

Perfecting Liquid-State Theories with Machine Intelligence

Jianzhong Wu^{*,†} and Mengyang Gu^{*,‡}

*[†]Department of Chemical and Environmental Engineering, University of California,
Riverside, CA 92521, USA*

*[‡]Department of Statistics and Applied Probability, University of California, Santa Barbara,
CA 93106, USA*

E-mail: jwu@engr.ucr.edu; mengyang@pstat.ucsb.edu

Abstract

Recent years have seen a significant increase in the use of machine intelligence for predicting electronic structure, molecular force fields, and the physicochemical properties of various condensed systems. However, substantial challenges remain in developing a comprehensive framework capable of handling a wide range of atomic compositions and thermodynamic conditions. This perspective discusses potential future developments in liquid-state theories leveraging on recent advancements of functional machine learning. By harnessing the strengths of theoretical analysis and machine learning techniques including surrogate models, dimension reduction and uncertainty quantification, we envision that liquid-state theories will gain significant improvements in accuracy, scalability and computational efficiency, enabling their broader applications across diverse materials and chemical systems.

Introduction

A conventional wisdom in the molecular theory of simple liquids is that the microscopic structure is determined by molecular excluded volume effects, whereas longer-ranged interactions make contributions to thermodynamic properties that can be adequately described by mean-field approximation.¹ This idea may trace its origin back to the pioneering work by Johannes van der Waals on the continuity of the gaseous and liquid states published in 1873. Because the intermolecular repulsion resembles the interaction between billiard balls, the hard-sphere model had been employed to characterize the microscopic structure of simple fluids long before the advent of Monte Carlo (MC) and molecular dynamics (MD) simulation methods. Figure 1 illustrates some major milestones in the development of modern liquid-state theories. It is probably not an exaggeration to state that molecular theories of liquid systems are rooted in the mathematical models for predicting the properties of hard spheres.

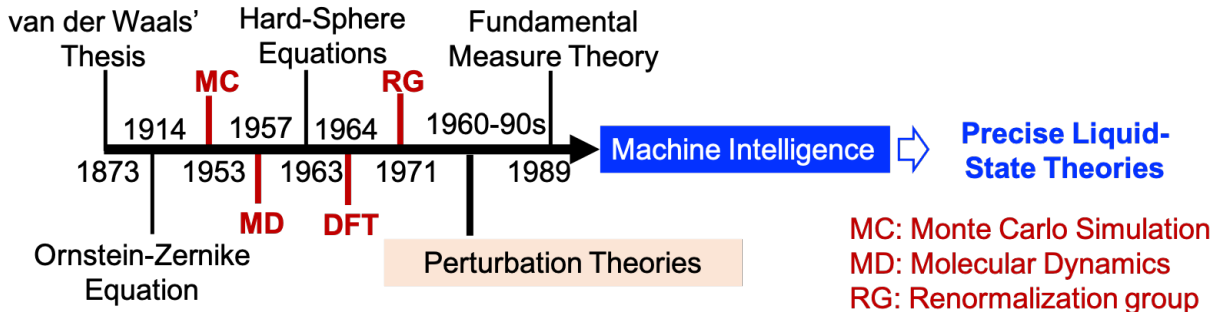


Figure 1: Milestones in the development of liquid-state theories and prospect of future advancement driven by machine intelligence.

Hard spheres do not interact with each other unless there is an overlap. As a result, the structure of a hard-sphere system cannot be described with mean-field approximation or typical perturbation methods. While the thermodynamic properties of hard-sphere systems can be obtained through experiment or simulation, analytical theories are mostly derived from geometric analysis, integral-equation theories, and various interpolation schemes.² The fundamental measure theory (FMT), originally formulated by Yaakov Rosenfeld, provides probably the most systematic and accurate description of the microscopic structure and

properties of hard-sphere systems at either uniform or inhomogeneous conditions.³

Given an analytical expression for the radial distribution functions of hard-sphere fluids and that for a pairwise additive potential describing intermolecular interactions, the van der Waals picture is capable of predicting the properties of simple liquids in the vicinity of the triple point.⁴ The mean-field approximation can also be used to account for long-range interactions in molecular fluids and electrolyte solutions through a judicious selection of a reference system that reproduces the liquid structure.⁵ While the distinct roles played by the intermolecular short- and long-range interactions in determining the properties of a dense fluid offer valuable insights, van der Waals-type theories encounter challenges when applied to associating fluids and liquid mixtures.⁶ As a matter of fact, its performance deteriorates even for one-component simple liquids as the thermodynamic condition departs from the triple point. The mean-field approximation ignores fluctuation effects associated with the variation of the local structure due to intermolecular attraction. Such effects can be adequately described with the Barker-Henderson perturbation theory⁷ or, more systematically, the hierarchical reference theory.⁸ Alternatively, the influence of long-range interactions on the microscopic structure can be explicitly accounted for by employing the functional Taylor expansion of the radial distribution function.⁹ The perturbation methods are naturally applicable to one-component fluids as well as mixtures. In combination with the renormalization group (RG) theory to incorporate critical fluctuations,¹⁰ they provide quantitative descriptions of thermodynamic properties and phase behavior over the entire range of the vapor-liquid coexistence. Importantly, an explicit consideration of long-range fluctuations is instrumental to understand the asymptotic features of phase transition in pure fluids and mixtures.

