of training points. The Supporting Information contains a detailed analysis of how variation in the grid spacing affects the accuracy of the predicted RDF. Moreover, extending the interpolative domain of the ML method in feature space can also be accomplished using a limited number of additional training points.

The hard-sphere system is a well-studied condensed-phase model that qualitatively, and in some cases quantitatively, describes the interactions and collective behaviors in isotropic fluids through a simple pairwise potential that is amenable to analytical examination. HS fluids are athermal, and therefore, in monodisperse HS systems the only pertinent feature is the volume fraction η. We generated RDF training data using MD simulations of the HS system at different η values with an equidistant spacing of $\Delta\eta = 0.01$ between points. The Supporting Information contains an analysis of how the accuracy of the ML procedure varies with the grid spacing and full details of the MD simulation methods used to generate the training data. The lower and upper bounds for η in the training data were $\eta = 0.1$ and $\eta = 0.46$, respectively, with the upper bound being below the fluid−solid phase transition value ($\eta_f \approx 0.492$).[64] In the ML procedure, the training data used in the local regression process in each feature subspace are taken from two nearest-neighbor points on the η line.

A comparison between the predicted and measured RDFs for the HS system is shown in Figure 4 as a function of r (given in units of σ) for volume fractions $\eta = 0.455$, $\eta = 0.311$, and $\eta = 0.1271$. Again, as for the LJ fluid, the functions predicted using ML are indistinguishable from those measured by simulation at the given level of visual fidelity. These results reiterate the effectiveness of using ML to predict structural correlation functions and also support the robust applicability of the present ML method.

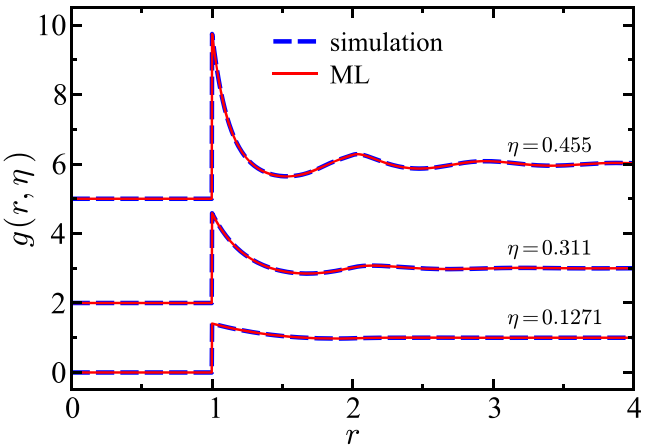


Figure 4. Radial distribution function of an HS fluid measured using MD (dashed blue) and predicted using ML (red) as a function of r (given in units of σ). Results are shown for the volume fractions labeled in the plot. Some of the plots have been vertically shifted for visual clarity.

Two prominent expressions to predict the HS RDF are the solution to the Percus−Yevick (PY) integral equation[21,22] and the Trokhymchuk−Nezbeda−Jirsák−Henderson (TNJH) expression,[32] of which the former is a purely analytical approach and the latter is a hybrid method that combines information from several sources into an empirically motivated physics-informed function. The PY solution is often corrected using the modification proposed by Verlet and Weis.[65−68] Here, we call this method the PYVW solution. Shown in Figure 5 are the results for $\Delta g(r)$ in the HS system using the ML, PY, PYVW, and TNJH methods. Each subplot corresponds to the results for a different volume fraction. In all panels, it can be observed that the PY solution systemically underestimates the value of the RDF at contact, $g(\sigma^+)$ while the ML prediction for $g(\sigma^+)$ is in excellent agreement with the MD result. The ML method

https://dx.doi.org/10.1021/acs.jpclett.0c00627
J. Phys. Chem. Lett. 2020, 11, 4372−4378