In contrast to simple liquids, molecular systems involve intra- and intermolecular interactions as characterized by bond connectivity, electrostatic forces as well as van der Waals interactions. The complexity of the potential energy necessitates the use of more sophisticated theoretical models and advanced computational techniques in order to accurately

describe the microscopic structure and thermodynamic properties.⁶ Such models often rely on semi-empirical force fields or coarse-grained representations of intermolecular interactions. While these classical models play a crucial role in molecular simulation, which yields exact results within the model, practical applications often rely on analytical equations to strike a balance between computational efficiency and accuracy. This balance is essential in various applied fields, such as materials science, drug development, and chemical engineering, where real-world applications demand a pragmatic approach.^{11–13}

From a fundamental perspective, one of the most enduring challenges in predicting the properties of liquid systems arises from the need for an accurate description of intermolecular interactions. Although *ab initio* molecular dynamics (AIMD) has the potential to predict the multi-body energy for any atomic system, its utility is severely limited by the computational expense, especially when conducting free-energy calculations as required for constructing the phase diagram of complex molecular systems. While semi-empirical force fields have been well-established to provide a systematic description of the atomic forces and total energies of molecular systems, their predictive power is constrained by specific functional forms and parameters that must be fitted with experimental data and/or first-principles predictions. Because the classical models frequently invoke the assumption of pairwise additivity for intermolecular interactions, there are also difficulties associated with transferring the model parameters between different molecular systems or thermodynamic conditions. The non-additive nature of many-body interactions implies that the classical models may not be able to provide an accurate description of the molecular energy in both low-density vapor and condensed states such as liquid and solid. Interestingly, the “short blanket dilemma”, *i.e.*, the difficulties for an accurate description of both two-body interactions and thermodynamic properties of condensed phases, continues to be a challenge in recent advancements of machine learning (ML) force fields.¹⁴

The potential energy of a molecular system can be highly complex and may not be easily scalable or transferable at the microscopic level. Fortunately, these issues tend to

diminish in significance as the system size approaches the thermodynamic limit. In this regard, liquid-state theories offer distinct advantages when compared to molecular simulation because they are naturally applicable to macroscopic systems and are not constrained by the box size used in simulations or the small time scales of atomic motions. Within the framework of a semi-empirical force field or coarse-grained model, a wide range of liquid-state theories are available (Figure 2). These theoretical techniques enable efficient analysis of the properties and phase behavior of molecular systems without the need to explicitly sample the microstates. Moreover, they provide insights into the intrinsic relationships between various thermodynamic properties and microscopic interactions, which can be challenging to extract from molecular simulation. For these reasons, liquid-state theories can serve as a powerful alternative to simulation methods for predicting the properties of molecular systems, particularly in dealing with large or complex systems where exhaustive simulation may be computationally prohibitive.

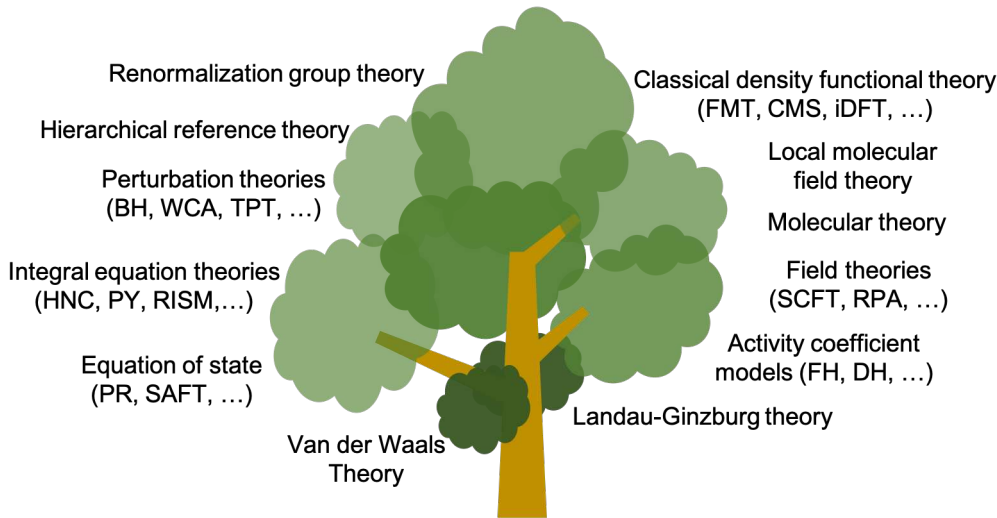


Figure 2: Various forms of liquid-state theories for predicting the thermodynamic properties of molecular systems. The list is incomplete and meant to be illustrative.

The shortcomings and limitations of liquid-state theories are also well known. Liquid-state theories often resort to perturbation expansions or mean-field approximation to describe thermodynamic non-ideality arising from correlation effects. These approximations may introduce unpredictable errors in comparison with exact results. While the analytical methods

offer advantages in terms of computational efficiency, their numerical implementation can still be demanding, particularly when dealing with highly inhomogeneous and/or asymmetric systems entailing both large and small molecules. To some extent, finding a delicate balance between accuracy and computational efficiency is akin to approaches taken in applying the Kohn-Sham density functional theory (KS-DFT) for electronic structure calculations. While there have been substantial endeavors to enhance the accuracy and computational efficiency of KS-DFT by incorporating ML methods,¹⁵ comparable efforts are only in their nascent stages in the realm of liquid-state theories. In this article, we highlight some of the recent advancements in this field by using integral-equation theories and classical density functional theory (cDFT) of simple fluids as illustrative examples. We outline a few future research directions aimed at improving the accuracy and computational efficiency of liquid-state calculations through incorporating recent advancements in machine intelligence, such as surrogate models, uncertainty quantification, and dimension reduction techniques.