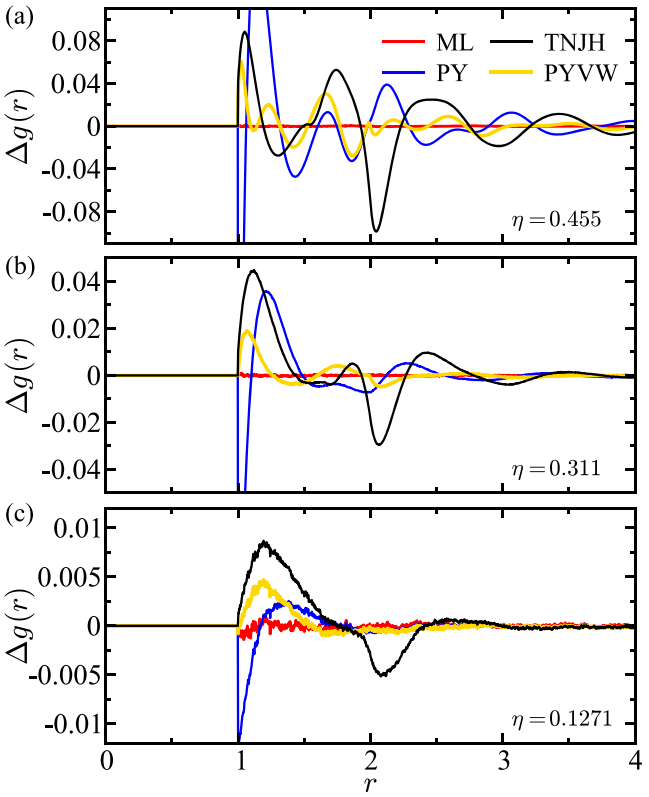


Figure 5. Difference between the predicted and measured RDF, $\Delta g(r)$, of an HS fluid as a function of r (units of σ) for the volume fractions labeled in the individual subplots. Each subplot contains the results for four theoretical methods applied to predict the RDF: the ML method developed in this Letter (red), the TNJH expression (black), the PY solution (blue), and the PYVW solution (yellow).

results in significant error decreases across all volume fractions with respect to the other methods. Also, just as in the LJ case, the error from the traditional methods is quasi-oscillatory while the ML error presents as noise.

In order to quantitatively compare how the RDF error generated by each of the analytical expressions discussed above compares with that for ML, we computed the MAEs generated by the four methods over the interval $1 \leq r \leq 4$ and then calculated the ratios of the MAEs at each volume fraction shown in Figure 5. For $\eta = 0.455$, the error ratios are $MAE_{PY}/MAE_{ML} \approx MAE_{TNJH}/MAE_{ML} \approx 150$ and $MAE_{PYVW}/MAE_{ML} \approx 50$. These correspond to error reductions by a factor greater than 2 orders of magnitude in comparison with PY and TNJH and greater than an order of magnitude in comparison with PYVW when the present ML approach is used. For $\eta = 0.311$, the error ratios are $MAE_{PY}/MAE_{ML} \approx MAE_{TNJH}/MAE_{ML} \approx 50$ and $MAE_{PYVW}/MAE_{ML} \approx 20$. Applying ML therefore results in error reductions of a factor greater than an order of magnitude in comparison with the other examined methods at intermediate HS fluid densities. In the low-density system with $\eta = 0.1271$, the error ratios are smaller ($MAE_{PY}/MAE_{ML} \approx MAE_{PYVW}/MAE_{ML} \approx 4$ and $MAE_{TNJH}/MAE_{ML} \approx 10$), but the error reductions generated using ML are still significant.

We further examined the performance of the ML method using a test set of 100 HS RDFs calculated using MD at random and uniformly distributed volume fractions over the interval [0.1, 0.47]. Using a grid spacing of $\Delta\eta = 0.01$ in the training data, the same as used above, results in more than an order of magnitude decrease in the average MAE of the test set in comparison with the other three theoretical methods. Additionally, using a grid spacing as large as $\Delta\eta = 0.12$ yields an average MAE that is lower than those for the PY and TNJH expressions, and a spacing as large as $\Delta\eta = 0.06$ yields an average MAE lower than that for the PYVW expression. A detailed analysis of these calculations is provided in the Supporting Information. These results illustrate that using ML can decrease the predicted RDF error for an HS fluid by greater than an order of magnitude in comparison with other theoretical approaches.

In conclusion, we have shown that an *ex machina* method can be used to predict structural correlation functions with significantly increased accuracy in comparison with integral equation methods and human-guided fitting procedures. Applying the developed method to determine structural correlations in anisotropic systems, complex molecular fluids, and coarse-grained systems is straightforward and, provided that the training data used in the procedure are correspondingly dense, should lead to errors similar to those generated here. The main advantages of the present approach include the following: it is simple to implement; it provides accurate estimates; and from the standpoint of training the process, the computational cost to apply it is negligible. The main disadvantages are that generating training data using molecular simulations can be computationally expensive and that scaling the method into high-dimensional feature spaces could be problematic, particularly with respect to generating sufficient training data. We posit that these disadvantages can be mitigated by augmenting the present method with active learning and deep learning approaches. The methodology developed in this Letter should serve as motivation for the design and implementation of data-driven approaches to compute structural correlation functions across diverse condensed-phase systems.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpclett.0c00627.

Details of the molecular dynamics simulation methods used to generate the training data and analysis of how variation in the grid spacing affects the accuracy of the predicted RDF (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Galen T. Craven − *Theoretical Division and Center for Nonlinear Studies (CNLS), Los Alamos National Laboratory, Los Alamos, New Mexico 87544, United States;* orcid.org/0000-0001-5117-2345; Email: gcraven@lanl.gov

Authors

Nicholas Lubbers − *Computer, Computational, and Statistical Sciences Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87544, United States*

Kipton Barros − *Theoretical Division and Center for Nonlinear Studies (CNLS), Los Alamos National Laboratory, Los Alamos, New Mexico 87544, United States*

Sergei Tretiak − *Theoretical Division, Center for Nonlinear Studies (CNLS), and Center for Integrated Nanotechnologies (CINT), Los Alamos National Laboratory, Los Alamos, New*

https://dx.doi.org/10.1021/acs.jpclett.0c00627
J. Phys. Chem. Lett. 2020, 11, 4372−4378

Mexico 87544, United States; orcid.org/0000-0001-5547-3647

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpclett.0c00627

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We acknowledge support from the Los Alamos National Laboratory (LANL) Directed Research and Development (LDRD) Funds. This research was performed in part at the Center for Nonlinear Studies (CNLS) and the Center for Integrated Nanotechnologies (CINT) at LANL. The computing resources used to perform this research were provided by the LANL Institutional Computing Program.