Integral-equation theories

Integral-equation theories provide formal mathematical relationships among various correlation functions underlying the non-ideal behavior of thermodynamic properties of a many-body system.² One prime example is the Ornstein-Zernike (OZ) equation, which describes the spatial correlations of local particle densities at different positions. For a uniform system of spherical particles with average density ρ , the OZ equation can be written as

$$h(r) = c(r) + \rho \int d\mathbf{r}' c(|\mathbf{r} - \mathbf{r}'|) h(r') \quad (1)$$

where $h(r)$ and $c(r)$ are the total and direct correlation functions, respectively. The total correlation function $h(r)$ is a dimensionless quantity depending on the radial distance r from the position of any tagged particle in the system. It depicts the (normalized) local density of other particles relative to the average density. Intuitively, Eq.(1) suggests that

$h(r)$ comprises a direct contribution $c(r)$ accounting for the correlated particle densities at two positions separated by distance r , and indirect contributions due to correlation with particle densities at all other positions $\mathbf{r}'$ in space. The 3-dimensional Fourier transforms of the total and direct correlation functions, $\hat{h}(\mathbf{k})$ and $\hat{c}(\mathbf{k})$, are related to the static structure factor $\hat{\mathcal{S}}(\mathbf{k})$

$$\hat{\mathcal{S}}(\mathbf{k}) = 1 + \rho \hat{h}(\mathbf{k}) = \frac{1}{1 - \rho \hat{c}(\mathbf{k})} \quad (2)$$

The static structure factor is directly proportional to the scattering intensity measured in neutron and X-ray scattering experiments, rendering it indispensable for analyzing the experimental data. Assuming a pairwise additive potential $u(r)$ for particle-particle interactions, one can readily predict all thermodynamic properties of the one-component system from the total and direct correlation functions using standard statistical-mechanical equations.

Integral-equation theories are theoretically rigorous and broadly applicable to mixtures as well as complex fluids.⁶ Because the correlation functions enable the prediction of both microscopic structure and thermodynamic properties without the need to explicitly consider microstates, integral-equation theories have the unique capability to describe the properties of macroscopic systems from a molecular perspective. However, one caveat is that, in practical applications, additional information is needed for describing the intermolecular potential and its connection with the correlation functions. For example, the OZ equation for one-component systems of spherical particles discussed above requires an additional relation linking the total and direct correlation functions to the pair potential $u(r)$

$$\ln[h(r) + 1] = -\beta u(r) + h(r) - c(r) + b(r) \quad (3)$$

where $\beta = 1/(k_B T)$ with k_B being the Boltzmann constant and T absolute temperature, and $b(r)$ denotes the bridge function. The latter is defined by the high-order terms in the diagrammatic expansion of the total correlation function when it is expressed in terms of the

ensemble average or, equivalently, the functional expansion of the free energy with respect to the local particle density.² Without $b(r)$, Eq.(3) is known as the hypernetted chain (HNC) approximation.

In integral-equation theories, the additional equation relating the total and/or direct correlation functions to the intermolecular potential, as given by Eq.(3) for one-component systems, is commonly known as the closure. Whereas a formal relation among these functions can be derived for any molecular system, an exact closure involves a bridge function that is not analytically tractable even for simple fluids. The practical application of the OZ equation mostly relies on approximate closures derived through heuristic analysis or the universality *ansatz* for the bridge function. Over the years, highly accurate closures have been developed for various systems, applicable not only to simple fluids but also to complex systems such as aqueous solutions of electrolytes, molten salts, and polymeric materials.^{16–18} Importantly, analytical results derived from the Percus-Yevick (PY) equation and the mean-spherical approximation (MSA) play a pivotal role in the development of modern liquid-state theories for both bulk and inhomogeneous fluids.

While the importance of integral-equation theories cannot be overstated, their widespread application is impeded by challenges in numerically solving the correlation functions and the inherent inaccuracies of approximate closures. These limitations may result in thermodynamic inconsistencies and nonphysical predictions, hindering their extensive utilization. These issues have been recognized for decades yet with only limited progress. It was only recently demonstrated that both numerical and theoretical challenges can be mitigated by leveraging on machine intelligence.

OZ surrogates

Statistical and machine learning techniques prove invaluable in overcoming computational bottlenecks associated with solving high-dimensional nonlinear equations. We anticipate that similar methodologies can be employed to alleviate numerical difficulties in solving the

OZ equation.

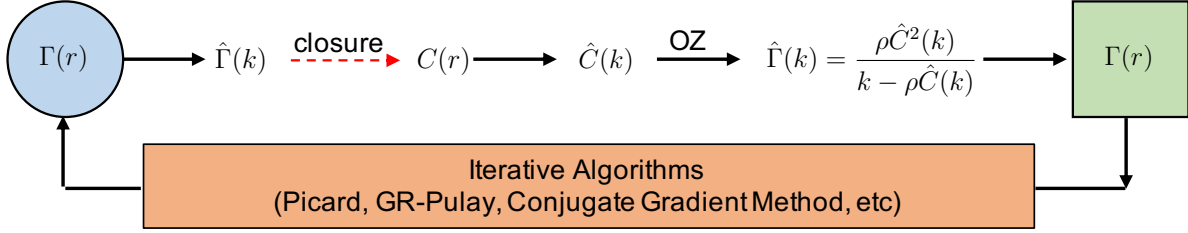


Figure 3: Solving the Ornstein-Zernike (OZ) equation with iterative methods is computationally demanding and frequently beset by convergence. Here, $\Gamma(r) \equiv r[h(r) - c(r)]$, $C(r) \equiv rc(r)$, $\hat{\Gamma}(k)$ and $\hat{C}(k)$ represent the Fourier transforms of $\Gamma(r)$ and $C(r)$.

As shown schematically in Figure 3, conventional methods for solving the OZ equation start with an initial guess for the indirect correlation function $\Gamma(r) \equiv r[h(r) - c(r)]$, followed by the calculation of the direct correlation function from a closure. Next, the OZ equation is used to generate a new estimate for $\Gamma(r)$ through the Fourier transform, and the procedure iterates with an appropriate numerical scheme until the solution reaches convergence. While the computational cost is dependent on the complexity of the system under consideration and numerical details, one major obstacle in practical applications is that the OZ equation may not have a numerical solution, or the iterative process fails to converge.

Recently, Carvalho and Braga demonstrated that, for the Lennard-Jones fluids, the OZ equation can be alternatively solved by using Physics Informed Neural Networks (PINN).¹⁹ Regrettably, the authors reported that the ML model took longer computational time than the direct iterative methods for solving the OZ equation with the PY and HNC closures. Nevertheless, PINN has advantages when considering additional physical information such as the asymptotic behavior of the correlation functions near the critical point. Besides, PINN can be used to retrieve correlation functions from experimental data for the static structure factor. More efficient ML techniques have been proposed to solve the OZ equation by combining physics-informed neural operator networks with Fourier feature embedding and a self-adaptive weighting method.²⁰ The surrogate models for solving the OZ equation can be similarly applied to other integral-equation theories including those for complex molecu-

lar systems. By establishing ML correlations between input and output functions directly, the integral-equation theories could facilitate rapid predictions of the equilibrium structure, solvent-mediated interactions and phase behavior of polymeric materials and biochemical systems.^{21–23}

Neural network approaches like PINN integrate both the fit of simulation runs and the governing equations for learning the latent function. However, they do not directly provide uncertainty quantification as the parameters are optimized using a loss function. Another strategy of building surrogate models²⁴ is to learn the mapping from inputs, such as the pair potential and average density in the OZ equation, through Bayesian inference. Bayesian models, such as Gaussian process (GP) emulators,^{25,26} have been widely used for predicting computationally expensive simulations in different disciplines. One advantage of Bayesian modeling is that the uncertainty of the predictions can be quantified from the predictive intervals, either through posterior sampling or approximation from the posterior mode estimation, whereas a neural network approach may not directly provide assessment of the uncertainty in predictions. Furthermore, a GP emulator has a closed-form predictive distribution and it does not have many tuning parameters as neural networks. These properties make the GP emulator less data-demanding and easier to train. However, the computational complexity of a GP emulator increases cubically with the number of samples in the training data set. When dealing with a large number of training samples, techniques such as the inducing point approach²⁷ are often used to approximate the likelihood function for training the GP emulator with a reduced computational cost.

ML closure

The total and direct correction functions of a molecular system can be extracted from molecular simulation.²⁸ Subsequently, the simulation data can be utilized to calculate and train the bridge function based on the pair potential and the exact closure relation, *e.g.*, Eq.(3) for one-component systems of spherical particles. Goodall and Lee carried out MD simulation

for thirteen one-component systems of spherical particles over a broad range of thermodynamic conditions and inferred the bridge functions using a class of neural networks called multi-layer perceptrons (MLP).²⁹ As illustrated schematically in Figure 4, the bridge function can be regressed in terms of a set of structure features including not only the total and direct correlation functions $h(r)$ and $c(r)$ but also the pair fluctuation function $\chi(r)$ and the gradient of the indirect correlation function $\gamma'(r) \equiv h'(r) - c'(r)$. The ML closure was tested by applying it to solve the so-called inverse problem of statistical mechanics, *i.e.*, extracting the pair potential from the simulation data for the static structure factor. These authors found that the ML approach yields better predictions of the pair potential than conventional closures such as PY and HNC. It was suggested that even more robust and reliable predictive capability may be achieved through Bayesian neural networks.

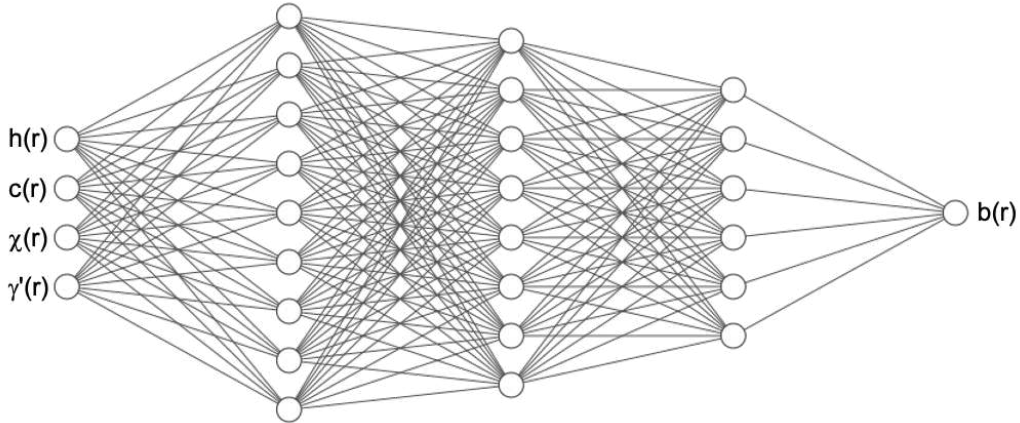


Figure 4: Schematic of a multilayer perceptrons neuron network (MPNN) for inferring $b(r)$ from $h(r)$, $c(r)$, $\chi(r)$ and $\gamma'(r)$. Reproduced from Goodall and Lee.²⁹

The ML closures derived by Goodall and Lee corroborate the universality *ansatz* for the bridge function, *i.e.*, the bridge function of a simple fluid is mainly determined by the short-range repulsion. However, it was not clear whether the learned bridge functions could be represented by effective hard-sphere systems as commonly assumed in the literature. Further development of ML closures is needed to correlate the bridge function directly with the potential energy and for systems exhibiting more complicated intermolecular interactions. It is probably worth noting that the ML closure approximates a complex functional whereby

accurate interpolation in high-dimensional space can be challenging for test inputs that are far away from the training data set. The predictive accuracy of ML closures can be improved by reducing the dimension of the input or utilizing quantified uncertainty from a probabilistic model to control the predictive error, which has been demonstrated for emulating classical density functional theory.³⁰

Structure-centered modeling

Typically, classical force fields employ pairwise additive potentials to describe intermolecular interactions. Despite the apparent deviations of the effective pair potential from the interaction between two isolated molecules in a vacuum, classical models can achieve remarkable accuracy in predicting the microscopic structure and thermodynamic properties of condensed-matter systems, given that the model parameters have been carefully calibrated. In this context, integral-equation theories can be used to solve the inverse problem mentioned above. With the ground truth on microscopic structure obtained from AIMD or scattering experiments, ML methods can be employed to improve not only the closure relations but also the quality of force-field parameters for representing diverse properties of the system under consideration.