■ REFERENCES

(1) Hansen, J. P.; McDonald, I. R. *Theory of Simple Liquids*; Academic Press: San Diego, 1986.
(2) Weeks, J. D.; Chandler, D.; Andersen, H. C. Role of Repulsive Forces in Determining the Equilibrium Structure of Simple Liquids. *J. Chem. Phys.* **1971**, *54*, 5237.
(3) Grünwald, M.; Geissler, P. L. Patterns without Patches: Hierarchical Self-Assembly of Complex Structures from Simple Building Blocks. *ACS Nano* **2014**, *8*, 5891−5897.
(4) Craven, G. T.; Popov, A. V.; Hernandez, R. Structure of a Tractable Stochastic Mimic of Soft Particles. *Soft Matter* **2014**, *10*, 5350−5361.
(5) Das, M.; Green, J. R. Self-Averaging Fluctuations in the Chaoticity of Simple Fluids. *Phys. Rev. Lett.* **2017**, *119*, 115502.
(6) Torquato, S. *Random Heterogenous Materials: Microstructure and Macroscopic Properties*; Springer: New York, 2002.
(7) Barkan, K.; Engel, M.; Lifshitz, R. Controlled Self-Assembly of Periodic and Aperiodic Cluster Crystals. *Phys. Rev. Lett.* **2014**, *113*, 098304.
(8) Craven, G. T.; Popov, A. V.; Hernandez, R. Effective Surface Coverage of Coarse-Grained Soft Matter. *J. Phys. Chem. B* **2014**, *118*, 14092−14102.
(9) Jadrich, R. B.; Bollinger, J. A.; Lindquist, B. A.; Truskett, T. M. Equilibrium Cluster Fluids: Pair Interactions via Inverse Design. *Soft Matter* **2015**, *11*, 9342−9354.
(10) Giri, N.; Del Pópolo, M. G.; Melaugh, G.; Greenaway, R. L.; Rätzke, K.; Koschine, T.; Pison, L.; Gomes, M. F. C.; Cooper, A. I.; James, S. L. Liquids with Permanent Porosity. *Nature* **2015**, *527*, 216−220.
(11) Lindquist, B. A.; Jadrich, R. B.; Truskett, T. M. Assembly of Nothing: Equilibrium Fluids with Designed Structured Porosity. *Soft Matter* **2016**, *12*, 2663−2667.
(12) Gaillac, R.; Pullumbi, P.; Beyer, K. A.; Chapman, K. W.; Keen, D. A.; Bennett, T. D.; Coudert, F.-X. Liquid Metal-Organic Frameworks. *Nat. Mater.* **2017**, *16*, 1149−1154.
(13) Bennett, T. D.; Horike, S. Liquid, Glass and Amorphous Solid States of Coordination Polymers and Metal−Organic Frameworks. *Nat. Rev. Mater.* **2018**, *3*, 431−440.
(14) Kirkwood, J. G. Statistical Mechanics of Fluid Mixtures. *J. Chem. Phys.* **1935**, *3*, 300−313.
(15) Kirkwood, J. G.; Buff, F. P. The Statistical Mechanical Theory of Solutions. I. *J. Chem. Phys.* **1951**, *19*, 774−777.
(16) Yarnell, J. L.; Katz, M. J.; Wenzel, R. G.; Koenig, S. H. Structure Factor and Radial Distribution Function for Liquid Argon at 85K. *Phys. Rev. A: At., Mol., Opt. Phys.* **1973**, *7*, 2130−2144.
(17) Soper, A. The Radial Distribution Functions of Water and Ice from 220 to 673 K and at Pressures up to 400 MPa. *Chem. Phys.* **2000**, *258*, 121−137.
(18) Izvekov, S.; Voth, G. A. Multiscale Coarse Graining of Liquid-State Systems. *J. Chem. Phys.* **2005**, *123*, 134105.
(19) Ratkova, E. L.; Palmer, D. S.; Fedorov, M. V. Solvation Thermodynamics of Organic Molecules by the Molecular Integral Equation Theory: Approaching Chemical Accuracy. *Chem. Rev.* **2015**, *115*, 6312−6356.
(20) Martin, T. B.; Gartner, T. E.; Jones, R. L.; Snyder, C. R.; Jayaraman, A. pyPRISM: A Computational Tool for Liquid-State Theory Calculations of Macromolecular Materials. *Macromolecules* **2018**, *51*, 2906−2922.
(21) Percus, J. K.; Yevick, G. J. Analysis of Classical Statistical Mechanics by Means of Collective Coordinates. *Phys. Rev.* **1958**, *110*, 1−13.