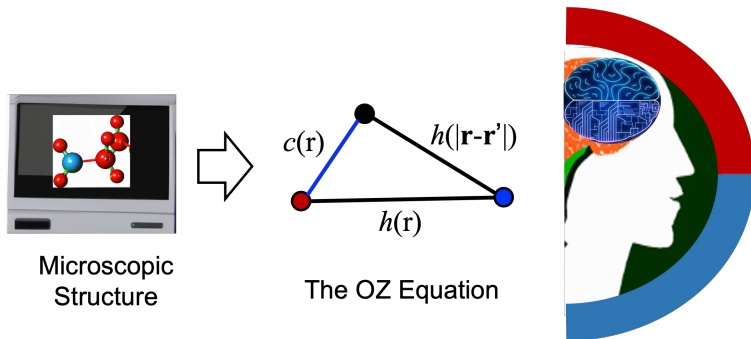


Figure 5: Predicting the properties of condensed-matter systems through pairwise potentials derived from the microscopic structure.

Figure 5 shows schematically the “structure-centered” approach for predicting the properties of condensed-matter systems. Rather than seeking multibody or ML potentials,

structure-property relationships can be established by assuming *a priori* pairwise additivity for the effective pair potential. While the state-dependent potential may have its own limitations in describing multi-body interactions, it can be optimized to reproduce microscopic structure as well as experimentally observed thermodynamic properties. Importantly, the effective potential eschews problems associated with force field transferability (and scalability) because the molecular model is constructed for the specific system under consideration. Recently, there have been significant advances in using ML methods to extract information on the microscopic structure from molecular simulation and scattering experiment, making it easier to calculate the pair distribution functions and the effective pair potentials.³¹ Impressively, the so-called point cloud-based deep learning strategy is able to rapidly predict the equilibrium structure of a liquid system from a single molecular configuration.³²

When estimating the parameters or functions in these ML methods, one often fits the model by minimizing the difference between the modeled output and transformed experimental data at a selected regime or reweighted data based on a loss function. However, the uncertainty of the transformed data from measurement noise or model discrepancy may not be always homogeneous or independent, particularly for scattering measurements, which could make some seemingly sensible choices of loss functions, such as the squared error loss, not optimal in estimation. Bayesian inference, on the other hand, often starts with a probabilistic generative model,^{33,34} which can improve the loss-minimization estimation through deriving more efficient estimators based on the probabilistic model. Having a generative model also enables uncertainty estimation to be quantified, and thus enhancing our understanding of implicit model assumptions made in the loss function.

Classical density functional theory

While integral-equation theories focus on total and direct correlation functions, classical density functional theory (cDFT) aims to predict thermodynamic properties directly from

one-body density profiles.³⁵ These methods are formally interconnected through variational calculus. In principle, all expressions in integral-equation theories can be derived from cDFT, but not *vice versa*, because the latter deals with the free energy instead of pair correlation functions. Although integral-equation theories have also been utilized for inhomogeneous systems, they do not rely on the minimization of a free energy and are mostly concerned with the structure of bulk systems. In contrast, the functional dependence of free energy on one-body density profiles and external potentials is an essential component of cDFT, making it naturally suitable for studying inhomogeneous systems.²

Conceptually, cDFT shares the mathematical foundation with KS-DFT and they face similar challenges in practical applications. Both methods are formally exact but constrained by the imprecise formulation of the density functional and the numerical complexity in solving the DFT equations. While KS-DFT predicts the ground-state energy of an electronic system by mapping the density into that of ideal fermions, cDFT is mainly concerned with the grand potential of a thermodynamic system as a functional of the density profiles of the underlying species. Minimization of the grand potential leads to the Euler-Lagrange equation, which for a one-component system of spherical particles, is given by

$$\rho(\mathbf{r}) = \frac{1}{\Lambda^3} \exp \left\{ -\beta\varphi(\mathbf{r}) + c_1(\mathbf{r}) \right\} \quad (4)$$

where Λ stands for the de Broglie thermal wavelength, $\varphi(\mathbf{r}) \equiv V^{ext}(\mathbf{r}) - \mu$ is defined as the external potential $V^{ext}(\mathbf{r})$ minus the chemical potential μ , and $c_1(\mathbf{r})$ represents the one-body direct correlation function. Eq.(4) is analogous to the KS equation, *viz.*, it predicts the density profile of a non-interacting system under an effective one-body potential. Similar cDFT equations can be written for molecular and polymeric systems.^{36–41}

Ottel and coworkers first explored the feasibility of using neural networks to construct free-energy functionals for one-dimensional hard rods and Lennard–Jones fluids.^{42,43} With the density functionals expressed in terms of a polynomial or artificial neural networks, they

found that the ML functionals are able to reproduce the density profiles at conditions beyond the training region. Yatsyshin *et al.* introduced a Bayesian method for predicting the external potential of classical particles from the density profiles or configurations generated through MC simulation.^{44,45} The predictions from the ML regressions were found in good agreement with the exact results for both the grand potential and density profiles of one-dimensional hard-rod systems. Cats *et al.* combined the cDFT formalism with a stochastic optimization method to regress the non-mean-field component of free-energy functional for the Lennard-Jones fluids.⁴⁶ They expressed the non-mean-field free energy in terms of the density profiles in quadratic and cubic forms with the corresponding kernels obtained from fitting with density profiles obtained from grand-canonical MC simulation. The ML functional also provides good predictions of the thermodynamic properties and pair correlation functions of bulk fluids.

In the aforementioned publications, the density profiles are calculated through an exact functional or molecular simulation under different external potentials. Subsequently, ML methods are employed to regress the excess Helmholtz energy as a functional of the one-body density profile with a predetermined form. More recently, Sammüller *et al.* demonstrated that a deep neural network can be constructed to regress the functional dependence of the one-body direct correlation function with respect to the density profile.⁷ Subsequently, the functional relation between $c_1(\mathbf{r})$ and $\rho(\mathbf{r})$ can be utilized for the calculation of the grand potential by functional line integration

$$\beta\Omega = \int d\mathbf{r} \rho(\mathbf{r}) \left\{ \ln[\rho(\mathbf{r})\Lambda^3] - 1 + \beta\varphi(\mathbf{r}) \right\} - \int_0^1 d\lambda \int d\mathbf{r} \rho(\mathbf{r}) c_1(\mathbf{r}; \lambda\rho(\mathbf{r})). \quad (5)$$

Impressively, the neural functional was found outperforming the state-of-the-art cDFT methods for inhomogeneous systems of hard-sphere fluids.

Whereas the ML techniques discussed above were tested only with simple systems, we expect that similar improvements can be achieved for all cDFT calculations by formulating

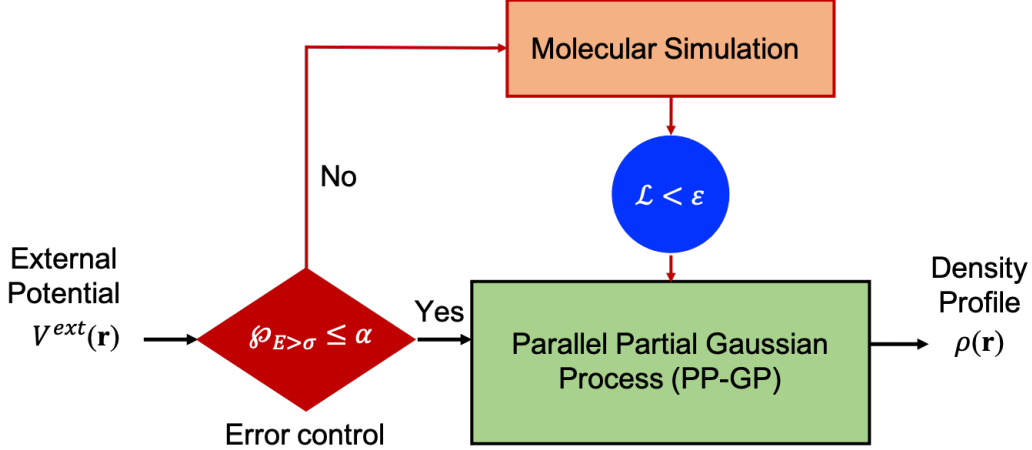


Figure 6: Schematic of active functional learning with error control. Here, the functional learning approach is applied to regress the dependence of the one-body density profile on the external potentials, $\wp_{E>\sigma} \leq \alpha$ stands for the probability of the predictive error larger than δ is less than α , and $\mathcal{L} < \varepsilon$ means the loss function is less than ε .

suitable neutral functionals. In the application of cDFT to molecular systems, the numerical procedure will be necessarily more complicated and requires more advanced ML methods such as active learning techniques and dimensionality reduction. Although cDFT calculations are faster than molecular simulation, minimizing the grand potential or solving the Euler-Lagrange equation is computationally demanding for inhomogeneous systems containing polymeric species or molecules interacting through fine-detailed potentials. With an expression for the grand potential, which can be obtained either from theoretical derivation or machine learning, surrogate ML models can be established to predict density profiles with the external potential as the input. For example, artificial neural network was used as a surrogate for the iterative solution of the Euler-Lagrange equation for liquid-crystal systems under confinement.⁴⁷ More recently, Fang *et al.* proposed an active learning framework that utilizes a parallel partial Gaussian process (PP-GP) emulator⁴⁸ to predict $\rho(\mathbf{r})$ from external potential $V^{ext}(\mathbf{r})$, and demonstrated that the ML approach is more computationally scalable than direct cDFT calculations.³⁰ As shown schematically in Figure 6, the active learning approach is able to discern the functional inputs that can be reliably predicted based on quantified uncertainty from the emulator, thereby controlling the predictive error

in a high-dimensional input space for new applications.

Summary and Perspectives

Liquid-state theories play an instrumental role in predicting the thermodynamic properties and phase behavior of liquid mixtures and solutions important for diverse scientific fields. Whereas approximations are inevitable in conventional practice, their integration with ML methods holds the potential for more scalable and accurate calculations over a wide range of conditions. In particular, the theoretically-guided functional learning methods will eventually offer a valuable alternative to molecular simulations for mapping the high-dimensional potential energies of thermodynamic systems into equilibrium properties of practical concern.

Machine learning (ML) methods such as Gaussian Processes (GP) and Physics-Informed Neural Networks (PINN) have been widely used for emulating complex functions or functionals, and for solving differential and integral equations. However, similar to other molecular-level simulations, a fundamental bottleneck arises from the high-dimensionality of input space when approximating liquid-state theories. Two promising directions are possible for improving the application of ML methods. First, reducing the dimension of the input space has the potential to significantly lower the computational cost and improve predictive accuracy. Some physics equations intrinsically depend on a low dimensional space of the entire input space. For example, the one-body direct correlation function $c_1(\mathbf{r})$ of a hard-sphere system is a short-ranged function of one-body density profile $\rho(\mathbf{r})$. In this case, we can build a ML surrogate model for $c_1(\mathbf{r})$ using the local density instead of the entire density profile as the input. The input dimension can be further reduced by using the active subspace (AC) approach,⁴⁹ which projects the high-dimensional input vector to a few principal axes that maintain the largest deviation from the output gradients. The AC approach is often more accurate to capture the input-output map than unsupervised approaches such as the principal component analysis, as the output variation characterized by the gradients are used for

building principal axes.