(22) Wertheim, M. Exact Solution of the Percus−Yevick Integral Equation for Hard Spheres. *Phys. Rev. Lett.* **1963**, *10*, 321−323.
(23) Coslovich, D.; Ikeda, A. Cluster and Reentrant Anomalies of Nearly Gaussian Core Particles. *Soft Matter* **2013**, *9*, 6786−6795.
(24) López de Haro, M.; Santos, A.; Yuste, S. B. On the Radial Distribution Function of a Hard-Sphere Fluid. *J. Chem. Phys.* **2006**, *124*, 236102.
(25) Matteoli, E.; Mansoori, G. A. A Simple Expression for Radial Distribution Functions of Pure Fluids and Mixtures. *J. Chem. Phys.* **1995**, *103*, 4672−4677.
(26) Goldman, S. An Explicit Equation for the Radial Distribution Function of a Dense Lennard-Jones Fluid. *J. Phys. Chem.* **1979**, *83*, 3033−3037.
(27) Morsali, A.; Goharshadi, E. K.; Mansoori, G. A.; Abbaspour, M. An Accurate Expression for Radial Distribution Function of the Lennard-Jones Fluid. *Chem. Phys.* **2005**, *310*, 11−15.
(28) Bamdad, M.; Alavi, S.; Najafi, B.; Keshavarzi, E. A New Expression for Radial Distribution Function and Infinite Shear Modulus of Lennard-Jones Fluids. *Chem. Phys.* **2006**, *325*, 554−562.
(29) Emampour, J. S.; Morsali, A.; Sarvghadi, M.; Jafari, G. R.; Beyzaie, N.; Beyramabadi, S. A. The Changes in Free Energies and Entropies from Analytical Radial Distribution Functions. *Phys. Chem. Liq.* **2012**, *50*, 187−198.
(30) Mueller, E. A.; Gubbins, K. E. An Equation of State for Water from a Simplified Intermolecular Potential. *Ind. Eng. Chem. Res.* **1995**, *34*, 3662−3673.
(31) Abbaspour, M.; Akbarzadeh, H.; Abroodi, M. A New and Accurate Expression for the Radial Distribution Function of Confined Lennard-Jones Fluid in Carbon Nanotubes. *RSC Adv.* **2015**, *5*, 95781−95787.
(32) Trokhymchuk, A.; Nezbeda, I.; Jirsák, J.; Henderson, D. Hard-Sphere Radial Distribution Function Again. *J. Chem. Phys.* **2005**, *123*, 024501.
(33) Alder, B.; Frankel, S.; Lewinson, V. Radial Distribution Function Calculated by the Monte-Carlo Method for a Hard Sphere Fluid. *J. Chem. Phys.* **1955**, *23*, 417.
(34) Wood, W. W.; Parker, F. R. Monte Carlo Equation of State of Molecules Interacting with the Lennard-Jones Potential. I. A Supercritical Isotherm at about Twice the Critical Temperature. *J. Chem. Phys.* **1957**, *27*, 720−733.
(35) Alder, B. J.; Wainwright, T. E. Studies in Molecular Dynamics. I. General Method. *J. Chem. Phys.* **1959**, *31*, 459−466.
(36) Frenkel, D.; Smit, B. *Understanding Molecular Simulation: From Algorithms to Application*; Academic Press: New York, 1996.
(37) VandeVondele, J.; Mohamed, F.; Krack, M.; Hutter, J.; Sprik, M.; Parrinello, M. The Influence of Temperature and Density Functional Models in Ab Initio Molecular Dynamics Simulation of Liquid Water. *J. Chem. Phys.* **2005**, *122*, 014515.
(38) Chen, M.; Ko, H.-Y.; Remsing, R. C.; Calegari Andrade, M. F.; Santra, B.; Sun, Z.; Selloni, A.; Car, R.; Klein, M. L.; Perdew, J. P.; et al. Ab Initio Theory and Modeling of Water. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, 10846−10851.
(39) Carrasquilla, J.; Melko, R. G. Machine Learning Phases of Matter. *Nat. Phys.* **2017**, *13*, 431−434.
(40) Carleo, G.; Troyer, M. Solving the Quantum Many-Body Problem with Artificial Neural Networks. *Science* **2017**, *355*, 602−606.