In addition to dimension reduction, having reliable uncertainty assessment from ML techniques can be useful to improve the predictive accuracy, particularly for extrapolating the input space. The predictive intervals generated by Bayesian models, such as the parallel partial Gaussian process (PP-GP) emulator⁴⁸ and corresponding software package,⁵⁰ can be used to determine whether particle density and grand potential in cDFT calculations can be reliably predicted for a new external potential.³⁰ These intervals help control the predictive errors in a probabilistic manner, offering valuable insights into the accuracy of the predictions. In ML techniques such as artificial neural networks, predictive uncertainty may not be directly available because parameters are estimated by minimizing a loss function, instead of defining a generative model of the data. Building a probabilistic generative model enables the uncertainty to be quantified. Moreover, the generative model can motivate the derivation of more efficient estimators that improve estimation from loss-minimization.³³ Therefore, dimension reduction and uncertainty quantification from probabilistic generative models are two critical directions to enhance the performance of liquid-state theories for computationally challenging scenarios, such as 3D systems with long-range intermolecular correlations.

Addressing the computational challenges associated with bridging molecular and macroscopic scale phenomena has been a fundamental pursuit in theoretical research within the realms of chemistry and materials science. Surrogate models and functional learning methods have proved valuable for both quantum and statistical mechanical calculations and are expected to play an increasingly important role in advancing this endeavor.

References

- (1) Evans, R.; Frenkel, D.; Dijkstra, M. From simple liquids to colloids and soft matter. *Physics Today* **2019**, *72*, 38–39.

- (2) Hansen, J.-P.; McDonald, I. R. *Theory of Simple Liquids: with Applications of Soft Matter*; Academic Press, 2013.
- (3) Roth, R. Fundamental measure theory for hard-sphere mixtures: a review. *Journal of Physics: Condensed Matter* **2010**, *22*, 063102.
- (4) Chandler, D.; Weeks, J. D.; Andersen, H. C. Van der Waals Picture of Liquids, Solids, and Phase Transformations. *Science* **1983**, *220*, 787–794.
- (5) Gao, A.; Remsing, R. C.; Weeks, J. D. Local Molecular Field Theory for Coulomb Interactions in Aqueous Solutions. *The Journal of Physical Chemistry B* **2023**, *127*, 809–821.
- (6) Gray, C. G.; Gubbins, K. E.; Joslin, C. G. *Theory of Molecular Fluids*; Oxford University Press, 2011.
- (7) Barker, J. A.; Henderson, D. What is 'liquid'? Understanding the states of matter. *Reviews of Modern Physics* **1976**, *48*, 587–671.
- (8) Parola, A.; Reatto, L. Recent developments of the hierarchical reference theory of fluids and its relation to the renormalization group. *Molecular Physics* **2012**, *110*, 2859–2882.
- (9) Tang, Y. P.; Lu, B. C. Y. First-Order Radial Distribution Functions Based on the Mean Spherical Approximation for Square-Well, Lennard-Jones, and Kihara Fluids. *Journal of Chemical Physics* **1994**, *100*, 3079–3084.
- (10) Behnejad, H.; Sengers, J. V.; Anisimov, M. A. Thermodynamic behaviour of fluids near critical points. *Applied Thermodynamics of Fluids* **2010**, *8*, 321–367.
- (11) Arora, A.; Qin, J.; Morse, D. C.; Delaney, K. T.; Fredrickson, G. H.; Bates, F. S.; Dorfman, K. D. Broadly Accessible Self-Consistent Field Theory for Block Polymer Materials Discovery. *Macromolecules* **2016**, *49*, 4675–4690.

- (12) Yoshida, N. Role of Solvation in Drug Design as Revealed by the Statistical Mechanics Integral Equation Theory of Liquids. *Journal of Chemical Information and Modeling* **2017**, *57*, 2646–2656.
- (13) Esper, T.; Bauer, G.; Rehner, P.; Gross, J. PCP-SAFT Parameters of Pure Substances Using Large Experimental Databases. *Industrial & Engineering Chemistry Research* **2023**, *62*, 15300–15310.
- (14) Zhai, Y.; Caruso, A.; Bore, S. L.; Luo, Z.; Paesani, F. A “short blanket” dilemma for a state-of-the-art neural network potential for water: Reproducing experimental properties or the physics of the underlying many-body interactions? *The Journal of Chemical Physics* **2023**, *158*, 084111.
- (15) Pederson, R.; Kalita, B.; Burke, K. Machine learning and density functional theory. *Nature Reviews Physics* **2022**, *4*, 357–358.
- (16) Pellicane, G.; Lee, L. L.; Caccamo, C. Integral-equation theories of fluid phase equilibria in simple fluids. *Fluid Phase Equilibria* **2020**, *521*, 112665.
- (17) Gray, J. G.; Giambaşu, G. M.; Case, D. A.; Luchko, T. Integral equation models for solvent in macromolecular crystals. *The Journal of Chemical Physics* **2022**, *156*, 014801.
- (18) Ratkova, E. L.; Palmer, D. S.; Fedorov, M. V. Solvation Thermodynamics of Organic Molecules by the Molecular Integral Equation Theory: Approaching Chemical Accuracy. *Chemical Reviews* **2015**, *115*, 6312–6356.
- (19) Carvalho, F. S.; Braga, J. P. Physics Informed Neural Networks applied to liquid state theory. *Journal of Molecular Liquids* **2022**, *367*, 120504.
- (20) Chen, W.; Gao, P.; Stinis, P. Physics-informed machine learning of the correlation functions in bulk fluids. *arXiv preprint arXiv:2309.00767* **2023**,