(41) Biamonte, J.; Wittek, P.; Pancotti, N.; Rebentrost, P.; Wiebe, N.; Lloyd, S. Quantum Machine Learning. *Nature* **2017**, *549*, 195−202.
(42) Deng, D.-L.; Li, X.; Das Sarma, S. Quantum Entanglement in Neural Network States. *Phys. Rev. X* **2017**, *7*, 021021.
(43) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: An Extensible Neural Network Potential with DFT Accuracy at Force Field Computational Cost. *Chem. Sci.* **2017**, *8*, 3192−3203.
(44) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine Learning for Molecular and Materials Science. *Nature* **2018**, *559*, 547−555.
(45) Lubbers, N.; Smith, J. S.; Barros, K. Hierarchical Modeling of Molecular Energies using a Deep Neural Network. *J. Chem. Phys.* **2018**, *148*, 241715.
(46) Moradzadeh, A.; Aluru, N. R. Transfer-Learning-Based Coarse-Graining Method for Simple Fluids: Toward Deep Inverse Liquid-State Theory. *J. Phys. Chem. Lett.* **2019**, *10*, 1242−1250.
(47) Moradzadeh, A.; Aluru, N. R. Molecular Dynamics Properties without the Full Trajectory: A Denoising Autoencoder Network for Properties of Simple Liquids. *J. Phys. Chem. Lett.* **2019**, *10*, 7568−7576.
(48) Mitchell, M. Artificial Intelligence Hits the Barrier of Meaning. *Information* **2019**, *10*, 51.
(49) Moore, T. C.; Iacovella, C. R.; McCabe, C. Derivation of Coarse-Grained Potentials via Multistate Iterative Boltzmann Inversion. *J. Chem. Phys.* **2014**, *140*, 224104.
(50) Mashayak, S. Y.; Jochum, M. N.; Koschke, K.; Aluru, N. R.; Rühle, V.; Junghans, C. Relative Entropy and Optimization-Driven Coarse-Graining Methods in VOTCA. *PLoS One* **2015**, *10*, e0131754.
(51) Nicolas, J.; Gubbins, K.; Streett, W.; Tildesley, D. Equation of State for the Lennard-Jones Fluid. *Mol. Phys.* **1979**, *37*, 1429−1454.
(52) Johnson, J. K.; Zollweg, J. A.; Gubbins, K. E. The Lennard-Jones Equation of State Revisited. *Mol. Phys.* **1993**, *78*, 591−618.
(53) Wu, G.-W.; Sadus, R. J. Hard Sphere Compressibility Factors for Equation of State Development. *AIChE J.* **2005**, *51*, 309−313.
(54) Thol, M.; Rutkai, G.; Köster, A.; Lustig, R.; Span, R.; Vrabec, J. Equation of State for the Lennard-Jones Fluid. *J. Phys. Chem. Ref. Data* **2016**, *45*, 023101.
(55) Nie, X. B.; Chen, S. Y.; E, W. N.; Robbins, M. O. A Continuum and Molecular Dynamics Hybrid Method for Micro- and Nano-Fluid Flow. *J. Fluid Mech.* **1999**, *500*, 55−64.
(56) Scukins, A.; Nerukh, D.; Pavlov, E.; Karabasov, S.; Markesteijn, A. Multiscale Molecular Dynamics/Hydrodynamics Implementation of Two Dimensional "Mercedes Benz" Water Model. *Eur. Phys. J.: Spec. Top.* **2015**, *224*, 2217−2238.
(57) Karimi, M.; Marchisio, D.; Laurini, E.; Fermeglia, M.; Pricl, S. Bridging the Gap Across Scales: Coupling CFD and MD/GCMC in Polyurethane Foam Simulation. *Chem. Eng. Sci.* **2018**, *178*, 39−47.
(58) Hura, G.; Russo, D.; Glaeser, R. M.; Head-Gordon, T.; Krack, M.; Parrinello, M. Water Structure as a Function of Temperature from X-ray Scattering Experiments and Ab Initio Molecular Dynamics. *Phys. Chem. Chem. Phys.* **2003**, *5*, 1981−1991.
(59) Chen, M.; Ko, H.-Y.; Remsing, R. C.; Calegari Andrade, M. F.; Santra, B.; Sun, Z.; Selloni, A.; Car, R.; Klein, M. L.; Perdew, J. P.; et al. Ab Initio Theory and Modeling of Water. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, 10846−10851.
(60) Friedman, J.; Hastie, T.; Tibshirani, R. *The Elements of Statistical Learning*; Springer: New York, 2001.
(61) Smit, B. Phase Diagrams of Lennard-Jones Fluids. *J. Chem. Phys.* **1992**, *96*, 8639−8640.
(62) Loscar, E. S.; Ferrara, C. G.; Grigera, T. S. Spinodals and Critical Point using Short-Time Dynamics for a Simple Model of Liquid. *J. Chem. Phys.* **2016**, *144*, 134501.
(63) Pieprzyk, S.; Brańka, A. C.; Maćkowiak, S.; Heyes, D. M. Comprehensive Representation of the Lennard-Jones Equation of State Based on Molecular Dynamics Simulation Data. *J. Chem. Phys.* **2018**, *148*, 114505.
(64) Robles, M.; López de Haro, M.; Santos, A. Note: Equation of State and the Freezing Point in the Hard-Sphere Model. *J. Chem. Phys.* **2014**, *140*, 136101.
(65) Verlet, L.; Weis, J.-J. Equilibrium Theory of Simple Liquids. *Phys. Rev. A: At., Mol., Opt. Phys.* **1972**, *5*, 939−952.
(66) Henderson, D.; Grundke, E. W. Direct Correlation Function: Hard Sphere Fluid. *J. Chem. Phys.* **1975**, *63*, 601−607.
(67) Smith, W. R.; Henderson, D. J.; Leonard, P. J.; Barker, J. A.; Grundke, E. W. Fortran Codes for the Correlation Functions of Hard Sphere Fluids. *Mol. Phys.* **2008**, *106*, 3−7.
(68) McQuarrie, D. A. *Statistical Mechanics*; Harper & Row: New York, 1976.

https://dx.doi.org/10.1021/acs.jpclett.0c00627
J. Phys. Chem. Lett. 2020, 11, 4372−4378