- (21) Hall, L. M.; Jayaraman, A.; Schweizer, K. S. Molecular theories of polymer nanocomposites. *Current Opinion in Solid State and Materials Science* **2010**, *14*, 38–48.
- (22) Osaki, K.; Ekimoto, T.; Yamane, T.; Ikeguchi, M. 3D-RISM-AI: A Machine Learning Approach to Predict Protein–Ligand Binding Affinity Using 3D-RISM. *The Journal of Physical Chemistry B* **2022**, *126*, 6148–6158.
- (23) Roy, D.; Kovalenko, A. Multiscale Methods Framework with the 3D-RISM-KH Molecular Solvation Theory for Supramolecular Structures, Nanomaterials, and Biomolecules: Where Are We Going? *Thermo* **2023**, *3*, 375–395.
- (24) Santner, T. J.; Williams, B. J.; Notz, W. I.; Williams, B. J. *The design and analysis of computer experiments*; Springer, 2003; Vol. 1.
- (25) Deringer, V. L.; Bartók, A. P.; Bernstein, N.; Wilkins, D. M.; Ceriotti, M.; Csányi, G. Gaussian process regression for materials and molecules. *Chemical Reviews* **2021**, *121*, 10073–10141.
- (26) Gramacy, R. B. *Surrogates: Gaussian process modeling, design, and optimization for the applied sciences*; CRC press, 2020.
- (27) Snelson, E.; Ghahramani, Z. Sparse Gaussian processes using pseudo-inputs. *Advances in neural information processing systems* **2005**, *18*.
- (28) Sheng, S.; Wu, J. Direct correlation functions for three-site and four-site water models. *Molecular Physics* **2016**, *114*, 2351–2363.
- (29) Goodall, R. E. A.; Lee, A. A. Data-driven approximations to the bridge function yield improved closures for the Ornstein-Zernike equation. *Soft Matter* **2021**, *17*, 5393–5400.
- (30) Fang, X.; Gu, M.; Wu, J. Reliable emulation of complex functionals by active learning with error control. *The Journal of Chemical Physics* **2022**, *157*, 214109.

- (31) Moradzadeh, A.; Aluru, N. R. Transfer-Learning-Based Coarse-Graining Method for Simple Fluids: Toward Deep Inverse Liquid-State Theory. *The Journal of Physical Chemistry Letters* **2019**, *10*, 1242–1250.
- (32) Li, C.; Gilbert, B.; Farrell, S.; Zarzycki, P. Rapid Prediction of a Liquid Structure from a Single Molecular Configuration Using Deep Learning. *Journal of Chemical Information and Modeling* **2023**, *63*, 3742–3750.
- (33) Gu, M.; He, Y.; Liu, X.; Luo, Y. Ab initio uncertainty quantification in scattering analysis of microscopy. *arXiv* **2023**, 2309.02468.
- (34) Gu, M.; Lin, Y.; Lee, V. C.; Qiu, D. Y. Probabilistic forecast of nonlinear dynamical systems with uncertainty quantification. *Physica D: Nonlinear Phenomena* **2024**, *457*, 133938.
- (35) Evans, R. The nature of the liquid-vapour interface and other topics in the statistical mechanics of non-uniform, classical fluids. *Advances in Physics* **1979**, *28*, 143–200.
- (36) Valiev, M.; Chuev, G. N.; Fedotova, M. V. CDFTPY: A python package for performing classical density functional theory calculations for molecular liquids. *Computer Physics Communications* **2022**, *276*, 108338.
- (37) Yagi, T.; Sato, H. Density functional theory for molecular liquids based on interaction site model and self-consistent integral equations for site–site pair correlation functions. *The Journal of Chemical Physics* **2020**, *153*, 164102.
- (38) Borgis, D.; Luukkonen, S.; Belloni, L.; Jeanmairet, G. Accurate prediction of hydration free energies and solvation structures using molecular density functional theory with a simple bridge functional. *The Journal of Chemical Physics* **2021**, *155*, 024117.
- (39) Pizio, O.; Sokołowski, S. Density functional approach to wetting behavior of water on

- solid surfaces modified by grafting of chains and their mixtures. *Journal of Molecular Liquids* **2023**, *390*, 123009.
- (40) Ginzburg, V. V. In *Theory and Modeling of Polymer Nanocomposites*; Ginzburg, V. V., Hall, L. M., Eds.; Springer International Publishing: Cham, 2021; pp 23–44.
 - (41) Forsman, J.; Woodward, C. E. In *Variational Methods in Molecular Modeling*; Wu, J., Ed.; Springer Singapore: Singapore, 2017; pp 101–136.
 - (42) Lin, S.-C.; Oettel, M. A classical density functional from machine learning and a convolutional neural network. *SciPost Phys.* **2019**, *6*, 025.
 - (43) Lin, S.-C.; Martius, G.; Oettel, M. Analytical classical density functionals from an equation learning network. *The Journal of Chemical Physics* **2020**, *152*, 021102.
 - (44) Malpica-Morales, A.; Yatsyshin, P.; Durán-Olivencia, M. A.; Kalliadasis, S. Physics-informed Bayesian inference of external potentials in classical density-functional theory. *The Journal of Chemical Physics* **2023**, *159*, 104109.
 - (45) Yatsyshin, P.; Kalliadasis, S.; Duncan, A. B. Physics-constrained Bayesian inference of state functions in classical density-functional theory. *The Journal of Chemical Physics* **2022**, *156*, 074105.
 - (46) Cats, P.; Kuipers, S.; de Wind, S.; van Damme, R.; Coli, G. M.; Dijkstra, M.; van Roij, R. Machine-learning free-energy functionals using density profiles from simulations. *APL Materials* **2021**, *9*, 031109.
 - (47) Santos-Silva, T.; Teixeira, P. I. C.; Anquetil-Deck, C.; Cleaver, D. J. Neural-network approach to modeling liquid crystals in complex confinement. *Phys. Rev. E* **2014**, *89*, 053316.
 - (48) Gu, M.; Berger, J. O. Parallel partial Gaussian process emulation for computer models with massive output. *The Annals of Applied Statistics* **2016**, *10*, 1317–1347.

- (49) Constantine, P. G.; Dow, E.; Wang, Q. Active subspace methods in theory and practice: applications to kriging surfaces. *SIAM Journal on Scientific Computing* **2014**, *36*, A1500–A1524.
- (50) Gu, M.; Palomo, J.; Berger, J. O. RobustGaSP: Robust Gaussian Stochastic Process Emulation in R. *The R Journal* **2019**, *11*, 112–136.

Graphical TOC Entry

